



revision: r519, 2012-11-09

Programming in **D**

*Learn to program in the
D programming language*

Ali Çehreli



D.ershane Series

0.1 Copyright © 2009-2012 Ali Çehreli

The contents of the ddili.org Forum and Wiki pages belong to all of the members.



Ddili.org icerigi by [Ali Cehreli](#) is licensed under a [Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License](#).

Based on a work at [digitalmars.com](#).

Permissions beyond the scope of this license may be available at <http://ddili.org/copyright.html>.

0.2 The PDF version of "Programming in D"

The latest PDF version of the book can be downloaded at the following address:

 http://ddili.org/ders/d.en/Programming_in_D.pdf

0.2.1 The revisions of the PDF file

- r519, 2012-11-09: Replace toHash() algorithms with more efficient ones.
- r514, 2012-11-02: The 'Classes', 'Inheritance', 'Object', 'Interfaces', 'destroy and scoped', 'Modules and Libraries', and 'Encapsulation and Protection Attributes' chapters.
- r506, 2012-09-15: The 'Operator Overloading' chapter and previous chapters.

The updates to the book can be seen at [the ddili.org project page](#).

1 Acknowledgments

I would like to thank my wife and daughter, who endured the long hours as much as I did while writing the original Turkish version as well as the English translation of this book.

These chapters have originally been reviewed by the members of the Turkish [Ddili Forum](#). I am grateful to the Turkish D community for keeping my excitement and motivation high.

Mert Ataol, Zafer Çelenk, and Salih Dinçer have given feedback on almost every line of the book. Can Alpay Çiftçi and Faruk Erdem Öncel have been instrumental in the development of both this book and of [ddili.org](#).

Thanks to the following people for their very valuable corrections, recommendations, and ideas: Ergin Güney, Jordi Sayol, David Herberth, Andre Tampubolon, Gour-Gadadhara Dasa, Raphaël Jakse.

This book has been proofread by Ergin Güney to be elevated to English from my Turkish.

2 Introduction

This book aims to teach the D language to readers who are new to computer programming. Although having experience in other programming languages would certainly be helpful, this book starts from the basics. If you are interested in learning how to program, I hope you will find this book helpful.

In order for this book to be useful, you will need an environment to write, compile, and run your D programs. This *development environment* must include at least the following:

- a text editor
- a D compiler

Instead of obtaining and installing those two separately, you may consider using an *integrated development environment* (IDE). You can find information about D IDEs at [wiki4d](#). You cannot learn to program in D without a text editor or a compiler. We will see how to install the **dmd** compiler and how to compile programs soon in later chapters.

Each chapter of this book introduces as few new concepts as possible. Most chapters have a number of exercises at the end. These exercises are accompanied by their solutions, so that you can compare your solutions to mine. On the web version of this book, the solutions appear in a new browser window. On the PDF version, they are all at the end of the book.

I suggest that you do not skip chapters. If you come across chapters that you find to be particularly difficult, it may be because the book fails to introduce all of the necessary concepts. Please e-mail the author about such problems to help make this book more useful.

This book does not cover *graphical user interface* (GUI) programming. Although many programs are more usable with a GUI, GUIs are not directly related to programming languages. Furthermore, the design choices and programming styles of GUIs may interfere with those of the programming language and its standard library, and make it more difficult to learn the language. For that reason, this book includes only *console* programs. Once you learn D and its standard library, Phobos, you will be able to use any GUI library that you like.

The chapters of this book are made available online as they are translated from their Turkish originals. You can use [the RSS feed](#) to be notified about the availability of new chapters.

Learning to program is more fun when shared with others. Follow the D.learn newsgroup at <http://forum.dlang.org/> to follow conversations and to ask and answer questions.

Please e-mail the author with any comments and corrections regarding this book. Thank you.

Author: Ali Çehreli

- Résumé: <http://acehrelil.org/resume.html>
- Interviews with him at [Ceviz.net](#) and at [İzmir'de Sanat](#)
- You can find him at [Ceviz.net's C and C++ forums](#), at [the Ddili Forum](#), and at [the D newsgroups](#)
- You can e-mail him at acehrelil@yahoo.com

3 Practice of Programming

It is somewhat debatable what the practice of programming exactly is, but the *craft* aspect of it is very strong. Some thoughts about programming:

- It is the task of creating programs that make the computer behave the way we want.
- As it requires tools and the use of those tools guided by the experience of master programmers, it is a craft.
- As it involves solving problems under constraints, it is engineering.
- It is very fun and satisfying.
- It is not a fine art, but as much as it's possible in anything that humans do, programs can be works of art.
- It is not a science, but the methods that it uses are created by computer science.

3.1 Programming can be very hard to learn and teach

Programming has been taught since the 1950s. There hasn't been an efficient or successful teaching method developed yet.

Unfortunately, learning to program can be a difficult task for roughly half of all students. According to [a research paper](#), the ones who can learn programming easily are those who are able to create *consistent models* to describe the unknown situations that they face.

Some of the difficulty in programming comes from the amount of technical details that need to be learned.

4 The Hello World Program

The first program to show in most programming language books is the *hello world* program. This very short and simple program merely writes "hello world" and finishes. This program is important because it includes some of the essential concepts of that language.

Here is a *hello world* program in D:

```
import std.stdio;

void main()
{
    writeln("Hello world!");
}
```

The *source code* above needs to be compiled by a D compiler to produce an executable program.

4.1 Compiler installation

At the time of writing this chapter, there are three D compilers to choose from: **dmd**, the Digital Mars compiler; **gdc**, the D compiler of GCC; and **ldc**, the D compiler that targets the LLVM compiler infrastructure.

dmd has been the reference compiler for the D programming language during the design and development of the language over the years. All of the examples in this book have been tested with **dmd**. For that reason, it would be the easiest for you to start with **dmd** and try other compilers only if you have a specific need to.

To install the latest version of **dmd**, go to the [download page at Digital Mars](#) and select the compiler build that matches your computer environment. You must select the **dmd** build that is for your operating system and package management system, and whether you have a 32-bit or a 64-bit CPU. Do not install a D1 compiler. This book covers only *D version two*.

The installation steps are different on different environments but it should be as easy as following simple on-screen instructions and by clicking a couple of buttons.

4.2 Source file

The file that the programmer writes for the D compiler to compile is called the *source file*. Since D is a compiled language, the source file itself is not an executable program. The source file must be converted to an executable program by the compiler.

As with any file, the source file must have a name. Although the name can be anything that is legal on the file system, it is customary to use the *.d file extension* for D source files because development environments, programming tools, and programmers all expect this to be the case. For example, **test.d**, **game.d**, **invoice.d**, etc. are appropriate D source file names.

4.3 Compiling the hello world program

Copy or type the hello world program above into a text file and save it under the name **hello.d**.

The compiler will soon check that the syntax of this source code is correct (i.e. it is valid according to the language rules) and make a program out of it by translating it into machine code. To compile the program follow these steps:

1. Open a console window.
2. Go to the directory where you saved **hello.d**.
3. Enter the following command. (Do not type the **\$** character; it is there to indicate the command line prompt.)

```
$ dmd hello.d
```

If you did not make any mistake, you may think that nothing have happened. To the contrary, it means that everything went well. There should be an executable file named **hello** (or **hello.exe** under Windows) that has just been created by the compiler.

If the compiler has instead printed some messages, you probably have made a mistake when copying the program code. Try to identify the mistake, correct it, and retry compiling. You will routinely make many mistakes when programming, so the process of correcting and compiling will become familiar to you.

Once the program has been created successfully, type the name of the executable program to run it. You should see that the program prints "Hello world!":

```
$ ./hello      ← running the program
Hello world!  ← the message that it prints
```

Congratulations! Your first D program works as expected.

4.4 Compiler switches

The compiler has many command line switches that are used for influencing how it compiles the program. To see a list of compiler switches enter only the name of the compiler:

```
$ dmd      ← enter only the name
DMD64 D Compiler v2.059
Copyright (c) 1999-2012 by Digital Mars written by Walter Bright
Documentation: http://www.dlang.org/index.html
Usage:
  dmd files.d ... { -switch }

  files.d      D source files
  ...
  -unittest    compile in unit tests
  ...
  -w           enable warnings
  ...
```

The abbreviated output above shows only the command line switches that I recommend that you always use. Although it makes no difference with the hello world program in this chapter, the following command line would compile the program by enabling warnings and unit tests. We will see these and other switches in more detail in later chapters:

```
$ dmd hello.d -w -unittest
```

The complete list of **dmd** command line switches can be found at [DMD Compiler documentation](http://www.dlang.org/index.html).

4.5 IDE

In addition to the compiler, you may also consider installing an IDE (integrated development environment). IDEs are designed to make program development easier by simplifying the steps of writing, compiling, and debugging.

If you do install an IDE, compiling and running the program will be as simple as pressing a key or clicking a button on the IDE. I still recommend that you familiarize yourself with compiling programs manually in a console window.

If you decide to install an IDE, go to the [Editor Support](#) page at Wiki4D to see a list of available IDEs.

4.6 Contents of the hello world program

Here is a quick list of the many D concepts that have appeared in this small program:

Core feature: Every language defines its syntax, fundamental types, keywords, rules, etc. All of these make the *core features* of that language. The parentheses, semicolons, words like **main** and **void**, are all according to the rules of D. These are similar to the rules of English: subject, verb, punctuation, sentence structure, etc.

Keyword: Special words that are a part of the core features of the language are *keywords*. There are two keywords in this program: **import**, which is used to introduce a module to the program; and **void**, which here means "not returning anything".

Library and function: The core features define only the structure of the language. They are used for defining functions and user types, and those in turn are used for building libraries. Libraries have a weak analogy in the English language: English has its special rules, sentence structure, etc., but it does not define the conversations that can be carried in that language. Scientific conversation, sport conversation, etc. are all within the rules of English, but they use it in special ways according to special conventions. Programming libraries are somewhat similar to conversations of natural languages.

writeln above is a *function* in D's standard *library*. It is used for printing a line of text, as its name suggests: write line.

Module: Libraries are introduced to D programs in the form of *modules*. The only module that this program uses is **std.stdio**, D's standard input and output module.

Character and string: Expressions like "Hello world!" are called *strings*, and the elements of strings are called *characters*. The only string in this program contains characters 'H', 'e', '!', and others.

Order of operations: Programs complete their tasks by executing operations in a certain order. These tasks start with the operations that are written in the function named **main**. The only operation in this program writes "Hello world!".

Significance of uppercase and lowercase letters: You can choose to type any character inside strings, but you must type the other characters exactly as they appear in the program. This is because lowercase vs. uppercase is significant in D programs. For example, **writeln** and **Writeln** are two different names.

We will cover all of the features of D in more detail in the following chapters.

4.7 Exercises

1. Make the program output something else.
2. Change the program to output more than one line.
3. Try to compile the program after making other changes; e.g. remove the semicolon at the end of the line with `writeln` and observe a compilation error.

... the solutions

5 writeln and write

In the previous chapter we have seen that **writeln** takes a string within parentheses and prints the string.

The parts of programs that actually do work are called *functions* and the information that they need to complete their work are called *parameters*. The act of giving such information to functions is called *passing parameter values* to them. Parameters are passed to functions within parentheses, separated by commas.

Note: *The word parameter describes the information that is passed to a function at the conceptual level. The concrete information that is actually passed during the execution of the program is called an argument. Although incorrect, these terms are sometimes used interchangeably in the software industry.*

writeln can take more than one argument. It prints them one after the other on the same line:

```
import std.stdio;

void main()
{
    writeln("Hello world!", "Hello fish!");
}
```

Sometimes, all of the information that is to be printed on the same line may not be readily available to be passed to **writeln**. In such cases, the first parts of the line may be printed by **write** and the last part of the line may be printed by **writeln**.

writeln advances to the next line, **write** stays on the same line:

```
import std.stdio;

void main()
{
    // Let's first print what we have available:
    write("Hello");

    // ... let's assume more operations at this point ...

    write("world!");

    // ... and finally:
    writeln();
}
```

Calling **writeln** without any parameter merely completes the current line.

Lines that start with *//* are called *comment lines* or briefly *comments*. A comment is not a part of the program code in the sense that it doesn't affect the behavior of the program. Its only purpose is to explain what the code does in that particular section of the program. The audience of a comment is anybody who may be reading the program code later, including the programmer who wrote the comment in the first place.

5.1 Exercises

- Both of the programs in this chapter print the strings without any spaces between them. Change the programs so that there is space between the arguments as in "Hello world!".
- Try calling **write** with more than one parameter as well.

... the solutions

6 Compiler

We have seen that the two tools that are used most in D programming are *the text editor* and *the compiler*. D programs are written in text editors.

The concept of compilation and the function of the compiler must also be understood when using *compiled* languages like D.

6.1 Machine code

The brain of the computer is the microprocessor (or the CPU, short for *central processing unit*). Telling the CPU what to do is called *coding*, and the instructions that are used when doing so are called *machine code*.

Most CPU architectures use machine code specific to that particular architecture. These machine code instructions are determined under hardware constraints during the design stage of the architecture. At the lowest level these machine code instructions are implemented as electrical signals. Because the ease of coding is not a primary consideration at this level, writing programs directly in the form of the machine code of the CPU is a very difficult task.

These machine code instructions are special numbers, which represent various operations supported by the CPU. For example, for an imaginary 8-bit CPU, the number 4 might represent the operation of loading, the number 5 might represent the operation of storing, and the number 6 might represent the operation of incrementing. Assuming that the leftmost 3 bits are the operation number and the rightmost 5 bits are the value that is used in that operation, a sample program in machine code for this CPU might look like the following:

Operation	Value	Meaning	
100	11110	LOAD	11110
101	10100	STORE	10100
110	10100	INCREMENT	10100
000	00000	PAUSE	

Being so close to hardware, machine code is not suitable for representing higher level concepts like *a playing card* or *a student record*.

6.2 Programming language

Programming languages are designed as efficient ways of programming a CPU, capable of representing higher-level concepts. Programming languages do not have to deal with hardware constraints; their main purposes are ease of use and expressiveness. Programming languages are easy for humans to understand, somewhat similar to natural languages:

```
if (a_card_has_been_played()) {
    display_the_card();
}
```

6.3 Compiled languages

In some programming languages, the program instructions must be compiled before becoming a program to be executed. Such languages produce fast executing programs but the development process involves two main steps: writing the program and compiling it. D is a compiled language.

6.4 Interpreted languages

Some programming languages don't require compilation. Such languages are called *interpreted languages*. The programs can be executed right from the hand-written program. Some examples of interpreted languages are Python, Ruby, and Perl. Because there is no compilation step, program development is easier using these languages. On the other hand, as the instructions of the program must be parsed to be interpreted every time the program is executed, programs that are written in these languages are slower than their equivalents written in compiled languages.

6.5 Compiler

The purpose of a compiler is translation: It translates the programming language into machine code. This is a translation from the language of the programmer to the language of the CPU. This translation is called *compilation*. Every compiler understands a particular programming language and is described as a compiler of that language, as in "a D compiler".

6.6 Compilation error

As the compiler compiles a program according to the rules of the language, it stops the compilation as soon as it comes across *illegal* instructions. Illegal instructions are the ones that are outside the specifications of the language. Problems like a mismatched parenthesis, a missing semicolon, a misspelled keyword, etc. all cause compilation errors.

The compiler may also emit a *compilation warning* when it sees a suspicious piece of code that may cause concern but not necessarily an error.

7 Fundamental Types

We have seen that the brain of a computer is the CPU. Most of the tasks of a program are performed by the CPU and the rest are dispatched to other parts of the computer.

The smallest unit of data in a computer is called *a bit*, which can have the value 0 or 1. At the hardware level, a bit is implemented by a few transistors.

Since a type of data that can hold only the values 0 and 1 would have very limited use, the CPU defines larger data types that are combinations of more than one bit. As an example, a *byte* consists of 8 bits. The most efficient data type of a CPU makes that CPU an *N-bit CPU*: as in 32-bit CPU, 64-bit CPU, etc.

The data types that the CPU defines are still not sufficient: they can't represent higher level concepts like *name of a student* or *a playing card*. D provides many useful data types, but even these types are not sufficient to represent many higher level concepts. Such concepts must be defined by the programmer as *structs* and *classes*, which we will see in later chapters.

D's *fundamental types* are very similar to the fundamental types of many other languages, as seen in the following table. The terms that appear in the table are explained below:

D's Fundamental Data Types

Type	Definition	Initial Value
bool	Boolean type	false
byte	signed 8 bits	0
ubyte	unsigned 8 bits	0
short	signed 16 bits	0
ushort	unsigned 16 bits	0
int	signed 32 bits	0
uint	unsigned 32 bits	0
long	signed 64 bits	0L
ulong	unsigned 64 bits	0L
float	32-bit floating point	float.nan
double	64-bit floating point	double.nan
real	either the largest floating point type that the hardware supports, or double; whichever is larger	real.nan
ifloat	imaginary value type of float	float.nan * 1.0i
idouble	imaginary value type of double	double.nan * 1.0i
ireal	imaginary value type of real	real.nan * 1.0i
cfloat	complex number type made of two floats	float.nan + float.nan * 1.0i
cdouble	complex number type made of two doubles	double.nan + double.nan * 1.0i
creal	complex number type made of two reals	real.nan + real.nan * 1.0i
char	UTF-8 code unit	0xFF
wchar	UTF-16 code unit	0xFFFF
dchar	UTF-32 code unit and Unicode code point	0x0000FFFF

In addition to the above, the keyword **void** represents *having no type*. The keywords **cent** and **ucent** are reserved for future use to represent signed and unsigned 128 bit values.

Unless there is a specific reason not to, you can use **int** to represent whole values. To represent concepts that can have fractional values, consider **double**.

The following are the terms that appeared in the table:

Boolean: The type of logical expressions, having the value **true** for truth and **false** for falsity.

Signed type: A type that can have negative and positive values. For example **byte** can have values from -128 to 127. The names of these types come from the negative *sign*.

Unsigned type: A type that can have only positive values. For example **ubyte** can have values from 0 to 255. The **u** at the beginning of the name of these types comes from *unsigned*.

Floating point: The type that can represent values with decimal fractions as in 1.25. The precision of floating point calculations are directly related to the bit count of the type: higher the bit count, more precise the results.

Only floating point types can represent decimal fractions; integer types like **int** can only represent whole values like 1 and 2.

Complex number type: The type that can represent the complex numbers of mathematics.

Imaginary number type: The type that represents only the imaginary part of complex numbers. The **i** that appears in the Initial Value column is the square root of -1 in mathematics.

nan: Short for "not a number", representing *invalid floating point value*.

7.1 Attributes of types

D types have *properties*. Properties are accessed with a dot after the name of the type. For example the **sizeof** property of **int** is accessed as **int.sizeof**. We will see only four of these attributes in this chapter:

- **stringof** is the name of the type
- **sizeof** is the length of the type in terms of bytes. (In order to determine the bit count, this value must be multiplied by 8, the number of bits in a byte.)
- **min** is short for "minimum"; this is the smallest value that the type can have
- **max** is short for "maximum"; this is the largest value that the type can have

Here is a program that prints these properties for **int**:

```
import std.stdio;

void main()
{
    writeln("Type           : ", int.stringof);
    writeln("Length in bytes: ", int.sizeof);
    writeln("Minimum value  : ", int.min);
    writeln("Maximum value  : ", int.max);
}
```

7.2 size_t

You will come across the `size_t` type as well. `size_t` is not a separate type but an alias of an existing unsigned type, e.g. `ulong`. Its name comes from "size type". It is the most suitable type to represent concepts like *size* or *count*.

You can use the `.stringof` property to see what `size_t` is an alias of on your system:

```
import std.stdio;

void main()
{
    writeln(size_t.stringof);
}
```

The output of the program is the following on my system:

```
ulong
```

7.3 Exercise

- Print the properties of other types.
Note: You can't use the reserved types `cent` and `ucent` in any program; and as an exception, `void` does not have the properties `.min` and `.max`.

[... the solution](#)

8 Assignment and Order of Evaluation

The first two difficulties that most students face when learning to program involve the assignment operation and the order of evaluation.

8.1 The assignment operation

You will see lines similar to the following in almost every program in almost every programming language:

```
a = 10;
```

The meaning of that line is "make **a**'s value become 10". Similarly, the following line means "make **b**'s value become 20":

```
b = 20;
```

Based on that information, what can be said about the following line?

```
a = b;
```

Unfortunately, that line is not about the equality concept of mathematics that I assume we all know. The expression above **does not** mean "a is equal to b"! When we apply the same logic from the earlier two lines, the expression above must mean "make **a**'s value become **b**". Making **a**'s value become **b** further means "make **a**'s value become the same as **b**'s value".

Do you see the difference? The very well known = symbol of mathematics has a completely different meaning in programming: make the left side's value the same as the right side's value.

8.2 Order of evaluation

The operations of a program are applied step by step in a specific order. We may see the previous three expressions in a program in that order:

```
a = 10;
b = 20;
a = b;
```

The meaning of those three lines altogether is this: "make **a**'s value become 10, *then* make **b**'s value become 20, *then* make **a**'s value become the same as **b**'s value". Accordingly, after those three operations are performed, both **a**'s and **b**'s values would be 20.

8.3 Exercise

Observe that the following three operations swap the values of **a** and **b**. If at the beginning their values are 1 and 2 respectively, after the operations the values become 2 and 1:

```
c = a;
a = b;
b = c;
```

... *the solution*

9 Variables

Concrete concepts that are represented in a program are called *variables*. A value like *air temperature* or a more complicated object like *a car engine* can be variables of a program.

Every variable is of a certain type and has a certain value. Most variables have names as well, but some variables are anonymous.

As an example of a variable, we can think of the concept of *the number of students* at a school. Since the number of students is a whole number, **int** is a suitable type, and **studentCount** would be a sufficiently descriptive name.

According to D's syntax rules, a variable is introduced by its type followed by its name. Introducing a variable to the program is called its *definition*. Once a variable is defined, its name starts representing its value.

```
import std.stdio;

void main()
{
    // The definition of the variable; this definition
    // specifies that the type of studentCount is int:
    int studentCount;

    // The name of the variable becomes its value:
    writeln("There are ", studentCount, " students.");
}
```

The output of this program is the following:

```
There are 0 students.
```

As seen from that output, the value of **studentCount** is 0. This is according to the fundamental types table from the previous chapter: the initial value of **int** is 0.

Note that **studentCount** does not appear in the output as its name. In other words, the output of the program is not "There are studentCount students".

The values of variables are changed by the = operator. The = operator assigns new values to variables, and for that reason is called the *assignment operator*:

```
import std.stdio;

void main()
{
    int studentCount;
    writeln("There are ", studentCount, " students.");

    // Assigning the value 200 to the studentCount variable:
    studentCount = 200;
    writeln("There are now ", studentCount, " students.");
}
```

```
There are 0 students.
There are now 200 students.
```

When the value of a variable is known at the time of the variable's definition, the variable can be defined and assigned at the same time. This is an important guideline; it makes it impossible to use a variable before assigning its intended value:

```
import std.stdio;

void main()
{
    // Definition and assignment at the same time:
    int studentCount = 100;

    writeln("There are ", studentCount, " students.");
}
```

```
There are 100 students.
```

9.1 Exercise

- Define two variables to print "I have exchanged 20 Euros at the rate of 2.11". You can use **double** for the floating point value.

... the solution

10 Standard Input and Output Streams

So far, the printed output of our programs has been appearing on the *console window* (or *screen*). Although this is actually the case with many uses of many programs, in reality the characters are printed to *standard output streams* of programs.

The standard output is character based; everything to be printed are first converted to their character representations and then sent to the output one by one as characters. For example, the integer value 100 that we've printed in the last chapter is not sent to the output as the value 100, but as the three characters **1**, **0**, and **0**.

Similarly, what we normally perceive as the *keyboard* is actually the *standard input stream* of a program and is also character based. The information always comes as characters to be converted to data. For example the integer value 42 actually comes through the standard input as the characters **4** and **2**.

These conversions happen automatically.

This concept of consecutive characters is called a *character stream*. As D's standard input and standard output fit this description, they are character streams.

The names of the standard input and output streams in D are **stdin** and **stdout**, respectively.

Operations on these streams normally require the name of the stream, a dot, and the operation; as in **stream.operation()**. Because **stdin** and **stdout** are used very commonly, as a convenience, standard operations on them can be called without the need of the name and the dot, as in **operation()**.

writeln that we've been using in the previous chapters is actually the short for **stdout.writeln**. Similarly, **write** is the short for **stdout.write**. Accordingly, the *hello world* program can also be written as the following:

```
import std.stdio;

void main()
{
    stdout.writeln("Hello world!");
}
```

10.1 Exercise

- Observe that **stdout.write** works the same as **write**.

... the solution

11 Reading from the Standard Input

Any data that is read by the program must first be stored in a variable. For example, a program that reads the number of students from the input must store this information in a variable. The type of this specific variable can be `int`.

As we've seen in the previous chapter, we don't need to type `stdout` when printing to the output, because it is implied. Further, what is to be printed is specified as the argument. So, `write(studentCount)` is sufficient to print the value of `studentCount`. To summarize:

```
stream:    stdout
operation: write
data:     the value of the studentCount variable
target:    commonly the console window
```

The reverse of `write` is `readf`; it reads from the standard input. The "f" in its name comes from "formatted" as what it reads must always be presented in a certain format.

We've also seen in the previous chapter that the standard input stream is `stdin`.

In the case of reading, one piece of the puzzle is still missing: where to store the data. To summarize:

```
stream:    stdin
operation: readf
data:     some information
target:    ?
```

The location of where to store the data is specified by the address of a variable. In D, the `&` character that is typed before a name is the address of what that name represents. For example, the address of `studentCount` is `&studentCount`. Here, `&studentCount` can be read as "the address of `studentCount`" and is the missing piece to replace the question mark above:

```
stream:    stdin
operation: readf
data:     some information
target:    the location of the studentCount variable
```

Typing a `&` in front of a name means *pointing* at what that name represents. This concept is the foundation of references and pointers that we will see in later chapters.

I will leave one peculiarity about the use of `readf` for later; for now, let's accept as a rule that the first argument to `readf` must be `"%s"`:

```
readf("%s", &studentCount);
```

Note: As I explain below, in most cases there must also be a space: `" %s"`.

`"%s"` indicates that the data should automatically be converted in a way that is suitable to the type of the variable. For example, when the '4' and '2' characters are read to a variable of type `int`, they would be converted as the integer value 42.

The program below asks the user to enter the number of students. You must press the Enter key after typing the input:

```
import std.stdio;

void main()
```

```

{
    write("How many students are there? ");

    /*
     * The definition of the variable that will be used to
     * store the information that is read from the input.
     */
    int studentCount;

    // Storing the input data to that variable
    readf("%s", &studentCount);

    writeln("Got it: There are ", studentCount, " students.");
}

```

11.1 Skipping the whitespace characters

Even the Enter key that we press after typing the data is stored as a special code and is placed into the `stdin` stream. This is useful to the programs to detect whether the information has been input on a single line or multiple lines.

Although sometimes useful, such special codes are mostly not important for the program and must be filtered out from the input. Otherwise they *block* the input and prevent reading other data.

To see this *problem* in a program, let's read also the number of teachers from the input:

```

import std.stdio;

void main()
{
    write("How many students are there? ");
    int studentCount;
    readf("%s", &studentCount);

    write("How many teachers are there? ");
    int teacherCount;
    readf("%s", &teacherCount);

    writeln("Got it: There are ", studentCount, " students",
           " and ", teacherCount, " teachers.");
}

```

Unfortunately, now the program gets *stuck* when reading the second `int`:

```

How many students are there? 100
How many teachers are there? 20
← The program gets stuck here

```

Although the user enters the number of teachers as 20, the special code(s) that represents the Enter key that has been pressed when entering the previous 100 is still in the input stream and is blocking it. The characters that appeared in the input stream are similar to the following:

```
100[EnterCode]20[EnterCode]
```

I have shown the Enter code that is blocking the input in red.

The solution is to use a space character before `%s` to indicate that the Enter code that appears before reading the number of teachers is not important: " `%s`". Spaces that are in the format strings are used to read and ignore zero or more invisible characters that would

otherwise appear in the input. Such characters include the actual space character, the code(s) that represent the Enter key, the Tab character, etc. and are called the *whitespace characters*.

As a general rule, you can use "`%s`" for every data that is read from the input. The program above works as expected with the following changes:

```
// ...
    readf(" %s", &studentCount);
// ...
    readf(" %s", &teacherCount);
// ...
```

The output:

```
How many students are there? 100
How many teachers are there? 20
Got it: There are 100 students and 20 teachers.
```

11.2 Additional information

- Lines that start with `//` are useful for single lines of comments. To write multiple lines as a single comment, enclose the lines within `/*` and `*/` markers. In order to be able to comment even other comments, use `/*+` and `+*/`:

```
/*+
// A single line of comment

/*
A comment that spans
multiple lines
*/

A comment block that includes other comments
+*/
```

- Most of the whitespace in the source code are insignificant. It is good practice to write longer expressions as multiple lines or add extra whitespace to make the code more readable. Still, as long as the syntax rules of the language are observed, the programs can be written without any extra whitespace:

```
import std.stdio;void main(){writeln("Hard to read!");}
```

It is hard to read source code that has so few whitespace.

11.3 Exercises

1. Let the user enter the Euro-USD exchange rate (you can assume 1 euro is 1.4 dollars) and the amount of Euros. Then print the equivalent number of USD to the output.
2. Enter non-numeral characters when the program is expecting integer values and observe that the program does not work correctly.

[... the solutions](#)

12 Logical Expressions

Actual work that a program performs are accomplished by *expressions*. Any part of a program that produces a value or a side effect is called an expression. It has a very wide definition, because even a constant value like **42** and a string like **"hello"** are expressions because they produce the respective constant values 42 and "hello".

Note: *Don't confuse producing a value with defining a variable. Values need not be associated with variables.*

Function calls like **writeln** are expressions as well, because they have side effects. In the case of **writeln**, the effect is on the output stream by the placement of characters on it. Another example from the programs that we have written so far would be the assignment operation, which effects the variable on its left-hand side.

Because of producing values, expressions can take part in other expressions. This allows us to form more complex expressions from simpler ones. For example, assuming that there is a function named **currentTemperature()** that produces the value of the current air temperature, the value that it produces may directly be used in a **writeln** expression:

```
writeln("It's ", currentTemperature(),
       " degrees at the moment.");
```

That line consists of four expressions:

1. **"It's "**
2. **currentTemperature()**
3. **" degrees at the moment."**
4. the **writeln** expression that makes use of the other three

In this chapter we will cover the particular type of expression that is used in conditional statements.

Before going further though, I would like to repeat the assignment operator once more, this time emphasizing the two expressions that appear on its left and right sides: the assignment operator (=) assigns the value of the expression on its right-hand side to the expression on its left-hand side (e.g. to a variable).

```
temperature = 23 // temperature's value becomes 23
```

12.1 Logical Expressions

Logical expressions are the expressions that are used in Boolean arithmetic. Logical expressions are what makes computer programs make decisions like "if the answer is yes, I will save the file".

Logical expressions can have one of only two values: **false** that indicates falsity, and **true** that indicates truth. For example, the value of the expression "Mediterranean is a sea" would be **true**, and the value of the expression "Mediterranean is a lake" would be **false**.

I will use **writeln** expressions in the following examples. If a line has **true** printed at the end, it will mean that what is printed on the line is true. Similarly, **false** will mean that what is on the line is false. For example, if the output of a program is the following,

```
There is coffee: true
```


then it will mean that "there is coffee". Similarly,

There is coffee: false

will mean that "there isn't coffee". Note that the fact that "is" appears on the left-hand side does not mean that coffee exists. I use the "... is ...: false" construct to mean "is not" or "is false".

Logical expressions are used extensively in *conditional statements, loops, function parameters*, etc. It is essential to understand how they work. Luckily, logical expressions are very easy to explain and use.

The logical operators that are used in logical expressions are the following:

- The `==` operator answers the question "is equal to?". Compares the two expressions on its left and right sides and produces **true** if they are equal and **false** if they are not. By definition, the value that `==` produces is a logical expression.

As an example, let's assume that we have the following two variables:

```
int daysInWeek = 7;
int monthsInYear = 12;
```

The following are two logical expressions that use those values:

```
daysInWeek == 7           // true
monthsInYear == 11        // false
```

- The `!=` operator answers the question "is not equal to?" Compares the two expressions on its sides and produces the opposite of `==`.

```
daysInWeek != 7           // false
monthsInYear != 11        // true
```

- The `||` operator means "or", and produces **true** if any one of the logical expressions is true.

If the value of the left-hand expression is **true**, it produces **true** without even looking at the expression that is on the right-hand side. If the left-hand side is **false**, then it produces the value of the right-hand side. This operator is similar to the "or" in English: if the left one, the right one, or both are **true**, then it produces **true**.

All of the possible values for both sides of this operator and its result are as in the following table:

Left expression	Right expression	Result
false	false	false
false	true	true
true	false (not evaluated)	true
true	true (not evaluated)	true

```
import std.stdio;

void main()
{
    // false means "no", true means "yes"

    bool existsCoffee = false;
    bool existsTea = true;
```

```
writeln("There is warm drink: ",
        existsCoffee || existsTea);
}
```

Because at least one of the two expressions is **true**, the logical expression above produces **true**.

- The **&&** operator means "and", and produces **true** if both of the expressions are true. If the value of the left-hand expression is **false**, it produces **false** without even looking at the expression that is on the right-hand side. If the left-hand side is **true**, then it produces the value of the right-hand side. This operator is similar to the "and" in English: if the left value and the right value are **true**, then it produces **true**.

Left expression	Right expression	Result
false	false (not evaluated)	false
false	true (not evaluated)	false
true	false	false
true	true	true

```
writeln("I will drink coffee: ",
        wantToDrinkCoffee && existsCoffee);
```

Note: The fact that the **||** and **&&** operators may not evaluate the right-hand expression is called their shortcut behavior. The only other operator that has this behavior is the ternary operator **?:**, which we will see in a later chapter. All of the other operators always evaluate and use all of their expressions.

- The **^** operator answers the question "is one or the other?" This operator produces **true** if only one expression is **true**, but not both.

Left expression	Right expression	Result
false	false	false
false	true	true
true	false	true
true	true	false

For example, the logic that represents my playing chess if *only one* of my two friends shows up can be coded like this:

```
writeln("I will play chess: ", jimShowedUp ^ bobShowedUp);
```

- The **<** operator answers the question "is less than?" (or "does come before in sort order?").

```
writeln("We beat: ", theirScore < ourScore);
```

- The **>** operator answers the question "is greater than?" (or "does come after in sort order?").

```
writeln("They beat: ", theirScore > ourScore);
```

- The **<=** operator answers the question "is less than or equal to?" (or "does come before or the same in sort order?"). This operator is the opposite of the **>** operator.

```
writeln("We were not beaten: ", theirScore <= ourScore);
```

- The `>=` operator answers the question "is greater than or equal to?" (or "does come after or the same in sort order?"). This operator is the opposite of the `<` operator.

```
writeln("We did not beat: ", theirScore >= ourScore);
```

- The `!` operator means "the opposite of". Different from the other logical operators, it takes just one expression and produces **true** if that expression is **false**, and **false** if that expression is **true**.

```
writeln("I will walk: ", !existsBicycle);
```

12.2 Grouping expressions

The order in which the expressions are evaluated can be specified by using parentheses to group them. When parenthesized expressions appear in more complex expressions, the parenthesized expressions are evaluated before they can be used in the expressions that they appear in. For example, the expression "if there is coffee or tea, and also cookie or scone; then I am happy" can be coded like the following:

```
writeln("I am happy: ",
(existsCoffee || existsTea) && (existsCookie || existsScone));
```

If the sub expressions were not parenthesized, the expressions would be evaluated according to *operator precedence* rules of D (which have been inherited from the C language). Since in these rules `&&` has a higher precedence than `||`, writing the expression without parentheses would not be evaluated as intended:

```
writeln("I am happy: ",
existsCoffee || existsTea && existsCookie || existsScone);
```

The `&&` operator would be evaluated first and the whole expression would be the semantic equivalent of the following expression:

```
writeln("I am happy: ",
existsCoffee || (existsTea && existsCookie) || existsScone);
```

That has a totally different meaning: "if there is coffee, or tea and cookie, or scone; then I am happy".

12.3 Reading bool input

All of the **bool** values above are automatically printed as "false" or "true". It is not the case in the opposite direction: the strings "false" and "true" are not automatically read as the values **false** and **true**. For that reason, the input must be first read as a string and then be converted to a **bool** value.

Since one of the exercises below requires you to enter "false" and "true" from the standard input, I have been forced to use D features that I haven't explained to you yet. I will have to define a function below that will convert the input string to a **bool** value. This function will achieve its task by calling `to`, which is defined in the module `std.conv`. (You may see **ConvException** errors if you enter anything other than "false" or "true".)

I am hoping that all of the code that are in the **main** functions of the following programs are clear at this point. **read_bool()** is the function that contains new features. Although I have inserted comments to explain what it does, you can ignore that function for now. Still, it must be a part of the source code for the program to compile and work correctly.

12.4 Exercises

1. We've seen above that the **<** and the **>** operators are used to determine whether a value is less than or greater than another value; but there is no operator that answers the question "is between?" to determine whether a value is between two other values.

Let's assume that a programmer has written the following code to determine whether **value** is between 10 and 20. Observe that the program cannot be compiled as written:

```
import std.stdio;

void main()
{
    int value = 15;

    writeln("Is between: ",
           10 < value < 20);    // ← compilation ERROR
}
```

Try using parentheses around the whole expression:

```
writeln("Is between: ",
       (10 < value < 20));    // ← compilation ERROR
```

Observe that it still cannot be compiled.

2. While searching for a solution to this problem, the same programmer discovers that the following use of parentheses now enables the code to be compiled:

```
writeln("Is between: ",
       (10 < value) < 20);    // ← compiles but WRONG
```

Observe that the program now works as expected and prints "true". Unfortunately that output is misleading because the program has a bug. To see the effect of that bug, replace 15 with a value greater than 20:

```
int value = 21;
```

Observe that the program still prints "true" even though 21 is not less than 20.

Hint: Remember that the type of a logical expression is **bool**. It shouldn't make sense whether a **bool** value is less than 20.

3. The logical expression that answers the question "is between?" must instead be coded like this: "is greater than the lower value and less than the upper value?"
Change the expression in the program according to that logic and observe that it now prints "true" as expected. Additionally, test that the logical expression works correctly for other values as well: for example, when **value** is 50 or 1, the program should print "false"; and when it's 12, the program should print "true".
4. Let's assume that we can go to the beach when one of the following conditions is true:
 - If the distance to the beach is less than 10 and there is a bicycle for everyone

- If there is less than 6 of us, and we have a car, and one of us has a driver license

As written, the following program always prints "true". Construct a logical expression that will print "true" when one of the conditions above is true. (When trying the program, enter "false" or "true" for questions that start with "Is there a".). Don't forget to include the `read_bool()` function when testing the following program:

```
import std.stdio;
import std.conv;
import std.string;

void main()
{
    write("How many are we? ");
    int personCount;
    readf(" %s", &personCount);

    write("How many bicycles are there? ");
    int bicycleCount;
    readf(" %s", &bicycleCount);

    write("What is the distance to the beach? ");
    int distance;
    readf(" %s", &distance);

    bool existsCar = read_bool("Is there a car? ");
    bool existsLicense =
        read_bool("Is there a driver license? ");

    /*
     * Replace the 'true' below with a logical expression that
     * produces the value 'true' when one of the conditions
     * listed in the question is satisfied:
     */
    writeln("We are going to the beach: ", true);
}

/*
 * Please note that this function includes features that will
 * be explained in later chapters.
 */
bool read_bool(string message)
{
    // Print the message
    write(message, "(false or true) ");

    // Read the line as a string
    string input;
    while (input.length == 0) {
        input =.chomp(readln());
    }

    // Produce a 'bool' value from that string
    bool result = to!bool(input);

    // Return the result to the caller
    return result;
}
```

Enter various values and test that the logical expression that you wrote works correctly.

... the solutions

13 if Statement

We've learned that actual work in a program are performed by expressions. All of the expressions of all of the programs that we've seen so far have started with the `main` function and were executed until the end of `main`.

Statements on the other hand, are features that affect the execution of expressions. Statements don't produce values and don't have side effects themselves. They determine whether and in what order the expressions are executed. Statements sometimes use logical expressions when making such decisions.

Note: *Other programming languages may have different definitions for expression and statement, while some others may not have a distinction at all.*

13.1 The if block and its scope

The `if` statement determines whether one or more expressions would be executed. It makes this decision by evaluating a logical expression. It has the same meaning as the English word "if" as in the phrase "if there is coffee then I will drink coffee".

`if` takes a logical expression in parenthesis. If the value of that logical expression is `true`, then it executes the expressions that are within the following curly brackets. Conversely, if the logical expression is `false`, it does not execute the expressions within the curly brackets.

The area within the curly brackets is called a *scope* and all of the code that is in that scope are called a *block of code*.

Here is the syntax of the `if` statement:

```
if (a_logical_expression)
{
    // ... expression(s) to execute if true
}
```

For example, the program construct that represents "if there is coffee then drink coffee and wash the cup" can be written as in the following program:

```
import std.stdio;

void main()
{
    bool existsCoffee = true;

    if (existsCoffee) {
        writeln("Drink coffee");
        writeln("Wash the cup");
    }
}
```

If the value of `existsCoffee` is `false`, then the expressions that are within the block would be skipped and the program would not print anything.

Also note that it is a common D coding style to type the opening curly brackets on the same line as the `if` keyword.

13.2 The else block and its scope

Sometimes there are operations to execute for when the logical expression of the **if** statement is **false**. For example, there is always an operation to execute in a decision like "if there is coffee I will drink coffee, else I will drink tea".

The operations to execute in the **false** case are placed in a scope after the **else** keyword:

```
if (a_logical_expression)
{
    // ... expression(s) to execute if true
}
else
{
    // ... expression(s) to execute if false
}
```

For example, under the assumption that there is always tea:

```
if (existsCoffee) {
    writeln("Drink coffee");
} else {
    writeln("Drink tea");
}
```

In that example, either the first or the second string would be printed depending on the value of **existsCoffee**.

Also note the placement of the curly brackets around **else**; that is a common D coding style as well.

else itself is not a statement but an optional *clause* of the **if** statement; it cannot be used alone.

13.3 Always use the scope curly brackets

It is not recommended but is actually possible to omit the curly brackets if there is only one statement within a scope. As both the **if** and the **else** scopes have just one statement above, that code can also be written as the following:

```
if (existsCoffee)
    writeln("Drink coffee");
else
    writeln("Drink tea");
```

Most experienced programmers use curly brackets even for single statements. (One of the exercises of this chapter is about omitting them.) The reason for me to stress the guideline of always using curly brackets is because I will now show the only case where omitting the curly brackets is actually better.

13.4 The "if, else if, else" chain

One of the powers of statements and expressions is the ability to use them in more complex ways. In addition to expressions, scopes can contain other statements. For example, an **else** scope can contain an **if** statement. Connecting statements and expressions in different ways allows us to make programs behave intelligently according to their purposes.

The following is a more complex code written under the agreement that riding to a good coffee shop is preferred over walking to a bad one:

```

if (existsCoffee) {
    writeln("Drink coffee at home");
} else {
    if (existsBicycle) {
        writeln("Ride to the good place");
    } else {
        writeln("Walk to the bad place");
    }
}

```

The code above represents the phrase "if there is coffee, drink at home; else if there is a bicycle, ride to the good place; otherwise walk to the bad place".

Let's complicate this decision process further: instead of having to walk to the bad place, let's first try the neighbor:

```

if (existsCoffee) {
    writeln("Drink coffee at home");
} else {
    if (existsBicycle) {
        writeln("Ride to the good place");
    } else {
        if (neighborIsHome) {
            writeln("Have coffee at the neighbor");
        } else {
            writeln("Walk to the bad place");
        }
    }
}

```

Such decisions like "if this case, else if that other case, else if that even other case, etc." are common in programs. Unfortunately, when the guideline of always using curly brackets is followed obstinately, the code ends up having too much horizontal and vertical space: ignoring the empty lines, the 3 **if** statements and the 4 **writeln** expressions above occupy a total of 13 lines.

In order to write such constructs in a more compact way, when an **else** scope contains only an **if** statement, then the curly brackets of that **else** scope are omitted as an exception of this guideline.

I am leaving the following code untidy as an intermediate step before showing the better form of it. No code should be written in such an untidy way. (I am also leaving it black and white to indicate that it is unacceptable as written.)

The following is what the code looks like after removing the curly brackets of the two **else** scopes that contain just a single **if** statement:


```

if (existsCoffee) {
    writeln("Drink coffee at home");
} else
    if (existsBicycle) {
        writeln("Ride to the good place");
    } else
        if (neighborIsHome) {
            writeln("Have coffee at the neighbor");
        } else {
            writeln("Walk to the bad place");
        }
}

```

If we now move those **if** statements up to the same lines as their enclosing **else** clauses and tidy up the code, we end up with the following more readable format:

```

if (existsCoffee) {
    writeln("Drink coffee at home");
} else if (existsBicycle) {
    writeln("Ride to the good place");
} else if (neighborIsHome) {
    writeln("Have coffee at the neighbor");
} else {
    writeln("Walk to the bad place");
}

```

Removing the curly brackets allows the code to be more compact, and lines up all of the expressions for easier readability. The logical expressions, the order that they are evaluated, and the operations that are executed when they are true are now easier to see at a glance.

This common programming construct is called the "if, else if, else" chain.

13.5 Exercises

1. Since the logical expression below is **true**, we would expect this program to *drink lemonade and wash the cup*:

```

import std.stdio;

void main()
{
    bool existsLemonade = true;

    if (existsLemonade) {
        writeln("Drinking lemonade");
        writeln("Washing the cup");
    } else
        writeln("Eating pie");
        writeln("Washing the plate");
}

```

But when you run that program you will see that it *washes the plate* as well:

```

Drinking lemonade
Washing the cup
Washing the plate

```

Why? Correct the program to wash the plate only when the logical expression is **false**.

- Write a program that plays a game with the user (obviously with trust). The user throws a die and enters its value. Either the user or the program wins according to the value of the die:

Value of the die	Output of the program
1	You won
2	You won
3	You won
4	I won
5	I won
6	I won
Any other value	ERROR: Invalid value

Bonus: Have the program also mention the value when the value is invalid. For example:

```

ERROR: 7 is invalid

```

- Let's change the game by having the user enter a value from 1 to 1000. Now the user wins when the value is in the range 1-500 and the computer wins when the value is in the range 501-1000. Can the previous program be easily modified to work in this way?

[... the solutions](#)

14 while Loop

The **while** loop is similar to the **if** statement and essentially works as a repeated **if** statement. Just like **if**, **while** also takes a logical expression and evaluates the block when the logical expression is **true**. The difference is that the **while** statement evaluates the logical expression and executes the expressions in the block again, as long as the logical expression is **true**, not just once. Repeating a block of code this way is called *looping*.

Here is the syntax of the **while** statement:

```
while (a_logical_expression)
{
    // ... expression(s) to execute while true
}
```

For example, the code that represents *eat cookies as long as there is cookie* can be coded like this:

```
import std.stdio;
void main()
{
    bool existsCookie = true;
    while (existsCookie) {
        writeln("Take cookie");
        writeln("Eat cookie");
    }
}
```

That program would continue repeating the loop because the value of **existsCookie** never changes from **true**.

while is useful when the value of the logical expression changes during the execution of the program. To see this, let's write a program that takes a number from the user as long as that number is zero or greater. Remember that the initial value of **int** variables is 0:

```
import std.stdio;
void main()
{
    int number;
    while (number >= 0) {
        write("Please enter a number: ");
        readf(" %s", &number);
        writeln("Thank you for ", number);
    }
    writeln("Exited the loop");
}
```

The program thanks for the provided number and exits the loop only when the number is less than zero.

14.1 The continue statement

This statement starts the next iteration of the loop right away, instead of executing the rest of the expressions of the block.

Let's modify the program above to be a little picky: instead of thanking for any number, let's not accept 13. The following program does not thank for 13 because in that case the **continue** statement makes the program go to the beginning of the loop to evaluate the logical expression again:

```
import std.stdio;

void main()
{
    int number;

    while (number >= 0) {
        write("Please enter a number: ");
        readf(" %s", &number);

        if (number == 13) {
            writeln("Sorry, not accepting that one...");
            continue;
        }

        writeln("Thank you for ", number);
    }

    writeln("Exited the loop");
}
```

We can define the behavior of that program as *take numbers as long as they are greater than or equal to 0 but skip 13*.

14.2 The break statement

Sometimes it becomes obvious that there is no need to stay in the **while** loop any longer. **break** allows the program to exit the loop right away. The following program exits the loop as soon as it finds a special number:

```
import std.stdio;

void main()
{
    int number;

    while (number >= 0) {
        write("Please enter a number: ");
        readf(" %s", &number);

        if (number == 42) {
            writeln("FOUND IT!");
            break;
        }

        writeln("Thank you for ", number);
    }

    writeln("Exited the loop");
}
```

We can summarize this behavior as *take numbers as long as they are greater than or equal to 0 until a number is 42*.

14.3 Infinite loop

Sometimes the logical expression is intentionally made a constant **true**. The **break** statement is a common way of exiting such *infinite loops*.

The following program prints a menu in an infinite loop; the only way of exiting the loop is a **break** statement:

```
import std.stdio;
void main()
{
    /* Infinite loop, because the logical expression is always
     * true */
    while (true) {
        write("0:Exit, 1:Turkish, 2:English - Your choice? ");

        int choice;
        readf(" %s", &choice);

        if (choice == 0) {
            writeln("See you later...");
            break; // The only exit of this loop
        } else if (choice == 1) {
            writeln("Merhaba!");
        } else if (choice == 2) {
            writeln("Hello!");
        } else {
            writeln("I don't know that... :/");
        }
    }
}
```

Note: Exceptions can terminate an infinite loop as well. We will see exceptions in a later chapter.

14.4 Exercises

1. The following program is designed to stay in the loop as long as the input is 3, but there is a bug: it doesn't ask for any input:

```
import std.stdio;
void main()
{
    int number;

    while (number == 3) {
        write("Number? ");
        readf(" %s", &number);
    }
}
```

Fix the bug. The program should stay in the loop as long as the input is 3.

2. Make the computer help Anna and Bill play a game. First, the computer should take a number from Anna in the range from 1 to 10. The program should not accept any other number and ask again.

Once the program takes a valid number from Anna, it should start taking numbers from Bill until he guesses Anna's number correctly.

Note: The numbers that Anna enters obviously stays on the console and can be seen by Bill. :o) Let's ignore this fact and write the program as an exercise of the **while** statement.

... the solutions

15 Integers and Arithmetic Operations

We have seen that the **if** and **while** statements allow programs to make decisions by using the **bool** type in the form of logical expressions. In this chapter, we will see arithmetic operations on the *integer* types of D. These features will allow us to write much more useful programs.

Although arithmetic operations are a part of our daily lives and are actually simple, there are very important concepts that a programmer must be aware of in order to produce correct programs: *bit length of a type*, *overflow*, *underflow*, and *truncation*.

Before going further, I would like to summarize the arithmetic operations in the following table as a reference:

Operator	Effect	Sample
++	increments by one	++variable
--	decrements by one	--variable
+	the result of adding two values	first + second
-	the result of subtracting 'second' from 'first'	first - second
*	the result of multiplying two values	first * second
/	the result of dividing 'first' by 'second'	first / second
%	the remainder of dividing 'first' by 'second'	first % second
^^	the result of raising 'first' to the power of 'second' (multiplying 'first' by itself 'second' times)	first ^^ second

Most of those operators have counterparts that have an = sign attached: +=, -=, *=, /=, %=, and ^=. The difference with these operators is that they assign the result to the left-hand side:

```
variable += 10;
```

That expression adds the value of **variable** and 10 and assigns the result to **variable**. In the end, the value of **variable** would be increased by 10. It is the equivalent of the following expression:

```
variable = variable + 10;
```

I would like also to summarize two important concepts here before elaborating on them below.

Overflow: Not all values can fit in a type, in which case *the value overflows*. For example, a variable of type **ubyte** can have values only in the range of 0 to 255; so when assigned 260, the value overflows and becomes 4.

Underflow: Similarly, values may not be less than the minimum value that a type can have.

Truncation: Integer types cannot have values with fractional parts. For example, the value of the **int** expression $3/2$ is 1, not 1.5.

15.1 More Information

We encounter arithmetic operations daily without much surprises: if a bagel is \$1, two bagels are \$2; if four sandwiches are \$15, one sandwich is \$3.75, etc.

Unfortunately things are not as simple with arithmetic operations in computers. If we don't understand how values are stored in a computer, we may be surprised to see that a company's debt is *reduced* to \$1.7 billions when it borrows \$3 billions more on top of its existing debt of \$3 billions! Or when a box of ice cream serves 4 kids, an arithmetic operation may claim that 2 boxes would be sufficient for 11 kids!

Programmers must understand how integers are stored in computers.

15.1.1 Integer types

Integer types are the types that can have only whole values like -2, 0, 10, etc. These types cannot have fractional parts as in 2.5. All of the integer types that we have seen in the [Fundamental Types chapter](#) are the following:

Type	Number of Bits	Initial Value
byte	8	0
ubyte	8	0
short	16	0
ushort	16	0
int	32	0
uint	32	0
long	64	0L
ulong	64	0L

The **u** at the beginning of the type names stands for "unsigned" and indicates that such types cannot have values less than zero.

15.1.2 Number of bits of a type

In today's computer systems, the smallest unit of information is called a *bit*. At the physical level, a bit is represented by electrical signals at certain points in the circuitry of a computer. A bit can be in one of two states that correspond to the presence and absence of electrical signals at the point that defines that particular bit. These two states are arbitrarily defined to have the values 0 and 1. As a result, a bit can have one of these two values.

As there aren't many concepts that can be represented by just two states, bit is not a very useful type. It can only be useful for concepts with two states like heads or tails or whether a light switch is on or off.

If we consider two bits at a time, the total number of information that can be represented multiplies. Based on each bit having a value of 0 or 1 individually, there are a total of 4 possible states. Assuming that the left and right digits represent the first and second bit respectively, these states are 00, 01, 10, and 11. Let's add one more bit to see this effect better; three bits can be in 8 different states: 000, 001, 010, 011, 100, 101, 110, 111. As can be seen, each added bit doubles the total number of states that can be represented.

The values to which these eight states correspond to are defined by conventions. The following table shows these values for the signed and unsigned representations of 3 bits:

Bit State	Unsigned Value	Signed Value
000	0	0
001	1	1
010	2	2
011	3	3
100	4	-4
101	5	-3
110	6	-2
111	7	-1

We can produce the following table by adding more bits:

Bits	Number of Distinct Values	D Type	Minimum Value	Maximum Value
1	2			
2	4			
3	8			
4	16			
5	32			
6	64			
7	128			
8	256	byte	-128	127
		ubyte	0	255
...	...			
16	65,536	short	-32768	32767
		ushort	0	65535
...	...			
32	4,294,967,296	int	-2147483648	2147483647
		uint	0	4294967295
...	...			
64	18,446,744,073,709,551,616	long	-9223372036854775808	9223372036854775807
		ulong	0	18446744073709551615
...	...			

I have skipped many rows in the table, and indicated the signed and unsigned versions of the D types that have the same number of bits on the same row (e.g. **int** and **uint** are both on the 32-bit row).

15.1.3 Choosing a type

Since a 3-bit type can only have 8 distinct values, it can only represent concepts like the value of a die or the number of the day of the week. (This is just an example; there isn't a 3-bit type in D.)

On the other hand, although **uint** is a very large type, it cannot represent the concept of an ID number for each living person; as its maximum value is less than the world population of 7 billion. **long** and **ulong** would be more than enough to represent many concepts.

As a general rule, as long as there is no specific reason not to, you can use **int** for integer values.

15.1.4 Overflow

The fact that types can have a limited range of values may cause unexpected results. For example, although adding two `uint` variables with values of 3 billion each should produce 6 billion, because that sum is greater than the maximum value that a `uint` variable can hold (about 4 billion), this sum *overflows*. Without any warning, only the difference of 6 and 4 billion gets stored. (A little more accurately, 6 minus 4.3 billion.)

15.1.5 Truncation

Since integers cannot have values with fractional parts, they lose the part after the decimal point. For example, assuming that a box of ice cream serves 4 kids, although 11 kids would actually need 2.75 boxes, 2.75 can only be stored as 2 in an integer type.

I will show limited techniques to help reduce the effects of overflow, underflow, and truncation later in the chapter.

15.1.6 `.min` and `.max`

I will take advantage of the `.min` and `.max` properties below, which we have seen in the fundamental types chapter. These properties provide the minimum and maximum values that an integer type can have.

15.1.7 Increment: `++`

This operator is used with a single variable (more generally, with a single expression) and is written before the name of that variable. It increments the value of that variable by 1:

```
import std.stdio;

void main()
{
    int number = 10;
    ++number;
    writeln("New value: ", number);
}
```

New value: 11

The increment operator is the equivalent of using the *add-and-assign* operator with the value of 1:

```
number += 1;    // same as ++number
```

If the result of the increment operation is greater than the maximum value of that type, the result *overflows* and becomes the minimum value. We can see this effect by incrementing a variable that initially has the value `int.max`:

```
import std.stdio;

void main()
{
    writeln("minimum int value  : ", int.min);
    writeln("maximum int value  : ", int.max);

    int number = int.max;
    writeln("before the increment: ", number);
    ++number;
```

```
    writeln("after the increment : ", number);
}
```

The value becomes `int.min` after the increment:

```
minimum int value   : -2147483648
maximum int value   : 2147483647
before the increment: 2147483647
after the increment : -2147483648
```

This is a very important observation, because the value changes from the maximum to the minimum as a result of *incrementing* and without any warning! This effect is called *overflow*. We will see similar effects with other operations.

15.1.8 Decrement: --

This operator is similar to the increment operator; the difference is that the value is decreased by 1:

```
--number; // the value decreases by 1
```

The decrement operation is the equivalent of using the *subtract-and-assign* operator with the value of 1:

```
number -= 1; // same as --number
```

Similar to the `++` operator, if the value is the minimum value to begin with, it becomes the maximum value. This effect is called *underflow*.

15.1.9 Addition: +

This operator is used with two expressions and adds their values:

```
import std.stdio;

void main()
{
    int first = 12;
    int second = 100;

    writeln("Result: ", first + second);
    writeln("With a constant expression: ", 1000 + second);
}
```

```
Result: 112
With a constant expression: 1100
```

If the result is greater than the sum of the two expressions, it again overflows and becomes a value that is less than both of the expressions:

```
import std.stdio;

void main()
{
    // 3 billions each
    uint first = 3000000000;
    uint second = 3000000000;
```

```
writeln("maximum value of uint: ", uint.max);
writeln("          first: ", first);
writeln("          second: ", second);
writeln("          sum: ", first + second);
writeln("OVERFLOW! The result is not 6 billions!");
}
```

```
maximum value of uint: 4294967295
          first: 3000000000
          second: 3000000000
          sum: 1705032704
OVERFLOW! The result is not 6 billions!
```

15.1.10 Subtraction: -

This operator is used with two expressions and gives the difference between the first and the second:

```
import std.stdio;

void main()
{
    int number_1 = 10;
    int number_2 = 20;

    writeln(number_1 - number_2);
    writeln(number_2 - number_1);
}
```

```
-10
10
```

It is again surprising if the actual result is less than zero and is stored in an unsigned type. Let's rewrite the program using the `uint` type:

```
import std.stdio;

void main()
{
    uint number_1 = 10;
    uint number_2 = 20;

    writeln("PROBLEM! uint cannot have negative values:");
    writeln(number_1 - number_2);
    writeln(number_2 - number_1);
}
```

```
PROBLEM! uint cannot have negative values:
4294967286
10
```

It is a good guideline to use signed types to represent concepts that may ever be subtracted. As long as there is no specific reason not to, you can choose `int`.

15.1.11 Multiplication: *

This operator multiplies the values of two expressions; the result is again subject to overflow:

```
import std.stdio;
void main()
{
    uint number_1 = 6;
    uint number_2 = 7;

    writeln(number_1 * number_2);
}
```

42

15.1.12 Division: /

This operator divides the first expression by the second expression. Since integer types cannot have fractional values, the fractional part of the value is discarded. This effect is called *truncation*. As a result, the following program prints 3, not 3.5:

```
import std.stdio;
void main()
{
    writeln(7 / 2);
}
```

3

For calculations where fractional parts matter, *floating point types* must be used instead of integers. We will see floating point types in the next chapter.

15.1.13 Remainder: %

This operator divides the first expression by the second expression and produces the remainder of the division:

```
import std.stdio;
void main()
{
    writeln(10 % 6);
}
```

4

A common application of this operator is to determine whether a value is odd or even. Since the remainder of dividing an even number by 2 is always 0, comparing the result against 0 is sufficient to make that distinction:

```
if ((number % 2) == 0) {
    writeln("even number");
} else {
    writeln("odd number");
}
```

15.1.14 Power: ^^

This operator raises the first expression to the power of the second expression. For example, raising 3 to the power of 4 is multiplying 3 by itself 4 times:

```
import std.stdio;

void main()
{
    writeln(3 ^^ 4);
}
```

81

15.1.15 Arithmetic operations with assignment

All of the operators that take two expressions have *assignment* counterparts. These operators assign the result back to the expression that is on the left-hand side:

```
import std.stdio;

void main()
{
    int number = 10;

    number += 20; // same as number = number + 20; now 30
    number -= 5; // same as number = number - 5; now 25
    number *= 2; // same as number = number * 2; now 50
    number /= 3; // same as number = number / 3; now 16
    number %= 7; // same as number = number % 7; now 2
    number ^^= 6; // same as number = number ^^ 6; now 64

    writeln(number);
}
```

64

15.1.16 Negation: -

This operator converts the value of the expression from negative to positive or positive to negative:

```
import std.stdio;

void main()
{
    int number_1 = 1;
    int number_2 = -2;

    writeln(-number_1);
    writeln(-number_2);
}
```

-1
2

The type of the result of this operation is the same as the type of the expression. Since unsigned types cannot have negative values, the result of using this operator with unsigned types can be surprising:

```
uint number = 1;
writeln("negation: ", -number);
```

The type of `-number` is `uint` as well, which cannot have negative values:

```
negation: 4294967295
```

15.1.17 Plus sign: +

This operator has no effect and exists only for symmetry with the negation operator. Positive values stay positive and negative values stay negative:

```
import std.stdio;

void main()
{
    int number_1 = 1;
    int number_2 = -2;

    writeln(+number_1);
    writeln(+number_2);
}
```

```
1
-2
```

15.1.18 Post-increment: ++

Note: Unless there is a strong reason, always use the regular increment operator (which is sometimes called the pre-increment operator).

Contrary to the regular increment operator, it is written after the expression and still increments the value of the expression by 1. The difference is, the post-increment operation itself produces the old value of the expression. To see this difference, let's compare it with the regular increment operator:

```
import std.stdio;

void main()
{
    int incremented_regularly = 1;
    writeln(++incremented_regularly); // prints 2
    writeln(incremented_regularly);  // prints 2

    int post_incremented = 1;

    // Gets incremented, but its old value is used:
    writeln(post_incremented++);      // prints 1
    writeln(post_incremented);        // prints 2
}
```

```
2
2
1
2
```

The `writeln(post_incremented++)`; statement above is the equivalent of the following code:

```
int old_value = post_incremented;
++post_incremented;
writeln(old_value);           // prints 1
```

15.1.19 Post-decrement: --

Note: Unless there is a strong reason, always use the regular decrement operator (which is sometimes called the pre-decrement operator).

Behaves the same as the post-increment operator but decrements.

15.1.20 Operator precedence

The operators we've discussed above have all been used in operations on their own with only one or two expressions. However, similar to logical expressions, it is common to combine these operators to form more complex arithmetic expressions:

```
int value = 77;
int result = ((value + 8) * 3) / (value - 1) % 5;
```

As with logical operators, arithmetic operators also obey operator precedence rules. For example, the `*` operator has precedence over the `+` operator. For that reason, when parentheses are not used (e.g. in the `value + 8 * 3` expression), the `*` operator is evaluated before the `+` operator. As a result, that expression becomes the equivalent of `value + 24`, which is quite different from `(value + 8) * 3`.

Using parentheses is useful both for ensuring correct results and for communicating the intent of the code to programmers who may work on the code in the future.

15.1.21 Potential remedy against overflow

If the result of an operation cannot fit in the type of the result, then there is nothing that can be done. Sometimes, although the ultimate result would fit in a certain type, the intermediate calculations may overflow and cause incorrect results.

As an example, let's assume that we need to plant an apple tree per 1000 square meters of an area that is 40 by 60 kilometers. How many trees are needed?

When we solve this problem on paper, we see that the result is 40000 times 60000 divided by 1000, being equal to 2.4 million trees. Let's write a program that executes this calculation:

```
import std.stdio;

void main()
{
    int width = 40000;
    int length = 60000;
    int areaPerTree = 1000;

    int treesNeeded = width * length / areaPerTree;

    writeln("Number of trees needed: ", treesNeeded);
}
```

```
Number of trees needed: -1894967
```


Not to mention it's not even close, the result is also less than zero! In this case, the intermediate calculation `width * length` overflows and the subsequent calculation of `/ areaPerTree` produces an incorrect result.

One way of avoiding the overflow in this example is to change the order of operations:

```
int treesNeeded = width / areaPerTree * length ;
```

The result would now be correct:

Number of trees needed: 2400000

The reason this method works is the fact that all of the steps of the calculation now fit the `int` type.

Please note that this is not a complete solution because this time the intermediate value is prone to truncation, which may affect the result significantly in certain other calculations. Another solution might be to use a floating point type instead of an integer type: `float`, `double`, or `real`.

15.1.22 Potential remedy against truncation

Changing the order of operations may be a solution to truncation as well. An interesting example of truncation can be seen by dividing and multiplying a value with the same number. We would expect the result of $10/9*9$ to be 10, but it comes out as 9:

```
import std.stdio;
void main()
{
    writeln(10 / 9 * 9);
}
```

9

The result is correct when truncation is avoided by changing the order of operations:

```
writeln(10 * 9 / 9);
```

10

This too is not a complete solution: This time the intermediate calculation could be prone to overflow. Using a floating point type may be another solution to truncation in certain calculations.

15.2 Exercises

1. Write a program that takes two integers from the user, prints the integer quotient resulting from the division of the first by the second, and also prints the remainder. For example, when 7 and 3 are entered, have the program print the following equation:

```
7 = 3 * 2 + 1
```

2. Modify the program to print shorter output when the remainder is 0. For example, when 10 and 5 are entered, it should not print "10 = 5 * 2 + 0" but just the following:

```
10 = 5 * 2
```

3. Write a simple calculator that supports the four basic arithmetic operations. Have the program let the operation to be selected from a menu and apply that operation to the two values that are entered. You can ignore overflow and truncation in this program.
4. Write a program that prints the values from 1 to 10, each on a separate line, with the exception of value 7. Do not use repeated lines like the following:

```
import std.stdio;

void main()
{
    // Do not do this!
    writeln(1);
    writeln(2);
    writeln(3);
    writeln(4);
    writeln(5);
    writeln(6);
    writeln(8);
    writeln(9);
    writeln(10);
}
```

Instead, imagine a variable whose value is incremented in a loop. You may need to take advantage of the *is not equal to* operator `!=` here.

[... the solutions](#)

16 Floating Point Types

In the previous chapter we have seen that despite their ease of use, arithmetic operations on integers are prone to programming errors due to overflow, underflow, and truncation. We have also seen that integers cannot have values with fractional parts as in 1.25.

Floating point types are especially to support fractional parts. The "point" in their name comes from the *decimal point* and "floating" refers to an implementation detail of how these types are implemented: the decimal point *floats* left and right as appropriate. (This detail is not important when using these types.)

We must cover important details in this chapter as well. Before doing that, I would like to give a list of some of the interesting aspects of floating point types:

- Adding 0.001 a thousand times is not the same as adding 1.
- Using the logical operators `==` and `!=` with floating point types is erroneous in most cases.
- The initial values of floating point types is `.nan`, not 0. `.nan` may not be used in expressions. When used in comparison operations, `.nan` is not less than nor greater than any value.
- The overflow value is `.infinity` and the underflow value is negative `.infinity`.

Although floating point types are more useful in some cases, they have peculiarities that every programmer must know. Compared to integers, they are very good with avoiding truncation because their main purpose is to support fractional values. Like any other type, being based on certain numbers of bits, they too are prone to overflow and underflow, but compared to integers, the range of values that they can support is vast. Additionally, instead of being silent in the case of overflow and underflow, they get the special values of positive and negative *infinity*.

As a reminder, all of the floating point types are the following:

Type	Number of Bits	Initial Value
float	32	float.nan
double	64	double.nan
real	at least 64, maybe more (e.g. 80, depending on hardware support)	real.nan

16.1 Floating point type properties

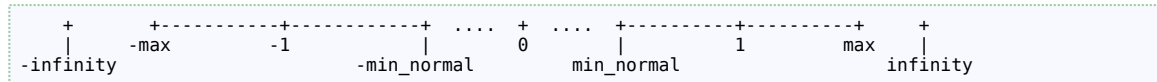
Floating point types have more properties than other types:

- `.stringof` is the name of the type.
- `.sizeof` is the length of the type in terms of bytes. (In order to determine the bit count, this value must be multiplied by 8, the number of bits in a byte.)
- `.max` is the short for "maximum" and is the maximum value that the type can have; the negative of this value is the minimum value that the type can have.
- `.min_normal` is the smallest normalized value that this type can have (the type can have smaller values than `.min_normal` but the precision of those values are less than the normal precision of the type).
- `.dig` is the short for "digits" and specifies the number of digits that signify the precision of the type.
- `.infinity` is the special value used to denote overflow and underflow.

Note that the minimum value of floating point types is not `.min` but the negative of `.max`, e.g. the minimum value of `double` is `-double.max`.

Other properties of floating point types are used less commonly. You can see all of them at [Properties for Floating Point Types](#).

The properties of floating point types and their relations can be shown on a number line like the following:



The dashed portions of the line above are to scale: the number of values that can be represented between `min_normal` and 1 is equal to the number of values that can be represented between 1 and `max`. This means that the precision of the fractional parts of the values that are between `min_normal` and 1 are very high. (The same is true for the negative side as well.)

16.2 `.nan`

We have already seen that this is the default value of floating point variables. `.nan` may appear as a result of meaningless floating point expressions as well. For example the floating point expressions in the following program all produce `double.nan`:

```
import std.stdio;

void main()
{
    double zero = 0;
    double infinity = double.infinity;

    writeln("any expression with nan: ", double.nan + 1);
    writeln("zero / zero          : ", zero / zero);
    writeln("zero * infinity       : ", zero * infinity);
    writeln("infinity / infinity    : ", infinity / infinity);
    writeln("infinity - infinity     : ", infinity - infinity);
}
```

16.3 Specifying floating point values

Floating point values can simply be specified as integers without a decimal point like 123, or with a decimal point like 12.3.

Floating point values can also be specified with the special floating point syntax as in `1.23e+4`. The `e+` part in that syntax can be read as "times 10 to the power of". According to that reading, the previous value is "1.23 times 10 to the power of 4", which is the same as "1.23 times 10⁴", which in turn is the same as 1.23x10000, being equal to 12300.

If the value after `e` is negative as in `5.67e-3`, then it is read as "divided by 10 to the power of". Accordingly, this example is "5.67 divided by 10³", which in turn is the same as 5.67/1000, being equal to 0.00567.

The values that are printed by the following program are all in the floating point format. The program prints the properties of the three types:

```
import std.stdio;

void main()
{
```

```

writeln("Type           : ", float.stringof);
writeln("Precision       : ", float.dig);
writeln("Minimum normalized value: ", float.min_normal);
writeln("Maximum value     : ", float.max);
writeln("Minimum value     : ", -float.max);
writeln();
writeln("Type           : ", double.stringof);
writeln("Precision       : ", double.dig);
writeln("Minimum normalized value: ", double.min_normal);
writeln("Maximum value     : ", double.max);
writeln("Minimum value     : ", -double.max);
writeln();
writeln("Type           : ", real.stringof);
writeln("Precision       : ", real.dig);
writeln("Minimum normalized value: ", real.min_normal);
writeln("Maximum value     : ", real.max);
writeln("Minimum value     : ", -real.max);
}

```

The output of the program is the following in my environment. Since **real** depends on the hardware, you may get a different output for it:

```

Type           : float
Precision       : 6
Minimum normalized value: 1.17549e-38
Maximum value   : 3.40282e+38
Minimum value   : -3.40282e+38

Type           : double
Precision       : 15
Minimum normalized value: 2.22507e-308
Maximum value   : 1.79769e+308
Minimum value   : -1.79769e+308

Type           : real
Precision       : 18
Minimum normalized value: 3.3621e-4932
Maximum value   : 1.18973e+4932
Minimum value   : -1.18973e+4932

```

16.3.1 Observations

As you will remember from the previous chapter, the maximum value of **ulong** has 20 digits: 18,446,744,073,709,551,616. That value looks small when compared to even the smallest floating point type: **float** can have values up to the 10^{38} range, e.g. 340,282,000,000,000,000,000,000,000,000,000.

The maximum value of **real** is in the range 10^{4932} , a value with more than 4900 digits!

As another observation, let's look at the minimum value that **double** can represent with 15-digit precision: 0.000...*there are 300 more zeroes here*...0000222507. (More precisely, the last part is ...222507385850720.)

16.4 Overflow and underflow are not ignored

Despite being able to take very large values, floating point types too are prone to overflow and underflow. The floating point types are safer than integer types in this regard because

overflow and underflow are not ignored. The values that overflow become **.infinity**, and the values that underflow become **-.infinity**. To see this, let's increment the value of **.max** by 10%. Since the value is already at the maximum, incrementing by 10% would overflow:

```
import std.stdio;

void main()
{
    real value = real.max;

    writeln("Before          : ", value);

    // Multiplying with 1.1 is the same as making 110%
    value *= 1.1;
    writeln("Added 10%       : ", value);

    // Let's try to make it less by dividing in half
    value /= 2;
    writeln("Divided in half: ", value);
}
```

Once the value overflows and becomes **real.infinity**, it remains that way even after being divided in half:

```
Before          : 1.18973e+4932
Added 10%       : inf
Divided in half: inf
```

16.5 Precision

Precision is a concept that we come across in daily life but not talk about much. Precision is the number of digits that is used when specifying a value. For example, when we say that the third of 100 is 33, the precision is 2 because 33 has 2 digits. When the value is specified more precisely like 33.33, then the precision is 4 digits.

The number of bits that each floating type has not only effects its maximum value, but also its precision. The greater the number of bits, the more precise the values.

16.6 There is no truncation in division

As we have seen in the previous chapter, integer division cannot preserve the fractional part of the result:

```
int first = 3;
int second = 2;
writeln(first / second);
```

Output:

```
1
```

Floating point types don't have this *truncation* problem; they are specifically for preserving the fractional parts:

```
double first = 3;
double second = 2;
writeln(first / second);
```

Output:

```
1.5
```

The accuracy of the fractional part depends on the precision of the type: **real** has the highest precision and **float** has the lowest precision.

16.7 Which type to use

Unless there is a specific reason otherwise, you can choose **double** for floating point values. **float** has very low precision but because of being smaller than the other types, it may be useful when storage space is limited. On the other hand, since the precision of **real** is higher than **double** on some hardware, it would be preferable for high precision calculations.

16.8 Cannot represent all values

We cannot represent certain values in our daily lives. In the decimal system that we use daily, the digits before the decimal point represent ones, tens, hundreds, etc. and the digits after the decimal point represent tenths, hundredths, thousandths, etc.

If a value is an exact combination of these values, it can be represented precisely. For example, because the value 0.23 consists of 2 tenths and 3 hundredths it is represented precisely. On the other hand, the value of $1/3$ cannot be represented in the decimal system because it is never sufficient no matter how many digits are specified: 0.33333...

It is very similar with the floating point types. Because these types are based on certain numbers of bits, they cannot represent every value.

The difference with the binary system that the computers use is that the digits before the decimal point are ones, twos, fours, etc. and the digits after the decimal point are halves, quarters, eighths, etc. Only the values that are exact combinations of those digits can be represented precisely.

A value that cannot be represented in computers is 0.1 as in 10 cents. Although this value can be represented precisely in the decimal system, its binary representation never ends and continuously repeats four digits as 0.0001100110011... (Note that it is written in binary system, not decimal.) It is always inaccurate at some level depending on the precision of the floating point type that is used.

The following program demonstrates this problem. The value of a variable is being incremented by 0.001 a thousand times in a loop. Surprisingly, the result is not 1:

```
import std.stdio;

void main()
{
    float result = 0;

    // We would expect the result to be 1 after looping 1000
    // times:
    while (result < 1) {
        result += 0.001;
    }

    // Let's validate
    if (result == 1) {
        writeln("As expected: 1");
    } else {
```

```

        writeln("DIFFERENT: ", result);
    }
}

```

```
DIFFERENT: 1.00099
```

Because 0.001 cannot be represented precisely, that inaccuracy affects the result multiple times. The output suggests that the loop has been repeated 1001 times.

16.9 Comparing floating point values

We have seen the following comparison operations for integers: equal to (`==`), not equal to (`!=`), less than (`<`), greater than (`>`), less than or equal to (`<=`), and greater than or equal to (`>=`). Floating point types have many more comparison operators.

Since the special value `.nan` represents invalid floating point values, it may not be compared with any other value. For example, it is meaningless to ask whether `.nan` or 1 is greater.

For that reason, floating point values introduce another comparison concept: unordered. Being unordered means that at least one of the values is `.nan`.

The following table includes all of the floating point comparison operators. All of the operators are binary operators (meaning that they take two operands) and used as in `left == right`. The columns that contain `false` and `true` are the results of the comparison operations.

The last column indicates whether the operation is meaningful if one of the operands is `.nan`. For example, even though the result of the expression `1.2 < real.nan` is `false`, that result is meaningless because one of the operands is `real.nan`. The result of the reverse comparison `real.nan < 1.2` would produce `false` as well. The abbreviation lhs stands for *left-hand side*, indicating the expression on the left-hand side of each operator.

Operator	Meaning	If lhs is greater	If lhs is less	If both are equal	If at least one is .nan	Meaningful with .nan
<code>==</code>	is equal to	false	false	true	false	yes
<code>!=</code>	is not equal to	true	true	false	true	yes
<code>></code>	is greater than	true	false	false	false	no
<code>>=</code>	is greater than or equal to	true	false	true	false	no
<code><</code>	is less than	false	true	false	false	no
<code><=</code>	is less than or equal to	false	true	true	false	no
<code>!<=></code>	is not less than, not greater than, nor equal to	false	false	false	true	yes
<code><></code>	is less than or greater than	true	true	false	false	no
<code><=></code>	is less than, greater than, or equal to	true	true	true	false	no
<code>!<=</code>	is not less than nor equal to	true	false	false	true	yes
<code>!<</code>	is not less than	true	false	true	true	yes
<code>!>=</code>	is not greater than nor equal to	false	true	false	true	yes
<code>!></code>	is not greater than	false	true	true	true	yes
<code>!<></code>	is not less than nor greater than	false	false	true	true	yes

Note that it makes sense to use `.nan` with any operator that contains *is not* in its meaning, and the result is always **true**. `.nan` not being a valid value makes the results of most comparisons to have *not* in their meanings.

16.9.1 `isnan()` for `.nan` equality comparison

Because the result would always be **false** according to the previous table, it is not possible to use the `==` operator to determine whether the value of a floating point variable is `.nan`:

```
if (variable == double.nan) {    // ← WRONG
    // ...
}
```

For that reason, the `isnan()` function from the `std.math` module must be used:

```
import std.math;
// ...
if (isnan(variable)) {    // ← correct
    // ...
}
```

Similarly, to determine whether a value is *not* `.nan`, one must use `!isnan()`, because otherwise the `!=` operator would always produce **true**.

16.10 Exercises

1. Modify the calculator from the previous chapter to support floating point types. The new calculator would work more accurately with that change. When trying the calculator, you can enter floating point values in various formats as in 1000, 1.23, and 1.23e4.
2. Write a program that reads 5 floating point values from the input. Make the program print first twice of those values and then one fifth of those values.

This exercise is a preparation for the array concept of the next chapter. If you write this program with what you have seen so far, you will understand arrays easier and appreciate them more.

[... the solutions](#)

17 Arrays

We have defined five variables in one of the exercises of the last chapter, and used them in certain calculations. The definitions of those variables were the following:

```
double value_1;
double value_2;
double value_3;
double value_4;
double value_5;
```

This method of defining variables individually does not scale to cases where even more variables are needed. Imagine needing a thousand values; it is almost impossible to define a thousand variables from **value_1** to **value_1000**.

Arrays are useful in such cases: The array feature allows the definition of many values together. Arrays are also the most common data structure when multiple values are used together as a collection.

This chapter covers only some of the features of arrays. More features will be introduced in a later chapter.

17.1 Definition

The definition of arrays is very similar to the definition of variables. The only difference is that the number of variables that are being defined at the same time is specified in square brackets. We can contrast the two definitions as follows:

```
int    singleVariable;
int[10] arrayOfTenVariables;
```

The first line above is the definition of a single variable, just like the variables that we have defined so far. The second line is the definition of an array consisting of 10 variables.

Accordingly, the equivalent of the five separate variables above can be defined as an array of five variables using the following syntax:

```
double[5] values;
```

That definition can be read as *5 double values*. Note that I have chosen the name of the array as plural to avoid confusing it with a single variable.

In summary, the definition of an array consists of the type of the variables, the number of variables, and the name of the array:

```
type_name[variable_count] array_name;
```

The type of the variables can also be a user-defined type. (We will see user-defined types later.) For example:

```
// An array that holds the weather information of all
// cities. Here, the bool values may mean
//   false: overcast
//   true : sunny
bool[cityCount] weatherConditions;

// An array that holds the weights of a hundred boxes
double[100] boxWeights;
```

```
// Information about the students of a school
StudentInformation[studentCount] studentInformations;
```

17.2 Containers and elements

Data structures that bring elements of a certain type together are called *containers*. According to this definition, arrays are containers. For example, an array that holds the air temperatures of the days in July can bring 31 **double** variables together and form *a container of elements of type double*.

The variables of a container are called *elements*. The number of elements of an array is called the *length* of the array.

17.3 Accessing the elements

In order to differentiate the variables in the exercise of the previous chapter, we had to append an underscore and a number to their names as in **value_1**. This is not possible nor necessary when the variables are defined together as a single array under a single name. Instead, the elements are accessed by specifying the *element number* within square brackets:

```
values[0]
```

That expression can be read as *the element with the number 0 of the array named values*. In other words, instead of typing **value_1** one must type **values[0]** with arrays.

There are two important points worth stressing here:

- **The numbers start with zero:** Although humans assign numbers to items starting with 1, the numbers in arrays start with 0. The values that we have numbered as 1, 2, 3, 4, and 5 before are numbered as 0, 1, 2, 3, and 4 in the array. This variation is a common cause of programming errors.
- **Two different uses of the [] characters:** Don't confuse the two separate uses of the [] characters. When defining arrays, the [] characters are written after the type of the elements and specify the number of elements. When accessing elements, the [] characters are written after the name of the array and specify the number of the element that is being accessed:

```
// This is a definition. It defines an array that consists
// of 12 elements. This array is used to hold the number
// of days in each month.
int[12] monthDays;

// This is an access. It accesses the element that
// corresponds to December and sets its value to 31.
monthDays[11] = 31;

// This is another access. It accesses the element that
// corresponds to January, the value of which is passed to
// writeln.
writeln("January has ", monthDays[0], " days.");
```

Reminder: The element numbers of January and December are 0 and 11 respectively; not 1 and 12.

17.4 Index

The number of an element is called its *index* and the act of accessing an element is called *indexing*.

An index need not be a constant value; the value of a variable can also be used as an index, making arrays even more useful. For example, the month is determined by the value of the `monthIndex` variable below:

```
writeln("This month has ", monthDays[monthIndex], " days.");
```

When the value of `monthIndex` is 2, the expression above would print the value of `monthDays[2]`, the number of days in March.

Only the index values between zero and one less than the length of the array are valid. For example, the valid indexes of a three-element array are 0, 1, and 2. Accessing an array with an invalid index causes the program to be terminated with an error.

Arrays are containers where the elements are placed side by side in the computer's memory. For example, the elements of the array holding the number of days in each month can be shown like the following (assuming a year when February has 28 days):

```
indexes →    0    1    2    3    4    5    6    7    8    9   10   11
elements → | 31 | 28 | 31 | 30 | 31 | 30 | 31 | 31 | 30 | 31 | 30 | 31 |
```

The element at index 0 has the value 31 (number of days in January); the element at index 1 has the value of 28 (number of days in February), etc.

17.5 Fixed-length arrays vs. dynamic arrays

When the length of an array is specified when the program is written, that array is a *fixed-length array*. When the length can change during the execution of the program, that array is a *dynamic array*.

Both of the arrays that we have defined above are fixed-length arrays because their element counts are specified as 5 and 12 at the time when the program is written. The lengths of those arrays cannot be changed during the execution of the program. To change their lengths, the source code must be modified and the program must be recompiled.

Defining dynamic arrays is simpler than defining fixed-length arrays because omitting the length makes a dynamic array:

```
int[] dynamicArray;
```

The length of such an array can increase or decrease during the execution of the program.

17.6 Array properties

As well as a type, an array has some properties as well. We will see only three of them here. Note that the `.length` property of dynamic arrays is more capable than that of fixed-length arrays:

- `.length`: Returns the number of elements of the array:

```
writeln("The array has ", array.length, " elements.");
```

Additionally, the length of dynamic arrays can be changed by assigning a value to this property:

```
int[] array;           // initially empty
array.length = 5;     // now has 5 elements
```

- **.sort**: Sorts the elements (in the case of integers, from the smallest value to the greatest value):

```
array.sort;
```

- **.reverse**: Sorts the elements in reverse order:

```
array.reverse;
```

Now let's revisit the exercise with the five values and write it again by using an array:

```
import std.stdio;

void main()
{
    // This variable is used as a loop counter
    int counter;

    // The definition of a fixed-length array of five
    // elements of type double
    double[5] values;

    // Reading the values in a loop
    while (counter < values.length) {
        write("Value ", counter + 1, ": ");
        readf(" %s", &values[counter]);
        ++counter;
    }

    writeln("Twice the values:");
    counter = 0;
    while (counter < values.length) {
        writeln(values[counter] * 2);
        ++counter;
    }

    // The loop that calculates the fifths of the values would
    // be written similarly
}
```

Observations: The value of **counter** determines how many times the loops are repeated (iterated). Iterating the loop while its value is less than **array.length** ensures that the loops are executed once per element. As the value of that variable is incremented at the end of each iteration, the **values[counter]** expression refers to the elements of the array one by one: **values[0]**, **values[1]**, etc.

To see how this program is better than the previous one, imagine needing to read 20 values. The program above would require a single change: replacing 5 with 20. On the other hand, the program that did not use an array would have to have 15 more variable definitions and the lines that they have been used in would have to be repeated 15 more times.

17.7 Initializing the elements

Like every variable in D, the elements of arrays are automatically initialized. The initial value of the elements depends on the type of the elements: 0 for **int**, **double.nan** for **double**, etc.

All of the elements of the **values** array above are initialized to **double.nan**:

```
double[5] values;    // elements are all double.nan
```

Obviously, the values of the elements can be changed later in the program. We have already seen this above when assigning to an element of an array:

```
monthDays[11] = 31;
```

And also when reading a value from the input:

```
readf(" %s", &values[counter]);
```

Sometimes the desired values of the elements are known at the time when the array is defined. In such cases, the initial values of the elements can be specified on the right-hand side of the assignment operator within square brackets. Let's see this in a program that reads the number of the month from the user, and prints the number of days in that month:

```
import std.stdio;

void main()
{
    // Assuming that February has 28 days
    int[12] monthDays =
        [ 31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31 ];

    write("Please enter the number of the month: ");
    int monthNumber;
    readf(" %s", &monthNumber);

    int index = monthNumber - 1;
    writeln("Month ", monthNumber, " has ",
           monthDays[index], " days.");
}
```

As you see, the **monthDays** array is defined and initialized at the same time. Also note that the number of the month, which is in the range 1-12 is converted to a valid array index in the range 0-11. Any value that is entered outside of the 1-12 range would cause the program to be terminated with an error.

When initializing arrays, it is possible to use a single value on the right-hand side. In that case all of the elements of the array are initialized to that value:

```
int[10] allOnes = 1;    // All of the elements are set to 1
```

17.8 Basic array operations

Arrays provide convenient operations that apply to all of their elements.

Copying fixed-length arrays: The assignment operator copies all of the elements from the right-hand side to the left-hand side:

```
int[5] source = [ 10, 20, 30, 40, 50 ];
int[5] destination;

destination = source;
```

Note: The meaning of the assignment operation is completely different for dynamic arrays. We will see this in a later chapter.

Adding elements to dynamic arrays: The `~=` operator adds a new element or a new array to the end of a dynamic array:

```
int[] array;           // empty
array ~= 7;           // has a single element
array ~= 360;         // has two elements
array ~= [ 30, 40 ]; // has four elements
```

It is not possible to add elements to fixed-length arrays:

```
int[10] array;
array ~= 7;           // ← compilation ERROR
```

Combining arrays: The `~` operator creates a new array by combining two arrays. Its `~=` counterpart combines the two arrays and assigns the result back to the left-hand side array:

```
import std.stdio;

void main()
{
    int[10] first = 1;
    int[10] second = 2;
    int[] result;

    result = first ~ second;
    writeln(result.length); // prints 20

    result ~= first;
    writeln(result.length); // prints 30
}
```

The `~=` operator cannot be used when the left-hand side array is a fixed-length array:

```
int[20] result;
// ...
result ~= first; // ← compilation ERROR
```

If the left-hand side array doesn't have room for the result, the program is terminated with an error during assignment:

```
int[10] first = 1;
int[10] second = 2;
int[21] result;

result = first ~ second;
```

object.Exception: lengths don't match for array copy

17.9 Exercises

1. Write a program that asks the user how many values will be entered and then reads all of them. Have the program sort the elements using `.sort` and `.reverse`.
2. Write a program that reads numbers from the input, and prints the odd and even ones separately but in order. Treat the value -1 specially to determine the end of the numbers; do not process that value.

For example, when the following numbers are entered,

```
1 4 7 2 3 8 11 -1
```

have the program print the following:

```
1 3 7 11 2 4 8
```

Hint: You may want to put the elements in separate arrays. You can determine whether a number is odd or even using the % (remainder) operator.

- The following is a program that does not work as expected. The program is written to read five numbers from the input and to place the squares of those numbers into an array. The program then attempts to print the squares to the output. Instead, the program terminates with an error.

Fix the bugs of this program and make it work as expected:

```
import std.stdio;

void main()
{
    int[5] squares;

    writeln("Please enter 5 numbers");

    int i = 0;
    while (i <= 5) {
        int number;
        write("Number ", i + 1, ": ");
        readf(" %s", &number);

        squares[i] = number * number;
        ++i;
    }

    writeln("=== The squares of the numbers ===");
    while (i <= squares.length)
    {
        write(squares[i], " ");
        ++i;
    }

    writeln();
}
```

[... the solutions](#)

18 Characters

Characters are the building blocks of strings. Any symbol of a writing system is called a character: letters of alphabets, numerals, punctuation marks, the space character, etc. Confusingly, the building blocks of characters themselves are called characters as well.

Arrays of characters make up *strings*. We have seen arrays in the previous chapter; strings will be covered two chapters later.

Like any other data, characters are also represented as integer values that are made up of bits. For example, the integer value of the lowercase **a** is 97 and the integer value of the numeral **1** is 49. These values have been assigned merely by conventions when the ASCII table has been designed.

In most languages, characters are represented by the **char** type, which can hold only 256 distinct values. If you are familiar with the **char** type from other languages, you may already know that it is not large enough to support the symbols of many writing systems. Before getting to the three distinct character types of D, let's first take a look at the history of characters in computer systems.

18.1 History

18.1.1 ASCII Table

The ASCII table was designed at a time when computer hardware was very limited compared to modern systems. Having been based on 7 bits, the ASCII table can have 128 distinct code values. That is sufficient to represent characters like the lowercase and uppercase versions of the 26 letters of the basic Latin alphabet, numerals, commonly used punctuation marks, and some terminal (console) control characters.

As an example, the ASCII codes of the characters of the string "hello" are the following (the commas are inserted just to make it easier to read):

```
104, 101, 108, 108, 111
```

Every code above represents a single letter of "hello". For example, there are two 108 values for the two 'l' letters. (**Note:** *The actual order of these characters depend on the platform and even the document that these values are a part of. The codes above are in the order that they appear in the string.*)

The codes of the ASCII table have later been increased to 8 bits to become the Extended ASCII table. Extended ASCII table has 256 distinct codes.

18.1.2 IBM Code Pages

IBM Corporation has defined a set of tables, each one of which assign the codes of the Extended ASCII table from 128 to 255 to one or more writing systems. These code tables allowed supporting the letters of many more alphabets. For example, the special letters of the Turkish alphabet are a part of IBM's code page 857.

Despite being much useful than ASCII, code pages have some problems and limitations: In order to display text correctly, it must be known what code page a given text has originally been written in. This is because the same code corresponds to a different character in most other tables. For example, the code that represents \tilde{G} in table 857 corresponds to ª in table 437.

Another problem is the limitation of the number of alphabets that can be supported in a single document. Additionally, alphabets that have more than 128 non-ASCII characters cannot be supported by an IBM table.

18.1.3 ISO/IEC 8859 Code Pages

These code pages are a result of international standardization efforts. They are similar to IBM's code pages in how they assign codes to characters. As an example, the special letters of the Turkish alphabet appear in code page 8859-9. These tables have the same problems and limitations of IBM's tables. For example, the Dutch digraph ij does not appear in any of these tables.

18.1.4 Unicode

Unicode solves all of the problems and limitations of previous solutions. Unicode includes more than 100 thousand characters and symbols of the writing systems of many human languages, current and old. (New ones are constantly under review for addition to the table.) Each of these characters has a unique code. Documents that are encoded in Unicode can include all of the characters of separate writing systems at the same time without any confusion or limitation.

18.2 Unicode encodings

Unicode assigns a unique code for each character. Since there are more Unicode characters than an 8-bit value can hold, some characters must be represented by at least two 8-bit values. For example, the Unicode character code of Ğ (286) is greater than the maximum value that an 8-bit type can hold (255).

Note: I will arbitrarily use 1 and 30 as the values of the 2 bytes that represent Ğ below. They are not valid in any Unicode encoding; but are useful when introducing Unicode encodings. The accurate encoding values are out of the scope of this chapter.

The way characters are represented in electronic mediums is called *encoding*. We have seen above how the string "hello" is encoded in ASCII. We will now see three Unicode encodings that correspond to D's character types.

UTF-32: This encoding uses 32 bits (4 bytes) for every Unicode character. The UTF-32 encoding of "hello" is similar to its ASCII encoding, but every character is represented by 4 bytes:

```
0, 0, 0, 104, 0, 0, 0, 101, 0, 0, 0, 108, 0, 0, 0, 108, 0, 0, 0, 111
```

As another example, the UTF-32 encoding of "aĞ" is:

```
0, 0, 0, 97, 0, 0, 1, 30
```

Note: The actual values of some of these bytes are different, and the actual order of the bytes may be different.

a and Ğ are represented by 1 and 2 significant bytes respectively, and the values of the other 5 bytes are all zeros. These zeros can be thought of as filler bytes to make every Unicode character occupy 4 bytes each.

For documents based on the basic Latin alphabet, this encoding always uses 4 times as many bytes as the ASCII encoding. When most of the characters of a given document have ASCII

equivalents, the 3 filler bytes for each of those characters would make this encoding more wasteful compared to other encodings.

On the other hand, there are benefits of every character being represented by an equal number of bytes.

UTF-16: This encoding uses 16 bits (2 bytes) to represent most of the Unicode characters. Since 16 bits can have about 65 thousand unique values, the other 35 thousand Unicode characters must be represented by more bytes.

As an example, "aĜ" is encoded by 4 bytes in UTF-16:

```
0, 97, 1, 30
```

Note: The actual values of some of these bytes are different, and the actual order of the bytes may be different.

Compared to UTF-32, this encoding takes less space for most documents, but because there are characters that are represented by more than 2 bytes, UTF-16 is more complicated to process.

UTF-8: This encoding uses 1 or more bytes for every character. If a character has an equivalent in the ASCII table, it is represented by 1 byte and by the same code as in the ASCII table. The rest of the Unicode characters are represented by 2, 3, or 4 bytes. Most of the special characters of the European writing systems are among the group of characters that are represented by 2 bytes.

For most documents, UTF-8 is the encoding that takes the least amount of space. Another benefit of UTF-8 is that the documents that have already been encoded in ASCII directly match their UTF-8 encodings. UTF-8 does not waste any space: every character is represented by significant bytes. As an example, the UTF-8 encoding of "aĜ" is:

```
97, 1, 30
```

Note: The actual values of some of these bytes are different, and the actual order of the bytes may be different.

18.3 The character types of D

There are three D types to represent characters. These characters correspond to the three Unicode encodings mentioned above. As you remember from the Fundamental Types chapter:

Type	Definition	Initial Value
char	UTF-8 code unit	0xFF
wchar	UTF-16 code unit	0xFFFF
dchar	UTF-32 code unit and Unicode code point	0x0000FFFF

Compared to some other programming languages, characters in D may not consist of equal number of bytes. For example, because Ĝ can be represented by at least 2 bytes in Unicode, it doesn't fit in a variable of type **char**. On the other hand, because **dchar** consists of 4 bytes, it can hold any Unicode character.

Despite providing these useful types, D does not support some esoteric features of Unicode. I will explain this below.

18.4 Character literals

Literals are constant values that are typed in the program as a part of the source code. In D, character literals are specified within single quotes:

```
char letter_a = 'a';
wchar letter_e_acute = 'é';
```

Double quotes are not valid for characters because double quotes are used when specifying *strings*, which we will see two chapters later. 'a' is a character literal and "a" is a string literal that consists of one character.

Variables of type **char** can only hold letters that are in the ASCII table.

There are many ways of inserting characters in code:

- Most naturally, typing them on the keyboard.
- Copying from another program or another text. For example, you can copy and paste from a web site, or from a program that is specifically for displaying Unicode characters. (One such program in most Linux environments is *Character Map*, *charmap* on the console.)
- Using short names of the characters. The syntax for this is `\&character_name;`. For example, the name of the Euro character is **euro** and it can be specified in the program like the following:

```
wchar currencySymbol = '\&euro;';
```

See [the list of named characters](#) for all of the characters that can be specified this way.

- Specifying characters by their integer Unicode values:

```
char a = 97;
wchar Ğ = 286;
```

- Specifying the codes of the characters of the ASCII table either by `\value_in_octal` or `\xvalue_in_hexadecimal` syntax:

```
char questionMarkOctal = '\77';
char questionMarkHexadecimal = '\x3f';
```

- Specifying the Unicode values of the characters by using the `\ufour_digit_value` syntax for **wchar**, and the `\Ueight_digit_value` syntax for **dchar**. (Note **u** vs. **U**.) The Unicode values must be specified in hexadecimal:

```
wchar Ğ_w = '\u011e';
dchar Ğ_d = '\U0000011e';
```

These methods can be used to specify the characters within strings as well. For example, the following two lines have the same string literals:

```
writeln("Résumé preparation: 10.25€");
writeln("\x52\xeacute;sum\u00e9 preparation: 10.25\&euro;");
```

18.5 Control characters

Some characters only affect the format of the text, they don't have a visual representation themselves. For example, the *new-line* character, which specifies that the output should

continue on a new line, does not have a visual representation. Such characters are called *control characters*. Control characters are specified with the `\control_character` syntax.

Syntax	Name	Definition
<code>\n</code>	new line	Moves the printing to a new line
<code>\r</code>	carriage return	Moves the printing to the beginning of the current line
<code>\t</code>	tab	Moves the printing to the next tab stop

For example, the `\n` character can make even `write` start new lines, which does not start a new line automatically. Every occurrence of the `\n` control character within the following literal represents the start of a new line:

```
write("first line\nsecond line\nthird line\n");
```

The output:

```
first line
second line
third line
```

18.6 Single quote and backslash

The single quote character itself cannot be written within single quotes because the compiler would take the second one as the closing character of the first one: `' '`. The first two would be the opening and closing quotes, and the third one would be left alone, causing a compilation error.

Similarly, since the backslash character has a special meaning in the control character and literal syntaxes, the compiler would take it as the start of such a syntax: `\'`. The compiler then would be looking for a closing single quote character, not finding one, and emitting a compilation error.

To avoid these confusions, the single quote and the backslash characters are *escaped* by a preceding backslash character:

Syntax	Name	Definition
<code>\'</code>	single quote	Allows specifying the single quote character: <code>'\''</code>
<code>\\</code>	backslash	Allows specifying the backslash character: <code>'\\'</code> or <code>"\\"</code>

18.7 The `std.uni` module

The `std.uni` module includes functions that are useful with Unicode characters. You can see these functions at [its documentation](#).

The functions that start with `is` answer certain questions about characters: the result is `false` or `true` depending on whether the answer is no or yes, respectively. These functions are useful in logical expressions:

- `isLower`: is a lowercase character?
- `isUpper`: is an uppercase character?
- `isAlpha`: is a Unicode alphanumeric character (generally, a letter or a numeral)?
- `isWhite`: is a whitespace character?

The functions that start with **to** produce new characters from existing characters:

- **toLower**: produces the lowercase of the given character
- **toUpper**: produces the uppercase of the given character

Here is a program that uses all of these functions:

```
import std.stdio;
import std.uni;

void main()
{
    writeln("Is ğ lowercase? ", isLower('ğ'));
    writeln("Is Ş lowercase? ", isLower('Ş'));

    writeln("Is İ uppercase? ", isUpper('İ'));
    writeln("Is ç uppercase? ", isUpper('ç'));

    writeln("Is z alphanumeric? ", isAlpha('z'));
    writeln("Is € alphanumeric? ", isAlpha('€'));

    writeln("Is new-line whitespace? ", isWhite('\n'));
    writeln("Is underline whitespace? ", isWhite('_'));

    writeln("The lowercase of Ğ: ", toLower('Ğ'));
    writeln("The lowercase of İ: ", toLower('İ'));

    writeln("The uppercase of ş: ", toUpper('ş'));
    writeln("The uppercase of ı: ", toUpper('ı'));
}
```

The output:

```
Is ğ lowercase? true
Is Ş lowercase? false
Is İ uppercase? true
Is ç uppercase? false
Is z alphanumeric? true
Is € alphanumeric? false
Is new-line whitespace? true
Is underline whitespace? false
The lowercase of Ğ: ğ
The lowercase of İ: i
The uppercase of ş: Ş
The uppercase of ı: I
```

18.8 Limited support for ı and i of the Turkish alphabet

The lowercase and uppercase versions of the letters ı and i are consistently dotted or undotted in the Turkish alphabet. Most other alphabets are inconsistent in this regard: the uppercase of the dotted i is undotted I.

Because the computer systems have started with the ASCII table, traditionally the uppercase of i is I, and the lowercase of I is i. For that reason, these two letters need special attention. The following program demonstrates this problem:

```
import std.stdio;
import std.uni;

void main()
```

```
{
    writeln("The uppercase of i: ", toUpper('i'));
    writeln("The lowercase of I: ", toLower('I'));
}
```

The output is according to the basic Latin alphabet:

```
The uppercase of i: I
The lowercase of I: i
```

18.8.1 Every alphabet has limited support

Characters are converted between their uppercase and lowercase versions normally by their Unicode character codes. This method is problematic for many alphabets. For example, the Azeri and Celt alphabets are subject to the same problem of producing the lowercase of I as i.

There are similar problems with sorting. Letters of many alphabets, like the Turkish ğ, are all sorted after z. Even accented characters like á are sorted after z even for the basic Latin alphabet.

18.8.2 The trileri library

A D project by the members of the Ddili Forum that addresses these problems is [the trileri library](#). At the time of writing this chapter, trileri supports the English, Turkish, Azeri, Turkmen, Kırım Tatar, Gagavuz, Kumuk, Hakas, Kurdish, and Celt alphabets. It sorts the letters of these alphabets and converts their letters to and from uppercase and lowercase correctly.

You can test the trileri library at the following URL:

http://www.ddili.org/cgi-bin/trileri_deneme

18.9 Problems with reading characters

The flexibility and power of D's Unicode characters may cause unexpected results when reading characters from an input stream. This contradiction is due to the multiple meanings of the term *character*. Before expanding on this further, let's look at a program that has this problem:

```
import std.stdio;

void main()
{
    char letter;
    write("Please enter a letter: ");
    readf(" %s", &letter);
    writeln("The letter that has been read: ", letter);
}
```

If you try that program in an environment that does not use Unicode, you may see that even the non-ASCII characters are read and printed correctly.

On the other hand, if you start the same program in a Unicode environment, for example in a Linux console, you may see that the character printed on the output is not the same character that has been entered. To see this, let's enter a non-ASCII character in a console that uses the UTF-8 encoding (like most Linux consoles):

```
Please enter a letter: ğ
The letter that has been read: ← no letter on the output
```

The reason for this problem is that the non-ASCII characters like ğ are represented by two codes, and reading a `char` from the input reads only the first one of those codes. Since that single `char` is not sufficient to represent the whole Unicode character, the console does not display the *incomplete* character.

To show that the UTF-8 codes that make up a character are read one `char` at a time, let's read two `char` variables and print them one after the other:

```
import std.stdio;

void main()
{
    char firstCode;
    char secondCode;

    write("Please enter a letter: ");
    readf(" %s", &firstCode);
    readf(" %s", &secondCode);

    writeln("The letter that has been read: ",
           firstCode, secondCode);
}
```

The program reads two `char` variables from the input and prints them in the same order that they are read. When those codes are sent to the console in that same order, they complete the UTF-8 encoding of the Unicode character on the console and this time the Unicode character is printed correctly:

```
Please enter a letter: ğ
The letter that has been read: ğ ← the Unicode character that
consists of two char codes
```

These results are also related to the fact that the standard inputs and outputs of programs are `char` streams.

We will see later in the Strings chapter that it is easier to read characters as strings, instead of dealing with UTF codes individually.

18.10 D's Unicode support

Unicode is a large and complicated standard. D supports a very useful subset of it.

A Unicode-encoded document consists of the following levels of concepts from the lowermost to the uppermost:

- **Code unit:** The values that make up the UTF encodings are called code units. Depending on the encoding and the characters themselves, Unicode characters are made up of one or more code units. For example, in the UTF-8 encoding the letter a is made up of a single code unit and the letter ğ is made up of two code units.
D's character types `char`, `wchar`, and `dchar` correspond to UTF-8, UTF-16, and UTF-32 code units, respectively.
- **Code point:** Every letter, numeral, symbol, etc. that the Unicode standard defines is called a code point. For example the Unicode code values of a and ğ are two distinct code points.

Depending on the encoding, code points are represented by one or more code units. As mentioned above, in the UTF-8 encoding a is represented by a single code unit, and ģ is represented by two code units. On the other hand, both a and ģ are represented by a single code unit in both UTF-16 and UTF-32 encodings.

The D type that supports code points is **dchar**. **char** and **wchar** can only be used as code units.

- **Character**: Any symbol that the Unicode standard defines and what we call "character" in daily talk is a character.

This definition of character is flexible in Unicode, and this brings a complication.

Some characters can be formed by more than one code point. For example, the letter ģ can be specified in two ways:

- as the single code point ģ
- as the two code points g and ˘ (combining breve)

D does not natively support the character concept of combined code points. In D, the single code point ģ is different from the two consecutive code points g and ˘.

18.11 Summary

- Unicode supports all of the characters of all of the writing systems.
- **char** is for the UTF-8 encoding; although it is not suitable to represent characters in general, it supports the ASCII table.
- **wchar** is for the UTF-16 encoding; although it is not suitable to represent characters in general, it can support letters of multiple alphabets.
- **dchar** is for the UTF-32 encoding; because it is 32 bits, it can also be used as code points.

19 Slices and Other Array Features

We have seen in the [Arrays chapter](#) how elements are grouped as a collection in an array. That chapter has intentionally been brief, leaving most of the features of arrays to this chapter.

Before going any further, here are brief definitions of some of the terms that happen to be close in meaning:

- **Array:** The general concept of a group of elements that are located side by side and are accessed by indexes.
- **Fixed-length array (static array):** Array with a fixed number of elements. This type of array owns its elements.
- **Dynamic array:** Array that can gain or lose elements. This type of array provides access to elements that are owned by the D runtime environment.
- **Slice:** Another name for *dynamic array*.

When I write *slice* I will specifically mean a slice; and when I write *array*, I will mean either a slice or a fixed-length array, with no distinction.

19.1 Slices

Slices are the same feature as dynamic arrays. They are called *dynamic arrays* for being used like arrays, and are called *slices* for providing access to portions of other arrays. They allow using those portions as if they are separate arrays.

Slices are defined by the *number range* syntax that correspond to the indexes that specify the beginning and the end of the range:

```
beginning_index .. one_beyond_the_end_index
```

In the number range syntax, the beginning index is a part of the range but the end index is outside of the range:

```
/* ... */ = monthDays[0 .. 3]; // 0, 1, and 2 are included; but not 3
```

Note: Number ranges are different from Phobos ranges. Phobos ranges are about struct and class interfaces. We will see these features in later chapters.

As an example, we can *slice* the `monthDays` array to be able to use its parts as four smaller arrays:

```
int[12] monthDays =
    [ 31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31 ];

int[] firstQuarter = monthDays[0 .. 3];
int[] secondQuarter = monthDays[3 .. 6];
int[] thirdQuarter = monthDays[6 .. 9];
int[] fourthQuarter = monthDays[9 .. 12];
```

The four variables in the code above are slices; they provide access to four parts of an already existing array. An important point worth stressing here is that those slices do not have their own elements. They merely provide access to the elements of the actual array. Modifying an element of a slice modifies the element of the actual array. To see this, let's modify the first elements of each slice and then print the actual array:

```

firstQuarter[0] = 1;
secondQuarter[0] = 2;
thirdQuarter[0] = 3;
fourthQuarter[0] = 4;

writeln(monthDays);

```

The output:

```
[1, 28, 31, 2, 31, 30, 3, 31, 30, 4, 30, 31]
```

Each slice modifies its first element, and the corresponding element of the actual array is affected.

We have seen earlier that valid array indexes are from 0 to one less than the length of the array. For example, the valid indexes of a 3-element array are 0, 1, and 2. Similarly, the end index in the slice syntax specifies one beyond the last element that the slice will be providing access to. For that reason, when the last element of an array needs to be included in a slice, the length of the array must be specified as the end index. For example, a slice of all of the elements of a 3-element array would be `array[0..3]`.

An obvious limitation is that the beginning index cannot be greater than the end index:

```

int[3] array = [ 0, 1, 2 ];
int[] slice = array[2 .. 1]; // ← runtime ERROR

```

It is legal to have the beginning and the end indexes to be equal. In that case the slice is empty. Assuming that `index` is valid:

```

int[] slice = anArray[index .. index];
writeln("The length of the slice: ", slice.length);

```

The output:

```
The length of the slice: 0
```

19.2 \$, instead of array.length

When indexing, `$` is a shorthand for the length of the array:

```

writeln(array[array.length - 1]); // the last element
writeln(array[$ - 1]);           // the same thing

```

19.3 .dup to copy

Short for "duplicate", the `.dup` property makes a new array from the copies of the elements of an existing array:

```

double[] array = [ 1.25, 3.75 ];
double[] theCopy = array.dup;

```

As an example, let's define an array that contains the number of days of the months of a leap year. A method is to take a copy of the non-leap-year array and then to increment the element that corresponds to February:

```
import std.stdio;

void main()
{
    int[12] monthDays =
        [ 31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31 ];

    int[] leapYear = monthDays.dup;

    ++leapYear[1]; // increments the days in February

    writeln("Non-leap year: ", monthDays);
    writeln("Leap year   : ", leapYear);
}
```

The output:

```
Non-leap year: [31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31]
Leap year   : [31, 29, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31]
```

19.4 Assignment

We have seen so far that the assignment operator *modifies* values of variables. It is the same with fixed-length arrays:

```
int[3] a = [ 1, 1, 1 ];
int[3] b = [ 2, 2, 2 ];

a = b; // the elements of 'a' become 2
writeln(a);
```

The output:

```
[2, 2, 2]
```

The assignment operation has a completely different meaning for slices: It makes the slice start providing access to new elements:

```
int[] odds = [ 1, 3, 5, 7, 9, 11 ];
int[] evens = [ 2, 4, 6, 8, 10 ];

int[] slice; // not providing access to any elements yet

slice = odds[2 .. $ - 2];
writeln(slice);

slice = evens[1 .. $ - 1];
writeln(slice);
```

Above, **slice** does not provide access to any elements when it is defined. It is then used to provide access to some of the elements of **odds**, and later to some of the elements of **evens**:

```
[5, 7]
[4, 6, 8]
```

19.5 Making a slice longer may terminate sharing

Since the length of a fixed-length array cannot be changed, the concept of *termination of sharing* is only about slices.

It is possible to access the same elements by more than one slice. For example, the first two of the eight elements below are being accessed through three slices:

```
import std.stdio;

void main()
{
    int[] slice = [ 1, 3, 5, 7, 9, 11, 13, 15 ];
    int[] half = slice[0 .. $ / 2];
    int[] quarter = slice[0 .. $ / 4];

    quarter[1] = 0;    // modify through one slice

    writeln(quarter);
    writeln(half);
    writeln(slice);
}
```

The effect of the modification to the second element of **quarter** is seen through all of the slices:

```
[1, 0]
[1, 0, 5, 7]
[1, 0, 5, 7, 9, 11, 13, 15]
```

When viewed this way, slices provide *shared* access to elements. This sharing opens the question of what happens when a new element is added to one of the slices. Since the slices provide access to the same actual elements, there may not be room to add elements to a slice without disturbing the others.

D answers this question by terminating the sharing relationship if there is no room for the new element: The slice that has no room to grow leaves the sharing. When this happens, all of the existing elements of that slice are copied to a new place automatically and the slice starts providing access to these new elements.

To see this in action, let's add an element to **quarter** before modifying its second element:

```
quarter ~= 42;    // this slice leaves the sharing because
                 // there is no room for the new element

quarter[1] = 0;  // for that reason this modification
                 // does not affect the other slices
```

The output of the program shows that the modification to the **quarter** slice does not affect the others:

```
[1, 0, 42]
[1, 3, 5, 7]
[1, 3, 5, 7, 9, 11, 13, 15]
```

Explicitly increasing the length of a slice makes it leave the sharing as well:

```
++quarter.length;    // leaves the sharing
```

or

```
quarter.length += 5;    // leaves the sharing
```

On the other hand, shortening a slice does not affect sharing. Shortening the slice merely means that the slice now provides access to less elements:

```
int[] a = [ 1, 11, 111 ];
int[] d = a;

d = d[1 .. $]; // shortening from the beginning
d[0] = 42;     // modifying the element through the slice

writeln(a);    // printing the other slice
```

As can be seen from the output, the modification through **d** is seen through **a**; the sharing is still in effect:

```
[1, 42, 111]
```

Reducing the length in different ways does not terminate sharing either:

```
d = d[0 .. $ - 1]; // shortening from the end
--d.length;       // same thing
d.length = d.length - 1; // same thing
```

Sharing of elements is still in effect.

19.5.1 Seemingly indeterministic sharing and capacity()

There is a confusing effect of the termination of sharing: It may seem to not happen for some of the slices, and sometimes. This is because when there is room at the end of the longest slice, the sharing does not terminate:

```
slice ~= 42; // adding to the longest slice
slice[1] = 0;
```

As seen in the output, although the added element increases the length of a slice, the sharing has not been terminated, and the modification is seen through all of the slices:

```
[1, 0]
[1, 0, 5, 7]
[1, 0, 5, 7, 9, 11, 13, 15, 42]
```

This seeming indeterminism can be removed by using the **capacity()** function. **capacity()** provides the number of elements that a slice can hold without needing to be copied. The number of new elements that a slice can hold can be calculated by subtracting the actual length of the slice from this capacity. There is no room for new elements when the length of the slice equals its capacity:

```
writeln("Capacity: ", capacity(slice));
writeln("In use: ", slice.length);
writeln("Extra room: ", capacity(slice) - slice.length);
```

The output:

Capacity: 15
In use: 8
Extra room: 7

As the output shows, the longest slice in the previous program can take seven more elements before needing to be copied to a larger place. In other words, the slice can take seven more elements before the sharing is terminated; adding one more element after that would terminate its sharing relationship with the other slices.

19.6 Operations on all of the elements

This feature is for both fixed-length arrays and slices.

The `[]` characters written after the name of an array means *all of the elements*. This feature simplifies the program when certain operations need to be applied to all of the elements of an array.

Note: *The D compiler that I have used when writing this chapter (dmd 2.056) does not support this feature fully. For that reason I have used only fixed-length arrays in some of the following examples.*

```
import std.stdio;

void main()
{
    double[3] a = [ 10, 20, 30 ];
    double[3] b = [ 2, 3, 4 ];

    double[3] result = a[] + b[];

    writeln(result);
}
```

The output:

```
[12, 23, 34]
```

The addition operation in that program is applied to the corresponding elements of both of the arrays in order: First the first elements are added, then the second elements are added, etc. A natural requirement is that the lengths of the two arrays must be equal.

The operator can be one of the arithmetic operators `+`, `-`, `*`, `/`, `%`, and `^^`; one of the binary operators `^`, `&`, and `|`; as well as the unary operators `-` and `~` that are typed in front of an array. We will see some of these operators in later chapters.

The assignment versions of these operators can also be used: `=`, `+=`, `-=`, `*=`, `/=`, `%=`, `^^=`, `=`, `&=`, and `|=`.

This feature is not only between two arrays; in addition to an array, an expression that is compatible with the elements can also be used. For example, the following operation divides all of the elements of an array by four:

```
double[3] a = [ 10, 20, 30 ];
a[] /= 4;

writeln(a);
```

The output:

```
[2.5, 5, 7.5]
```

To assign all of the elements to a specific value:

```
a[] = 42;
writeln(a);
```

The output:

```
[42, 42, 42]
```

This feature requires great attention when used with slices. Although there is no apparent difference in element values, the following two expressions have very different meanings:

```
slice2 = slice1;    // ← slice2 starts providing access
                   //   to the same elements that
                   //   slice1 provides access to

slice3[] = slice1;  // ← the values of the elements of
                   //   slice3 change
```

The assignment of `slice2` makes it share the same elements as `slice1`. On the other hand, as `slice3[]` means *all elements of slice3*, the values of its elements become the same as the values of the elements of `slice1`. The effect of presence or absence of the `[]` characters cannot be ignored.

We can see an example of this difference in the following program:

```
import std.stdio;

void main()
{
    double[] slice1 = [ 1, 1, 1 ];
    double[] slice2 = [ 2, 2, 2 ];
    double[] slice3 = [ 3, 3, 3 ];

    slice2 = slice1;    // ← slice2 starts providing access
                       //   to the same elements that
                       //   slice1 provides access to

    slice3[] = slice1;  // ← the values of the elements of
                       //   slice3 change

    writeln("slice1 before: ", slice1);
    writeln("slice2 before: ", slice2);
    writeln("slice3 before: ", slice3);

    slice2[0] = 42;    // ← the value of an element that
                     //   it shares with slice1 changes

    slice3[0] = 43;    // ← the value of an element that
                     //   only it provides access to
                     //   changes

    writeln("slice1 after : ", slice1);
    writeln("slice2 after : ", slice2);
    writeln("slice3 after : ", slice3);
}
```

The modification through `slice2` affects `slice1` too:


```

slice1 before: [1, 1, 1]
slice2 before: [1, 1, 1]
slice3 before: [1, 1, 1]
slice1 after  : [42, 1, 1]
slice2 after  : [42, 1, 1]
slice3 after  : [43, 1, 1]

```

The danger here is that the potential bug may not be noticed until after the value of a shared element is changed.

19.7 Multi-dimensional arrays

So far we have used arrays with only fundamental types like `int` and `double`. The element type can actually be any other type, including other arrays. This enables the programmer to define complex containers like *array of arrays*. Arrays of arrays are called *multi-dimensional arrays*.

All of the elements of all of the arrays that we have defined so far have been written in the source code from left to right. To help us understand the concept of a two-dimensional array, let's define an array from top to bottom this time:

```

int[] array = [
    10,
    20,
    30,
    40
];

```

As you remember, most spaces in the source code are to help with readability and do not change the meaning of the code. The array above could have been defined on a single line and would have the same meaning.

Let's now replace every element of that array with another array:

```

/* ... */ array = [
    [ 10, 11, 12 ],
    [ 20, 21, 22 ],
    [ 30, 31, 32 ],
    [ 40, 41, 42 ]
];

```

We have replaced elements of type `int` with elements of type `int[]`. To make the code conform to array definition syntax, we must now specify the type of the elements as `int[]` instead of `int`:

```

int[][] array = [
    [ 10, 11, 12 ],
    [ 20, 21, 22 ],
    [ 30, 31, 32 ],
    [ 40, 41, 42 ]
];

```

Such arrays are called *two-dimensional arrays* because they can be seen as having lines and columns.

Two-dimensional arrays are used the same way as any other array as long as we remember that each element is an array itself and is used in array operations:

```
array ~= [ 50, 51 ]; // adds a new element (i.e. a slice)
array[0] ~= 13;     // adds to the first element
```

The new state of the array:

```
[[10, 11, 12, 13], [20, 21, 22], [30, 31, 32], [40, 41, 42], [50, 51]]
```

Both the array and its elements can be fixed-length as well. The following is a three-dimensional array where all dimensions are fixed-length:

```
int[2][3][4] array; // 2 columns, 3 lines, 4 pages
```

The definition above can be seen as *four pages of three lines of two columns*. As an example, such an array can be used to represent a 4-story building in an adventure game, each consisting of $2 \times 3 = 6$ rooms.

For example, the number of items in the first room of the second floor can be incremented like the following:

```
// The index of the second floor is 1, and the first room
// of that floor is accessed by [0][0]
++itemCounts[1][0][0];
```

In addition to the syntax above, the **new** expression can also be used to create a *slice of slices*. The following example uses only two dimensions:

```
import std.stdio;

void main()
{
    int[][] s = new int[][](2, 3);
    writeln(s);
}
```

The **new** expression above creates 2 slices containing 3 elements each and returns a slice that provides access to those slices and elements. The output:

```
[[0, 0, 0], [0, 0, 0]]
```

19.8 Summary

- Fixed-length arrays own their elements; slices provide access to elements that don't belong to them.
- Within the `[]` operator, `$` is the equivalent of `array_name.length`.
- `.dup` makes a new array that consists of the copies of the elements of an existing array.
- With fixed-length arrays, the assignment operation changes the values of elements; with slices, it makes the slices start providing access to other elements.
- Slices that get longer *may* stop sharing elements and start providing access to newly copied elements. `capacity()` determines whether this will be the case.
- The syntax `array[]` means *all of the elements*; the operation that is applied to it is applied to each element individually.
- Arrays of arrays are called multi-dimensional arrays.

19.9 Exercise

- Iterate over the elements of an array of **double**s and halve the ones that are greater than 10. For example, given the following array:

```
double[] array = [ 1, 20, 2, 30, 7, 11 ];
```

Modify it as the following:

```
[1, 10, 2, 15, 7, 5.5]
```

Although there are many solutions of this problem, try to use only the features of slices. You can start with a slice that provides access to all of the elements. Then you can shorten the slice from the beginning and always use the first element.

The following expression shortens the slice from the beginning:

```
slice = slice[1 .. $];
```

... the solution

20 Strings

We have used strings in many programs that we have seen so far. Strings are a combination of the two features that we have covered in the last three chapters: characters and arrays. In the simplest definition, strings are nothing but arrays of characters. For example, `char[]` is a type of string.

This simple definition may be misleading. As we have seen in the [Characters chapter](#), D has three separate character types. Arrays of these character types lead to three separate string types, some of which may have surprising outcomes in some string operations.

20.1 `readln` and `chomp`, instead of `readf`

There are surprises even when reading strings from the console.

Being character arrays, strings can contain control characters like `'\n'` as well. When reading strings from the input, the control character that corresponds to the Enter key that is pressed at the end of console input becomes a part of the string as well. Further, because there is no way to tell `readf()` how many characters to read, it continues to read until the end of the entire input. For these reasons, `readf()` does not work as intended when reading strings:

```
import std.stdio;

void main()
{
    char[] name;

    write("What is your name? ");
    readf(" %s", &name);

    writeln("Hello ", name, "!");
}
```

The Enter key that the user presses after the name does not terminate the input. `readf()` continues to wait for more characters to add to the string:

```
What is your name? Mert
  ← The input is not terminated although Enter has been pressed
  ← (Let's assume that Enter is pressed a second time here)
```

One way of terminating the standard input stream in a console is pressing Ctrl-D under Unix-based systems and Ctrl-Z under Windows systems. If the user eventually terminates the input that way, we see that the new-line characters have been read as parts of the string as well:

```
Hello Mert
  ← new-line character after the name
! ← (one more before the exclamation mark)
```

The exclamation mark appears after those characters instead of being printed right after the name.

`readln()` is more suitable when reading strings. Short for "read line", `readln()` reads until the end of the line. It is used differently because the `" %s"` format string and the `&` operator are not needed:

```
import std.stdio;

void main()
{
    char[] name;

    write("What is your name? ");
    readln(name);

    writeln("Hello ", name, "!");
}
```

`readln()` stores the new-line character as well. This is so that the program has a way of determining whether the input consisted of a complete line or whether the end of input has been reached:

```
What is your name? Mert
Hello Mert
! ← new-line character before the exclamation mark
```

Such control characters that are at the ends of strings can be removed by `std.string.chomp`:

```
import std.stdio;
import std.string;

void main()
{
    char[] name;

    write("What is your name? ");
    readln(name);
    name = chomp(name);

    writeln("Hello ", name, "!");
}
```

The `chomp()` expression above returns a new string that does not contain the trailing control characters. Assigning that return value back to `name` produces the intended output:

```
What is your name? Mert
Hello Mert! ← no new-line character
```

`readln()` can be used without a parameter. In that case it *returns* the line that it has just read. Chaining the result of `readln()` to `chomp()` enables a shorter and more readable syntax:

```
string name = chomp(readln());
```

I will start using that form after introducing the `string` type below.

20.2 Double quotes, not single quotes

We have seen that single quotes are used to define character literals. String literals are defined with double quotes. `'a'` is a character; `"a"` is a string that contains a single character.

20.3 string, wstring, and dstring are immutable

There are three string types that correspond to the three character types: `char[]`, `wchar[]`, and `dchar[]`.

There are three *aliases* of the *immutable* versions of these types: `string`, `wstring`, and `dstring`. The characters of the variables that are defined by these aliases cannot be modified. For example, the characters of a `wchar[]` can be modified but the characters of a `wstring` cannot be modified. (We will see D's *immutability* concept in later chapters.)

For example, the following code that tries to capitalize the first letter of a `string` would cause a compilation error:

```
string canNotBeMutated = "hello";
canNotBeMutated[0] = 'H';           // ← compilation ERROR
```

We may think of defining the variable as a `char[]` instead of the `string` alias but that cannot be compiled either:

```
char[] a_slice = "hello"; // ← compilation ERROR
```

This time the compilation error is due to the combination of two facts:

1. The types of string literals like `"hello"` are `string`, not `char[]` i.e. they are immutable.
2. The `char[]` on the left-hand side is a slice, which, if the code compiled, would provide access to all of the characters of the right-hand side.

Since `char[]` is mutable and `string` is not, there is a mismatch. The compiler does not allow accessing characters of an immutable array by a mutable slice.

The solution here is to take a copy of the immutable string by the `.dup` property:

```
import std.stdio;

void main()
{
    char[] s = "hello".dup;
    s[0] = 'H';
    writeln(s);
}
```

The program can now be compiled and print the modified string:

```
Hello
```

Similarly, `char[]` cannot be used where a `string` is needed. In such cases, the `.idup` property can be used for producing an immutable `string` variable from a mutable `char[]` variable. For example, if `s` is a variable of type `char[]`, the following line cannot be compiled:

```
string result = s ~ '.';           // ← compilation ERROR
```

When the type of `s` is `char[]`, the type of the expression on the right-hand side of the assignment above is `char[]` as well. `.idup` is used for producing immutable strings from existing strings:

```
string result = (s ~ '.').idup;    // ← now compiles
```

20.4 Potentially confusing length of strings

We have seen that some Unicode characters are represented by more than one byte. For example, the letter `é` is represented by two bytes. This fact is reflected in the `.length` property of strings:

```
writeln("résumé".length);
```

Although "résumé" contains six *letters*, the length of the string is the number of *characters* that it contains:

```
8
```

The type of the elements of string literals like "hello" are `char` and `char` represents a UTF-8 code unit. A problem that this may cause is when we try to replace a two-code-unit letter with a single-code-unit letter:

```
char[] s = "résumé".dup;
writeln("Before: ", s);
s[1] = 'e';
s[5] = 'e';
writeln("After : ", s);
```

The two 'e' characters do not replace the two letters `é`; they replace single code units, resulting in an incorrect UTF-8 encoding:

```
Before: résumé
After : re@sueé ← INCORRECT
```

When dealing with letters, symbols, and other Unicode characters directly as in the code above, the correct type to use is `dchar`:

```
dchar[] s = "résumé"d.dup;
writeln("Before: ", s);
s[1] = 'e';
s[5] = 'e';
writeln("After : ", s);
```

The output:

```
Before: résumé
After : resume
```

Please note the two differences in the new code:

1. The type of the string is `dchar[]`.
2. There is a `d` at the end of the literal "résumé"`d`, specifying its type as an array of `dchars`.

20.5 String literals

The optional character that is specified after string literals determines the type of the elements of the string:

```
import std.stdio;

void main()
```

```

{
    string s = "résumé"c; // same as "résumé"
    wstring w = "résumé"w;
    dstring d = "résumé"d;

    writeln(s.length);
    writeln(w.length);
    writeln(d.length);
}

```

The output:

```

8
6
6

```

Because all of the letters of "résumé" can be represented by a single **wchar** or **dchar**, the last two lengths are equal to the number of letters.

20.6 String concatenation

Since they are actually arrays, all of the array operations can be applied to strings as well. `~` concatenates two strings and `~=` appends to an existing string:

```

import std.stdio;
import std.string;

void main()
{
    write("What is your name? ");
    string name =.chomp(readln());

    // Concatenate:
    string greeting = "Hello " ~ name;

    // Append:
    greeting ~= "! Welcome...";

    writeln(greeting);
}

```

The output:

```

What is your name? Can
Hello Can! Welcome...

```

20.7 Comparing strings

Note: Unicode does not define how the characters are ordered other than their Unicode codes. For that reason, you may get results that don't match with your expectations below. When alphabetical ordering is important, you may want to use a library like [triler](#) that supports the concept of an alphabet.

We have used comparison operators `<`, `>=`, etc. with integer and floating point values before. The same operators can be used with strings as well, but with a different meaning: strings are ordered *lexicographically*. This ordering takes each character's Unicode code to be its place in a hypothetical grand Unicode alphabet. The concepts of *less* and *greater* are replaced with *before* and *after* in this hypothetical alphabet:


```

import std.stdio;
import std.string;

void main()
{
    write("    Enter a string: ");
    string s1 =.chomp(readln());

    write("Enter another string: ");
    string s2 =.chomp(readln());

    if (s1 == s2) {
        writeln("They are the same!");
    } else {
        string former;
        string latter;

        if (s1 < s2) {
            former = s1;
            latter = s2;
        } else {
            former = s2;
            latter = s1;
        }

        writeln("", former, " comes before ", latter, ".");
    }
}

```

Because Unicode adopts the letters of the basic Latin alphabet from the ASCII table, the strings that contain only the letters of the ASCII table appear to be ordered correctly.

20.8 Lowercase and uppercase are different

Because each letter has a unique code, every letter is different from every other. For example, 'A' and 'a' are different letters.

Additionally, as a consequence of their ASCII code values, all of the uppercase letters are sorted before all of the lowercase letters. For example, 'B' is before 'a'. The `icmp()` function of the `std.string` module can be used when strings need to be compared regardless of lowercase and uppercase. You can see the functions of this module at [its online documentation](#).

Because strings are arrays (and as a corollary, *ranges*), the functions of the `std.array`, `std.algorithm`, and `std.range` modules are very useful with strings as well.

20.9 Exercises

1. Browse the documentations of the `std.string`, `std.array`, `std.algorithm`, and `std.range` modules.
2. Write a program that makes use of the `~` operator: The user enters the first name and the last name, all in lowercase letters. Produce the full name that contains the proper capitalization of the first and last names. For example, when the strings are "ebru" and "domates" the program should print "Ebru Domates".
3. Read a line from the input and print the part between the first and last 'e' letters of the line. For example, when the line is "this line has five words" the program should print "e has five".

You find the `indexOf()` and `lastIndexOf()` functions useful to get two indexes to produce a slice.

As it is indicated in their documentation, the return types of `indexOf()` and `lastIndexOf()` are not `int` nor `size_t`, but `sizediff_t`. You may have to define variables of that exact type:

```
sizediff_t first_e = indexOf(line, 'e');
```

It is possible to define variables shorter with the `auto` keyword, which we will see in a later chapter:

```
auto first_e = indexOf(line, 'e');
```

... the solutions

21 Redirecting Standard Input and Output Streams

All of the programs that we have seen so far have interacted through `stdin` and `stdout`, the standard input and output streams. Input and output functions like `readf` and `writeln` operate on these streams by default.

While using these streams, we have made an assumption that the standard input is from the keyboard and the standard output is to the screen.

We will start writing programs that deal with files in later chapters. We will see that just like standard input and output streams, files are character streams as well; so they are used in almost the same way as `stdin` and `stdout`.

Before seeing how files are accessed from within programs, I would like to show how standard inputs and outputs of programs can be redirected to files, and can be piped to other programs. Programs can print to files instead of the screen, and read from files instead of the keyboard. Although these features are not directly related to programming languages, they are useful tools that are available in all modern consoles.

21.1 Redirecting standard output to a file with >

When starting the program from the console, typing a `>` character and a file name at the end of the command line redirects the standard output of that program to the specified file. Everything that the program prints to its standard output would be written to that file instead.

Let's test this with a program that reads a floating point number from its input, multiplies that number by two, and prints the result to its standard output:

```
import std.stdio;

void main()
{
    double number;
    readf(" %s", &number);

    writeln(number * 2);
}
```

If the name of the program is `by_two`, its output would be written to a file named `by_two_result.txt` when the program is started on the command line as in the following line:

```
./by_two > by_two_result.txt
```

For example, if we enter `1.2` at the console, the result `2.4` would appear in `by_two_result.txt`. (*Note: Although the program does not display a prompt like "Please enter a number", it still expects a number to be entered.*)

21.2 Redirecting standard input from a file with <

Similar to redirecting standard output by `>`, standard input can be redirected from a file by `<`. In this case, the program reads from the specified file instead of from the keyboard.

To test this, let's use a program that calculates one tenth of a number:

```
import std.stdio;

void main()
{
    double number;
    readf(" %s", &number);

    writeln(number / 10);
}
```

Assuming that the file `by_two_result.txt` still exists and contains `2.4` from the previous output, and that the name of the new program is `one_tenth`, we can redirect the new program's standard input from that file as in the following line:

```
./one_tenth < by_two_result.txt
```

This time the program would read from `by_two_result.txt` and print the result to the console as `0.24`.

21.3 Redirecting both standard streams

> and < can be used at the same time:

```
./one_tenth < by_two_result.txt > one_tenth_result.txt
```

This time the standard input would be read from `by_two_result.txt` and the standard output would be written to `one_tenth_result.txt`.

21.4 Piping programs with |

Note that `by_two_result.txt` is an intermediary between the two programs; `by_two` writes to it and `one_tenth` reads from it.

The `|` character pipes the standard output of the program that is on its left-hand side to the standard input of the program that is on its right-hand side without the need for an intermediary file. For example, when the two programs above are piped together as in the following line, they collectively calculate *one fifth* of the input:

```
./by_two | ./one_tenth
```

First `by_two` runs and reads a number from its input. (Remember that although it does not prompt for one, it still waits for a number.)

The result of `by_two` would appear at the standard input of `one_tenth`, which in turn would calculate and print one tenth of that result.

21.5 Exercise

- Pipe more than one program:

```
./one | ./two | ./three
```

... [the solution](#)

22 Files

We have seen in the previous chapter that standard input and output streams can be redirected to and from files and other programs with the `>`, `<`, and `|` operators on the console. Despite being very powerful, these tools are not suitable in every situation because in many cases programs can not complete their tasks by reading from their input and writing to their output like a simple pipe.

For example, a program that deals with student records may use its standard output to display the program menu. Such a program would need to write the student records to an actual file instead of `stdout`.

In this chapter, we will cover reading from and writing to files of file systems.

22.1 Fundamental concepts

Files are represented by the `File` struct of the `std.stdio` module. Since I haven't introduced structs yet, we will have to accept the syntax of struct construction as-is for now.

Before getting to code samples we have to go through fundamental concepts about files.

22.1.1 The producer and the consumer

Files that are created on one platform may not be readily usable on other platforms. Merely opening a file and writing data in it may not be sufficient for that data to be available on the consumer's side. The producer and the consumer of the data must have already agreed on the format of the data that is in the file. For example, if the producer has written the id and the name of the student records in a certain order, the consumer must read the data back in the same order.

Additionally, the sample codes below do not write BOM to the beginning of the file. This may make your files incompatible with systems that require BOM. (BOM is short for *byte order mark*. It specifies in what order the UTF code units of characters are arranged in a file.)

22.1.2 Access rights

File systems present files to programs under certain access rights. Access rights are important for both data integrity and performance.

When it is about reading, allowing multiple programs read from the same file would improve performance because no program would be waiting for any other to finish their reading. On the other hand, when it is about writing, only one program should be allowed to write to a file, otherwise programs may overwrite each other's data.

22.1.3 Opening a file

The standard input and output streams `stdin` and `stdout` are already *open* when programs start running. They are ready to be used.

On the other hand, files must first be opened by specifying the name of the file and the access rights that are needed.

22.1.4 Closing a file

Any file that has been opened by a program must be closed when the program finishes using that file. In most cases the files need not be closed explicitly; they are closed automatically when **File** objects are terminated automatically:

```
if (aCondition) {
    // Assume a File object has been created and used here.
    // ...
} // ← The actual file would be closed automatically here
// when leaving this scope. No need to close explicitly.
```

In some cases a file object may need to be re-opened to access a different file or the same file with different access rights. In such cases the file must be closed and re-opened:

```
file.close();
file.open("student_records", "r");
```

22.1.5 Writing to and reading from files

Since files are character streams, input and output functions **writeln**, **readf**, etc. are used exactly the same way with them. The only difference is that the name of the **File** variable and a dot must be typed:

```
writeln("hello"); // writes to standard output
stdout.writeln("hello"); // same as above
file.writeln("hello"); // writes to the specified file
```

22.1.6 eof() to determine the end of a file

The **eof()** member function determines whether the end of a file has been reached while reading from a file. It returns **true** if the end of the file has been reached.

For example, the following loop will be active until the end of the file:

```
while (!file.eof()) {
    /* ... */
}
```

22.1.7 The std.file module

The **std.file** module contains functions and types that are useful when working with contents of directories. For example, **exists** can be used to determine whether a file or a directory exists on the file systems:

```
import std.file;
// ...
if (exists(fileName)) {
    // there is a file or directory under that name
} else {
    // no file or directory under that name
}
```

22.2 std.stdio.File struct

File uses the same mode characters that are used by **fopen** of the C programming language:

Mode	Definition
r	read access the file is opened to be read from the beginning
r+	read and write access the file is opened to be read from and written at the beginning
w	write access if the file does not exist, it is created as empty if the file already exists, its contents are cleared
w+	read and write access if the file does not exist, it is created as empty if the file already exists, its contents are cleared
a	append access if the file does not exist, it is created as empty if the file already exists, its contents are preserved and it is opened to be written at the end
a+	read and append access if the file does not exist, it is created as empty if the file already exists, its contents are preserved and the file is opened to be read from the beginning and written at the end

A 'b' character may be added to the mode string as in "rb". This may have an effect on platforms that support the *binary mode*, but it is ignored on all POSIX systems.

The **File** struct is included in the [std.stdio module](#).

22.2.1 Writing to a file

The file must have been opened in one of the write modes first:

```
import std.stdio;

void main()
{
    File file = File("student_records", "w");

    file.writeln("Name : ", "Zafer");
    file.writeln("Number: ", 123);
    file.writeln("Class : ", "1A");
}
```

As you remember from the [Strings chapter](#), the type of literals like "student_records" is **string**, consisting of immutable characters. For that reason, it is not possible to construct **File** objects by a mutable file name (e.g. **char[]**). When needed, it would be necessary to call the **.idup** property of the mutable string.

The program above creates or overwrites the contents of a file named **student_records** in the directory that it has been started under (in the program's *working directory*).

Note: File names can contain any character that is legal for that file system. To be portable, I will use only ASCII characters.

22.2.2 Reading from a file

The file must have been opened in one of the read modes first:

```
import std.stdio;
import std.string;

void main()
{
    File file = File("student_records", "r");

    while (!file.eof()) {
        string line =.chomp(file.readln());
        writeln("read line -> |", line);
    }
}
```

The program above reads all of the lines of the file named **student_records** and prints those lines to its standard output.

22.3 Exercise

- Write a program that takes a file name from the user, opens that file, and writes all of the non-empty lines of that file to another file. The name of the new file can be based on the name of the original file. For example, if the original file is **foo.txt**, the new file can be **foo.txt.out**.

[... the solution](#)

23 auto and typeof

23.1 auto

When defining **File** variables in the previous chapter, we have repeated the name of the type on both sides of the = operator:

```
File file = File("student_records", "w");
```

It feels redundant. It would also be cumbersome and error-prone especially when the type name were longer:

```
VeryLongTypeName var = VeryLongTypeName(/* ... */);
```

Fortunately, the type name on the left-hand side is not necessary because the compiler can infer the type of the left-hand side from the expression on the right-hand side. For the compiler to infer the type, the **auto** keyword can be used:

```
auto var = VeryLongTypeName(/* ... */);
```

auto can be used with any type even when the type is not spelled out on the right-hand side:

```
auto duration = 42;
auto distance = 1.2;
auto greeting = "Hello";
auto vehicle = BeautifulBicycle("blue");
```

Although "auto" is the abbreviation of *automatic*, it does not come from *automatic type inference*. It comes from *automatic storage class*, which is a concept about life times of variables. **auto** is used when there no other specifier is appropriate. For example, the following definition that does not use **auto**:

```
immutable i = 42;
```

The compiler infers the type of **i** as **immutable int** above. (We will see **immutable** in a later chapter.)

23.2 typeof

typeof provides the actual type of expressions (including single variables, objects, literals, etc.).

The following is an example of how **typeof** can be used to specify a type without explicitly spelling it out:

```
int value = 100;           // already defined as 'int'
typeof(value) value2;    // means "type of value"
typeof(100) value3;      // means "type of literal 100"
```

The last two variable definitions above are the equivalent of the following:

```
int value2;
int value3;
```

It is obvious that **typeof** is not needed in situations like above when actual types are known. This keyword is especially useful in template definitions, which we will see in later chapters.

23.3 Exercise

- As we have seen above, the type of literals like 100 is **int** (as opposed to **short**, **long**, or any other type). Write a program to determine the type of floating point literals like 1.2. **typeof** and **.stringof** would be useful in this program.

... the solution

24 Name Space

Any name is accessible from the point that it has been defined at to the point that its scope has been exited, as well as in all of the scopes that its scope includes. In this regard, every scope defines a *name space*.

Names are not available once their scope has been exited:

```
void main()
{
    int outer;

    if (aCondition) { // ← curly bracket starts a new scope
        int inner = 1;
        outer = 2;    // ← works; 'outer' is available here
    } // ← 'inner' is not available beyond this point

    inner = 3; // ← compilation ERROR
              // ← 'inner' is not available in the outer scope
}
```

Because **inner** is defined within the scope of the **if** condition it is available only in that scope. On the other hand, **outer** is available in both the outer scope and the inner scope.

It is not legal to define the same name in an inner scope:

```
size_t length = oddNumbers.length;

if (aCondition) {
    size_t length = primeNumbers.length; // ← compilation ERROR
}
```

24.1 Defining names closest to their first use

As we have been doing in all of the programs so far, variables must be defined before their first use:

```
writeln(number); // ← compilation ERROR
                // ← number is not known yet
int number = 42;
```

For that code to be accepted by the compiler, **number** must be defined before it is used with **writeln**. Although there is no restriction on how many lines earlier it should be defined, it is accepted as good programming practice that variables be defined closest to where they are first used.

Let's see this in a program that prints the average of the numbers that it takes from the user. Programmers who are experienced in some other programming languages may be used to defining variables tops of scopes:

```
int count; // ← HERE
int[] numbers; // ← HERE
double averageValue; // ← HERE

write("How many numbers are there? ");

readf(" %s", &count);

if (count >= 1) {
    numbers.length = count;
```

```

    // ... assume the calculation is here ...
} else {
    writeln("ERROR: You must enter at least one number!");
}

```

Contrast the code above to the one below that defines the variables later, as each variable actually starts taking part in the program:

```

write("How many numbers are there? ");

int count; // ← HERE
readf(" %s", &count);

if (count >= 1) {
    int[] numbers; // ← HERE
    numbers.length = count;

    double averageValue; // ← HERE

    // ... assume that the calculation is here ...
} else {
    writeln("ERROR: You must enter at least one number!");
}

```

Although defining all of the variables at top may look better structurally, there are several benefits of defining them as late as possible:

- **Speed:** Every variable definition has a speed cost in the program. As every variable is initialized in D, defining variables at the top would potentially waste their initialization cost if those variables would not actually be used later.
- **Risk of mistakes:** Every line that is between the definition and use of a variable carries a higher risk of programming mistakes. As an example of this, consider a variable named as common as **length**. It is possible to use that variable inadvertently for another length before its intended line. At the time its intended line has been reached, the variable may have an unintended value.
- **Readability:** As the number of lines in a scope are increased, it is more likely that the definition of a variable may become too far up in the source code, forcing the programmer to scroll back to look at its definition for any reason.
- **Code maintenance:** Source code is in constant modification and improvement: new features are added, old features are removed, bugs are fixed, etc. These changes sometimes make it necessary to move a group of lines altogether as a new function. When that happens, having all of the variables defined close to the lines that use them makes it easier to move them as a coherent bunch.

For example, in the latter code above that followed this guideline, all of the lines within the **if** statement can be moved to a new a function in the program.

On the other hand, when the variables are always defined at the top, if the lines ever need to be moved, the variables that are used in those lines must be identified one by one.

25 for Loop

The **for** loop is for the same purpose as [the while loop](#). **for** makes it possible to put the definitions and expressions about the loop's iteration on the same line.

25.1 Sections of the while loop

The **while** loop evaluates the loop condition and continues executing the loop as long as that condition is **true**. For example, a loop to print the numbers between 1 and 11 may check the condition *less than 11*:

```
while (number < 11)
```

Iterating the loop can be achieved by incrementing **number** at the end of the loop:

```
++number;
```

To be compilable as D code, **number** must have been defined before its first use:

```
int number = 1;
```

Finally, there is the actual work within the loop body:

```
writeln(number);
```

These four sections can be defined as in the comments below:

```
int number = 1;           // ← preparation
while (number < 11) {    // ← condition check
    writeln(number);     // ← actual work
    ++number;           // ← iteration
}
```

The sections of the **while** loop are executed in the following order during the iteration of the **while** loop:

```
preparation
condition check
actual work
iteration

condition check
actual work
iteration

...
```

A **break** statement or a thrown exception can terminate the loop as well.

25.2 Sections of the for loop

for brings three of these sections on a single line. They are written within the parentheses of the **for** loop, separated with semicolons. The loop body contains only the actual work:

```
for (/* preparation */; /* condition check */; /* iteration */) {
    /* actual work */
}
```

The same code is more coherent when written as a **for** loop:

```
for (int number = 1; number < 11; ++number) {
    writeln(number);
}
```

The benefit of the **for** loop is more obvious when the loop body has a large number of statements. The expression that increments the loop variable is visible on the **for** line instead of being mixed with the other statements of the loop.

The sections of the **for** loop are executed in the same order as the **while** loop.

The **break** and **continue** statements work exactly the same in the **for** loop.

The only difference between the **while** and **for** loops is the name space of the loop variable. This is explained below.

Although very common, the iteration variable need not be an integer, nor it is modified only by incrementing. For example, the following loop is used to print the halves of the previous floating point values:

```
for (double value = 1; value > 0.001; value /= 2) {
    writeln(value);
}
```

Note: The information above under this title is technically incorrect but does cover the majority of the use cases of the **for** loop, especially by programmers who have a C or C++ background. In reality, D's **for** loop does not have *three sections that are separated by semicolons*. It has two sections, first of which contains the preparation and the loop condition together. Without getting into the details of this syntax, here is how to define two variables of different types in the preparation section:

```
for ({ int i = 0; double d = 0.5; } i < 10; ++i) {
    writeln("i: ", i, ", d: ", d);
    d /= 2;
}
```

Note that the preparation section is the area within the yellow curly brackets and that there is not a semicolon between the preparation section and the condition section.

25.3 The sections may be empty

All three of the **for** loop sections may be left empty:

- Sometimes a special loop variable is not needed, possibly because an already-defined variable would be used.
- Sometimes the loop would be exited by a **break** statement instead of by relying on the loop condition.
- Sometimes the iteration expressions depend on certain conditions that would be checked within the loop body.

When all of the sections are empty, the **for** loop models *forever*:

```
for ( ; ; ) {
    // ...
}
```

Such a loop may be designed to never end or end with a **break** statement.

25.4 The name space of the loop variable

The only difference between the **for** and **while** loop is the name space of the variable defined during loop preparation: The variable is accessible only within the **for** loop, not outside:

```
for (int i = 0; i < 5; ++i) {
    // ...
}

writeln(i); // ← compilation ERROR
            // i is not accessible here
```

On the other hand, since the variable is defined in the name space that contains the **while** loop, the name is accessible even after the loop:

```
int i = 0;

while (i < 5) {
    // ...
    ++i;
}

writeln(i); // ← works; i is accessible here
```

We have seen the guideline of *defining names closest to their first use* in the previous chapter. Similar to the rationale for that guideline, the smaller the name space of a variable the better. In this regard, when the loop variable is not needed outside the loop, **for** is better than **while**.

25.5 Exercises

1. Print the following 9x9 table by using two **for** loops, one inside the other:

```
0,0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8
1,0 1,1 1,2 1,3 1,4 1,5 1,6 1,7 1,8
2,0 2,1 2,2 2,3 2,4 2,5 2,6 2,7 2,8
3,0 3,1 3,2 3,3 3,4 3,5 3,6 3,7 3,8
4,0 4,1 4,2 4,3 4,4 4,5 4,6 4,7 4,8
5,0 5,1 5,2 5,3 5,4 5,5 5,6 5,7 5,8
6,0 6,1 6,2 6,3 6,4 6,5 6,6 6,7 6,8
7,0 7,1 7,2 7,3 7,4 7,5 7,6 7,7 7,8
8,0 8,1 8,2 8,3 8,4 8,5 8,6 8,7 8,8
```

2. Use one or more **for** loops to print the * character as needed to produce geometrical patterns:

```
*  
**  
***  
****  
*****  
*****  
*****  
*****  
*****  
*****
```

```
*****  
*****  
*****  
*****  
*****  
*****  
*****  
*****  
*****  
*****
```

etc.

... the solutions

26 Ternary Operator ?:

The `?:` operator works very similarly to an `if-else` statement:

```
if (/* condition check */) {
    /* ... expression(s) to execute if true */
} else {
    /* ... expression(s) to execute if false */
}
```

The `if` statement executes either the block for the case of `true` or the block for the case of `false`. As you remember, being a statement, it does not have a value; `if` merely affects the execution of code blocks.

On the other hand, the `?:` operator is an expression. In addition to working similarly to the `if-else` statement, it produces a value. The equivalent of the above code is the following:

```
/* condition */ ? /* truth expression */ : /* falsity expression */
```

Because it uses three expressions, the `?:` operator is called the ternary operator.

The value that is produced by this operator is either the value of the truth expression or the value of the falsity expression. Because it is an expression, it can be used anywhere that expressions can be used.

The following examples contrast the `?:` operator to the `if-else` statement. The ternary operator is more concise for the cases that are similar to these examples.

- **Initialization**

To initialize with 366 if leap year, 365 otherwise:

```
int days = isLeapYear ? 366 : 365;
```

With `if`, one way is to define the variable without an explicit initial value and then assign the intended value:

```
int days;
if (isLeapYear) {
    days = 366;
} else {
    days = 365;
}
```

An alternative with `if` is to initialize the variable with the non-leap year value and then increment it if leap year:

```
int days = 365;
if (isLeapYear) {
    ++days;
}
```

- **Printing**

Printing a part of a message differently depending on a condition:

```
writeln("The glass is half ",
        isOptimistic ? "full." : "empty.");
```

With **if**, the first and last parts of the message may be printed separately:

```
write("The glass is half ");
if (isOptimistic) {
    writeln("full.");
} else {
    writeln("empty.");
}
```

Alternatively, the entire message can be printed separately:

```
if (isOptimistic) {
    writeln("The glass is half full.");
} else {
    writeln("The glass is half empty.");
}
```

- **Calculation**

Increasing the score of the winner in a backgammon game 2 or 1 points depending on whether the game has ended with gammon;

```
score += isGammon ? 2 : 1;
```

Straightforward with **if**:

```
if (isGammon) {
    score += 2;
} else {
    score += 1;
}
```

An alternative with **if** is to first increment by one and then increment again if gammon:

```
++score;
if (isGammon) {
    ++score;
}
```

As can be seen from the examples above, the code is more concise and more clear with the ternary operator in certain situations.

26.1 The types of the selection expressions must match

The value of the **?:** operator is either the value of the truth expression or the value of the falsity expression. The types of these two expressions need not be the same but they must match.

From the examples above, the values that would be selected depending on the value of **isLeapYear** were 366 and 365. Both of those values are of type **int** and do match.

To see a compilation error due to the values of the expressions not matching, let's look at composing a message that reports the number of items to be shipped. Let's print "A dozen" when the value equals 12: "A dozen items will be shipped." Otherwise, let's have the message include the exact number: "3 items will be shipped."

One might think that the varying part of the message can be selected with the `?:` operator:

```
writeln((count == 12) ? "A dozen" : count, // ← compilation ERROR
        " items will be shipped.");
```

The expressions do not match because the type of "A dozen" is **string** and the type of `count` is **int**.

A solution is to convert `count` to **string** as well. The function `toString` from the `std.conv` module produces a **string** value from the specified parameter:

```
import std.conv;
// ...
writeln((count == 12) ? "A dozen" : toString(count),
        " items will be shipped.");
```

Now as both of the selection expressions of the `?:` operator are of **string** type, the code compiles and prints the expected message.

26.2 Exercise

- Have the program read a single **int** value as *the net amount* where a positive value represents a gain and a negative value represents a loss.
The program should print a message that contains "gained" or "lost" depending on whether the amount is positive or negative. For example, "\$100 lost" or "\$70 gained".
Even though it may be more suitable, do not use the **if** statement in this exercise.

... the solution

27 Literals

Programs achieve their tasks by manipulating values of variables and objects. They produce new values and new objects by using them with functions and operators.

Some values need not be produced during the execution of the program; they are written directly into the source code. For example, the floating point value `0.75` and the `string` value `"Total price: "` below are not calculated:

```
discountedPrice = actualPrice * 0.75;
totalPrice += count * discountedPrice;
writeln("Total price: ", totalPrice);
```

Such values that are typed into the source code literally are called literals. We have used many literals in the programs that we have written so far. We will cover all of the types of literals and their syntax rules.

27.1 Integer literals

Integer literals can be written in one of four ways: the decimal system that we use in our daily lives; the hexadecimal and binary systems, which are more suitable in certain computing tasks; and the octal system, which may be needed in very rare cases.

In order to make the code more readable or for any other reason, it is possible to insert `_` characters anywhere among the characters of integer literals. For example, to separate the numbers at three digit intervals as in `1_234_567`. These characters are optional; they are ignored by the compiler.

In the decimal system: The literals are specified by the decimal numerals exactly the same way as we are used to in our daily lives, as in `12`. In this system, the first digit cannot be `0` as that digit is reserved for indicating the octal system in *most other languages*. In D, integer literals cannot start with the digit `0` to prevent bugs that are caused by this subtle difference. This does not preclude `0` on its own: `0` is zero.

In the hexadecimal system: The literals start with `0x` or `0X` and include the numerals of the hexadecimal system: "0123456789abcdef" and "ABCDEF" as in `0x12ab00fe`.

In the octal system: The literals are specified by the `octal` template from the `std.conv` module and include the numerals of the octal system: "01234567" as in `octal!576`.

In the binary system: The literals start with `0b` or `0B` and include the numerals of the binary system: 0 and 1 as in `0b01100011`.

27.1.1 The types of integer literals

Just like any other value, every literal is of a certain type. The types of literals are not specified explicitly as `int`, `double`, etc. The compiler infers the type from the value and syntax of the literal itself.

Although most of the time the types of literals are not important, sometimes the types may not match the expressions that they are used in. In such cases the type must be specified explicitly.

By default, integer literals are inferred to be of type `int`. When the value happens to be too large to be represented by an `int`, the compiler uses the following logic to decide on the type of the literal:

- If the value of the literal does not fit an **int** and it is specified in the decimal system then its type is **long**.
- If the value of the literal does not fit an **int** and it is specified in any other system then the type is tried to be **uint**, **long**, and **ulong**, in that order, depending on which type can accommodate the value.

To see this logic in action, let's try the following program that takes advantage of **typeof** and **stringof**:

```
import std.stdio;

void main()
{
    writeln("\n--- these are written in decimal ---");

    // fits an int, so the type is int
    writeln(      2_147_483_647, "\t\t",
            typeof(2_147_483_647).stringof);

    // does not fit an int and is decimal, so the type is long
    writeln(      2_147_483_648, "\t\t",
            typeof(2_147_483_648).stringof);

    writeln("\n--- these are NOT written in decimal ---");

    // fits an int, so the type is int
    writeln(      0x7FFF_FFFF, "\t\t",
            typeof(0x7FFF_FFFF).stringof);

    // does not fit an int and is not decimal, so the type is uint
    writeln(      0x8000_0000, "\t\t",
            typeof(0x8000_0000).stringof);

    // does not fit a uint and is not decimal, so the type is long
    writeln(      0x1_0000_0000, "\t\t",
            typeof(0x1_0000_0000).stringof);

    // does not fit a long and is not decimal, so the type is ulong
    writeln(      0x8000_0000_0000_0000, "\t\t",
            typeof(0x8000_0000_0000_0000).stringof);
}
```

The output:

```
--- these are written in decimal ---
2147483647          int
2147483648          long

--- these are NOT written in decimal ---
2147483647          int
2147483648          uint
4294967296          long
9223372036854775808      ulong
```

27.1.2 The L suffix

Regardless of the magnitude of the value, if it ends with **L** as in **10L**, the type is **long**.

27.1.3 The U suffix

Regardless of the magnitude of the value, if it ends with **U** as in **10U**, the type is **unsigned**. Lowercase **u** can also be used.

The **L** and **U** specifiers can be used together in any order. For example, **7UL** and **8LU** are both **ulong**.

27.2 Floating point literals

The floating point literals can be specified in either the decimal system as in **1.234** or in the hexadecimal system as in **0x9a.bc**.

In the decimal system: An exponent may be appended after the character **e** or **E** meaning "times 10 to the power of". For example, **3.4e5** means "3.4 times 10 to the power of 5". A **+** character may also be specified before the value of the exponent, but it has no effect. For example **5.6e2** and **5.6e+2** are the same.

The **-** character typed before the value of the exponent changes the meaning to be "divided by 10 to the power of". For example, **7.8e-3** means "7.8 divided by 10 to the power of 3".

In the hexadecimal system: The value starts with either **0x** or **0X** and the parts before and after the point are specified in the numerals of the hexadecimal system. Since **e** and **E** are valid numerals in this system, the exponent is specified by **p** or **P**.

Another difference is that the exponent does not mean "10 to the power of", but "2 to the power of". For example, the **P4** part in **0xabc.defP4** means "2 to the power of 4".

Floating point literals almost always have a point but it may be omitted if there is an exponent specified. For example, **2e3** is a floating point literal with the value 2000.

The value before the point may be omitted if zero. For example, **.25** is a literal having the value "quarter".

The optional **_** characters may be used with floating point literals as well, as in **1_000.5**.

27.2.1 The types of floating point literals

Unless explicitly specified, the type of a floating point literal is **double**. The **f** and **F** specifiers mean **float**, and the **L** specifier means **real**. For example; **1.2** is **double**, **3.4f** is **float**, and **5.6L** is **real**.

27.3 Character literals

Character literals are specified within single quotes as in **'a'**, **'\n'**, **'\x21'**, etc.

As the character itself: The character may be typed directly by the keyboard or copied from a separate text: **'a'**, **'ş'**, etc.

As the character specifier: The character literal may be specified by a backslash character followed by a special letter. For example, the backslash character itself can be specified by **'\'**. Character specifier letters are the following:

Syntax	Definition
\'	single quote
\"	double quote
\?	question mark
\\	backslash
\a	alert (bell sound on some terminals)
\b	delete character
\f	new page
\n	new-line
\r	carriage return
\t	tab
\v	vertical tab

As the extended ASCII character code: Character literals can be specified directly by their codes. The codes can be specified either in the hexadecimal system or in the octal system. When using the hexadecimal system, the literal must start with `\x` and must use two digits for the code, and when using the octal system the literal must start with `\` and have up to three digits. For example, the literals `\x21` and `\41` are both the exclamation point.

As the Unicode character code: When the literal is specified with `u` followed by 4 hexadecimal digits, then its type is `wchar`. When it's specified with `U` followed by 8 hexadecimal digits, then its type is `dchar`. For example, `\u011e` and `\U0000011e` are both the Ğ character, having the type `wchar` and `dchar`, respectively.

As named character entity: Characters that have entity names can be specified by that name by the syntax `&name;`. (See [the Named Character Entities table](#).) For example, `€` is €, `♥` is ♥, and `©` is ©.

27.4 String literals

String literals are a combination of character literals and can be specified in a variety of ways.

27.4.1 Double-quoted string literals

The most common way of specifying string literals is typing their characters within double quotes as in `"hello"`. Individual characters of string literals follow the rules of character literals. For example, the literal `"A4 ka\u011fıt: 3\½TL"` is the same as `"A4 kağıt: 3½TL"`.

27.4.2 Wysiwyg string literals

When string literals are specified within back-quotes, the individual characters of the string do not obey the special syntax rules of character literals. For example, the literal ``c:\nurten`` can be a directory name on the Windows operating system. If it were written within double quotes, the `\n` part would mean the *new-line* character:

```
writeln(`c:\nurten`);
writeln("c:\nurten");
```

```
c:\nurten ← wysiwyg (what you see is what you get)
c:        ← the character literal is taken as new-line
urten
```

Wysiwyg string literals can alternatively be specified within double quotes but prepended with the `r` character: `r"c:\nurten"` is also a wysiwyg string literal.

27.4.3 Hexadecimal string literals

In situations where every character in a string needs to be specified in hexadecimal system, instead of typing `\x` before every one of them, a single `x` character may be typed before the opening double quote. In that case, every character in the string literal is taken to be hexadecimal. Additionally, the string literal may contain spaces, which are ignored by the compiler. For example, `"\x44\x64\x69\x6c\x69"` and `x"44 64 69 6c 69"` are the same string literal.

27.4.4 Delimited string literals

The string literal may contain delimiters that are typed right inside the double quotes. These delimiters are not considered to be parts of the value of the literal. Delimited string literals start with a `q` before the opening double quote. For example, the value of `q".hello."` is `"hello"`; the dots are not parts of the value. As long as it ends with a new-line, the delimiter can have more than one character:

```
writeln(q"MY_DELIMITER
first line
second line
MY_DELIMITER");
```

`MY_DELIMITER` is not a part of the value:

```
first line
second line
```

27.4.5 Token string literals

String literals that start with `q` and that use `{` and `}` as delimiters can contain legal D code:

```
auto str = q{int number = 42; ++number;};
writeln(str);
```

The output:

```
int number = 42; ++number;
```

27.4.6 Types of string literals

By default the type of a string literal is `immutable(char)[]`. An appended `c`, `w`, or `d` character specifies the type of the string explicitly as `immutable(char)[]`, `immutable(wchar)[]`, or `immutable(dchar)[]`, respectively. For example, the characters of `"hello"d` are of type `immutable(dchar)`.

We have seen in the [Strings chapter](#) that these three string types are aliased as `string`, `wstring`, and `dstring` respectively.

27.5 Literals are calculated at compile time

It is possible to specify literals as expressions. For example, instead of writing the total number of seconds in January as `2678400` or `2_678_400`, it is possible to specify it by the terms that make up that value as `60 * 60 * 24 * 31`. The multiplication operations of that expressions do not affect the runtime speed of the program, as the program would be compiled as if `2678400` written instead.

The same applies to string literals. For example, the concatenation operation in `"hello " ~ "world"` would be executed at compile time, not at runtime. The program would be compiled as if the code contains the single string literal `"hello world"`.

27.6 Exercises

1. The following line causes a compilation error:

```
int amount = 10_000_000_000; // ← compilation ERROR
```

Change the program so that the line can be compiled and that `amount` equals ten billions.

2. Write a program that increases the value of a variable and prints it in an infinite loop. Make the value be printed on the same line:

```
Number: 25774 ← always on the same line
```

A special character literal other than `'\n'` may be useful here.

... the solutions

28 Formatted Output

This chapter is about features of Phobos's `std.format` module, not about the core features of the D language.

D's input and output format specifiers are similar to the ones in the C language.

Before going further, I would like to summarize the format specifiers and flags as a reference:

```

Flags (can be used together)
-      flush left
+      print sign
#      print in the alternative way
0      print zero-filled
space  print space-filled

Format Specifiers
s      default
b      binary
d      decimal
o      octal
x,X    hexadecimal
e,E    scientific floating point
f,F    dotted floating point
g,G    as e or f
a,A    hexadecimal floating point

```

We have been using functions like `writeln` with multiple parameters as necessary. The parameters would be converted to their string representations and then sent to the output.

Sometimes this is not sufficient. The output may have to be in a certain format. Let's look at the following code that is used to print items of an invoice:

```

items ~= 1.23;
items ~= 45.6;

for (int i = 0; i != items.length; ++i) {
    writeln("Item ", i + 1, ": ", items[i]);
}

```

The output:

```

Item 1: 1.23
Item 2: 45.6

```

Despite the information being correct, it may be required to be printed in a certain format. For example the dots may have to be lined up and that there must always be two digits after the dot as in the following output:

```

Item 1:    1.23
Item 2:   45.60

```

Formatted output is useful in such cases. The output functions that we have been using so far has counterparts that contain the letter `f` in their names: `writelnf()` and `writelnf()`. The letter `f` is short for *formatted*. The first parameter of these functions is a *format string* that describes how the other parameters should be printed.

For example, the format string that would have `writelnf()` produce the desired output above is the following:

```
writeln("Item %d:%9.02f", i + 1, items[i]);
```

The format string contains regular characters that are passed to the output as-is, as well as special format specifiers that correspond to each parameter that is to be printed. Format specifiers start with the % character and ends with a *format character*. The format string above has two format specifiers: %d and %9.02f.

Every specifier is associated to the following parameters in order. For example, %d is associated with `i + 1` and %9.02f is associated with `items[i]`. Every specifier specifies the format of the parameter that it corresponds to. (Format specifiers may have parameter numbers as well. This will be explained later in the chapter.)

All of the other characters of the format string that are not parts of format specifiers are printed as-is. Such *regular* characters of the format specifier above are highlighted as red as in "Item %d:%9.02f".

Format specifiers consist of six parts most of which are optional. The part named *position* will be explained later below. The other five are the following (**Note:** *The spaces between these parts are inserted here to help with readability; they are not parts of the specifiers.*):

```
% flags width precision format_character
```

The % character at the beginning and the format character at the end are required; the others are optional.

Because % has a special meaning in format strings, when a % needs to be printed as a regular character, it must be typed as %%.

28.1 *format_character*

b: Integer parameter is printed in the binary system.

o: Integer parameter is printed in the octal system.

x and **X:** Integer parameter is printed in the hexadecimal system; with lowercase letters for **x** and uppercase letters for **X**.

d: Integer parameter is printed in decimal system; a negative sign is also printed if it is a signed type and the value is less than zero.

```
writeln("Binary      : %b", value);
writeln("Octal       : %o", value);
writeln("Hexadecimal: %x", value);
writeln("Decimal     : %d", value);
```

```
Binary      : 1100
Octal       : 14
Hexadecimal: c
Decimal     : 12
```

e: Floating point parameter is printed according to the following rules.

- single digit before the dot
- dot if *precision* is non-zero
- digits after the dot, number of which is determined by *precision* (default precision is 6)
- the **e** character (meaning "10 to the power of")

- the - or + character depending on whether the exponent is less than or greater than zero
- the exponent consisting of at least two digits

E: Same as in **e** with the exception of the **E** character at the output instead of **e**.

f and **F**: Floating point parameter is printed in the decimal system; there is at least one digit before the dot and the default precision is 6.

g: Same as in **f** if the exponent is between -5 and *precision*; otherwise as in **e**. *precision* does not specify the number of digits after the dot, but the significant digits of the entire value. If there are no significant digits after the dot, then the dot is not printed. The rightmost zeros after the dot are not printed.

G: Same as in **g** with the exception of **E** or **F** at the output.

a: Floating point parameter is printed in the hexadecimal floating format:

- the characters **0x**
- single hexadecimal digit
- dot if *precision* is non-zero
- digits after the dot, number of which is determined by *precision*; if no *precision* is specified, then as many digits as necessary
- the **p** character (meaning "2 to the power of")
- the - or + character depending on whether the exponent is less than or greater than zero
- the exponent consisting of at least one digit (the exponent of the value 0 is 0)

A: Same as in **a** with the exception of the **0X** and **P** characters at the output.

```
double value = 123.456789;

writefln("with e: %e", value);
writefln("with f: %f", value);
writefln("with g: %g", value);
writefln("with a: %a", value);
```

```
with e: 1.234568e+02
with f: 123.456789
with g: 123.457
with a: 0x1.edd3c07ee0b0bp+6
```

s: The value is printed in the same way as in non-formatted output, according to the type of the parameter:

- **bool** values as **true** or **false**
- integer values same as **%d**
- floating point values same as **%g**
- strings in UTF-8 encoding; *precision* determines the maximum number of bytes to use (remember that in UTF-8 encoding, the number of bytes is not the same as the number of characters; for example the string "ağ" has 2 characters, consisting a total of 3 bytes)
- struct and class objects as the return value of the **toString()** member functions of their types; *precision* determines the maximum number of bytes to use
- arrays as their element values side by side

```
bool b = true;
int i = 365;
double d = 9.87;
string s = "formatted";
auto o = File("test_file", "r");
```

```
int[] a = [ 2, 4, 6, 8 ];

writefln("bool   : %s", b);
writefln("int    : %s", i);
writefln("double: %s", d);
writefln("string: %s", s);
writefln("object: %s", o);
writefln("array  : %s", a);
```

```
bool   : true
int    : 365
double: 9.87
string: formatted
object: File(55738FA0)
array  : [2, 4, 6, 8]
```

28.2 width

This part determines the width of the field that the parameter is printed in. If the width is specified as the character `*`, then the actual width value is read from the next parameter. If width is a negative value, then the `-` flag is assumed.

```
int value = 100;

writefln("In a field of 10 characters:%10s", value);
writefln("In a field of 5 characters :%5s", value);
```

```
In a field of 10 characters:      100
In a field of 5 characters :  100
```

28.3 precision

Precision is specified after a dot in the format specifier. For floating point types, it determines the precision of the printed representation of the values. If the precision is specified as the character `*`, then the actual precision is read from the next parameter (that parameter must be an `int`). Negative precision values are ignored.

```
double value = 1234.56789;

writefln("%.8g", value);
writefln("%.3g", value);
writefln("%.8f", value);
writefln("%.3f", value);
```

```
1234.5679
1.23e+03
1234.56789000
1234.568
```

```
auto number = 0.123456789;
writefln("Number: %.*g", 4, number);
```

```
Sayı: 0.1235
```

28.4 flags

More than one flag can be specified.

`-`: the value is printed left-aligned in its field; this flag cancels the `0` flag

```
int value = 123;

writefln("Normally right-aligned:|%10d|", value);
writefln("Left-aligned          :|%-10d|", value);
```

```
Normally right-aligned:|      123|
Left-aligned          :|123      |
```

`+`: if the value is positive, it is prepended with the `+` character; this flag cancels the `space` flag

```
writefln("No effect for negative values    : %+d", -50);
writefln("Positive value with the + flag   : %+d", 50);
writefln("Positive value without the + flag: %d", 50);
```

```
No effect for negative values    : -50
Positive value with the + flag   : +50
Positive value without the + flag: 50
```

`#`: prints the value in an *alternate* form depending on the *format_character*

- `o`: the first character of the octal value is always printed as `0`
- `x` and `X`: if the value is not zero, it is prepended with `0x` or `0X`
- floating points: a dot is printed even if there are no significant digits after the dot
- `g` and `G`: even the insignificant zero digits after the dot are printed

```
writefln("Octal starts with 0           : %#o", 1000);
writefln("Hexadecimal starts with 0x     : %#x", 1000);
writefln("Contains dot even when unnecessary: %#g", 1f);
writefln("Rightmost zeros are printed    : %#g", 1.2);
```

```
Octal starts with 0           : 01750
Hexadecimal starts with 0x     : 0x3e8
Contains dot even when unnecessary: 1.00000
Rightmost zeros are printed    : 1.20000
```

`0`: the field is padded with zeros (unless the value is `nan` or `infinity`); if *precision* is also specified, this flag is ignored

```
writefln("In a field of 8 characters: %08d", 42);
```

```
In a field of 8 characters: 00000042
```

`space` character: if the value is positive, a space character is prepended to align the negative and positive values

```
writefln("No effect for negative values: % d", -34);
writefln("Positive value with space    : % d", 56);
writefln("Positive value without space : %d", 56);
```

```
No effect for negative values: -34
Positive value with space    : 56
Positive value without space : 56
```

28.5 Positional parameters

We have seen above that the parameters are associated one by one with the specifiers in the format string. It is possible to use position numbers within format specifiers. This enables associating the specifiers with specific parameters. Parameters are numbered in increasing fashion, starting with 1. The parameter numbers are specified immediately after the % character, followed by a \$:

```
% position$ flags width precision format_character
```

An advantage of positional parameters is being able to use the same parameter in more than one places in the same format string:

```
writeln("%1$d %1$x %1$o %1$b", 42);
```

The format string above uses the parameter numbered 1 within four specifiers to print it in decimal, hexadecimal, octal, and binary formats:

```
42 2a 52 101010
```

Another application of positional parameters is supporting multiple human languages. When referred by position numbers, parameters can be moved anywhere within the specific format string for a given human language. For example, the number of students of a given classroom can be printed as in the following:

```
writeln("There are %s students in room %s.", count, room);
```

```
There are 20 students in room 1A.
```

Let's assume that the program must also support Turkish. In this case the format string needs be selected according to the active language. The following method takes advantage of the ternary operator:

```
auto format = (language == "en"
? "There are %s students in room %s."
: "%s sınıfında %s öğrenci var.");

writeln(format, count, room);
```

Unfortunately, when the parameters are associated one by one, the classroom and student count information appear in reverse order in the Turkish message; the room information is where the count should be and the count is where the room should be:

```
20 sınıfında 1A öğrenci var. ← Wrong: means "room 20", and "1A students"!
```

To avoid this, the parameters can be specified by numbers as **1\$** and **2\$** to associate each specifier with the exact parameter:

```

auto format = (language == "en"
? "There are %1$s students in room %2$s."
: "%2$s sınıfında %1$s öğrenci var.");

writeln(format, count, room);

```

Now the parameters appear in the proper order, regardless of the language selected:

```
There are 20 students in room 1A.
```

```
1A sınıfında 20 öğrenci var.
```

28.6 format

Formatted output is available by the `format()` function of the `std.string` module as well. `format()` works the same as `writeln()` but it *returns* the result as a `string` instead of printing it to the output:

```

import std.stdio;
import std.string;

void main()
{
    write("What is your name? ");
    auto name = chomp(readln());

    auto result = format("Hello %s!", name);
}

```

The program can make use of that result in later expressions.

28.7 Exercises

1. Write a program that reads a value and prints it in the hexadecimal system.
2. Write a program that reads a floating point value and prints it as percentage value with two digits after the dot. For example, if the value is 1.2345, it should print `%1.23`.

... the solutions

29 Formatted Input

It is possible to specify the format of the data that is expected at the input. The format specifies both the data that is to be read and the characters that should be ignored.

D's input format specifiers are similar to the ones that are in the C language.

As we have already been using in the previous chapters, the format specifier "`%s`" reads the data according to the type of the variable. For example, as the type of the following variable is **double**, the characters at the input would be read as floating point:

```
double number;
readf(" %s", &number);
```

The format string can contain three types of information:

- **The space character:** Indicates *zero* or more whitespace characters at the input and specifies that all of those characters should be read and ignored.
- **Format specifier:** Similar to the output format specifiers, input format specifiers start with the `%` character and determine the format of the data that is to be read.
- **Any other character:** Indicates the characters that are expected at the input as-is, which should be read and ignored.

The format string makes it possible to select specific information from the input and ignore the others.

Let's have a look at an example that uses all of the three types of information in the format string. Let's assume that the student number and the grade are expected to appear at the input in the following format:

```
number:123 grade:90
```

Let's further assume that the tags **number:** and **grade:** must be ignored. The following format string would *select* the values of number and grade and would ignore the other characters:

```
int number;
int grade;
readf("number:%s grade:%s", &number, &grade);
```

The characters that are indicated in orange in "`number:%s grade:%s`" must appear at the input exactly as specified; `readf()` reads and ignores them.

The single space character that appears in the format string above would cause all of the whitespace characters that appear exactly at that position to be read and ignored.

As the `%` character has a special meaning in format strings, when that character itself needs to be read and ignored, it must be written twice in the format string as `%%`.

Reading a single line of data from the input has been recommended as `chomp(readln())` in the [Strings chapter](#). Instead of that method, a `\n` character at the end of the format string can achieve a similar goal:

```
import std.stdio;
void main()
{
```

```

write("First name: ");
string firstName;
readf(" %s\n", &firstName);    // ← \n at the end

write("Last name : ");
string lastName;
readf(" %s\n", &lastName);    // ← \n at the end

write("Age      : ");
int age;
readf(" %s", &age);

writefln("%s %s (%s)", firstName, lastName, age);
}

```

The `\n` characters at the ends of the format strings when reading `firstName` and `lastName` would cause the new-line characters to be read from the input and to be ignored. However, potential whitespace characters at the ends of the strings may still need to be removed by `chomp()`.

29.1 Format specifier characters

Note: As of this writing, the version of Phobos that has been released with `dmd 2.057` does not support all of these specifiers.

d: Read an integer in the decimal system.

o: Read an integer in the octal system.

x: Read an integer in the hexadecimal system.

f: Read a floating point number.

s: Read according to the type of the variable.

c: Read a single character. This specifier allows reading whitespace characters as well (it cancels the ignore behavior).

For example, if the input contains "23 23 23", the values would be read differently according to the different format specifiers:

```

int number_d;
int number_o;
int number_x;

readf(" %d %o %x", &number_d, &number_o, &number_x);

writeln("Read with %d: ", number_d);
writeln("Read with %o: ", number_o);
writeln("Read with %x: ", number_x);

```

Although the input contains three sets of "23" characters, the values of the variables are different:

```

Read with %d: 23
Read with %o: 19
Read with %x: 35

```

Note: Very briefly, "23" is equal to $2 \times 8 + 3 = 19$ in the octal system and to $2 \times 16 + 3 = 35$ in the hexadecimal system.

29.2 Exercise

- Assume that the input contains the date in the format *year.month.day*. Write a program that prints the number of the month. For example, if the input is **2009.09.30**, the output should be **9**.

... the solution

30 do-while Loop

In the [for Loop chapter](#), we have seen the steps in which the **while loop** is executed:

```
preparation
condition check
actual work
iteration
condition check
actual work
iteration
...
```

The **do-while** loop is very similar to the **while** loop. The difference is that the *condition check* is performed at the end of each iteration of the **do-while** loop, so that the *actual work* is performed at least once:

```
preparation
actual work
iteration
condition check ← at the end of the iteration
actual work
iteration
condition check ← at the end of the iteration
...
```

For example, **do-while** may be more natural in the following program where the user guesses a number, as the user must guess at least once so that the number can be compared:

```
import std.stdio;
import std.random;

void main()
{
    int number = uniform(1, 101);
    writeln("I am thinking of a number between 1 and 100.");
    int guess;
    do {
        write("What is your guess? ");
        readf(" %s", &guess);
        if (number < guess) {
            write("My number is less than that. ");
        } else if (number > guess) {
            write("My number is greater than that. ");
        }
    } while (guess != number);
    writeln("Correct!");
}
```

The function **uniform()** that is used in the program is a part of the **std.random** module. It returns a random number in the specified range. The way it is used above, the second

number is considered to be outside of the range. In other words, **uniform()** would not return 101 for that call.

30.1 Exercise

- Write a program that plays the same game but have the program do the guessing. If the program is written correctly, it should guess the user's number in at most 7 tries.

... the solution

31 Associative Arrays

Associative arrays are a feature that is found in most of the modern high-level languages. They are very fast data structures that work like mini databases and are commonly used in many programs.

We have seen arrays in the [Arrays chapter](#) as containers that store their elements side-by-side and provide access to them by indexes. An array that stores the names of the days of the week can be defined like this:

```
string[] dayNames =
    [ "Monday", "Tuesday", "Wednesday", "Thursday",
      "Friday", "Saturday", "Sunday" ];
```

The name of a specific day can be accessed by its index in that array:

```
writeln(dayNames[1]); // prints "Tuesday"
```

The fact that arrays provide access to elements by index values can be described as an *association* of indexes with elements. In other words, arrays map indexes to elements. Arrays can use only integers as indexes.

Associative arrays allow indexing not only by integers but any type. They map the values of one type to the values of another type. The values of the type that associative arrays *map from* are called *keys*, rather than indexes. They provide access to elements by the key of each element.

31.1 Very fast but unordered

Associative arrays are a *hash table* implementation. Hash tables are among the fastest collections for storing and accessing elements. Other than rare pathological cases, the time it takes to store an element or to access an element is independent of the number of elements that are in the associative array.

The high performance of hash tables is at the expense of their storing the elements in an unordered way. Unlike arrays, the elements of hash tables are not side-by-side. They are not sorted in any way either.

31.2 Definition

The syntax of associative arrays is similar to the array syntax. The difference is that it is the type of the key that is specified within the square brackets, not the length of the array:

```
element_type[key_type] associative_array_name;
```

For example, an associative array that maps day names of type **string** to day numbers of type **int** can be defined like this:

```
int[string] dayNumbers;
```

The **dayNumbers** variable above is an associative array that can be used as a table that provides a mapping from day names to day numbers. In other words, it can be used as the reverse of the **dayNames** array from the beginning of this chapter. We will use the **dayNumbers** associative array in the examples below.

The keys of associative arrays can be of any type including user-defined **struct** and **class** types. We will see user-defined types in later chapters.

The length of associative arrays cannot be specified when defined. They grow automatically as elements are added.

31.3 Adding elements

Assignment operator is sufficient to build the association between a key and a value:

```
// associates element 0 with key "Monday"
dayNumbers["Monday"] = 0;

// associates element 1 with key "Tuesday"
dayNumbers["Tuesday"] = 1;
```

The table grows automatically with each association. For example, **dayNumbers** would have two elements after the operations above. This can be demonstrated by printing the entire table:

```
writeln(dayNumbers);
```

The output indicates that the element values 0 and 1 correspond to key values "Monday" and "Tuesday", respectively:

```
["Monday":0, "Tuesday":1]
```

There can be only one element per key. For that reason, when assigned to an existing element, the table does not grow; instead, the value of the existing element changes:

```
dayNumbers["Tuesday"] = 222; // changing the existing element
writeln(dayNumbers);
```

The output:

```
["Monday":0, "Tuesday":222]
```

31.4 Initialization

Sometimes some of the mappings between the keys and the values are already known at the time of the definition of the associative array. Associative arrays are initialized similarly to arrays, with the difference of a colon separating each key from its element:

```
// key : element
int[string] dayNumbers =
    [ "Monday"   : 0, "Tuesday" : 1, "Wednesday" : 2,
      "Thursday" : 3, "Friday"  : 4, "Saturday"  : 5,
      "Sunday"   : 6 ];

writeln(dayNumbers["Tuesday"]); // prints 1
```

31.5 Removing elements

The element that corresponds to a key is removed by **.remove()**:

```
dayNumbers.remove("Tuesday");
writeln(dayNumbers["Tuesday"]); // ← runtime ERROR
```

The first line above removes the element for the key "Tuesday". Since that element is not in the container anymore, the second line would cause an exception to be thrown and the program to be terminated if that exception is not caught. We will see exceptions in a later chapter.

It is possible to remove all of the elements at once as will be seen in the first exercise below.

31.6 Determining the presence of an element

The `in` operator determines whether the element for a given key exists in the associative array:

```
int[string] colorCodes = [ /* ... */ ];
if ("purple" in colorCodes) {
    // there is an element for key "purple"
} else {
    // no element for key "purple"
}
```

Sometimes it makes sense to use a default element value for keys that do not exist in the associative array. For example, the special value of -1 can be used as the code for colors that are not in `colorCodes`. `.get()` is useful in such cases: It returns the value of the element if the element for the specified key exists, the default value otherwise. The default value is specified as the second parameter of `.get()`:

```
int[string] colorCodes = [ "blue" : 10, "green" : 20 ];
writeln(colorCodes.get("purple", -1));
```

Since the array does not contain an element for the key "purple", `.get()` returns -1:

```
-1
```

31.7 Properties

- `.length` returns the number of elements in the array.
- `.keys` returns copies of all of the keys in the associative array as a dynamic array.
- `.byKey` provides access to the keys in the associative array without copying them; we will see how `.byKey` is used in `foreach` loops in the next chapter.
- `.values` returns copies of all of the values in the associative array as a dynamic array.
- `.byValue` provides access to the elements of the associative array without copying them.
- `.rehash` may make the array more efficient in some cases like after inserting a large number of elements and before actually starting to use the associative array.
- `.sizeof` is the size of the array *reference* (it has nothing to do with the number of elements in the table and is the same value for all associative arrays).
- `.get` returns the element if it exists, the default value otherwise.
- `.remove` removes the specified element from the array.

31.8 Example

Here is a program that prints the Turkish names of colors that are specified in English:

```
import std.stdio;
import std.string;

void main()
{
    string[string] colors = [ "black" : "siyah",
                             "white" : "beyaz",
                             "red"   : "kırmızı",
                             "green" : "yeşil",
                             "blue"  : "mavi"
                           ];

    writeln("I know the Turkish names of these %s colors: %s",
           colors.length, colors.keys);

    write("Please ask me one: ");
    string inEnglish =.chomp(readln());

    if (inEnglish in colors) {
        writeln("\'%s\' is \'%s\' in Turkish.",
              inEnglish, colors[inEnglish]);
    } else {
        writeln("I don't know that one.");
    }
}
```

31.9 Exercises

- How can all of the elements of an associative array be removed? There are at least three methods:

- Removing the elements one-by-one from the associative array.
- Assigning an empty associative array.
- Similar to the previous method, assigning the array's `.init` property.

Note: The `.init` property of any variable or type is the initial value of that type:

```
number = int.init;    // 0 for int
```

- Just like with arrays, there can be only one value for each key. This may be seen as a limitation for some applications.

Assume that an associative array is used for storing student grades. For example, let's assume that the grades 90, 85, 95, etc. are to be stored for the student named "emre".

Associative arrays make it easy to access the grades by the name of the student as in `grades["emre"]`. However, the grades cannot be inserted as in the following code because each grade would overwrite the previous one:

```
int[string] grades;
grades["emre"] = 90;
grades["emre"] = 85;    // ← Overwrites the previous grade!
```

How can you solve this problem? Define an associative array that can store multiple grades per student.

... the solutions

32 foreach Loop

One of the most common statements in D is the **foreach** loop. It is used for applying the same operations to every element of a container (or a *range*).

Operations that are applied to elements of containers are very common in programming. We have seen in the [for Loop chapter](#) that elements of an array are accessed in a **for** loop by an index value that is incremented at each iteration:

```
for (int i = 0; i != array.length; ++i) {
    writeln(array[i]);
}
```

The following are the steps that are involved in iterating over all of the elements:

- Defining a variable as a counter, which is conventionally named as **i**
- Iterating the loop up to the value of the **.length** property of the array
- Incrementing **i**
- Accessing the element

foreach has essentially the same behavior but it simplifies the code by handling those steps automatically:

```
foreach (element; array) {
    writeln(element);
}
```

Part of the power of **foreach** comes from the fact that it can be used the same way regardless of the type of the container. As we have seen in the previous chapter, one way of iterating over the values of an associative array in a **for** loop is by first calling the array's **.values** property:

```
auto values = aa.values;
for (int i = 0; i != values.length; ++i) {
    writeln(values[i]);
}
```

foreach does not require anything special for associative arrays; it is used exactly the same as with arrays:

```
foreach (value; aa) {
    writeln(value);
}
```

32.1 The foreach syntax

foreach consists of three sections:

```
foreach (names; container_or_range) {
    operations
}
```

- **container_or_range** specifies where the elements are.
- **operations** specifies the operations to apply to each element.
- **names** specifies the name of the element and potentially other variables depending on the type of the container or the range. Although the choice of names is up to the

programmer, the number of and the types of these names depend on the type of the container.

32.2 continue and break

These keywords have the same meaning as they have with the **for** loop: **continue** moves to the next iteration before completing the rest of the operations for the current element, and **break** terminates the loop altogether.

32.3 foreach with arrays

When there is a single name specified in the *names* section, it is the value of the element at each iteration:

```
foreach (element; array) {
    writeln(element);
}
```

When two names are specified in the *names* section, they are an automatic counter and the value of the element, respectively:

```
foreach (i, element; array) {
    writeln(i, ":", element);
}
```

The counter is incremented automatically by **foreach**. Its name is up to the programmer.

32.4 foreach with strings and std.range.stride

Since strings are arrays of characters, **foreach** works the same as it works with arrays: Single name is the character, two names are the counter and the character:

```
foreach (c; "hello") {
    writeln(c);
}

foreach (i, c; "hello") {
    writeln(i, ":", c);
}
```

However, being UTF code units, **char** and **wchar** iterate over UTF code units, not Unicode code points:

```
foreach (i, code; "abcçd") {
    writeln(i, ":", code);
}
```

The two UTF-8 code units that make up ç would be accessed as separate elements:

```
0: a
1: b
2: c
3:
4: ç
5: d
```

One way of iterating over Unicode characters of strings in a **foreach** loop is **stride** from the **std.range** module. **stride** presents the string as a container that consists of Unicode characters. It takes the length of its *strides* over the characters as its second parameter:

```
import std.range;

// ...

foreach (c; stride("abcçd", 1)) {
    writeln(c);
}
```

Regardless of the character type of the string, **stride** always presents its elements as Unicode characters:

```
a
b
c
ç
d
```

I will explain below why this loop could not include an automatic counter.

32.5 foreach with associative arrays

A single name specifies the value, two names specify the key and the value:

```
foreach (value; aa) {
    writeln(value);
}

foreach (key, value; aa) {
    writeln(key, ":", value);
}
```

Associative arrays can provide their keys and values as *ranges* as well. We will see ranges in a later chapter. **.byKey()** and **.byValue()** return efficient range objects that are also useful in other contexts. **.byValue()** does not bring any benefit in **foreach** loops over the regular value iteration above. On the other hand, **.byKey()** is the only effective way of iterating over just the keys of an associative array:

```
foreach (key; aa.byKey()) {
    writeln(key);
}
```

32.6 foreach with number ranges

We have seen number ranges in the [Slices and Other Array Features chapter](#) before. It is possible to specify a number range in the *container_or_range* section:

```
foreach (number; 10..15) {
    writeln(number);
}
```

Remember that 10 would be included in the range but 15 would not be.

32.7 foreach with structs, classes, and ranges

foreach can also be used with objects of user types that define their own iteration in **foreach** loops. As the type itself defines its iteration behavior, it is not possible to say anything here. Programmers must refer to the documentation of the particular type.

Structs and classes provide support for **foreach** iteration either by their **opApply()** member functions, or by a set of *range* member functions. We will see these features in later chapters.

32.8 The counter is automatic only for arrays

The automatic counter is provided only when iterating over arrays. When a counter is needed while iterating over other types of containers, the counter must be defined and incremented explicitly:

```
int i;
foreach (element; container) {
    // ...
    ++i;
}
```

Such a variable is also needed when counting a specific condition. For example, the following code counts only the values that are divisible by 10:

```
import std.stdio;

void main()
{
    auto numbers = [ 1, 0, 15, 10, 3, 5, 20, 30 ];

    int count;
    foreach (number; numbers) {
        if ((number % 10) == 0) {
            ++count;
            write(count);

        } else {
            write(' ');
        }

        writeln(": ", number);
    }
}
```

The output:

```
: 1
1: 0
 : 15
2: 10
 : 3
 : 5
3: 20
4: 30
```

32.9 The copy of the element, not the element itself

The **foreach** loop normally provides a copy of the element, not the actual element that is stored in the container. This may be a cause of bugs.

To see an example of this, let's have a look at the following program that is trying to double the values of the elements of an array:

```
import std.stdio;
void main()
{
    double[] numbers = [ 1.2, 3.4, 5.6 ];
    writeln("Before: %s", numbers);
    foreach (number; numbers) {
        number *= 2;
    }
    writeln("After : %s", numbers);
}
```

The output of the program indicates that the assignment made to each element inside the **foreach** body does not have any effect on the elements of the container:

```
Before: [1.2, 3.4, 5.6]
After : [1.2, 3.4, 5.6]
```

That is because **number** is not an actual element of the array, but a copy of each element. When the actual elements need to be operated on, the name must be defined as a *reference* of the actual element by the **ref** keyword:

```
foreach (ref number; numbers) {
    number *= 2;
}
```

The new output shows that the assignments now modify the actual elements of the array:

```
Before: [1.2, 3.4, 5.6]
After : [2.4, 6.8, 11.2]
```

The **ref** keyword makes **number** an *alias* of the actual element at each iteration. As a result, the modifications through **number** modifies that actual element of the container.

32.10 The integrity of the container must be preserved

Although it is fine to modify the elements of the container through **ref** variables, the structure of the container must not be changed. For example, elements must not be removed nor added to the container during a **foreach** loop.

Such modifications may confuse the inner workings of the loop iteration and result in incorrect program states.

32.11 Exercise

- We know that associative arrays provide a mapping from keys to values. This mapping is unidirectional: values are accessed by keys but not the other way around.

Assume that there is already the following associative array:

```
string[int] names = [ 1:"one", 7:"seven", 20:"twenty" ];
```

Use that associative array and a **foreach** loop to fill another associative array named **values**. The new associative array should provide values that correspond to names.

For example, the following line should print 20:

```
writeln(values["twenty"]);
```

... the solution

33 switch and case

switch is a statement that allows comparing the value of an expression against multiple special values. It is similar to but not the same as an "if, else if, else" chain. **case** is used for specifying the values that are to be compared with **switch**'s expression. **case** is a part of the **switch** statement, not a statement itself.

switch takes an expression within parentheses, compares the value of that expression to the **case** values, and executes the operations of the **case** that is equal to the value of the expression. Its syntax consists of a **switch** block that contains one or more **case** sections and a **default** section:

```
switch (expression) {
    case value_1:
        // operations to execute if the expression is equal to value_1
        // ...
        break;

    case value_2:
        // operations to execute if the expression is equal to value_2
        // ...
        break;

    // ... other cases ...

    default:
        // operations to execute if the expression is not equal to any case
        // ...
        break;
}
```

Although it is used in conditional checks, the expression that **switch** takes is not used as a logical expression. It is not evaluated as "if this condition is true" as in an **if** statement. The *value* of the **switch** expression is used in equality comparisons with the **case** values. It is similar to an "if, else if, else" chain that has only equality comparisons:

```
auto value = expression;

if (value == value_1) {
    // operations for value_1
    // ...
} else if (value == value_2) {
    // operations for value_2
    // ...
}

// ... other 'else if's ...

} else {
    // operations for other values
    // ...
}
```

However, the "if, else if, else" above is not the equivalent of the **switch** statement. The reasons will be explained in the following sections.

If a **case** value matches the value of the **switch** expression, then the operations that are under the **case** are executed. If no value matches, then the operations that are under the **default** are executed. The operations are executed until a **break** or a **goto** statement is reached.

33.1 The goto statement

Some uses of **goto** is advised against in most programming languages in general. However, **goto** is useful in **switch** statements in rare situations. The **goto** statement will be covered in more detail in a later chapter.

case does not introduce a *scope* as the **if** statement does. Once the operations within an **if** or **else** scope are finished, the evaluation of the entire **if** statement is also finished. It is not the same with the **case** sections; once a matching **case** is found, the execution of the program jumps to that case and executes the operations under that **case**. When needed in rare situations, **goto case** makes the program execution jump under the next **case**:

```
switch (value) {
case 5:
    writeln("five");
    goto case;    // continues with the next case

case 4:
    writeln("four");
    break;

default:
    writeln("unknown");
    break;
}
```

If **value** is 5, the execution continues under the **case 5** line and the program prints "five". Then the **goto case** statement causes the execution to continue to the next **case**, and as a result "four" is also printed:

```
five
four
```

goto can appear in three ways under **case** sections:

- **goto case** causes the execution to continue to the next **case**.
- **goto default** causes the execution to continue to the **default** section.
- **goto expression** causes the execution to continue to the **case** that matches that expression.

The following program demonstrates these three uses by taking advantage of a **foreach** loop:

```
import std.stdio;

void main()
{
    foreach (value; [ 1, 2, 3, 10, 20 ]) {
        writeln("--- value: %s ---", value);

        switch (value) {
            case 1:
                writeln("case 1");
                goto case;

            case 2:
                writeln("case 2");
                goto case 10;

            case 3:
                writeln("case 3");
                goto default;
        }
    }
}
```

```

        case 10:
            writeln("case 10");
            break;

        default:
            writeln("default");
            break;
    }
}

```

The output:

```

--- value: 1 ---
case 1
case 2
case 10
--- value: 2 ---
case 2
case 10
--- value: 3 ---
case 3
default
--- value: 10 ---
case 10
--- value: 20 ---
default

```

33.2 The expression type must be integer, string, or bool

Any type can be used in equality comparisons in **if** statements. On the other hand, the type of the **switch** expression is limited to integer types, string types, and **bool**.

```

string op = /* ... */;
// ...
switch (op) {
case "add":
    result = first + second;
    break;

case "subtract":
    result = first - second;
    break;

case "multiply":
    result = first * second;
    break;

case "divide":
    result = first / second;
    break;

default:
    throw new Exception(format("Unknown operation: %s", op));
}

```

Note: The code above throws an exception when the operation is not recognized by the program. We will see exceptions in a later chapter.

Although it is possible to use **bool** expressions as well, since **bool** has only two values, it may be more suitable to use an **if** statement or the ternary operator (**?:**) with that type.

33.3 case values must be known at compile time

With the **if** statement, the values that are on both sides of the equality operator can be calculated at runtime. For example, the following **if** statement may be using two values, one of which is calculated by the program and the other entered by the user:

```
if (guess == number) {
    writeln("Correct!");
}
```

On the other hand, although the value of the **switch** expression is normally calculated at runtime, the **case** values must be known at compile time.

33.4 Value ranges

Ranges of values can be specified by **..** between **cases**:

```
switch (dieValue) {
case 1:
    writeln("You won");
    break;

case 2: .. case 5:
    writeln("It's a draw");
    break;

case 6:
    writeln("I won");
    break;

default:
    // The program should never get here because the cases
    // above cover the entire range of valid die values.
    break;
}
```

The code above determines that the game ends in a draw when the die value is 2, 3, 4, or 5.

33.5 Distinct values

Let's assume that it is a draw for the values 2 and 4, rather than for the values that are in the range [2,5]. Distinct values of a **case** are separated by commas:

```
case 2, 4:
    writeln("It's a draw");
    break;
```

33.6 The final switch statement

This statement works similarly to the regular **switch** statement with the following differences:

- It cannot have a **default** section. Note that this section is meaningless when the **case** sections cover the entire range of values anyway, as has been with the six values of the die above.

- Value ranges cannot be used with **cases** (distinct values can be).
- If the expression is of an **enum** type, all of the values of the type must be covered by the **cases** (we will see **enum** types in the next chapter).

```
int dieValue = 1;

final switch (dieValue) {
case 1:
    writeln("You won");
    break;

case 2, 3, 4, 5:
    writeln("It's a draw");
    break;

case 6:
    writeln("I won");
    break;
}
```

33.7 When to use

switch is suitable for comparing the value of an expression against a set of values that are known at compile time.

When there are only two values to compare, an **if** statement may make more sense. For example, to check whether its heads or tails:

```
if (headsTailsResult == heads) {
    // ...
} else {
    // ...
}
```

As a general rule, **switch** is more suitable when there are three or more values to compare.

33.8 Exercises

1. Write a calculator program that supports arithmetic operations. Have the program first read the operation as a **string**, then two values of type **double** from the input. The calculator should print the result of the operation. For example, when the operation and values are "add" and "5 7", respectively, the program should print 12.

The input can be read as in the following code:

```
string op;
double first;
double second;

// ...

op =.chomp(readln());
readf(" %S %S", &first, &second);
```

2. Improve the calculator to support operators like "+" in addition to words like "add".
3. Have the program throw an exception for unknown operators.

... the solutions

34 enum

enum is the feature that enables defining named constant values.

34.1 Effects of magic constants on code quality

The following code appeared in the [exercise solutions](#) of the Integers and Arithmetic Operations chapter:

```
if (operation == 1) {
    result = first + second;
} else if (operation == 2) {
    result = first - second;
} else if (operation == 3) {
    result = first * second;
} else if (operation == 4) {
    result = first / second;
}
```

The integer literals **1**, **2**, **3**, and **4** in that piece of code are called *magic constants*. It is not easy to determine what each of those literals means in the program. One must examine the code in each scope to determine that **1** means *addition*, **2** means *subtraction*, etc. This task is relatively easy for the code above because all of the scopes contain just a single line. It would be considerably more difficult to decipher the meanings of magic constants in most other programs.

Magic constants must be avoided because they reduce two most important qualities of programs: readability and maintainability.

enum enables giving names to such constants and as a consequence making the code more readable and maintainable. Each condition would be readily understandable when the following **enum** values were used:

```
if (operation == Operation.add) {
    result = first + second;
} else if (operation == Operation.subtract) {
    result = first - second;
} else if (operation == Operation.multiply) {
    result = first * second;
} else if (operation == Operation.divide) {
    result = first / second;
}
```

The **enum** type **Operation** above that obviates the need for magic constants **1**, **2**, **3**, and **4** can be defined like this:

```
enum Operation { add = 1, subtract, multiply, divide }
```

34.2 The enum syntax

The simplest form of an **enum** definition is the following:

```
enum TypeName { ValueName_1, ValueName_2, /* etc. */ }
```

Sometimes it is necessary to specify the actual type (the *base type*) of the values as well:

```
enum TypeName : base_type { ValueName_1, ValueName_2, /* etc. */ }
```

We will see how this is used in the next section.

TypeName defines what the values collectively mean. All of the values of an **enum type** are listed within curly brackets. Here are some examples:

```
enum HeadsOrTails { heads, tails }
enum Suit { spades, hearts, diamonds, clubs }
enum Fare { regular, child, student, senior }
```

Each set of values becomes part of a separate type. For example, **heads** and **tails** become values of the type **HeadsOrTails**. The new type can be used like other fundamental types when defining variables:

```
HeadsOrTails result;           // default initialized
auto ht = HeadsOrTails.heads; // inferred type
```

As has been seen in the codes above, the values of **enum** types are always specified by the name of their **enum** type:

```
if (result == HeadsOrTails.heads) {
    // ...
}
```

34.3 Actual values and base types

The values of **enum** types are normally implemented in the background as **int** values. In other words, although they appear as named values as **heads** and **tails** in the code, they are actually **int** values. (**Note:** *It is possible to choose a type other than **int** when needed.*)

Unless explicitly specified by the programmer, the **int** values start by **0** and are incremented by one for each **enum** value. For example, the two values of the **HeadsOrTails** have the values 0 and 1:

```
writeln("heads is 0: ", (HeadsOrTails.heads == 0));
writeln("tails is 1: ", (HeadsOrTails.tails == 1));
```

The output:

```
heads is 0: true
tails is 1: true
```

It is possible to manually reset the values at any point. That has been the case when specifying the value of **Operation.add** as 1 above. The following example resets the values twice:

```
enum Test { a, b, c, ç = 100, d, e, f = 222, g, ğ }
writeln("%d %d %d", Test.b, Test.ç, Test.ğ);
```

The expected output is the following (**Note:** *Outputting **enum** values according to format specifiers like "%d" is expected to be available as of dmd 2.059*):

```
1 100 224
```

If `int` is not suitable as the base type of the `enum` values, the base type can be specified explicitly after the name of the `enum`:

```
enum NaturalConstant : double { pi = 3.14, e = 2.72 }
enum TemperatureUnit : string { C = "Celcius", F = "Fahrenheit" }
```

34.4 enum values that are not of an enum type

We have discussed that it is important to avoid magic constants and instead to take advantage of the `enum` feature.

However, sometimes it may not be natural to come up with `enum` type names just to use named constants. Let's assume that a named constant is needed to represent the number of seconds per day. It should not be necessary to also define an `enum type` for this constant value. All that is needed is a constant value that can be referred to by its name. In such cases, the type of the `enum` and the value parentheses are not specified:

```
enum secondsPerDay = 60 * 60 * 24;
```

The type of the value can be specified explicitly, which would be required if the type cannot be inferred from the right hand side:

```
enum int secondsPerDay = 60 * 60 * 24;
```

Since there is no `enum` type to refer to, such named constants can be used in code simply by their names:

```
totalSeconds = totalDays * secondsPerDay;
```

`enum` can be used for defining named constants of other types as well. For example, the type of the following constant would be `string`:

```
enum fileName = "list.txt";
```

34.5 Properties

The `.min` and `.max` properties are the minimum and maximum values of an `enum` type. When the values of the `enum` type are consecutive, they can be iterated over in a `for` loop within these limits:

```
enum Suit { spades, hearts, diamonds, clubs }

for (auto suit = Suit.min; suit <= Suit.max; ++suit) {
    writefln("%s: %d", suit, suit);
}
```

Format specifiers `"%s"` and `"%d"` have different outcomes (*expected to be available as of dmd 2.059*):

```
spades: 0
hearts: 1
diamonds: 2
clubs: 3
```

Note that a `foreach` loop over that range would leave the `.max` value out of the iteration:


```
foreach (suit; Suit.min .. Suit.max) {
    writeln("%s: %d", suit, suit);
}
```

The output:

```
spades: 0
hearts: 1
diamonds: 2
      ← clubs is missing
```

34.6 Converting from the base type

As has been seen in the formatted outputs above, an **enum** value can automatically be converted to its base type (e.g. to **int**). The reverse conversion is not automatic:

```
Suit suit = 1;      // ← compilation ERROR
```

The reason for this is to avoid ending up with invalid **enum** values:

```
suit = 100;        // ← would be an invalid enum value
```

The values that are known to correspond to valid **enum** values of a particular **enum** type can still be converted to that type by an explicit *type cast*:

```
suit = cast(Suit)1; // now hearts
```

It would be the programmer's responsibility to ensure the validity of the values when an explicit cast is used. We will see type casting and type conversions in later chapters.

34.7 Exercise

- Modify the calculator program from the exercises of the [Integers and Arithmetic Operations chapter](#) by having the user select the arithmetic operation from a menu. This program should be different from the previous one at least in the following areas:
 - Use **enum** values, not magic constants.
 - Use **double** instead of **int**.
 - Use a **switch** statement instead of an "if, else if, else" chain.

... the solution

35 Functions

Similar to how fundamental types are building blocks of program data, functions are building blocks of program behavior.

Functions are also closely related to the craft aspect of programming. The functions that are written by experienced programmers are succinct, simple, and clear. This goes both ways: The mere act of trying to identify and write smaller building blocks of a program makes a better programmer.

We have covered basic statements and expressions in previous chapters. Although there will be many more that we will see in later chapters, what we have seen so far are commonly-used features of D. Still, they are not sufficient on their own to write large programs. The programs that we have written so far have all been very short, each demonstrating just a simple feature of the language. Trying to write a program with any level of complexity without functions would be very difficult and open to bugs.

Functions are features that put statements and expressions together as units of program execution. Such statements and expression are given a name that describes what they collectively achieve. They can then be *called* (or *executed*) by that name.

The concept of *giving names to a group of steps* is common in daily lives. For example, the act of cooking an omelet can be described in some level of detail by the following steps:

- get the pan
- get butter
- get the egg
- turn on the stove
- put the pan on the fire
- put butter into the pan when it is hot
- put the egg into butter when it is melted
- remove the pan from the fire when the egg is cooked
- turn off the stove

Since that much detail is obviously excessive, steps that are related together would be combined under a single name:

- **make preparations** (get the pan, butter, and the egg)
- turn on the stove
- **cook the egg** (put the pan on the fire, etc.)
- turn off the stove

Going further, there can be a single name for all of the steps:

- **make a one-egg omelet** (all of the steps)

Functions are based on the same concept: Steps that can collectively be named as a whole are put together to form a function. As an example, let's start with the following lines of code that achieve the task of printing a menu:

```
writeln(" 0 Exit");
writeln(" 1 Add");
writeln(" 2 Subtract");
writeln(" 3 Multiply");
writeln(" 4 Divide");
```

Since it would make sense to name those lines altogether as **printMenu**, they can be put together to form a function by the following syntax:

```

void printMenu()
{
    writeln(" 0 Exit");
    writeln(" 1 Add");
    writeln(" 2 Subtract");
    writeln(" 3 Multiply");
    writeln(" 4 Divide");
}

```

The contents of that function can now be executed from within `main()` simply by its name:

```

void main()
{
    printMenu();

    /* ... */
}

```

It may be obvious from the similarities of the definitions of `printMenu()` and `main()` that `main()` is a function as well. The execution of D programs start with the function named `main()` and branch out to other functions from there.

35.1 Parameters

Some of the powers of functions come from the fact that their behaviors are adjustable through parameters.

Let's continue with the omelet example by modifying it to make an omelet of five eggs instead of always one. The steps would exactly be the same, the only difference being the number of eggs to use. We can change the more general description above accordingly:

- make preparations (get the pan, butter, and **five eggs**)
- turn on the stove
- cook **the eggs** (put the pan on the fire, etc.)
- turn off the stove

Likewise, the most general single step would become the following:

- make a **five-egg** omelet (all of the steps)

This time there is an additional information that concerns some of the steps: "get five eggs", "cook the eggs", and "make a five-egg omelet".

The behaviors of functions can be adjusted similarly to the omelet example. The information that functions use to adjust their behavior are called *parameters*. Parameters are specified in the *function parameter list*, separated by commas from each other. The parameter list is inside of the parentheses that comes after the name of the function.

The `printMenu()` function above has been defined with an empty parameter list because that function always printed the same menu. Let's assume that sometimes the menu will need to be printed differently in different contexts. For example, it may make more sense to print the first entry as "Return" instead of "Exit" depending on the part of the program that is being executed at that time.

In such a case, the first entry of the menu can be *parameterized* by having been defined in the parameter list. The function then uses the value of that parameter instead of the literal `"Exit"`:

```
void printMenu(string firstEntry)
{
    writeln(" 0 ", firstEntry);
    writeln(" 1 Add");
    writeln(" 2 Subtract");
    writeln(" 3 Multiply");
    writeln(" 4 Divide");
}
```

Notice that since the information that the **firstEntry** parameter conveys is a piece of text, its type has been specified as **string** in the parameter list. This function can now be *called* with different parameter values to print menus having different first entries. All that needs to be done is to use appropriate **string** values depending on where the function is being called from:

```
// At some place in the program:
printMenu("Exit");
// ...
// At some other place in the program:
printMenu("Return");
```

Note: When you write your own functions, you may encounter compilation errors with parameters of type **string**. The **printMenu()** function above has such a problem. As written, it cannot be called by parameter values of type **char[]**. For example, the following code would cause a compilation error:

```
char[] anEntry;
anEntry ~= "Take square root";
printMenu(anEntry); // ← compilation ERROR
```

On the other hand, if **printMenu()** were defined to take its parameter as **char[]**, then it could not be called with **strings** like "Exit". This is related to the concept of immutability and the **immutable** keyword, both of which will be covered in the next chapter.

Let's continue with the menu function and assume that it is not appropriate to always start the menu selection numbers with zero. In that case the starting number can also be passed to the function as its second parameter. The parameters of the function must be separated by commas:

```
void printMenu(string firstEntry, int firstNumber)
{
    writeln(' ', firstNumber, ' ', firstEntry);
    writeln(' ', firstNumber + 1, " Add");
    writeln(' ', firstNumber + 2, " Subtract");
    writeln(' ', firstNumber + 3, " Multiply");
    writeln(' ', firstNumber + 4, " Divide");
}
```

It is now possible to tell the function what number to start from:

```
printMenu("Return", 1);
```

35.2 Calling a function

Starting a function so that it achieves its task is called *calling a function*. The function call syntax is the following:

```
function_name(parameter_values)
```

The actual parameter values that are passed to functions are called *function arguments*. Although the terms *parameter* and *argument* are sometimes used interchangeably in the literature, they signify different concepts.

The arguments are matched to the parameters one by one in the order that the parameters are defined. For example, the last call of `printMenu()` above uses the *arguments* "Return" and `1`, which correspond to the *parameters* `firstEntry` and `firstNumber`, respectively.

The type of each argument must match the type of the corresponding parameter.

35.3 Doing work

In previous chapters, we have defined expressions as entities that do work. Function calls are expressions as well: they do some work. Doing work can mean a combination of two things:

- **Having side effects:** Side effects are any change in the state of the program or its environment. Some operations have only side effects. An example is how the `printMenu()` function above changes `stdout` by printing to it. As another example, a function that adds a `Student` object to a student container would also have a side effect: it would be causing the container to grow.

In summary, operations that cause a change in the state of the program have side effects.

- **Producing a value:** Some operations only produce values. For example, a function that adds numbers would be producing the result of that addition. As another example, a function that makes a `Student` object by using the student's name and address would be producing a `Student` object.
- **Having side effects and producing a value:** Some operations do both. For example, a function that reads two values from `stdin` and calculates their sum would be having side effects due to changing the state of `stdin` and also producing the sum of the two values.
- **No operation:** Although every function is of one of the three categories above, sometimes some functions end up doing no work at all, depending on certain conditions at compile time or at runtime.

35.4 The return value

The value that a function produces as a result of its work is called its *return value*. This term comes from the observation that once the program execution branches into a function, it eventually *returns* back to where the function has been called. Functions get *called* and they *return* values.

Just like any other value, return values have types. The type of the return value is specified right before the name of the function at the point where the function is defined. For example, a function that adds two values of `int` and returns their sum also as `int` would be defined like the following:

```
int add(int first, int second)
{
    // ... the actual work of the function ...
}
```

The value that a function returns takes place of the function call itself. For example, assuming that the function call `add(5, 7)` produces the value `12`, then the following two lines would be equivalent:

```
writeln("Result: ", add(5, 7));
writeln("Result: ", 12);
```

At the first line above, the **add()** function is called with the arguments **5** and **7** before **writeln()** gets called. The value **12** that the function returns is in turn passed to **writeln()** as its second argument.

This allows passing the return values of functions to other functions to form complex expressions:

```
writeln("Result: ", add(5, divide(100, studentCount())));
```

In the line above, the return value of **studentCount()** is passed to **divide()** as its second argument, the return value of **divide()** is passed to **add()** as its second argument, and eventually the return value of **add()** is passed to **writeln()** as its second argument.

35.5 The return statement

The return value of a function is specified by the **return** keyword:

```
int add(int first, int second)
{
    int result = first + second;
    return result;
}
```

A function produces its return value by taking advantage of statements, expressions, and potentially calling other functions. The function would then return that value by the **return** keyword, at which point the execution of the function ends.

It is possible to have more than one **return** statement in a function. The value of the first **return** statement that gets executed determines the return value of the function for a particular call:

```
int complexCalculation(int aParameter, int anotherParameter)
{
    if (aParameter == anotherParameter) {
        return 0;
    }
    return aParameter * anotherParameter;
}
```

The function above returns **0** when the two parameters are equal, and the product of their values when they are different.

35.6 void functions

The return types of functions that do not produce values are specified as **void**. We have seen this many times with the **main()** function so far, as well as the **printMenu()** function above. Since they do not return any value to the caller, their return values have been defined as **void**. (*Note: **main()** can also be defined as returning **int**. We will see this in a later chapter.*)

35.7 The name of the function

The name of a function must be chosen to communicate the purpose of the function clearly. For example, the names `add` and `printMenu` has been appropriate because their purposes have been to add two values, and to print a menu, respectively.

A common guideline for function names is that they contain a verb like *add* or *print*. According to this guideline names like `addition()` and `menu()` would be less than ideal.

However, it is acceptable to name functions simply as nouns, if those functions do not have any side effects. For example, a function that returns the current temperature can be named as `currentTemperature()` instead of `getCurrentTemperature()`.

Coming up with names that are clear, short, and consistent is more on the art side of programming.

35.8 Code quality through functions

Functions can improve the quality of code. Smaller functions with fewer responsibilities lead to programs that are easier to maintain.

35.8.1 Code duplication is harmful

One of the practices that is highly detrimental to program quality is code duplication. Code duplication is when there are more than one piece of code in the program that achieve the same task.

Although this sometimes happens by copying lines of code to a new place, it may also happen by ending up with similar lines of code unintentionally.

One of the problems with code duplication is that potential bugs would naturally occur in all of the copies, requiring to be fixed in more than one place. Conversely, when the code appears in only one place in the program, then fixing it at that one place would get rid of the bug once and for all.

As I have mentioned above, functions are closely related to the craft aspect of programming. Experienced programmers are always on the lookout for code duplication. They always try to identify commonalities in code and move common pieces of code to separate functions (or to common structs, classes, templates, etc. as we will see in later chapters).

Let's start with a program that contains some code duplication. Let's see how that duplication can be removed by moving code into functions (i.e. by *refactoring* the code). The following program reads numbers from the input and prints them first in the order that they have arrived and then in numerical order:

```
import std.stdio;

void main()
{
    int[] numbers;

    int count;
    write("How many numbers are you going to enter? ");
    readf(" %s", &count);

    // Read the numbers
    foreach (i; 0 .. count) {
        int number;
        write("Number ", i, "? ");
        readf(" %s", &number);
    }
}
```

```

    numbers ~= number;
}

// Print the numbers
writeln("Before sorting:");
foreach (i, number; numbers) {
    writefln("%3d:%5d", i, number);
}

numbers.sort;

// Print the numbers
writeln("After sorting:");
foreach (i, number; numbers) {
    writefln("%3d:%5d", i, number);
}
}

```

Some of the duplicated lines of code are obvious in that program. The last two **foreach** loops that are used for printing the numbers are exactly the same. Defining a function that might appropriately be named as **print()** would remove that duplication. The function could take a slice as a parameter and print it:

```

void print(int[] slice)
{
    foreach (i, element; slice) {
        writefln("%3s:%5s", i, element);
    }
}

```

Notice that the name of the parameter is defined to be more generally as **slice** instead of **numbers**. The reason for that is the fact that the function would not know what the elements of the slice would specifically represent. That can only be known at the place where the function has been called from. The elements maybe student IDs, parts of a password, etc. Since that cannot be known in the **print()** function, general names like **slice** and **element** are used in its implementation.

The new function can be called from the two places where the slice needs to be printed:

```

import std.stdio;

void print(int[] slice)
{
    foreach (i, element; slice) {
        writefln("%3s:%5s", i, element);
    }
}

void main()
{
    int[] numbers;

    int count;
    write("How many numbers are you going to enter? ");
    readf(" %s", &count);

    // Read the numbers
    foreach (i; 0 .. count) {
        int number;
        write("Number ", i, "? ");
        readf(" %s", &number);

        numbers ~= number;
    }
}

```



```

// Print the numbers
writeln("Before sorting:");
print(numbers);

numbers.sort;

// Print the numbers
writeln("After sorting:");
print(numbers);
}

```

There is more to do. Notice that there is always a title line printed right before printing the elements of the slice. Although the title is different, the task is the same. If printing the title can be seen as a part of printing the slice, the title too can be passed as a parameter. Here are the new changes:

```

void print(string title, int[] slice)
{
    writeln(title, ":");

    foreach (i, element; slice) {
        writefln("%3s:%5s", i, element);
    }
}

// ...

// Print the numbers
print("Before sorting", numbers);

// ...

// Print the numbers
print("After sorting", numbers);

```

This step has the added benefit of obviating the comments that appear right before the two `print()` calls. Since the name of the function already clearly communicates what it does, those comments are unnecessary:

```

print("Before sorting", numbers);
numbers.sort;
print("After sorting", numbers);

```

Although subtle, there is more code duplication in this program: The values of `count` and `number` are read in exactly the same way. The only difference is the message that is printed to the user and the name of the variable:

```

int count;
write("How many numbers are you going to enter? ");
readf(" %s", &count);

// ...

int number;
write("Number ", i, "? ");
readf(" %s", &number);

```

The code would become even better if it took advantage of a new function that might be named appropriately as `read_int()`. The new function can take the message as a parameter, print that message, read an `int` from the input, and return that `int`:

```
int read_int(string message)
{
    int result;
    write(message, "? ");
    readf(" %s", &result);
    return result;
}
```

count can now be initialized directly by the return value of a call to this new function:

```
int count = read_int("How many numbers are you going to enter");
```

number cannot be initialized in as straightforward a way because the loop counter **i** happens to be a part of the message that is displayed when reading **number**. This can be overcome by taking advantage of **format**:

```
import std.string;
// ...
int number = read_int(format("Number %s", i));
```

Further, since **number** is used in only one place in the **foreach** loop, its definition can be eliminated altogether and the return value of **read_int()** can directly be used in its place:

```
foreach (i; 0 .. count) {
    numbers ~= read_int(format("Number %s", i));
}
```

Let's make a final modification to this program by moving the lines that read the numbers to a separate function. This would also eliminate the need for the "Read the numbers" comment because the name of the new function would already carry that information.

The new **readNumbers()** function does not need any parameter to complete its task. It reads some numbers and returns them as a slice. The following is the final version of the program:

```
import std.stdio;
import std.string;

void print(string title, int[] slice)
{
    writeln(title, ":");

    foreach (i, element; slice) {
        writefln("%3s:%5s", i, element);
    }
}

int read_int(string message)
{
    int result;
    write(message, "? ");
    readf(" %s", &result);
    return result;
}

int[] readNumbers()
{
    int[] result;

    int count = read_int("How many numbers are you going to enter");

    foreach (i; 0 .. count) {
        result ~= read_int(format("Number %s", i));
    }
}
```

```

    return result;
}

void main()
{
    int[] numbers = readNumbers();
    print("Before sorting", numbers);
    numbers.sort();
    print("After sorting", numbers);
}

```

Compare this version of the program to the first one. The major steps of the program are very clear in the `main()` function of the new program. In contrast, the `main()` function of the first program had to be examined to understand the purpose of that program.

Although the total numbers of nontrivial lines of the two versions of the program ended up being equal, functions make programs shorter in general. This effect is not apparent in this simple program. For example, before the `read_int()` function has been defined, reading an `int` from the input involved three lines of code. After the definition of `read_int()`, the same goal is achieved by a single line of code. Further, the definition of `read_int()` allowed removing the definition of the variable `number` altogether.

35.8.2 Commented lines of code as functions

Sometimes the need to write a comment to describe the purpose of a group of lines of code is an indication that those lines could better be moved to a newly defined function. If the name of the function is descriptive enough then there would be no need for the comment either.

The three commented groups of lines of the first version of the program have been used for defining new functions that achieved the same tasks.

Another important benefit of removing comment lines is that comments tend to become outdated as the code gets modified over the time. Comments are sometimes forgotten to be updated along with the code and become either useless or even worse, misleading. For that reason, it is beneficial to try to write programs without the need for comments.

35.9 Exercises

1. Modify the `printMenu()` function to take the entire set of menu items as a parameter. For example, the menu items can be passed to the function as in the following code:

```

string[] items =
    [ "Black", "Red", "Green", "Blue", "White" ];
printMenu(items, 1);

```

Have the program produce the following output:

```

1 Black
2 Red
3 Green
4 Blue
5 White

```

2. The following program uses a two dimensional array as a canvas. Start with that program and improve it by adding more functionality to it:

```

import std.stdio;

enum totalLines = 20;
enum totalColumns = 60;

/*
 * The 'alias' in the next line makes 'Line' an alias of
 * dchar[totalColumns]. Every 'Line' that is used in the rest
 * of the program will mean dchar[totalColumns] from this
 * point on.
 *
 * Also note that 'Line' is a fixed-length array.
 */
alias dchar[totalColumns] Line;

/*
 * A dynamic array of Lines is being aliased as 'Canvas'.
 */
alias Line[] Canvas;

/*
 * Prints the canvas line by line.
 */
void print(Canvas canvas)
{
    foreach (line; canvas) {
        writeln(line);
    }
}

/*
 * Places a dot at the specified location on the canvas. In a
 * sense, "paints" the canvas.
 */
void putDot(Canvas canvas, int line, int column)
{
    canvas[line][column] = '#';
}

/*
 * Draws a vertical line of the specified length from the
 * specified position.
 */
void drawVerticalLine(Canvas canvas,
                     int line,
                     int column,
                     int length)
{
    foreach (lineToPaint; line .. line + length) {
        putDot(canvas, lineToPaint, column);
    }
}

void main()
{
    Line emptyLine = '.';

    /* An empty canvas */
    Canvas canvas;

    /* Constructing the canvas by adding empty lines */
    foreach (i; 0 .. totalLines) {
        canvas ~= emptyLine;
    }

    /* Using the canvas */
    putDot(canvas, 7, 30);
    drawVerticalLine(canvas, 5, 10, 4);

    print(canvas);
}

```

... the solutions

36 Immutability

We have seen that variables represent concepts in programs. The interactions of these concepts are achieved by expressions that change the values of those variables:

```
// Pay the bill
totalPrice = calculateAmount(itemPrices);
moneyInWallet -= totalPrice;
moneyAtMerchant += totalPrice;
```

Modifying a variable is called *mutating* that variable. The concept of mutability is essential for most tasks. However, there are some cases where mutability is not suitable:

- Some concepts are immutable by definition. For example, there are always seven days in a week, the math constant π (π) is constant, a program may be supporting only a short list of human languages (e.g. only English and Turkish), etc.
- If every variable were modifiable as we have seen so far, then every piece of code that used that variable could potentially have modified it. Even if there was no reason to modify a variable in an operation, there would be no guarantee that this would not happen either. Programs are difficult to read or maintain when there is no guarantee of immutability.

For example, it may be clear that the function call `retire(office, worker)` would retire a worker of an office. If every variable were mutable, it would not be clear which of the two variables would be modified after that function call. It may be expected that the number of active employees of `office` would be decreased, but would the function call also modify `worker` in some way?

The concept of immutability helps with understanding parts of programs by guaranteeing that certain operations do not change certain variables. It also reduces the risk of certain types of program errors.

The *immutability* concept of D is represented by the `const` and `immutable` keywords. Although the two words themselves are close in meaning, their responsibilities in programs are different and they are sometimes incompatible.

36.1 Immutable variables

Both of the terms "immutable variable" and "constant variable" are nonsensical when the word "variable" is taken literally to mean *something that changes*. The word "variable" means any concept of a program which may be mutable or immutable.

There are three ways of defining variables that can never be mutated.

36.1.1 enum constants

We have seen earlier in the [enum chapter](#) that `enum` defines named constant values:

```
enum fileName = "list.txt";
```

As long as their values can be determined at compile time, `enum` variables can be initialized by return values of functions as well:

```
int totalLines()
{
    return 42;
}
```

```

int totalColumns()
{
    return 7;
}

string name()
{
    return "list";
}

void main()
{
    enum fileName = name() ~ ".txt";
    enum totalSquares = totalLines() * totalColumns();
}

```

As expected, values of **enum** constants cannot be modified:

```

++totalSquares;    // ← compilation ERROR

```

Although it is a very effective way of representing immutable values, **enum** can only be used for compile-time values.

An **enum** constant is a *manifest constant*, meaning that the program is compiled as if it has been replaced by its value. As an example, let's consider the following **enum** definition and the two expressions that make use of it:

```

enum i = 42;
writeln(i);
foo(i);

```

The code above would be the same as replacing **i** with its value **42**:

```

writeln(42);
foo(42);

```

Although that replacement makes sense for simple types like **int** and makes no difference in the program, **enum** constants bring a hidden cost when they are used for arrays or associative arrays:

```

enum a = [ 42, 100 ];
writeln(a);
foo(a);

```

After replacing **a** with its value, the equivalent code that the compiler would be compiling is the following:

```

writeln([ 42, 100 ]);    // an array is created at runtime
foo([ 42, 100 ]);      // another array is created at runtime

```

The hidden cost here is that there would be two separate arrays created for the two expressions above. For that reason, it may make more sense to define arrays and associative arrays as **immutable** variables if they are going to be used more than once in the program.

36.1.2 immutable variables

Like **enum**, this keyword specifies that the value of a variable will never change. Its difference from **enum** is that the values of **immutable** variables can be calculated during the execution of the program.

The following program compares the uses of **enum** and **immutable**. The program waits for the user to guess a number that has been picked randomly. Since the random number cannot be determined at compile time, it cannot be defined as **enum**. Still, since the randomly picked value must never be changed after having being decided, it is suitable to specify that variable as **immutable**.

The program takes advantage of the **read_int()** that has been defined in the previous chapter:

```
import std.stdio;
import std.random;

int read_int(string message)
{
    int result;
    write(message, "? ");
    readf("%s", &result);
    return result;
}

void main()
{
    enum min = 1;
    enum max = 10;

    immutable number = uniform(min, max + 1);

    writeln("I am thinking of a number between %s and %s.",
           min, max);

    auto isCorrect = false;
    while (!isCorrect) {
        immutable guess = read_int("What is your guess");
        isCorrect = (guess == number);
    }

    writeln("Correct!");
}
```

Observations:

- **min** and **max** are integral parts of the behavior of this program and their values are known at compile time. For that reason they are defined as **enum** constants.
- **number** is specified as **immutable** because it would not be appropriate to modify it during the execution of the program. It is the same with each user guess: Once read, the guess should not be modified.
- Observe that the types of those variables are not specified explicitly. As with **auto**, the types of **enum** and **immutable** variables can be inferred from the expression on the right hand side.

Although it is not necessary to write the type fully as e.g. **immutable(int)**, **immutable** normally takes the actual type within parentheses. The output of the following program demonstrates that the full names of the types of the three variables are in fact the same:

```
import std.stdio;

void main()
```



```

{
    immutable    inferredType = 0;
    immutable int explicitType = 1;
    immutable(int) wholeType   = 2;

    writeln(typeof(inferredType).stringof);
    writeln(typeof(explicitType).stringof);
    writeln(typeof(wholeType).stringof);
}

```

The actual name of the type includes **immutable**:

```

immutable(int)
immutable(int)
immutable(int)

```

The type that is specified within the parentheses has significance. We will see this below when discussing the immutability of the whole slice vs. its elements.

36.1.3 const variables

This keyword has the same effect as **immutable** on variables. **const** variables cannot be modified:

```

const half = total / 2;
half = 10; // ← compilation ERROR

```

I recommend that you prefer **immutable** over **const** when defining variables. The reason is that **immutable** variables can be passed to functions as their **immutable** parameters. We will see this below.

36.2 Immutable parameters

It is possible for functions to promise that they do not modify certain parameters that they take. The compiler guarantees that this promise is enforced. Before seeing how this is achieved, let's first see that functions can indeed modify the elements of slices that are passed as arguments to those function.

As you would remember from the [Slices and Other Array Features](#) chapter, slices do not own elements but provide access to them. There may be more than one slice at a given time that provide access to the same elements.

Although the examples in this section focus only on slices, this topic is applicable to associative arrays and classes as well because they too are *reference types*.

A slice that is passed as a function argument is not the slice that the function is called with. The argument is a copy of the actual slice:

```

import std.stdio;

void main()
{
    int[] slice = [ 10, 20, 30, 40 ]; // 1
    halve(slice);
    writeln(slice);
}

void halve(int[] numbers) // 2
{

```

```

foreach (ref number; numbers) {
    number /= 2;
}

```

When program execution enters the `halve()` function, there are two slices that provide access to the same four elements:

1. The slice named `slice` that is defined in `main()`, which is passed to `halve()` as its argument
2. The slice named `numbers` that `halve()` receives as its argument, which provides access to the same elements as `slice`

Also due to the `ref` keyword in the `foreach` loop, the values of the original (and the only) elements get halved:

```
[5, 10, 15, 20]
```

It is useful for functions to be able to modify the elements of the slices that are passed as arguments. Some functions exist just for that purpose as has been seen in this example.

The compiler does not allow passing `immutable` variables as arguments to such functions, because it is impossible to modify an immutable variable:

```

immutable int[] slice = [ 10, 20, 30, 40 ];
halve(slice); // ← compilation ERROR

```

The compilation error indicates that a variable of type `immutable(int[])` cannot be used as an argument of type `int[]`:

```

Error: function deneme.halve (int[] numbers) is not callable
using argument types (immutable(int[]))

```

36.2.1 const parameters

It is important and natural that `immutable` variables be prevented from being passed to functions like `halve()`, which modify their arguments. However, it would be a limitation if they could not be passed to functions that did not modify their arguments in any way:

```

import std.stdio;

void main()
{
    immutable int[] slice = [ 10, 20, 30, 40 ];
    print(slice); // ← compilation ERROR
}

void print(int[] slice)
{
    writeln("%s elements: ", slice.length);

    foreach (i, element; slice) {
        writeln("%s: %s", i, element);
    }
}

```

It does not make sense above that a slice is prevented from being printed just because it is `immutable`. The proper way of dealing with this situation is `const` parameters.

The **const** keyword specifies that a variable would not be modified through *that particular reference* (e.g. a slice) of that variable. Specifying a parameter as **const** guarantees that the elements of the slice would not be modified inside the function. Once **print()** provides this guarantee, the program can now be compiled:

```
    print(slice);    // now compiles
// ...
void print(const int[] slice)
```

This guarantee allows passing both mutable and **immutable** variables as arguments:

```
immutable int[] slice = [ 10, 20, 30, 40 ];
print(slice);          // compiles

int[] mutableSlice = [ 7, 8 ];
print(mutableSlice);  // compiles
```

A parameter that is not modified in a function but is not specified as **const** reduces the usability of that function. Additionally, **const** parameters provide useful information to the programmer. Knowing that a variable will not be modified when passed to a function makes the code easier to understand. It also prevents potential errors because the compiler prevents modifications to **const** parameters:

```
void print(const int[] slice)
{
    slice[0] = 42;    // ← compilation ERROR
```

The programmer would either realize the mistake in the function or would rethink the design and perhaps remove the **const** specifier.

The fact that **const** parameters can accept both mutable and **immutable** variables has an interesting consequence. This is explained in the "Should a parameter be **const** or **immutable**?" section below.

36.2.2 immutable parameters

As we have seen above, both mutable and **immutable** variables can be passed to functions as their **const** parameters. In a way, **const** parameters are welcoming.

In contrast, **immutable** parameters bring a strong requirement: Only **immutable** variables can be passed to functions as their **immutable** parameters:

```
void func(immutable int[] slice)
{
    /* ... */
}

void main()
{
    immutable int[] immSlice = [ 1, 2 ];
    int[] slice = [ 8, 9 ];

    func(immSlice);    // compiles
    func(slice);      // ← compilation ERROR
}
```

For that reason, the **immutable** specifier should be used only when this requirement is necessary. We have indeed been using the **immutable** specifier indirectly through certain string types. This will be covered below.

We have seen that the parameters that are specified as **const** or **immutable** promise not to modify *the actual variable* that is passed as an argument. This is relevant only for reference types because only then there is *the actual variable* to talk about the immutability of.

Reference types and *value types* will be covered in later chapters. Among the types that we have seen so far, only slices and associative arrays are reference types; the others are value types.

36.2.3 Should a parameter be const or immutable?

The two sections above may give the impression that being more flexible, **const** parameters should be preferred over **immutable** parameters. This is not always true.

const erases the information about whether the original variable is mutable or **immutable**. This information is hidden even from the compiler.

A consequence of this fact is that **const** parameters cannot be passed as arguments to functions that take **immutable** parameters. For example, **foo()** below cannot pass its **const** parameter to **bar()**:

```
void main()
{
    /* The original variable is immutable */
    immutable int[] slice = [ 10, 20, 30, 40 ];
    foo(slice);
}

/* A function that takes its parameter as const, in order to
 * be more useful. */
void foo(const int[] slice)
{
    bar(slice);    // ← compilation ERROR
}

/* A function that takes its parameter as immutable for a
 * plausible reason. */
void bar(immutable int[] slice)
{
    /* ... */
}
```

bar() requires that the parameter is **immutable**. However, it is not known whether the original variable that **foo()**'s **const** parameter references is **immutable** or not.

Note: It is clear in the code above that the original variable in **main()** is **immutable**. However, the compiler compiles functions individually without regard to all of the places that function is called from. To the compiler, the **slice** parameter of **foo()** may be referring to a mutable variable or an **immutable** one.

A solution would be to call **bar()** with an immutable copy of the parameter:

```
void foo(const int[] slice)
{
    bar(slice.idup);
}
```

Although that is a sensible solution, it does have the cost of copying, which would be wasteful in the case where the original variable has been **immutable** to begin with.

After this analysis, it should be clear that taking parameters always as **const** does not seem to be the best approach in every situation. After all, if **foo()**'s parameter has been defined as **immutable**, there would not be any need to copy it before calling **bar()**:

```
void foo(immutable int[] slice) // This time immutable
{
    bar(slice);    // Copying is not needed anymore
}
```

Although the code compiles, defining the parameter as **immutable** has a similar cost: This time an immutable copy of the original variable is needed when calling **foo()**, if that variable were not immutable to begin with:

```
foo(mutableSlice.idup);
```

Templates can provide help. (We will see templates in later chapters.) Although I don't expect you to fully understand the following function at this point in the book, I will present it as a solution to this problem. The following function template **foo()** can be called both by mutable and by **immutable** variables. The parameter would be copied only if the original variable has been mutable; no copying would take place if it has been **immutable**:

```
import std.conv;
// ...

/* Because it is a template, foo() can be called by mutable
 * and immutable variables. */
void foo(T)(T[] slice)
{
    /* 'to()' does not make a copy if the original variable is
     * already immutable. */
    bar(to!(immutable T[])(slice));
}
```

36.3 Immutability of the slice versus the elements

We have seen above that the type of an **immutable** slice has been printed as **immutable(int[])**. As the parentheses after **immutable** indicate, it is the entire slice that is **immutable**. Such a slice cannot be modified in any way: elements may not be added or removed, their values may not be modified, and the slice may not start providing access to a different set of elements:

```
immutable int[] immSlice = [ 1, 2 ];
immSlice ~= 3;           // ← compilation ERROR
immSlice[0] = 3;        // ← compilation ERROR
immSlice.length = 1;    // ← compilation ERROR

immutable int[] immOtherSlice = [ 10, 11 ];
immSlice = immOtherSlice; // ← compilation ERROR
```

Taking immutability to that extreme may not be suitable in every case. In most cases, what is important is the immutability of just the elements. Since a slice is just an access tool to elements, it should not matter to make changes to the slice itself as long as the elements are not modified.

To specify that only the elements are immutable, only the element type is written within parentheses. When the code is modified accordingly, now only the elements are immutable, not the slice itself:

```
immutable(int)[] immSlice = [ 1, 2 ];
immSlice ~= 3;           // can add elements
immSlice[0] = 3;        // ← compilation ERROR
immSlice.length = 1;    // can lose elements
```

```
immutable int[] immOtherSlice = [ 10, 11 ];
immSlice = immOtherSlice;    // can provide access to other elements
```

Although the two syntaxes are very similar, they have different meanings. To summarize:

```
immutable int[] a = [1]; /* Neither the elements nor the
                          * slice can be modified */

immutable(int[]) b = [1]; /* The same meaning as above */

immutable(int)[] c = [1]; /* The elements cannot be
                          * modified but the slice can be */
```

This distinction has been in effect in some of the programs that we have written so far. As you may remember, the three string aliases involve immutability:

- **string** is an alias for **immutable(char)[]**
- **wstring** is an alias for **immutable(wchar)[]**
- **dstring** is an alias for **immutable(dchar)[]**

Likewise, string literals are immutable as well:

- The type of literal **"hello" c** is **string**
- The type of literal **"hello" w** is **wstring**
- The type of literal **"hello" d** is **dstring**

According to these definitions, D strings are normally arrays of *immutable characters*.

36.3.1 .dup and .idup

There may be mismatches in immutability when strings are passed to functions as parameters. The **.dup** and **.idup** properties make copies of arrays with the desired immutability:

- **.dup** makes a mutable copy of the array; its name comes from "duplicate"
- **.idup** makes an immutable copy of the array

For example, a function that insists on the immutability of a parameter may have to be called with an immutable copy of a mutable string:

```
void foo(string s)
{
    // ...
}

void main()
{
    char[] salutation;
    foo(salutation);           // ← compilation ERROR
    foo(salutation.idup);     // ← this compiles
}
```

36.4 How to use

- As a general rule, prefer immutable variables over mutable ones.
- Define constant values as **enum** if their values can be calculated at compile time. For example, the constant value of *seconds per minute* can be an **enum**:

```
enum int secondsPerMinute = 60;
```

There is no need to specify the type explicitly if it can be inferred from the right hand side:

```
enum secondsPerMinute = 60;
```

- Consider the hidden cost of **enum** arrays and **enum** associative arrays. Define them as **immutable** variables if the arrays are large and they are used more than once in the program.
- Specify variables as **immutable** if their values will never change but cannot be known at compile time. Again, the type can be inferred:

```
immutable guess = read_int("What is your guess");
```

- If a function does not modify a parameter, specify that parameter as **const**. This would allow both mutable and **immutable** variables to be passed as arguments:

```
void foo(const char[] s)
{
    // ...
}

void main()
{
    char[] mutableString;
    string immutableString;

    foo(mutableString);    // ← compiles
    foo(immutableString); // ← compiles
}
```

- Following from the previous guideline, consider that **const** parameters cannot be passed to functions taking **immutable**. See the section titled "Should a parameter be **const** or **immutable**?" above.
- If the function modifies a parameter, leave that parameter as mutable (**const** or **immutable** would not allow modifications anyway):

```
import std.stdio;

void reverse(dchar[] s)
{
    foreach (i; 0 .. s.length / 2) {
        immutable temp = s[i];
        s[i] = s[$ - 1 - i];
        s[$ - 1 - i] = temp;
    }
}

void main()
{
    dchar[] salutation = "hello"d.dup;
    reverse(salutation);
    writeln(salutation);
}
```

The output:

```
olleh
```

36.5 Summary

- **enum** variables represent immutable concepts that are known at compile time.
- **immutable** variables represent immutable concepts that must be calculated at runtime.
- **const** parameters are the ones that functions do not modify. Both mutable and **immutable** variables can be passed as arguments of **const** parameters.
- **immutable** parameters are the ones that functions specifically require them to be so. Only **immutable** variables can be passed as arguments of **immutable** parameters.
- **immutable(int[])** specifies that neither the slice nor its elements can be modified.
- **immutable(int)[]** specifies that only the elements cannot be modified.
- Immutability is a very powerful tool in programming. It is helpful to know that variables and parameters do not change in specific contexts and during specific operations.

37 Function Parameters

This chapter covers different ways of defining function parameters.

Some of the concepts of this chapter have already appeared in previous chapters. For example, the **ref** keyword that we have seen in the [foreach Loop chapter](#) was making actual elements available in **foreach** loops as opposed to copies of those elements.

Additionally, we have covered the **const** and **immutable** keywords in the previous chapter.

We have written functions that produced results by making use of their parameters. For example, the following function uses its parameters in a calculation:

```
double weightedAverage(double quizGrade, double finalGrade)
{
    return quizGrade * 0.4 + finalGrade * 0.6;
}
```

That function calculates the average grade by taking 40% of the quiz grade and 60% of the final grades. Here is how it may be used:

```
int quizGrade = 76;
int finalGrade = 80;

writeln("Weighted average: %2.0f",
        weightedAverage(quizGrade, finalGrade));
```

37.1 Most parameters are copied

In the code above, the two variables are passed as arguments to **weightedAverage()** and the function *uses* its parameters. This fact may give the false impression that the function uses the actual variables that have been passed as arguments. In reality, what the function uses are *copies* of those variables.

This distinction is important because modifying a parameter changes only the copy. This can be seen in the following function that is trying to modify its parameter (i.e. making a side effect). Let's assume that the following function is written for reducing the energy of a game character:

```
void reduceEnergy(double energy)
{
    energy /= 4;
}
```

Here is a program that tests **reduceEnergy()**:

```
import std.stdio;

void reduceEnergy(double energy)
{
    energy /= 4;
}

void main()
{
    double energy = 100;

    reduceEnergy(energy);
    writeln("New energy: ", energy);
}
```

The output:

```
New energy: 100      ← Not changed!
```

Although `reduceEnergy()` drops the value of its parameter to a quarter of its original value, the variable `energy` in `main()` does not change. The reason for this is that the `energy` variable in `main()` and the `energy` parameter of `reduceEnergy()` are separate; the parameter is a copy of the variable in `main()`.

To observe this more closely, let's insert some `writeln()` expressions:

```
import std.stdio;

void reduceEnergy(double energy)
{
    writeln("Entered the function      : ", energy);
    energy /= 4;
    writeln("Leaving the function     : ", energy);
}

void main()
{
    double energy = 100;

    writeln("Calling the function      : ", energy);
    reduceEnergy(energy);
    writeln("Returned from the function: ", energy);
}
```

The output:

```
Calling the function      : 100
Entered the function      : 100
Leaving the function     : 25  ← the parameter changes,
Returned from the function: 100 ← the variable remains the same
```

37.2 Objects of reference types are not copied

Elements of slices and associative arrays, and class objects are not copied when passed as parameters. Such variables are passed to functions as *references*. In effect, the parameter becomes a reference to the actual object; modifications made through the reference modifies the actual object.

Being slices, strings are passed as references as well:

```
import std.stdio;

void makeFirstLetterDot(dchar[] str)
{
    str[0] = '.';
}

void main()
{
    dchar[] str = "abc".dup;
    makeFirstLetterDot(str);
    writeln(str);
}
```

The change made to the first element of the parameter effects the actual element in `main()`:

.bc

37.3 Parameter qualifiers

Parameters are passed to functions according to the general rules described above:

- Value types are copied
- Reference types are passed as references

Those are the default rules that are applied when parameter definitions have no qualifiers. The following qualifiers change the way parameters are passed and what operations are allowed on them.

37.3.1 in

We have seen that functions are a facility that produces values and can have side effects. The **in** keyword specifies that a parameter is going to be used only as an input data by this facility. Such parameters cannot be modified by the function. **in** implies **const**:

```
import std.stdio;

double weightedTotal(in double currentTotal,
                    in double weight,
                    in double addend)
{
    return currentTotal + (weight * addend);
}

void main()
{
    writeln(weightedTotal(1.23, 4.56, 7.89));
}
```

in parameters cannot be modified:

```
void foo(in int value)
{
    value = 1;    // ← compilation ERROR
}
```

37.3.2 out

We have seen that the functions return the values that they produce as their return values. The fact that there is only one return value is sometimes limiting as some functions may need to produce more than one result. (**Note:** *It is actually possible to return more than one result by defining the return type as a **Tuple** or a **struct**. We will see these features in later chapters.*)

The **out** keyword makes it possible for functions to return results through their parameters. When **out** parameters are modified within the function, those modifications effect the actual variable that has been passed to the function. In a sense, the assigned value goes *out* of the function through **out** parameters.

Let's have a look at a function that divides two numbers and produces both the quotient and the remainder. The return value can be used for the quotient and the remainder can be returned through an **out** parameter:

```

import std.stdio;

int divide(in int dividend, in int divisor, out int remainder)
{
    remainder = dividend % divisor;
    return dividend / divisor;
}

void main()
{
    int remainder;
    int result = divide(7, 3, remainder);

    writeln("result: ", result, ", remainder: ", remainder);
}

```

Modifying the **remainder** parameter of the function modifies the **remainder** variable in **main()** (their names need not be the same):

```
result: 2, remainder: 1
```

Regardless of their values at the call site, **out** parameters are first automatically assigned to the **.init** value of their types:

```

import std.stdio;

void foo(out int parameter)
{
    writeln("After entering the function      : ", parameter);
}

void main()
{
    int variable = 100;

    writeln("Before calling the function      : ", variable);
    foo(variable);
    writeln("After returning from the function: ", variable);
}

```

Even though there is no explicit assignment to the parameter in the function, the value of the parameter automatically becomes the initial value of **int**, affecting the variable in **main()**:

```

Before calling the function      : 100
After entering the function      : 0 ← the value of int.init
After returning from the function: 0

```

As this demonstrates, **out** parameters cannot pass values into functions; they are strictly for passing values out of functions.

We will see in later chapters that returning **Tuple** or **struct** types may be better alternatives to **out** parameters.

37.3.3 const

As we have seen in the previous chapter, **const** guarantees that the parameter will not be modified inside the function. It is helpful for the programmers to know that certain variables will not be changed by the function. **const** also makes functions more useful by allowing *mutable*, **const** and **immutable** variables to be passed as parameters:

```
import std.stdio;

dchar lastLetter(const dchar[] str)
{
    return str[$ - 1];
}

void main()
{
    writeln(lastLetter("constant"));
}
```

37.3.4 immutable

As we have seen in the previous chapter, **immutable** makes functions require that certain arguments must consist of **immutable** elements. Because of such a requirement, the following function can only be called with **immutable** strings (e.g. with string literals):

```
import std.stdio;

dchar[] mix(immutable dchar[] first,
            immutable dchar[] second)
{
    dchar[] result;
    int i;

    for (i = 0; (i < first.length) && (i < second.length); ++i) {
        result ~= first[i];
        result ~= second[i];
    }

    result ~= first[i..$];
    result ~= second[i..$];

    return result;
}

void main()
{
    writeln(mix("HELLO", "world"));
}
```

Since it brings a requirement on the parameter, **immutable** parameters should be used only when really necessary. Because it is more welcoming, **const** is more useful.

37.3.5 ref

This keyword allows passing a parameter by reference even though it would normally be passed as a copy.

For the **reduceEnergy()** function above to be able to modify the actual variable that is passed as its argument, it must take the parameter as **ref**:

```
import std.stdio;

void reduceEnergy(ref double energy)
{
    energy /= 4;
}

void main()
{
    double energy = 100;

    reduceEnergy(energy);
}
```

```

    writeln("New energy: ", energy);
}

```

This time, the modification that is made to the parameter changes the actual variable that is passed to the function in `main()`:

```
New energy: 25
```

As can be seen, **ref** parameters can be used both as input and output. **ref** parameters can also be thought of as aliases of actual variables. The function parameter **energy** above is an alias of the variable **energy** in `main()`.

Similar to **out** parameters, **ref** parameters allow functions make side effects as well. In fact, **reduceEnergy()** does not return a value; it only makes a side effect through its only parameter.

The programming style called *functional programming* favors return values over side effects, so much so that some functional programming languages do not allow side effects at all. This is because functions that produce results *purely* through their return values are easier to understand, to write correctly, and to maintain.

The same function can be written in functional programming style by returning the result, instead of making a side effect. The parts of the program that has changed are highlighted in yellow:

```

import std.stdio;
double reducedEnergy(double energy)
{
    return energy / 4;
}
void main()
{
    double energy = 100;
    energy = reducedEnergy(energy);
    writeln("New energy: ", energy);
}

```

Note the change in the name of the function as well. Now it is a noun as opposed to a verb.

37.3.6 inout

Despite its name consisting of **in** and **out**, this keyword does not mean both input and output; we have already seen that the keyword that achieves both input and output is **ref**.

inout carries the *mutability* of the parameter to the return type. If the parameter is *mutable*, **const**, or **immutable**, then the return value is *mutable*, **const**, or **immutable**, respectively.

To see how **inout** helps in programs, let's look at a function that returns a slice having one less element from both the front and the back of the original slice:

```

import std.stdio;
int[] trimmed(int[] slice)
{
    if (slice.length) {
        --slice.length;           // trim from the end
        if (slice.length) {

```

```

        slice = slice[1 .. $];    // trim from the beginning
    }
}
return slice;
}
void main()
{
    int[] numbers = [ 5, 6, 7, 8, 9 ];
    writeln(trimmed(numbers));
}

```

The output:

```
[6, 7, 8]
```

According to the previous chapter, in order for the function to be more useful, its parameter better be `const(int)[]` because the parameter is not being modified inside the function. (Note that there is no harm in modifying the parameter slice itself as it is a copy of the original argument.)

However, when the function is defined that way, there would a compilation error:

```

int[] trimmed(const(int)[] slice)
{
    // ...
    return slice;    // ← compilation ERROR
}

```

The compilation error indicates that a slice of `const(int)` may not be returned as a slice of mutable `int`:

```
Error: cannot implicitly convert expression (slice) of type
const(int)[] to int[]
```

One may think that specifying the return type also as `const(int)[]` may be the solution:

```

const(int)[] trimmed(const(int)[] slice)
{
    // ...
    return slice;    // now compiles
}

```

Although the code can now be compiled, it brings a limitation: Even if the function is called with a slice of *mutable* elements, the returned slice ends up consisting of `const` elements. To see how limiting this would be, let's look at the following code that is trying to modify the elements of a slice other than the ones at the front and at the back:

```

int[] numbers = [ 5, 6, 7, 8, 9 ];
int[] middle = trimmed(numbers);    // ← compilation ERROR
middle[] *= 10;

```

It would be expected that the returned slice of type `const(int)[]` cannot be assigned to a slice of type `int[]`:

```
Error: cannot implicitly convert expression (trimmed(numbers))
of type const(int)[] to int[]
```

However, since the original slice is of mutable elements to begin with, this limitation can be seen as artificial and unfortunate. **inout** solves this mutability problem about parameters and return values. It is specified both on the parameter and on the return type and carries the mutability of the former to the latter:

```
inout(int)[] trimmed(inout(int)[] slice)
{
    // ...
    return slice;
}
```

With that change, the same function can now be called with *mutable*, **const**, and **immutable** slices:

```
{
    int[] numbers = [ 5, 6, 7, 8, 9 ];
    // The return type is slice of mutable elements
    int[] middle = trimmed(numbers);
    middle[] *= 10;
    writeln(middle);
}

{
    immutable int[] numbers = [ 10, 11, 12 ];
    // The return type is slice of immutable elements
    immutable int[] middle = trimmed(numbers);
    writeln(middle);
}

{
    const int[] numbers = [ 13, 14, 15, 16 ];
    // The return type is slice of const elements
    const int[] middle = trimmed(numbers);
    writeln(middle);
}
```

37.3.7 lazy

It is natural to expect that arguments are evaluated *before* entering functions that use those arguments. For example, the function **add()** below is called with the return values of two other functions:

```
result = add(anAmount(), anotherAmount());
```

In order for **add()** to be called, first **anAmount()** and **anotherAmount()** must be called. Otherwise, the values that **add()** needs would not be available.

Evaluating arguments before calling a function is *eager*.

However, some parameters may not get a chance to be used in the function at all depending on certain conditions. In such cases, the eager evaluations of the arguments would be wasteful.

Let's look at such a program that uses one of its parameters only when needed. The following function tries to get the required number of eggs first from the refrigerator. When there is sufficient number of eggs in there, it doesn't need to know the number of eggs that the neighbors have:

```
void makeOmelet(in int requiredEggs,
               in int eggsInFridge,
               in int eggsAtNeighbors)
{
```



```

writefln("Need to make a %s-egg omelet", requiredEggs);

if (requiredEggs <= eggsInFridge) {
    writeln("Take all of the eggs from the fridge");
} else if (requiredEggs <= (eggsInFridge + eggsAtNeighbors)) {
    writefln("Take %s eggs from the fridge"
            " and %s eggs from the neighbors",
            eggsInFridge, requiredEggs - eggsInFridge);
} else {
    writefln("Cannot make a %s-omelet", requiredEggs);
}
}

```

Additionally, let's assume that there is a function that calculates and returns the total number of eggs available at the neighbors. For demonstration purposes, the function prints some information as well:

```

int countEggs(in int[string] availableEggs)
{
    int result;

    foreach (neighbor, count; availableEggs) {
        writeln(neighbor, ": ", count, " eggs");
        result += count;
    }

    writefln("A total of %s eggs available at the neighbors",
            result);

    return result;
}

```

That function iterates over the elements of an associative array and adds up all of the egg counts.

The `makeOmelet()` function can be called with the return value of `countEggs()` as in the following program:

```

import std.stdio;

void main()
{
    int[string] atNeighbors = [ "Jane":5, "Jim":3, "Bill":7 ];

    makeOmelet(2, 5, countEggs(atNeighbors));
}

```

As seen in the output of the program, first `countEggs()` function is executed and then `makeOmelet()` is called:

```

Jane: 5 eggs      }
Bill: 7 eggs     } counting the eggs at the neighbors
Jim: 3 eggs      }
A total of 15 eggs available at the neighbors
Need to make a 2-egg omelet
Take all of the eggs from the fridge

```

Although it is possible to make the two-egg omelet with only the eggs in the fridge, the eggs at the neighbors have already been counted eagerly.

The **lazy** keyword specifies that an expression that has been passed to a function as a parameter will be evaluated only if and when needed:

```
void makeOmelet(in int requiredEggs,
               in int eggsInFridge,
               lazy int eggsAtNeighbors)
{
    // ... the body of the function is the same as before ...
}
```

As seen in the new output, when the number of eggs in the fridge satisfy the required eggs, the counting of the eggs at the neighbors does not happen anymore:

```
Need to make a 2-egg omelet
Take all of the eggs from the fridge
```

That count would still happen if needed. For example, let's take a look at a case where the required eggs are more than the eggs in the fridge:

```
makeOmelet(9, 5, countEggs(atNeighbors));
```

This time the total number of eggs at the neighbors is really needed:

```
Need to make a 9-egg omelet
Jane: 5 eggs
Bill: 7 eggs
Jim: 3 eggs
A total of 15 eggs available at the neighbors
Take 5 eggs from the fridge and 4 eggs from the neighbors
```

The values of **lazy** parameters are evaluated every time that they are used in the function.

For example, because the **lazy** parameter of the following function is used three times in the function, the expression that provides its value is evaluated three times:

```
import std.stdio;

int valueOfArgument()
{
    writeln("Calculating...");
    return 1;
}

void functionWithLazyParameter(lazy int value)
{
    int result = value + value + value;
    writeln(result);
}

void main()
{
    functionWithLazyParameter(valueOfArgument());
}
```

The output

```
Calculating
Calculating
Calculating
3
```

37.3.8 scope

This keyword specifies that a parameter will not be used beyond the scope of the function:

```
int[] globalSlice;

int[] foo(scope int[] parameter)
{
    globalSlice = parameter;    // ← compilation ERROR
    return parameter;          // ← compilation ERROR
}

void main()
{
    int[] slice = [ 10, 20 ];
    int[] result = foo(slice);
}
```

That function breaks the promise of **scope** in two places: It assigns the parameter to a global variable and returns it. Both of those would make it possible for the parameter to be accessed after the function finishes.

(**Note:** *dmd 2.058*, the compiler that has been used when compiling the examples in this chapter does not support the **scope** keyword.)

37.3.9 shared

This keyword requires that the parameter is sharable between threads:

```
void foo(shared int[] i)
{
    // ...
}

void main()
{
    int[] numbers = [ 10, 20 ];
    foo(numbers);    // ← compilation ERROR
}
```

The program above cannot be compiled because the argument is not **shared**. The program can be compiled with the following change:

```
shared int[] numbers = [ 10, 20 ];
foo(numbers);    // now compiles
```

We will use the **shared** keyword later in the [Data Sharing Concurrency chapter](#).

37.4 Summary

- *Parameter* is what the function takes from its caller to accomplish its task.
- *Argument* is an expression (e.g. a variable) that is passed to a function as a parameter.
- Arguments of value types are passed by-copy; arguments of reference types are passed by-reference. (We will see this topic again in later chapters.)

- **in** specifies that the parameter is only for data input.
- **out** specifies that the parameter is only for data output.
- **ref** specifies that the parameter is for data input and data output.
- **const** guarantees that the parameter is not modified inside the function.
- **immutable** requires that the argument is **immutable**.
- **inout** appears both at the parameter and the return type, and transfers the *mutability* of the parameter to the return type.
- **lazy** evaluates the parameter when (and every time) its value is actually used.
- **scope** guarantees that no reference to the parameter will be leaked from the function.
- **shared** requires that the parameter is **shared**.

37.5 Exercise

- The following program is trying to swap the values of two arguments:

```
import std.stdio;

void swap(int first, int second)
{
    int temp = first;
    first = second;
    second = temp;
}

void main()
{
    int a = 1;
    int b = 2;

    swap(a, b);

    writeln(a, ' ', b);
}
```

The program does not have any effect on **a** or **b**:

```
1 2 ← not swapped
```

Fix the function so that the values of **a** and **b** are swapped.

[... the solution](#)

38 Lazy Operators

Lazy evaluations is delaying the execution of expressions until their results are needed. Lazy evaluations are among the fundamental features of some of the programming languages.

Naturally, delaying execution of expressions until their results are really necessary may make programs run faster.

A concept that is similar to lazy evaluations is the shortcut behavior of three operators:

- **||** (or) operator: The second expression is evaluated only if the first expression is **false**.

```
if (anExpression() || mayNotBeEvaluated()) {
    // ...
}
```

If the result of **anExpression()** is **true**, the result of the **||** expression is also **true** and the second expression is not evaluated.

- **&&** (and) operator: The second expression is evaluated only if the first expression is **true**.

```
if (anExpression() && mayNotBeEvaluated()) {
    // ...
}
```

If the result of **anExpression()** is **false**, the result of the **&&** expression is also **false** and the second expression is not evaluated.

- **?:** (ternary) operator: Either the first or the second expression is evaluated depending on whether the condition is **true** or **false**, respectively.

```
int i = condition() ? eitherThis() : orThis();
```

The laziness of these operators is not only about performance. Sometimes, evaluating one of the expressions can be an error.

For example, the *if the first letter is A* condition check below would be an error when the string is empty:

```
dstring s;
// ...
if (s[0] == 'A') {
    // ...
}
```

In order to access the zero-indexed element of **s**, it must first be ensured that the string does have such an element. For that reason, the following condition check moves that potentially erroneous logical expression to the right-hand side of the **&&** operator to ensure that it will be evaluated only when it is safe to do so:

```
dstring s;
// ...
if ((s.length >= 1) && (s[0] == 'A')) {
    // ...
}
```

Lazy evaluations can be achieved by function pointers, delegates, and ranges as well. We will see these in later chapter.

39 Program Environment

We have seen that `main()` is a function. Program execution starts with `main()` and branches to other functions from there. The definition of `main()` that we have used so far has been the following:

```
void main()
```

According to that definition `main()` does not take any parameters and does not return a value. It is actually not possible to not return a value, because environments that start programs expect return values from them. Although it is possible to specify the return type of `main()` as `void`, it does indeed return a value.

39.1 The return value of `main()`

Programs are always started by an entity in a particular environment. The entity that starts the program may be the shell where the user types the name of the program and presses the Enter key, or it may be a development environment where the programmer clicks the [Run] button, etc.

The program communicates its exit status to its environment by the return value of `main()`.

A return value of 0 means successful completion and any other value means some type of failure. Although the return value is up to the programmer, by convention the value 0 always means success.

Note: Only values in the range [0,127] are portable; not all environments support other values.

Values other than 0 can have different meanings depending on each program. For example, the common Unix program `ls`, which is used for listing contents of directories, returns 1 for minor errors and 2 for serious ones. In many environments, the return value of the program that has been executed most recently in the console can be seen through the `$?` environment variable. For example, when we ask `ls` to list a file that does not exist, its non-zero return value can be observed with `$?` as seen below.

Note: In the command line interactions below, the lines that start with `#` indicate the lines that the user types. If you want to try the same commands, you must enter the contents of those lines except the `#` character. Also, the commands below start a program named `deneme`; replace that name with the name of your test program.

Additionally, although the following examples show interactions in a Linux console, they would be similar but not exactly the same in consoles in other operating systems.

```
# ls a_file_that_does_not_exist
ls: cannot access a_file_that_does_not_exist: No such file or directory
# echo $?
2      ← the return value of ls
```

39.1.1 `main()` always returns a value

Some of the programs that we have written so far have thrown exceptions when they could not continue with their tasks. As much as we have seen so far, when an exception is thrown, the program ends with an `object.Exception` error message.

When that happens, even if `main()` has been defined as returning `void`, a value of 1 is automatically returned to the program's environment. Let's see this in action by the following simple program that terminates with an exception:

```
void main()
{
    throw new Exception("There has been an error.");
}
```

Although the return type is specified as `void`, the return value is 1:

```
# ./deneme
object.Exception: There has been an error.
...
# echo $?
1
```

Similarly, `void main()` functions that terminate successfully also automatically return 0 as their return values. Let's see this with the following program that terminates *successfully*:

```
import std.stdio;
void main()
{
    writeln("Done!");
}
```

The program returns 0:

```
# ./deneme
Done!
# echo $?
0
```

39.1.2 Specifying the return value

Returning a specific value from `main()` is the same as returning a value from any other function. The return type must be specified as `int` and the value must be returned by the `return` statement:

```
import std.stdio;
int main()
{
    int number;
    write("Please enter a number between 3 and 6: ");
    readf(" %s", &number);

    if ((number < 3) || (number > 6)) {
        stderr.writefln("ERROR: %s is not valid!", number);
        return 111;
    }

    writefln("Thank you for %s.", number);

    return 0;
}
```

When the entered number is within the valid range, the return value of the program is 0:

```
# ./deneme
Please enter a number between 3 and 6: 5
Thank you for 5.
# echo $?
0
```

When the number is outside of the valid range, the return value of the program is 111:

```
# ./deneme
Please enter a number between 3 and 6: 10
ERROR: 10 is not valid!
# echo $?
111
```

The value of 111 in the above program is arbitrary; normally 1 is suitable as the failure code.

39.2 Standard error stream `stderr`

The program above uses the stream `stderr`. That stream is the third of the standard streams. It is used for writing error messages:

- `stdin`: standard input stream
- `stdout`: standard output stream
- `stderr`: standard error stream

When a program is started in a console, normally the messages that are written to `stdout` and `stderr` both appear on the console. When needed, it is possible to redirect these outputs individually. This subject is outside of the focus of this chapter and the details may vary for each shell program.

39.3 Parameters of `main()`

It is common for programs to take parameters from the environment that started them. For example, we have already passed a file name as a command line option to `ls` above. There are two command line options in the following line:

```
# ls -l deneme
-rwxr-xr-x 1 acehrel1 users 460668 Nov  6 20:38 deneme
```

The set of command line parameters and their meanings are defined entirely by the program. Every program documents its usage, including what every parameter means.

The arguments that are used when starting a D program are passed to that program's `main()` as a slice of `strings`. Defining `main()` as taking a parameter of type `string[]` is sufficient to have access to program arguments. The name of this parameter is commonly abbreviated as `args`. The following program prints all of the arguments that the program has been started with:

```
import std.stdio;

void main(string[] args)
{
    foreach (i, arg; args) {
        writeln("Argument %-3s: %s", i, arg);
    }
}
```



```
}
}
```

Let's start the program with arbitrary arguments:

```
# ./deneme some arguments on the command line 42 --an-option
Argument 0 : ./deneme
Argument 1 : some
Argument 2 : arguments
Argument 3 : on
Argument 4 : the
Argument 5 : command
Argument 6 : line
Argument 7 : 42
Argument 8 : --an-option
```

The first argument is always the name of the program, in the way it has been entered by the user. The other arguments appear in the same order that they have been entered.

It is completely up to the program how it makes use of the arguments. The following program prints its two arguments in reverse order:

```
import std.stdio;

int main(string[] args)
{
    if (args.length != 3) {
        stderr.writeln("ERROR! Correct usage:\n"
            " %s word1 word2", args[0]);
        return 1;
    }

    writeln(args[2], ' ', args[1]);

    return 0;
}
```

The program also shows its correct usage if there are not two words exactly:

```
# ./deneme
ERROR! Correct usage:
 ./deneme word1 word2
# ./deneme world hello
hello world
```

39.4 Command line options and the `std.getopt` module

That is all there is to know about the parameters and the return value of `main()`. However, parsing the arguments is a repetitive task. The `std.getopt` module is designed to help with parsing the command line options of programs.

Some parameters like "world" and "hello" above are purely data for the program to use. Other kinds of parameters are called *command line options*, and are for changing behaviors of programs. An example of a command line option has been `-l` that has been passed to `ls` above.

Command line options make programs more useful by removing the need for a human user to interact with the program to have it behave in a certain way. With command line options, programs can be started from script programs and their behaviors can be specified through command line options.

Although the syntax and meanings of command line arguments of every program is specific to that program, their format is somewhat standard. For example, in POSIX, command line options start with `--` followed by the name of the option, and values come after `=` characters:

```
# ./deneme --an-option=17
```

The `std.getopt` module simplifies parsing such options. It has more capabilities than what is covered in this section.

Let's design a program that prints random numbers. Let's take the minimum, maximum, and total number of these numbers as program arguments. Let's expect the following syntax to get these values from the command line:

```
# ./deneme --count=7 --minimum=10 --maximum=15
```

The `getopt()` function parses and assigns those values to variables. Similar to `readf()`, the addresses of variables must be specified by the `&` operator:

```
import std.stdio;
import std.getopt;
import std.random;

void main(string[] args)
{
    int count;
    int minimum;
    int maximum;

    getopt(args,
           "count", &count,
           "minimum", &minimum,
           "maximum", &maximum);

    foreach (i; 0 .. count) {
        write(uniform(minimum, maximum + 1), ' ');
    }

    writeln();
}
```

```
# ./deneme --count=7 --minimum=10 --maximum=15
11 11 13 11 14 15 10
```

Most command line options of most programs have shorter syntax as well. For example, `-c` may have the same meaning as `--count`. Such alternative syntax for each option is specified after a `|` character. There may be more than one shortcut for each option:

```
getopt(args,
       "count|c", &count,
       "minimum|n", &minimum,
       "maximum|x", &maximum);
```

It is common to use a single dash for the short versions and the `=` character is usually omitted:

```
# ./deneme -c7 -n10 -x15
11 13 10 15 14 15 14
```

`getopt()` converts the arguments from **string** to the type of each variable. For example, since `count` above is **int**, `getopt()` converts the value specified for the `--count` argument to **int**. When needed, such conversions may also be performed explicitly by `to`.

So far we have used `std.conv.to` only when converting to **string**. `to` can in fact convert from any type to any type as long as that conversion is possible. For example, the following program takes advantage of `to` when converting its argument to **size_t**:

```
import std.stdio;
import std.conv;

void main(string[] args)
{
    // The default count is 10
    size_t count = 10;

    if (args.length > 1) {
        // There is an argument
        count = to!size_t(args[1]);
    }

    foreach (i; 0 .. count) {
        write(i * 2, ' ');
    }

    writeln();
}
```

The program produces 10 numbers when there is no argument specified:

```
# ./deneme
0 2 4 6 8 10 12 14 16 18
# ./deneme 3
0 2 4
```

39.5 Environment variables

The environment that a program is started in provides some variables that the program may make use of. The environment variables can be accessed by `std.process.getenv`. The following program is accessing and printing the **PATH** environment variable:

```
import std.stdio;
import std.process;

void main()
{
    writeln(getenv("PATH"));
}
```

The output:

```
# ./deneme
/usr/local/bin:/usr/bin:/home/acehrelidmd/linux/bin
```

`std.process.environment` provides access to the environment variables through the associative array syntax:

```
import std.process;
// ...
writeln(environment["PATH"]);
```

However, **environment** itself is not an associative array. When needed, the environment variables can be converted to an associative array by **toAA()**:

```
string[string] envVars = environment.toAA();
```

39.6 Starting other programs

Programs may start other programs and become the *environment* for those programs. Two functions that enable this are **system** and **shell** from the **std.process** module.

system executes its parameter as if the command has been typed at the console. It then returns the return value of that command:

```
import std.stdio;
import std.process;

void main()
{
    immutable retVal = system("ls -l deneme");
    writeln("ls has returned %s.", retVal);
}
```

The output includes the output of **ls** followed by the line that this program prints:

```
# ./deneme
-rwxr-xr-x 1 acehrelt users 461107 Nov  7 17:33 deneme
ls has returned 0.
```

shell executes its parameter in a shell environment and returns the standard output of that command. It throws an exception if the program could not be started or finished with a non-zero error code:

```
import std.stdio;
import std.process;

void main()
{
    immutable output = shell("ls -l deneme*");
    writeln("The output of ls:");
    writeln(output);
}
```

The output:

```
# ./deneme
The output of ls:
-rwxrwxr-x 1 acehrelt acehrelt  923117 2012-03-25 11:19 deneme
-rwxr-xr-x 1 acehrelt acehrelt 1091303 2012-03-25 11:19 deneme.d
```

39.7 Summary

- Even when it is defined with return type `void`, `main()` automatically returns 0 for success and 1 for failure.
- `stderr` is suitable to print error messages.
- `main` can take parameters as `string[]`.
- `std.getopt` helps with parsing command line options.
- `std.process` helps with accessing environment variables and starting other programs.

39.8 Exercises

1. Write a calculator program that takes an operator and two operands as command line arguments. Have the program support the following usage:

```
# ./deneme 3.4 x 7.8
26.52
```

Note: Because the `*` character has a special meaning on most consoles, I have used `x`. You may still use `*` as long as it is escaped as `*`.

2. Write a program that asks the user which program to start, starts that program, and returns its return value.

[... the solutions](#)

40 Exceptions

Unexpected situations are parts of programs: user mistakes, programming errors, changes in the program environment, etc. Programs must be written in ways to avoid producing incorrect results when faced with such *exceptional* conditions.

Some of these conditions may be severe enough to stop the execution of the program. For example, a required piece of information may be missing or invalid, or a device may not be functioning correctly. The exception handling mechanism of D helps with stopping program execution when necessary, and to recover from the unexpected situations when possible.

As an example of a severe condition, we can think of passing an unknown operator to a function that knows only the four arithmetic operators, as we have seen in the exercises of the previous chapter:

```
switch (operator) {
case "+":
    writeln(first + second);
    break;

case "-":
    writeln(first - second);
    break;

case "x":
    writeln(first * second);
    break;

case "/":
    writeln(first / second);
    break;

default:
    throw new Exception(format("Invalid operator: %s", operator));
}
```

The **switch** statement above does not know what to do with operators that are not listed on the **case** statements; so throws an exception.

There are many examples of thrown exceptions in Phobos. For example, **to!int**, which can be used to convert a string representation of an integer to an **int** value throws an exception when that representation is not valid:

```
import std.conv;

void main()
{
    const int value = to!int("hello");
}
```

The program terminates with an exception that is thrown by **to!int**:

```
std.conv.ConvException@std/conv.d(38): std.conv(1157): Can't
convert value `hello' of type const(char)[] to type int
```

std.conv.ConvException at the beginning of the message is the type of the thrown exception object. We can tell from the name that the type is **ConvException** that is defined in the **std.conv** module.

40.1 The throw statement to throw exceptions

We've seen the **throw** statement both in the examples above and in the previous chapters.

throw throws an *exception object* and this terminates the current operation of the program. The expressions and statements that are written after the **throw** statement are not executed. This behavior is according to the nature of exceptions: they must be thrown when the program cannot continue with its current task.

Conversely, if the program could continue then the situation would not warrant throwing an exception. In such cases the function would find a way and continue.

40.1.1 The exception types **Exception** and **Error**

Only the types that are inherited from the **Throwable** class can be thrown. **Throwable** is almost never used directly in programs. The types that are actually thrown are types that are inherited from **Exception** or **Error**, which themselves are the types that are inherited from **Throwable**. For example, all of the exceptions that Phobos throws are inherited from either **Exception** or **Error**.

Error represents unrecoverable conditions and is not recommended to be *caught*. For that reason, most of the exceptions that a program throws are the types that are inherited from **Exception**. (**Note:** *Inheritance is a topic related to classes. We will see classes in a later chapter.*)

Exception objects are constructed with a **string** value that represents an error message. You may find it easy to create this message with the **format()** function from the **std.string** module:

```
import std.stdio;
import std.random;
import std.string;

int[] randomDiceValues(int count)
{
    if (count < 0) {
        throw new Exception(
            format("Invalid dice count: %s", count));
    }

    int[] values;

    foreach (i; 0 .. count) {
        values ~= uniform(1, 7);
    }

    return values;
}

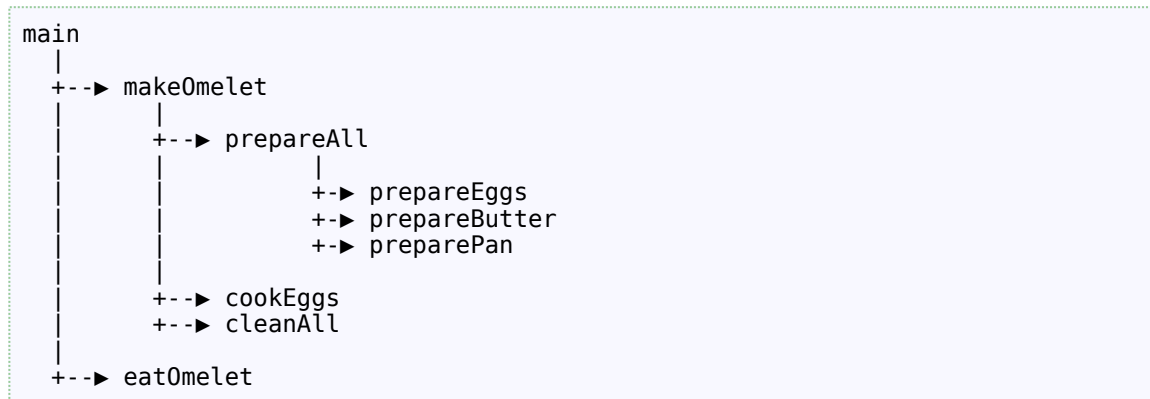
void main()
{
    writeln(randomDiceValues(-5));
}
```

```
object.Exception...: Invalid dice count: -5
```

40.1.2 Throw exception terminates all scopes

We have seen that the program execution starts from the `main` function and branches into other functions from there. This layered execution of going deeper into functions and eventually returning from them can be seen as the branches of a tree.

For example, `main` may call a function named `makeOmelet`, which in turn may call another function named `prepareAll`, which in turn may call another function named `prepareEggs`, etc. Assuming that the arrows indicate function calls, the branching of such a program can be shown as in the following function call tree:



The following program demonstrates the branching above by using different levels of indentation in its output. The program doesn't do anything useful other than producing an output suitable to our purposes:

```

import std.stdio;

void indent(in int level)
{
    foreach (i; 0 .. level * 2) {
        write(' ');
    }
}

void entering(in char[] functionName, in int level)
{
    indent(level);
    writeln("▶ ", functionName, "'s first line");
}

void exiting(in char[] functionName, in int level)
{
    indent(level);
    writeln("◀ ", functionName, "'s last line");
}

void main()
{
    entering("main", 0);
    makeOmelet();
    eatOmelet();
    exiting("main", 0);
}

void makeOmelet()
{
    entering("makeOmelet", 1);
    prepareAll();
    cookEggs();
    cleanAll();
    exiting("makeOmelet", 1);
}

```



```
void eatOmelet()
{
    entering("eatOmelet", 1);
    exiting("eatOmelet", 1);
}

void prepareAll()
{
    entering("prepareAll", 2);
    prepareEggs();
    prepareButter();
    preparePan();
    exiting("prepareAll", 2);
}

void cookEggs()
{
    entering("cookEggs", 2);
    exiting("cookEggs", 2);
}

void cleanAll()
{
    entering("cleanAll", 2);
    exiting("cleanAll", 2);
}

void prepareEggs()
{
    entering("prepareEggs", 3);
    exiting("prepareEggs", 3);
}

void prepareButter()
{
    entering("prepareButter", 3);
    exiting("prepareButter", 3);
}

void preparePan()
{
    entering("preparePan", 3);
    exiting("preparePan", 3);
}
```

The program produces the following output:

```

▶ main, first line
  ▶ makeOmelet, first line
    ▶ prepareAll, first line
      ▶ prepareEggs, first line
        ◁ prepareEggs, last line
      ▶ prepareButter, first line
        ◁ prepareButter, last line
      ▶ preparePan, first line
        ◁ preparePan, last line
    ◁ prepareAll, last line
  ▶ cookEggs, first line
    ◁ cookEggs, last line
  ▶ cleanAll, first line
    ◁ cleanAll, last line
◁ makeOmelet, last line
▶ eatOmelet, first line
  ◁ eatOmelet, last line
◁ main, last line

```

The functions **entering** and **exiting** are used to indicate the first and last lines of functions with the help of the ▶ and ◁ characters. The program starts with the first line of **main**, branches down to other functions, and finally ends with the last line of **main**.

Let's modify the **prepareEggs** function to take the number of eggs as a parameter. Since certain values of this parameter would be an error, let's have this function throw an exception when the number of eggs is less than one:

```

void prepareEggs(int count)
{
    entering("prepareEggs", 3);

    if (count < 1) {
        throw new Exception(
            format("Cannot take %s eggs from the fridge", count));
    }

    exiting("prepareEggs", 3);
}

```

In order to be able to compile the program, we must modify other lines of the program to be compatible with this change. The number of eggs to take out of the fridge can be handed down from function to function, starting with **main**. The parts of the program that need to change are the following. The invalid value of -8 is intentional to show how the output of the program will be different from the previous output when an exception is thrown:

```

// ...

void main()
{
    entering("main", 0);
    makeOmelet(-8);
    eatOmelet();
    exiting("main", 0);
}

void makeOmelet(int eggCount)
{
    entering("makeOmelet", 1);
    prepareAll(eggCount);
}

```

```

    cookEggs();
    cleanAll();
    exiting("makeOmelet", 1);
}

// ...

void prepareAll(int eggCount)
{
    entering("prepareAll", 2);
    prepareEggs(eggCount);
    prepareButter();
    preparePan();
    exiting("prepareAll", 2);
}

// ...

```

When we start the program now, we see that the lines that used to be printed after the point where the exception is thrown are missing:

```

▶ main, first line
  ▶ makeOmelet, first line
    ▶ prepareAll, first line
      ▶ prepareEggs, first line
object.Exception: Cannot take -8 eggs from the fridge

```

When the exception is thrown, the program execution exits the **prepareEggs**, **prepareAll**, **makeOmelet** and **main** functions in that order, from the bottom level to the top level. No additional steps are executed as the program exits these functions.

The rationale for such a drastic termination is that a failure in a lower level function would mean that the higher level functions that needed it should also be considered as failed.

The exception object that is thrown from a lower level function is transferred to the higher level functions one level at a time and causes the program to finally exit the **main** function. The path that the exception takes can be shown as the red path in the following tree:



The point of the exception mechanism is precisely this behavior of exiting all of the layers of function calls right away. Sometimes it makes sense to *catch* the thrown exception to find a way to continue the execution of the program. I will introduce the **catch** keyword below.

40.1.3 When to use throw

Use **throw** in situations when it is not possible to continue. For example, a function that reads the number of students from a file may throw an exception if this information is not available or incorrect.

On the other hand, if the problem is caused by some user action like entering invalid data, it may make more sense to validate the data instead of throwing an exception. Displaying an error message and asking the user to re-enter the data is more appropriate in many cases.

40.2 The try-catch statement to catch exceptions

As we've seen above, a thrown exception causes the program execution to exit all functions and this finally terminates the whole program.

The exception object can be caught by a **try-catch** statement at any point on its path as it exits the functions. The **try-catch** statement models the phrase "try to do something, and catch exceptions that may be thrown." Here is the syntax of **try-catch**:

```
try {
    // the code block that is being executed, where an
    // exceptions may potentially be thrown
} catch (an_exception_type) {
    // expressions to execute if an exception of this
    // type is caught
} catch (another_exception_type) {
    // expressions to execute if an exception of this
    // other type is caught

// ... more catch blocks as appropriate ...

} finally {
    // expressions to execute regardless of whether an
    // exception is thrown
}
```

Let's start with the following program that does not use a **try-catch** statement at this state. The program reads the value of a die from a file and prints it to the standard output:

```
import std.stdio;

int readDieFromFile()
{
    auto file = File("the_file_that_contains_the_value", "r");

    int die;
    file.readf(" %s", &die);

    return die;
}

void main()
{
    const int die = readDieFromFile();

    writeln("Die value: ", die);
}
```

Note that the **readDieFromFile** function is written in a way that ignores error conditions, expecting that the file and the value that it contains are available. In other words, the function is dealing only with its own task instead of paying attention to error conditions. This

is a benefit of exceptions: many functions can be written in ways that focus on their actual tasks, rather than focusing on error conditions.

Let's start the program when `the_file_that_contains_the_value` is missing:

```
std.exception.ErrnoException@std/stdio.d(286): Cannot open
file `the_file_that_contains_the_value' in mode `r' (No such
file or directory)
```

An exception of type `ErrnoException` is thrown and the program terminates without printing "Die value: ".

Let's add an intermediate function to the program that calls `readDieFromFile` from within a `try` block and let's have `main` call this new function:

```
import std.stdio;

int readDieFromFile()
{
    auto file = File("the_file_that_contains_the_value", "r");

    int die;
    file.readf(" %s", &die);

    return die;
}

int tryReadingFromFile()
{
    int die;

    try {
        die = readDieFromFile();
    } catch (std.exception.ErrnoException exc) {
        writeln("(Could not read from file; assuming 1)");
        die = 1;
    }

    return die;
}

void main()
{
    const int die = tryReadingFromFile();

    writeln("Die value: ", die);
}
```

When we start the program again when `the_file_that_contains_the_value` is still missing, this time the program does not terminate with an exception:

```
(Could not read from file; assuming 1)
Die value: 1
```

The new program *tries* executing `readDieFromFile` in a `try` block. If that block executes successfully, the function ends normally with the `return die;` statement. If the execution of the `try` block ends with the specified `std.exception.ErrnoException`, then the program execution enters the `catch` block.

The following is a summary of events when the program is started when the file is missing:

- like in the previous program, a `std.exception.ErrnoException` object is thrown, (it is thrown by `File()`, not by our code)
- this exception is caught by `catch`,
- the value of 1 is assumed during the normal execution of the `catch` block,
- and the program continues its normal operations.

`catch` is to catch thrown exceptions presumably to find a way to continue executing the program.

As another example, let's go back to the omelet program and add a `try-catch` statement to its `main` function:

```
void main()
{
    entering("main", 0);

    try {
        makeOmelet(-8);
        eatOmelet();
    } catch (Exception exc) {
        write("Failed to eat omelet: ");
        writeln("", exc.msg, "");
        writeln("Shall eat at the neighbor's...");
    }

    exiting("main", 0);
}
```

That `try` block contains two lines of code. Any exception thrown from either of those lines would be caught by the `catch` block.

```
► main, first line
  ► makeOmelet, first line
    ► prepareAll, first line
      ► prepareEggs, first line
Failed to eat omelet: "Cannot take -8 eggs from the fridge"
Shall eat at the neighbor's...
◀ main, last line
```

As can be seen from the output, the program doesn't terminate because of the thrown exception anymore. It recovers from the error condition and continues executing normally till the end of the `main` function.

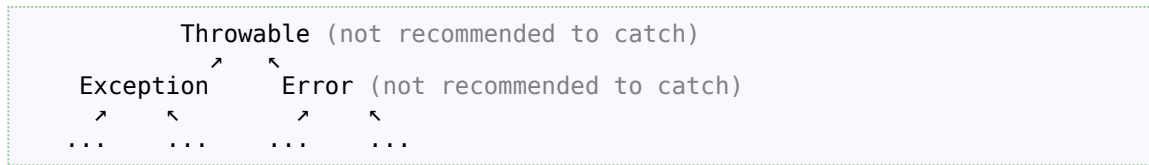
40.2.1 catch blocks are looked up sequentially

The type `Exception`, which we have used so far in the examples is a general exception type. This type merely specifies that an error occurred in the program. It also contains a message that can explain the error further, but it does not contain information about the *type* of the error.

`ConvException` and `ErrnoException` that we have seen earlier in this chapter are more specific exception types: the former is about a conversion error, and the latter is about a system error. Like many other exception types in Phobos and as their names suggest, `ConvException` and `ErrnoException` are both inherited from the `Exception` class.

`Exception` and its sibling `Error` are further inherited from `Throwable`, the most general exception type.

Although possible, it is not recommended to catch objects of type **Error** and objects of types that are inherited from **Error**. Since it is more general than **Error**, it is not recommended to catch **Throwable** either. What should normally be caught are the types that are under the **Exception** hierarchy.



Note: I will explain the hierarchy representations later in the *Inheritance* chapter. The tree above indicates that **Throwable** is the most general and **Exception** and **Error** are more specific.

It is possible to catch exception objects of a particular type. For example, it is possible to catch an **ErrnoException** object specifically to detect and handle a system error.

Exceptions are caught only if they match a type that is specified in a **catch** block. For example, a catch block that is trying to catch a **SpecialExceptionType** would not catch an **ErrnoException**.

The type of the exception object that is thrown during the execution of a **try** block is matched to the types that are specified by the **catch** blocks, in the order that the **catch** blocks are written. If the type of the object matches the type of the **catch** block, then the exception is considered to be caught by that **catch** block, and the code that is within that block is executed. Once a match is found, the remaining **catch** blocks are ignored.

Because the **catch** blocks are matched in order from the first to the last, the **catch** blocks must be ordered from the most specific exception types to the most general exception types. Accordingly, and if it makes sense to catch that type of exceptions, the **Exception** type must be specified at the last **catch** block.

For example, a **try-catch** statement that is trying to catch several specific types of exceptions about student records may order the **catch** blocks from the most specific to the most general as in the following code:

```

try {
    // operations about student records that may throw ...
} catch (StudentIdDigitException exc) {
    // an exception that is specifically about errors with
    // the digits of student ids
} catch (StudentIdException exc) {
    // a more general exception about student ids but not
    // necessarily about their digits
} catch (StudentRecordException exc) {
    // even more general exception about student records
} catch (Exception exc) {
    // the most general exception that may not be related
    // to student records
}
  
```

40.2.2 The finally block

finally is an optional block of the **try-catch** statement. It includes expressions that should be executed regardless of whether an exception is thrown or not.

To see how **finally** works, let's look at a program that throws an exception 50% of the time:

```
import std.stdio;
import std.random;

void throwsHalfTheTime()
{
    if (uniform(0, 2) == 1) {
        throw new Exception("the error message");
    }
}

void foo()
{
    writeln("the first line of foo()");

    try {
        writeln("the first line of the try block");
        throwsHalfTheTime();
        writeln("the last line of the try block");

        // ... there may be one or more catch blocks here ...
    } finally {
        writeln("the body of the finally block");
    }

    writeln("the last line of foo()");
}

void main()
{
    foo();
}
```

The output of the program is the following when the function does not throw:

```
the first line of foo()
the first line of the try block
the last line of the try block
the body of the finally block
the last line of foo()
```

The output of the program is the following when the function does throw:

```
the first line of foo()
the first line of the try block
the body of the finally block
object.Exception@deneme.d: the error message
```

As can be seen, although "the last line of the try block" and "the last line of foo()" are not printed, the content of the **finally** block is still executed when an exception is thrown.

40.2.3 When to use the try-catch statement

The **try-catch** statement is useful to catch exceptions to do something special about them.

For that reason, the **try-catch** statement should be used only when there is something special to be done. Do not catch exceptions otherwise and leave them to higher level functions that may want to catch them.

40.3 Types of errors

We have seen how useful the exception mechanism is. It enables both the lower and higher level operations to be aborted right away, instead of the program continuing with incorrect or missing data, or behaving in any other incorrect way.

This does not mean that every error condition warrants throwing an exception. There may be better things to do depending on the types of errors.

40.3.1 User errors

Some of the errors are caused by the user. As we have seen above, the user may have entered a string like "hello" even though the program has been expecting a number. It may be more appropriate to display an error message and ask the user to enter appropriate data again.

Even so, it may be fine to accept and use the data directly without validating the data up front; as long as the code that uses the data would throw anyway. What is important is to be able to notify the user why the data is not suitable.

For example, let's look at a program that takes a file name from the user. There are at least two ways of dealing with potentially invalid file names:

- **Validating the data before use:** We can determine whether the file with the given name exists by calling `exists()` of the `std.file` module:

```
if (exists(fileName)) {
    // yes, the file exists
} else {
    // no, the file doesn't exist
}
```

This gives us the chance to be able to open the data only if it exists. Unfortunately, it is still possible that the file cannot be opened even if `exists()` returns `true`, if for example another process on the systems deletes or renames the file before this program actually opens it.

For that reason, the following method may be more useful.

- **Using the data without first validating it:** We can assume that the data is valid and start using it right away, because `File` would throw an exception if the file cannot be opened anyway.

```
import std.stdio;
import std.string;

void useTheFile(string fileName)
{
    auto file = File(fileName, "r");
    // ...
}

string read_string(in char[] prompt)
{
    write(prompt, ": ");
```

```

    return chomp(readln());
}

void main()
{
    bool is_fileUsed = false;

    while (!is_fileUsed) {
        try {
            useTheFile(
                read_string("Please enter a file name"));

            /*
             * If we are at this line, it means that
             * useTheFile() function has been completed
             * successfully. This indicates that the file
             * name has been valid.
             *
             * We can now set the value of the loop flag to
             * terminate the while loop.
             */
            is_fileUsed = true;
            writeln("The file has been used successfully");
        } catch (std.exception.ErrnoException exc) {
            stderr.writeln("This file could not be opened");
        }
    }
}

```

40.3.2 Programmer errors

Some errors are caused by programmer errors. For example the programmer may think that a function that has just been written will always be called with a value greater than or equal zero, and this may be true according to the design of the program. The function having still been called with a value less than zero would indicate either a mistake in the design of the program or in the implementation of that design. Both of these can be thought of programming errors.

It is more appropriate to use **assert** instead of the exception mechanism for errors that are caused by programmer errors. (**Note:** We will cover **assert** in a later chapter.)

```

void processMenuSelection(int selection)
{
    assert(selection >= 0);
    // ...
}

void main()
{
    processMenuSelection(-1);
}

```

The program terminates with an **assert** failure:

```
core.exception.AssertError@deneme.d(3): Assertion failure
```

assert validates program state and prints the file name and line number of the validation if it fails. The message above indicates that the assertion at line 3 of deneme.d has failed.

40.3.3 Unexpected situations

For unexpected situations that are outside of the two general cases above, it is still appropriate to throw exceptions. If the program cannot continue its execution, there is nothing else to do but to throw.

It is up to the higher layer functions that call this function to decide what to do with thrown exceptions. They may catch the exceptions that we throw to remedy the situation.

40.4 Summary

- When faced with a user error either warn the user right away or ensure that an exception is thrown; the exception may be thrown anyway by another function when using incorrect data, or you may throw directly.
- Use **assert** to validate program logic and implementation. (**Note:** *I will explain **assert** in the next chapter.*)
- When in doubt, **throw** an exception.
- Catch exceptions if and only if you can do something useful about that exception. Otherwise, do not encapsulate code with a **try-catch** statement; instead, leave the exceptions to higher layers of the code that may do something about them.
- Order the **catch** blocks from the most specific to the most general.
- Put the expressions that must always be executed when leaving a scope, in **finally** blocks.

41 scope

As we have seen, expressions that must always be executed are written in the **finally** block, and expressions that must be executed when there are error conditions are written in **catch** blocks.

We can make the following observations about the use of these blocks:

- **catch** and **finally** cannot be used without a **try** block.
- Some of the variables that these blocks need, may not be accessible within these blocks:

```
void foo(ref int r)
{
    try {
        int addend = 42;

        r += addend;
        mayThrow();

    } catch (Exception exc) {
        r -= addend;           // ← compilation ERROR
    }
}
```

That function first modifies the reference parameter, and then reverts this modification when an exception is thrown. Unfortunately, **addend** is accessible only in the **try** block, where it is defined. (**Note:** This is related to name spaces; as well as object lifetimes that will be explained in a later chapter.)

- Writing all of potentially unrelated expressions in the single **finally** block at the bottom separates those expressions from the actual code that they are related to.

The **scope** statements have similar functionality with the **catch** and **finally** scopes but they are better in many respects. The three **scope** statements are also about expressions that should be executed when leaving scopes:

- **scope(exit)**: the expression is always executed when exiting the scope, regardless of whether successfully or due to an exception
- **scope(success)**: the expression is executed only if the scope is being exited successfully
- **scope(failure)**: the expression is executed only if the scope is being exited due to an exception

Although these statements are closely related to exceptions, they can be used without a **try-catch** block.

As an example, let's write the function above with a **scope(failure)** statement:

```
void foo(ref int r)
{
    int addend = 42;

    r += addend;
    scope(failure) r -= addend;

    mayThrow();
}
```

The **scope(failure)** statement above ensures that the **r -= addend** expression will be executed if the function's scope is exited due to an exception. A benefit of **scope(failure)** is the fact that the expression that reverts another expression is written close to it.

scope statements can be specified as blocks as well:

```
scope(exit) {
    // ... expressions ...
}
```

Here is another function that tests all three of these statements:

```
void test()
{
    scope(exit) writeln("when exiting 1");

    scope(success) {
        writeln("if successful 1");
        writeln("if successful 2");
    }

    scope(failure) writeln("if thrown 1");
    scope(exit) writeln("when exiting 2");
    scope(failure) writeln("if thrown 2");

    throwsHalfTheTime();
}
```

If no exception is thrown, the output of the function includes only the **scope(exit)** and **scope(success)** expressions:

```
when exiting 2
if successful 1
if successful 2
when exiting 1
```

If an exception is thrown, the output includes the **scope(exit)** and **scope(failure)** expressions:

```
if thrown 2
when exiting 2
if thrown 1
when exiting 1
object.Exception@...: the error message
```

As seen in the outputs, the blocks of the **scope** statements are executed in reverse order. This is because later code may depend on previous variables. Executing the **scope** statements in reverse order enables undoing side effects of earlier expressions in a consistent order.

42 assert and enforce

In the previous two chapters we have seen how exceptions and **scope** statements are used towards program correctness. **assert** is another powerful tool to achieve the same goal by ensuring that certain assumptions that the program is based on are valid.

It may sometimes be difficult to decide whether to throw an exception or to call **assert**. I will use **assert** in all of the examples below without much justification. I will explain the differences later in the chapter.

Although not always obvious, programs are full of assumptions. For example, the following function is written under the assumption that both age parameters are greater than or equal to zero:

```
double averageAge(double first, double second)
{
    return (first + second) / 2;
}
```

Although it may be invalid for the program to ever have an age value that is negative, the function would still produce an average, which may be used in the program unnoticed, resulting in the program's continuing with incorrect data.

As another example, the following function assumes that it will always be called with two commands: "sing" or "dance":

```
void applyCommand(string command)
{
    if (command == "sing") {
        robotSing();
    } else {
        robotDance();
    }
}
```

Because of that assumption, the **robotDance()** function would be called for every command other than "sing", valid or invalid.

When such assumptions are kept only in the programmer's mind, the program may end up working incorrectly. **assert** checks assumptions and terminates programs immediately when they are not valid.

42.1 Syntax

assert can be used in two ways:

```
assert(logical_expression);
assert(logical_expression, message);
```

The logical expression represents an assumption about the program. **assert** evaluates that expression to validate that assumption. If the value of the logical expression is **true** then the assumption is considered to be valid. Otherwise the assumption is invalid and an **AssertError** is thrown.

As its name suggests, this exception is inherited from **Error**, and as you may remember from the [Exceptions chapter](#), exceptions that are inherited from **Error** must never be caught. It is

important for the program to be terminated right away instead of continuing under invalid assumptions.

The two implicit assumptions of `averageAge()` above may be spelled out by two `assert` calls as in the following function:

```
double averageAge(double first, double second)
{
    assert(first >= 0);
    assert(second >= 0);

    return (first + second) / 2;
}

void main()
{
    auto result = averageAge(-1, 10);
}
```

Those `assert` checks carry the meaning "assuming that both of the ages greater than or equal to zero". It can also be thought of as meaning "this function can work correctly only if both of the ages are greater than or equal to zero".

`assert` checks those assumptions and terminates the program with an `AssertError` when they are not valid:

```
core.exception.AssertError@deneme(3): Assertion failure
```

The part after the `@` character in the message indicates the source file and the line number of the `assert` check that has failed. According to the output above, the `assert` that has failed is on line 3 of file `deneme.d`.

The other syntax of `assert` allows printing a custom message when the `assert` check fails:

```
assert(first >= 0, "Age cannot be negative.");
```

The output:

```
core.exception.AssertError@deneme.d(3): Age cannot be negative.
```

Sometimes it is thought to be impossible for the program to ever enter a code path. In such cases it is common to use the literal `false` as the logical expression to fail an `assert` check. For example, to indicate that `applyCommand()` function is never expected to be called with a command other than "sing" and "dance", and to guard against such a possibility, an `assert(false)` can be inserted into the *impossible* branch:

```
void applyCommand(string command)
{
    if (command == "sing") {
        robotSing();
    } else if (command == "dance") {
        robotDance();
    } else {
        assert(false);
    }
}
```

The function is guaranteed to work with the only two commands that it knows about. (**Note:** An alternative choice here would be to use a **final switch** statement.)

42.2 static assert

Since **assert** checks are for correct execution of programs, they are applied when the program is actually running. Other checks are about the structure of the program and can be applied even at compile time.

static assert is the counterpart of **assert** that is applied at compile time. The advantage is that it does not allow even compiling a program that would have otherwise run incorrectly. A natural requirement is that it must be possible to evaluate the logical expression at compile time.

For example, assuming that the title of a menu will be printed on an output device that has limited width, the following **static assert** ensures that it will never be wider than that limit:

```
enum dstring menuTitle = "Command Menu";
static assert(menuTitle.length <= 16);
```

Note that the string is defined as **enum** so that its length can be evaluated at compile time.

Let's assume that a programmer changes that title to make it more descriptive:

```
enum dstring menuTitle = "Directional Commands Menu";
static assert(menuTitle.length <= 16);
```

The **static assert** check prevents compiling the program:

```
Error: static assert (25u <= 16u) is false
```

This would remind the programmer of the limitation of the output device.

static assert is even more useful when used in templates. We will see templates in later chapters.

42.3 assert even if *absolutely true*

I emphasize "absolutely true" because assumptions about the program are never expected to be false anyway. A large set of program errors are based on assumptions that are thought to be absolutely true.

For that reason, take advantage of **assert** checks even if they feel unnecessary. Let's look at the following function that returns the days of months in a given year:

```
int[] monthDays(in int year)
{
    int[] days = [
        31, februaryDays(year),
        31, 30, 31, 30, 31, 31, 30, 31, 30, 31
    ];

    assert((sum(days) == 365) ||
           (sum(days) == 366));

    return days;
}
```


That **assert** check may be seen as unnecessary because the function would naturally return either 365 or 366. However, those checks are guarding against potential mistakes even in the **februaryDays()** function. For example, the program would be terminated if **februaryDays()** returned 30.

Another seemingly unnecessary check can ensure that the length of the slice would always be 12:

```
assert(days.length == 12);
```

That way, deleting or adding elements to the slice unintentionally would also be caught. Such checks are important tools towards program correctness.

assert is also the basic tool that is used in *unit testing* and *contract programming*, both of which will be covered in later chapters.

42.4 No value nor side effect

We have seen that expressions produce values or make side effects. **assert** checks do not have values nor they *should* have any side effects.

The D language requires that the evaluation of the logical expression must not have any side effect. **assert** must remain as a passive observer of program state.

42.5 Disabling assert checks

Since **assert** is about program correctness, they can be seen as unnecessary once the program has been tested sufficiently. Further, since **assert** checks produce no values nor they have side effects, removing them from the program should not make any difference.

The compiler switch **-release** causes the **assert** checks to be ignored as if they have never been included in the program:

```
dmd deneme.d -release
```

This would allow programs run faster by not evaluating potentially slow logical expressions of the **assert** checks.

As an exception, the **assert** checks that have the literal **false** (or 0) as the logical expression are not disabled even when the program is compiled with **-release**. This is because **assert(false)** is for ensuring that a block of code is never reached, and that should be prevented even for **-release** builds.

42.6 enforce for throwing exceptions

Not every unexpected situation is an indication of a program error. Programs may also experience unexpected inputs and unexpected environmental state. For example, the data that is entered by the user should not be validated by an **assert** check because invalid data has nothing to do with the correctness of the program itself. In such cases it is appropriate to throw exceptions like we have been doing in previous programs.

std.exception.enforce is a convenient way of throwing exceptions. For example, let's assume that an exception must be thrown when a specific condition is not met:

```
if (count < 3) {
    throw new Exception("Must be at least 3.");
}
```

enforce() is a wrapper around the condition check and the **throw** statement. The following is the equivalent of the previous code:

```
import std.exception;
// ...
enforce(count >= 3, "Must be at least 3.");
```

Note how the logical expression is negated compared with the **if** statement, as it now spells out what is being *enforced*.

42.7 How to use

assert is for catching programmer errors. The conditions that **assert** guards against in the **monthDays()** function and the **menuTitle** variable above are all about programmer mistakes.

Sometimes it is difficult to decide whether to rely on an **assert** check or to throw an exception. The decision should be based on whether the unexpected situation is due to a problem with how the program has been coded.

Otherwise, the program must throw an exception when it is not possible to accomplish a task. **enforce()** is expressive and convenient when throwing exceptions.

Another point to consider is whether the unexpected situation can be remedied in any way. If the program can do anything special, even by simply printing an error message about the problem with some input data, then it is appropriate to throw an exception. That way, callers of the code that threw the exception can catch it to do something special to recover from the error condition.

42.8 Exercises

- The following program includes a number of **assert** checks. Compile and run the program to discover its bugs that are revealed by those **assert** checks.
The program takes a start time and a duration from the user and calculates the end time by adding the duration to the start time:

```
10 hours and 8 minutes after 06:09 is 16:17.
```

Note that this problem can be written in a much cleaner way by defining **struct** types. We will refer to this program in later chapters.

```
import std.stdio;
import std.string;
import std.exception;

/* Reads the time as hour and minute after printing a
 * message. */
void readTime(in string message, out int hour, out int minute)
{
    write(message, "? (HH:MM) ");

    readf(" %s:%s", &hour, &minute);

    enforce((hour >= 0) && (hour <= 23) &&
           (minute >= 0) && (minute <= 59),
```

```

        "Invalid time!");
    }

    /* Returns the time in string format. */
    string timeToString(in int hour, in int minute)
    {
        assert((hour >= 0) && (hour <= 23));
        assert((minute >= 0) && (minute <= 59));

        return format("%02s:%02s", hour, minute);
    }

    /* Adds duration to start time and returns the result as the
     * third pair of parameters. */
    void addDuration(in int startHour, in int startMinute,
                   in int durationHour, in int durationMinute,
                   out int resultHour, out int resultMinute)
    {
        resultHour = startHour + durationHour;
        resultMinute = startMinute + durationMinute;

        if (resultMinute > 59) {
            ++resultHour;
        }
    }

    void main()
    {
        int startHour;
        int startMinute;
        readTime("Start time", startMinute, startHour);

        int durationHour;
        int durationMinute;
        readTime("Duration", durationHour, durationMinute);

        int endHour;
        int endMinute;
        addDuration(startHour, startMinute,
                   durationHour, durationMinute,
                   endHour, endMinute);

        writefln("%s hours and %s minutes after %s is %s.",
                 durationHour, durationMinute,
                 timeToString(startHour, startMinute),
                 timeToString(endHour, endMinute));
    }

```

Run the program and enter **06:09** as the start time and **1:2** as the duration. Observe that the program terminates normally.

Note: You may notice a problem with the output. Ignore that problem for now as you will discover it by the help of **assert** checks soon.

2. This time enter **06:09** and **15:2**. Observe that the program is terminated by an **AssertError**. Go to the line of the program that is indicated in the assert message and see which one of the **assert** checks have failed. It may take a while to discover the cause of this particular failure.
3. Enter **06:09** and **20:0** and observe that the same **assert** check still fails and fix that bug as well.
4. Modify the program to print the times in 12-hour format with an "am" or "pm" indicator.

... the solutions

43 Unit Testing

As it should be known by most people, any device that runs some piece of computer program contains software bugs. Software bugs plague computer devices from the simplest to the most complex. Debugging and fixing software bugs is among the less favorable daily activities of a programmer.

43.1 Causes of bugs

There are many reasons why software bugs occur. The following is an incomplete list roughly from the design stage of a program through the actual coding of it:

- The requirements and the specifications of the program may not be clear. What the program should actually do may not be known at the design stage.
- The programmer may misunderstand some of the requirements of the program.
- The programming language may not be expressive enough. Considering that there are confusions even between native speakers of human languages, the unnatural syntax and rules of a programming language may be cause of mistakes.
- Certain assumptions of the programmer may be incorrect. For example, the programmer may be assuming that 3.14 would be precise enough to represent π .
- The programmer may have incorrect information on a topic or none at all. For example, the programmer may not know that using a floating point variable in a particular logical expression would not be reliable.
- The program may encounter an unforeseen situation. For example, one of the files of a directory may be deleted or renamed while the program is using the files of that directory in a **foreach** loop.
- The programmer may make silly mistakes. For example, the name of a variable may be mistyped and accidentally matched the name of another variable.
- etc.

Unfortunately, there is still no software development methodology that ensures that a program will always work correctly. This is still a hot software engineering topic where promising solutions emerge every decade or so.

43.2 Discovering the bugs

Software bugs are discovered at various stages of the lifetime of the program by various types of tools and people. The following is a partial list of when a bug may be discovered, from the earliest to the latest:

- When writing the program
 - By the programmer
 - By another programmer during *pair programming*
 - By the compiler through compiler messages
 - By **unit tests** as a part of building the program
- When reviewing the code
 - By tools that analyze the code at compile time
 - By other programmers during *code reviews*
- When running the program
 - By tools that analyze the execution of the program at runtime (e.g. by *valgrind*)

- During QA testing, either by the failure of **assert** checks or by the observed behavior of the program
- By the *beta* users before the release of the program
- By the end users after the release of the program

Detecting the bug as early as possible reduces loss of money, time, and in some cases human lives. Additionally, identifying the causes of bugs that have been discovered by the end users are harder than identifying the causes of bugs that are discovered early, during development.

43.3 Unit testing for catching bugs

Since programs are written by programmers and D is a compiled language, the programmers and the compiler will always be there to discover bugs. Those two aside, the earliest and partly for that reason the most effective way of catching bugs is unit testing.

Unit testing is an indispensable part of modern programming. It is the most effective method of reducing coding errors. According to some development methodologies, code that is not guarded by unit tests is buggy code.

Unfortunately, the opposite is not true: Unit tests do not guarantee that the code is free of bugs. Although they are very effective, they can only reduce the risk of bugs.

Unit testing also enables refactoring the code (i.e. making improvements to it) with ease and confidence. Otherwise, it is common to accidentally break some of the existing functionality of a program when adding new features to it. Bugs of this type are called *regressions*. Without unit testing, regressions are sometimes discovered as late as during the QA testing of future releases, or worse, by the end users.

Risk of regressions discourage programmers from refactoring the code, sometimes preventing them from performing the simplest of improvements like correcting the name of a variable. This in turn causes *code rot*, a condition where the code becomes more and more unmaintainable. For example, although some lines of code would better be moved to a newly defined function in order to be called from more than one place, fear of regressions make programmers copy and paste the existing lines to other places instead, leading to the problem of *code duplication*.

Phrases like "if it isn't broken, don't fix it" are related to fear of regressions. Although they seem to be conveying wisdom, such guidelines cause the code to rot slowly and become an untouchable mess.

Modern programming rejects such "wisdom". To the contrary, to prevent it from becoming a source of bugs, the code is supposed to be "refactored mercilessly". The most powerful tool of this modern approach is unit testing.

Unit testing involves testing the smallest units of code independently. When units of code are tested independently, it is less likely that there are bugs in higher-level code that use those units. When the parts work correctly, it is more likely that the whole will work correctly as well.

Units tests are provided as library solutions in other languages (e.g. JUnit, CppUnit, unittest++, etc.) In D, unit testing is a core feature of the language. It is debatable whether a library solution or a language feature is better for unit testing. Because D does not provide some of the features that are commonly found in unit testing libraries, it may be worthwhile to consider library solutions as well.

The unit testing features of D are as simple as inserting **assert** checks into **unittest** blocks.

43.4 Activating the unit tests

Unit tests are not a part of the actual execution of the program. They should be activated only during program development when explicitly requested.

The **dmd** compiler switch that activates unit tests is **-unittest**.

Assuming that the program is written in a single source file named **deneme.d**, its unit tests can be activated by the following command:

```
dmd deneme.d -w -unittest
```

When a program that is built by the **-unittest** switch is started, its unit test blocks are executed first. Only if all of the unit tests pass that the program execution continues with **main()**.

43.5 unittest blocks

The lines of code that involve unit tests are written inside **unittest** blocks. These blocks do not have any significance for the program other than containing the unit tests:

```
unittest
{
    /* ... the tests and the code that support them ... */
}
```

Although **unittest** blocks can appear anywhere, it is convenient to define them right after the code that they test.

As an example, let's test a function that returns the ordinal form of the specified number as in "1st", "2nd", etc. A **unittest** block of this function can simply contain **assert** statements that compare the return values of the function to the expected values. The following function is being tested with the four distinct expected outcomes of the function:

```
string ordinal(size_t number)
{
    // ...
}

unittest
{
    assert(ordinal(1) == "1st");
    assert(ordinal(2) == "2nd");
    assert(ordinal(3) == "3rd");
    assert(ordinal(10) == "10th");
}
```

The four tests above test that the function works correctly at least for the values of 1, 2, 3, and 10 by making four separate calls to the function and comparing the returned values to the expected ones.

Although unit tests are based on **assert** checks, **unittest** blocks can contain any D code. This allows for preparations before actually starting the tests or any other supporting code that the tests may need. For example, the following block first defines a variable to reduce code duplication:

```
dstring toFront(dstring str, in dchar letter)
{
    // ...
}
```

```

}

unittest
{
    immutable str = "hello"d;

    assert(toFront(str, 'h') == "hello");
    assert(toFront(str, 'o') == "ohell");
    assert(toFront(str, 'l') == "llheo");
}

```

The three **assert** checks above test that **toFront()** works according to its specification.

As these examples show, unit tests are also useful as examples of how particular functions should be called. Usually it is easy to get an idea on what a function does just by reading its unit tests.

43.6 Test driven development

Test driven development (TDD) is a software development methodology that prescribes writing unit tests before implementing functionality. In TDD, the focus is on unit testing. Coding is a secondary activity that makes the tests pass.

In accordance to TDD, the **ordinal()** function above can first be implemented intentionally incorrectly:

```

import std.string;

string ordinal(size_t number)
{
    return "";    // ← intentionally wrong
}

unittest
{
    assert(ordinal(1) == "1st");
    assert(ordinal(2) == "2nd");
    assert(ordinal(3) == "3rd");
    assert(ordinal(10) == "10th");
}

void main()
{}

```

Although the function is obviously wrong, the next step would be to run the unit tests to see that the tests do indeed catch problems with the function:

```

$ dmd deneme.d -w -0 -unittest
$ ./deneme
core.exception.AssertError@deneme(10): unittest failure

```

The function should be implemented only *after* seeing the failure, and only to make the tests pass. Here is just one implementation that passes the tests:

```

import std.string;

string ordinal(size_t number)
{
    string suffix;

    switch (number) {
        case 1: suffix = "st"; break;

```

```

    case 2: suffix = "nd"; break;
    case 3: suffix = "rd"; break;
    default: suffix = "th"; break;
}

return format("%S%S", number, suffix);
}

unittest
{
    assert(ordinal(1) == "1st");
    assert(ordinal(2) == "2nd");
    assert(ordinal(3) == "3rd");
    assert(ordinal(10) == "10th");
}

void main()
{}

```

Since the implementation above does pass the unit tests, there is reason to trust that the `ordinal()` function is correct. Under the assurance that the tests bring, the implementation of the function can be changed in many ways with confidence.

43.6.1 Unit tests before bug fixes

Unit tests are not a panacea; there will always be bugs. If a bug is discovered when actually running the program, it can be seen as an indication of a limitation of the unit tests. For that reason, it is better to *first* write a unit test that reproduces the bug and only *then* to fix the bug to pass the new test.

Let's have a look at the following function that returns the spelling of the ordinal form a number specified as a `dstring`:

```

import std.exception;
import std.string;

dstring ordinalSpelled(dstring number)
{
    enforce(number.length, "number cannot be empty");

    dstring[dstring] exceptions = [
        "one": "first", "two": "second", "three": "third",
        "five": "fifth", "eight": "eighth", "nine": "ninth",
        "twelve": "twelfth"
    ];

    dstring result;

    if (number in exceptions) {
        result = exceptions[number];
    } else {
        result = number ~ "th";
    }

    return result;
}

unittest
{
    assert(ordinalSpelled("one") == "first");
    assert(ordinalSpelled("two") == "second");
    assert(ordinalSpelled("three") == "third");
    assert(ordinalSpelled("ten") == "tenth");
}

```



```
void main()
{}
```

The function takes care of exceptional spellings and even includes a unit test for that. Still, the function has a bug yet to be discovered:

```
import std.stdio;

void main()
{
    writeln("He came the %s in the race.",
           ordinalSpelled("twenty"));
}
```

The spelling error in the output of the program is due to a bug in `ordinalSpelled()`, which its unit tests have unfortunately fail to catch:

```
He came the twentyth in the race.
```

Although it is easy to see that the function does not produce the correct spelling for numbers that end with the letter y, TDD prescribes that first a unit test must be written to reproduce the bug before actually fixing it:

```
unittest
{
    // ...
    assert(ordinalSpelled("twenty") == "twentieth");
}
```

With that improvement to the tests, now the bug in the function is being caught during development:

```
core.exception.AssertError@deneme(3274338): unittest failure
```

The function should be fixed only then:

```
dstring ordinalSpelled(dstring number)
{
    // ...
    if (number in exceptions) {
        result = exceptions[number];
    } else {
        if (number[$-1] == 'y') {
            result = number[0..$-1] ~ "ieth";
        } else {
            result = number ~ "th";
        }
    }
    return result;
}
```

43.7 Exercise

- Implement `toFront()` according to TDD. Start with the intentionally incomplete implementation below. Observe that the unit tests fail and provide an implementation that passes the tests.

```
dstring toFront(dstring str, in dchar letter)
{
    dstring result;
    return result;
}

unittest
{
    immutable str = "hello"d;

    assert(toFront(str, 'h') == "hello");
    assert(toFront(str, 'o') == "ohell");
    assert(toFront(str, 'l') == "llheo");
}

void main()
{}
```

... the solution

44 Contract Programming

Contract programming is a software design approach that treats parts of software as individual entities that provide services to each other. This approach realizes that software can work according to its specification as long as the provider and the consumer of the service both obey a *contract*.

D's contract programming features involve functions as the units of software services. Like in unit testing, contract programming is also based on **assert** checks.

Contract programming in D are implemented by three types of code blocks:

- Function **in** blocks
- Function **out** blocks
- Struct and class **invariant** blocks

We will see **invariant** blocks and *contract inheritance* in a later chapter after covering structs and classes.

44.1 in blocks for entry conditions

Correct execution of functions usually depend on whether the values of their parameters are valid. For example, a square root function may require that its parameter cannot be negative. A function that deals with dates may require that the number of the month must be between 1 and 12.

We have already seen such condition checks in the [assert and enforce chapter](#). Conditions on parameter values can be enforced by **assert** checks within function definitions:

```
string timeToString(in int hour, in int minute)
{
    assert((hour >= 0) && (hour <= 23));
    assert((minute >= 0) && (minute <= 59));

    return format("%02s:%02s", hour, minute);
}
```

In contract programming, the same checks are written inside the **in** blocks of functions. When an **in** or **out** block is used, the actual body of the function must be specified as a **body** block:

```
import std.stdio;
import std.string;

string timeToString(in int hour, in int minute)
in
{
    assert((hour >= 0) && (hour <= 23));
    assert((minute >= 0) && (minute <= 59));
}
body
{
    return format("%02s:%02s", hour, minute);
}

void main()
{
    writeln(timeToString(12, 34));
}
```

A benefit of an **in** block is that all of the entry conditions can be kept together and separate from the actual body of the function. This way, the function body would be free of **assert** checks about the entry conditions. As needed, it is still possible and advisable to have other **assert** checks inside the function body as unrelated checks that guard against potential programming errors in the function body.

The code that is inside the **in** block is executed automatically every time the function is called. The actual execution of the function starts only if all of the **assert** checks inside the **in** block pass. This prevents executing the function with invalid entry conditions and as a consequence, avoids producing incorrect results.

An **assert** check that fails inside the **in** block indicates that the contract has been violated by the caller.

44.2 out blocks for exit conditions

The other side of the contract involves guarantees that the function provides. An example of a function with an exit condition would be a function that returns the number of days in February; the returned value would always be either 28 or 29.

The exit conditions are checked inside the **out** blocks of functions.

Because the value that a function returns by the **return** statement need not be defined as a variable inside the function, there is usually no name to refer to the return value. This can be seen as a problem because the **assert** checks inside the **out** block cannot refer to the returned variable by name.

D solves this problem by providing a way of naming the return value right after the **out** keyword. That name represents the very value that the function is in the process of returning:

```
int daysInFebruary(in int year)
out (result)
{
    assert((result == 28) || (result == 29));
}
body
{
    return isLeapYear(year) ? 29 : 28;
}
```

Although **result** is a reasonable name for the returned value, other valid names may also be used.

Some functions do not have return values or the return value need not be checked. In that case the **out** block does not specify a name:

```
out
{
    // ...
}
```

Similar to **in** blocks, the **out** blocks are executed automatically after the body of the function is executed.

An **assert** check that fails inside the **out** block indicates that the contract has been violated by the function.

As it has been obvious, **in** and **out** blocks are optional. Considering the **unittest** blocks as well, which are also optional, D functions may consist of up to four blocks of code:

- **in**: Optional
- **out**: Optional
- **body**: Mandatory but the **body** keyword may be skipped if no **in** or **out** block is defined.
- **unittest**: Optional and technically not a part of a function's definition but commonly defined right after the function.

Here is an example that uses all of these blocks:

```
import std.stdio;

/*
 * Distributes the sum between two variables.
 *
 * Distributes to the first variable first, but never gives
 * more than 7 to it. The rest of the sum is distributed to
 * the second variable.
 */
void distribute(in int sum, out int first, out int second)
in
{
    assert(sum >= 0);
}
out
{
    assert(sum == (first + second));
}
body
{
    first = (sum >= 7) ? 7 : sum;
    second = sum - first;
}

unittest
{
    int first;
    int second;

    // Both must be 0 if the sum is 0
    distribute(0, first, second);
    assert(first == 0);
    assert(second == 0);

    // If the sum is less than 7, then all of it must be given
    // to first
    distribute(3, first, second);
    assert(first == 3);
    assert(second == 0);

    // Testing a boundary condition
    distribute(7, first, second);
    assert(first == 7);
    assert(second == 0);

    // If the sum is more than 7, then the first must get 7
    // and the rest must be given to second
    distribute(8, first, second);
    assert(first == 7);
    assert(second == 1);

    // A random large value
    distribute(1_000_007, first, second);
    assert(first == 7);
    assert(second == 1_000_000);
}
```

```

void main()
{
    int first;
    int second;

    distribute(123, first, second);
    writeln("first: ", first, " second: ", second);
}

```

The program can be compiled and run on the console by the following commands:

```

$ dmd deneme.d -w -unittest
$ ./deneme
first: 7 second: 116

```

Although the actual work of the function consists of only two lines, there are a total of 19 nontrivial lines that support its functionality. It may be argued that so much extra code may be too much for such a short function. However, bugs are never intentional. The programmer always writes code that is *expected to work correctly*, which ends up containing various types of bugs.

When expectations are layed out explicitly by unit tests and contracts, functions that are initially correct have a greater chance of staying correct. I recommend that you take full advantage of any feature that improves program correctness. Both unit tests and contracts are effective tools towards that goal. They help reduce time spent for debugging, effectively increasing time spent for actually writing code.

44.3 Disabling contract programming

Contrary to unit testing, contract programming features are enabled by default. The **-release** compiler switch disables contract programming:

```
dmd deneme.d -w -release
```

When the program is compiled with the **-release** switch, the contents of **in**, **out**, and **invariant** blocks are ignored.

44.4 in blocks versus enforce checks

We have seen in the [assert and enforce chapter](#) that sometimes it is difficult to decide whether to use **assert** or **enforce** checks. Similarly, sometimes it is difficult to decide whether to use **assert** checks within the **in** blocks versus **enforce** checks within function bodies.

The fact that it is possible to disable contract programming is an indication that contract programming is for protecting against programmer errors as well. For that reason, the decision here should also be based on the same guidelines that we have seen in the **assert** and **enforce** chapter:

- If the check is guarding against a coding error, then it should be in the **in** block. For example, if the function is called only from other parts of the program, likely to help with achieving a functionality of it, then the parameter values are entirely the responsibility of the programmer. For that reason, the entry conditions of such a function should be checked in its **in** block.

- If the function cannot achieve some task for any other reason, including invalid parameter values, then it must throw an exception, conveniently by **enforce**.

To see an example of this, let's define a function that returns a slice of the middle of another slice. Let's assume that this function is for the consumption of the users of the module, as opposed to being an internal function used by the module itself. Since the users of this module can call this function by various and potentially invalid parameter values, it would be appropriate to check the parameter values every time the function is called. It would be insufficient to only check them at program development time, after which contracts can be disabled by **-release**.

For that reason, the following function validates its parameters by calling **enforce** in the function body instead of an **assert** check in the **in** block:

```
import std.exception;

inout(int)[] middle(inout(int)[] originalSlice, size_t width)
out (result)
{
    assert(result.length == width);
}
body
{
    enforce(originalSlice.length >= width);

    immutable start = (originalSlice.length - width) / 2;
    immutable end = start + width;

    return originalSlice[start .. end];
}

unittest
{
    auto slice = [1, 2, 3, 4, 5];

    assert(middle(slice, 3) == [2, 3, 4]);
    assert(middle(slice, 2) == [2, 3]);
    assert(middle(slice, 5) == slice);
}

void main()
{}
```

There isn't a similar problem with the **out** blocks. Since the return value of every function is the responsibility of the programmer, exit conditions must always be checked in the **out** blocks. The function above follows this guideline.

- Another criterion to consider when deciding between **in** blocks versus **enforce** is to consider whether the condition is recoverable. If it is recoverable by the higher layers of code, then it may be more appropriate to throw an exception, conveniently by **enforce**.

44.5 Exercise

- Write a program that increases the total points of two football (soccer) teams according to the result of a game.

The first two parameters of this function are the goals that each team has scored. The other two parameters are the points of each team before the game. This function should adjust the points of the teams according to the goals that they have scored. As a reminder, the winner takes 3 points and the loser takes no point. In the event of a draw, both teams get 1 point each.

Additionally, the function should indicate which team has been the winner: 1 if the first team has won, 2 if the second team has won, and 0 if the game has ended in a draw.

Start with the following program and fill in the four blocks of the function appropriately. Do not remove the `assert` checks in `main()`; they demonstrate how this function is expected to work.

```
int addPoints(in int goals1,
             in int goals2,
             ref int points1,
             ref int points2)
in
{
  // ...
}
out (result)
{
  // ...
}
body
{
  int winner;

  // ...

  return winner;
}

unittest
{
  // ...
}

void main()
{
  int points1 = 10;
  int points2 = 7;
  int winner;

  winner = addPoints(3, 1, points1, points2);
  assert(points1 == 13);
  assert(points2 == 7);
  assert(winner == 1);

  winner = addPoints(2, 2, points1, points2);
  assert(points1 == 14);
  assert(points2 == 8);
  assert(winner == 0);
}
```

Note: It may be better to return an `enum` value from this function:

```
enum GameResult
{
  firstWon, secondWon, draw
}

GameResult addPoints(in int goals1,
                    in int goals2,
                    ref int points1,
                    ref int points2)
// ...
```

I chose to return an `int` for this exercise, so that the return value can be checked against the valid values of 0, 1, and 2.

... the solution

45 Lifetimes and Fundamental Operations

We will soon cover structs, the basic feature that allows the programmer to define application-specific types. Structs are for combining fundamental types and other structs together to define higher-level types that behave according to special needs of programs. After structs, we will learn about classes, which are the basis of the object oriented programming features of D.

Before getting to structs and classes, it will be better to talk about some important concepts first. These concepts will help understand structs and classes and some of their differences.

We have been calling any piece of data that represented a concept in a program a *variable*. In a few places we have called struct and class variables specifically as *objects*. I will continue calling both of these concepts variables in this chapter.

Although this chapter includes only fundamental types, slices, and associative arrays; these concepts apply to user-defined types as well.

45.1 Lifetime of a variable

The time between when a variable is defined and when it is *finalized* is the lifetime of that variable. Although it is the case for many types, *becoming unavailable* and *being finalized* need not be at the same time.

You would remember from the [Name Space chapter](#) how variables become unavailable. In simple cases, exiting the scope where a variable has been defined would make that variable unavailable.

Let's consider the following example as a reminder:

```
void speedTest()
{
    int speed;           // Single variable ...

    foreach (i; 0 .. 10) {
        speed = 100 + i; // ... takes 10 different values.
        // ...
    }
} // ← 'speed' is unavailable beyond this point.
```

The lifetime of the **speed** variable in that code ends upon exiting the **speedTest()** function. There is a single variable in the code above, which takes ten different values from 100 to 109.

When it comes to variable lifetimes, the following code is very different compared to the previous one:

```
void speedTest()
{
    foreach (i; 0 .. 10) {
        int speed = 100 + i; // Ten separate variables.
        // ...
    } // ← Lifetime of each variable ends here.
}
```

There are ten separate variables in that code, each taking a single value. Upon every iteration of the loop, a new variable starts its life, which eventually ends at the end of each iteration.

45.2 Lifetime of a parameter

The lifetime of a parameter depends on its qualifiers:

ref: The parameter is just an alias of the actual variable that is specified when calling the function. **ref** parameters do not affect the lifetimes of actual variables.

in: For *value types*, the lifetime of the parameter starts upon entering the function and ends upon exiting it. For *reference types*, the lifetime of the parameter is the same as with **ref**. (We will see value types and reference types in the next chapter.)

out: Same with **ref**, the parameter is just an alias of the actual variable that is specified when calling the function. The only difference is that the variable is set to its **.init** value automatically upon entering the function.

lazy: The life of the parameter starts when the parameter is actually used and ends right then.

The following example uses these four types of parameters and explains their lifetimes in program comments:

```
void main()
{
    int main_in;      /* The value of main_in is copied to the
                     * parameter. */

    int main_ref;    /* main_ref is passed to the function as
                     * itself. */

    int main_out;    /* main_out is passed to the function as
                     * itself. Its value is set to int.init
                     * upon entering the function. */

    foo(main_in, main_ref, main_out, aCalculation());
}

void foo(
    in int p_in,      /* The lifetime of p_in starts upon
                     * entering the function and ends upon
                     * exiting the function. */

    ref int p_ref,    /* p_ref is an alias of main_ref. */

    out int p_out,    /* p_out is an alias of main_out. Its
                     * value is set to int.init upon
                     * entering the function. */

    lazy int p_lazy) /* The lifetime of p_lazy starts when it
                     * is used and ends when its use
                     * ends. Its value is calculated by
                     * calling aCalculation() every time
                     * p_lazy is used in the function. */
{
    // ...
}

int aCalculation()
{
    int result;
    // ...
    return result;
}
```

45.3 Fundamental operations

Regardless of its type, there are three fundamental operations throughout the lifetime of a variable:

- **Initialization:** The start of its life.
- **Finalization:** The end of its life.
- **Assignment:** Changing its value at once.

To be considered an object, it must first be initialized. There may be final operations for some types. The value of a variable may change during its lifetime.

45.3.1 Initialization

Every variable must be initialized before being used. Initialization involves two steps:

1. **Reserving space for the variable:** This space is where the value of the variable is stored in memory.
2. **Construction:** Setting the first value of the variable on that space (or the first values of the members of structs and classes).

Every variable lives in a place in memory that is reserved for it. Some of the code that the compiler generates is about reserving space for each variable.

Let's consider the following variable:

```
int speed = 123;
```

That variable takes up space the size of an `int` in the program's memory. If we visualize the memory as a ribbon going from left to right, we can imagine the variable living on some part of it:

```
--+-----+-----+-----+--
|         | 123 |         |
--+-----+-----+-----+--
```

The memory location that a variable is placed at is called its address. In a sense, the variable lives at that address. When the value of a variable is changed, the new value is stored at the same place:

```
++speed;
```

The new value would be at the same place where the old value has been:

```
--+-----+-----+-----+--
|         | 124 |         |
--+-----+-----+-----+--
```

Construction is necessary to prepare variables for use. Since a variable cannot be used reliably before being constructed, it is performed by the compiler automatically.

Variables can be constructed in three ways:

- **By their default value:** when the programmer does not specify a value explicitly
- **By copying:** when the variable is constructed as a copy of another variable of the same type
- **By a specific value:** when the programmer specifies a value explicitly

When a value is not specified, the value of the variable would be the *default* value of its type, i.e. its `.init` value.

```
int speed;
```

The value of `speed` above is `int.init`, which happens to be zero. Naturally, a variable that is constructed by its default value may have other values during its lifetime (unless it is `immutable`).

```
File file;
```

With the definition above, the variable `file` is a `File` object that is not yet associated with an actual file on the file system. It is not usable until it is modified to be associated with a file.

Variables are sometimes constructed as a copy of another variable:

```
int speed = otherSpeed;
```

`speed` above is constructed by the value of `otherSpeed`.

As we will see in later chapters, this operation has a different meaning for class variables:

```
auto classVariable = otherClassVariable;
```

Although `classVariable` starts its life as a copy of `otherClassVariable`, there is a fundamental difference with classes: Although `speed` and `otherSpeed` are distinct values, `classVariable` and `otherClassVariable` both provide access to the same value. This is the fundamental difference between value types and reference types. We will cover this topic in the next chapter.

Finally, variables can be constructed by the value of an expression of a compatible type:

```
int speed = someCalculation();
```

`speed` above would be constructed by the return value of `someCalculation()`.

45.3.2 Finalization

Finalizing is the final operations that are executed for a variable and reclaiming its memory:

1. **Destruction:** The final operations that must be executed for the variable.
2. **Reclaiming the variable's memory:** Reclaiming the piece of memory that the variable has been living on.

For simple fundamental types, there are no final operations to execute. For example, the value of a variable of type `int` is not set back to zero. For such variables there is only reclaiming their memory, so that it will be used for other variables later.

On the other hand, some types of variables require special operations during finalization. For example, a `File` object would need to write the characters that are still in its output buffer to disk and notify the file system that it no longer uses the file. These operations are the destruction of a `File` object.

Final operations of arrays are at a little higher-level: Before finalizing the array, first its elements are destructed. If the elements are of a simple fundamental type like `int`, then there are no special final operations for them. If the elements are of a struct or a class type that needs finalization, then those operations are executed for each element.

Associative arrays are similar to arrays. Additionally, the keys may also be finalized if they are of a type that needs destruction.

The garbage collector: D is a *garbage-collected* language. In such languages finalizing an object need not be initiated explicitly by the programmer. When a variable's lifetime ends, its finalization is automatically handled by the garbage collector. We will cover the garbage collector and special memory management in a later chapter.

Variables can be finalized in two ways:

- **When the lifetime ends:** The finalization happens upon the end of life of the variable.
- **Some time in the future:** The finalization happens at an indeterminate time in the future by the garbage collector.

Which of the two ways a variable will be finalized depends primarily on its type. Some types like arrays, associative arrays and classes are normally destroyed by the garbage collector some time in the future.

45.3.3 Assignment

The other fundamental operation that a variable experiences during its lifetime is assignment.

For simple fundamental types assignment is merely changing the value of the variable. As we have seen above on the memory representation, an **int** variable would start having the value 124 instead of 123. However, more generally, assignment consists of two steps, which are not necessarily executed in the following order:

- **Destructing the old value**
- **Constructing the new value**

These two steps are not important for simple fundamental types that don't need destruction. For types that need destruction, it is important to remember that assignment is a combination of the two steps above.

46 Value Types and Reference Types

This chapter introduces the concepts of value types and reference types. These concepts are particularly important to understand the differences between structs and classes.

This chapter also gets into more detail with the **&** operator.

The chapter ends with a table that contains the outcomes of the following two concepts for different types of variables:

- Value comparison
- Address comparison

46.1 Value types

Value types are easy to describe: Variables of value types carry values. For example, all of the integer and floating point types are value types. Although not immediately obvious, fixed-length arrays are value types as well.

For example, a variable of type **int** has an integer value:

```
int speed = 123;
```

Carrying on from the previous chapter and representing the memory as a ribbon going from left to right, such a variable would be living on some part of the memory:

```

speed
---+-----+---
  | 123 |
---+-----+---
```

When variables of value types are copied, they get their own values:

```
int newSpeed = speed;
```

The new variable would have a place and a value of its own:

```

speed          newSpeed
---+-----+---  ---+-----+---
  | 123 |         | 123 |
---+-----+---  ---+-----+---
```

Naturally, modifications that are made to these variables are independent:

```
speed = 200;
```

The value of the other variable does not change:

```

speed          newSpeed
---+-----+---  ---+-----+---
  | 200 |         | 123 |
---+-----+---  ---+-----+---
```

46.1.1 The use of assert checks below

The following examples contain **assert** checks to indicate that their conditions are true. In other words, they are not checks in the normal sense, rather my way of telling to the reader that "this is true".

For example, the check **assert(speed == newSpeed)** below means "speed is equal to newSpeed".

46.1.2 Value identity

As the memory representations above indicate, there are two types of equality that concern variables:

- **Value equality:** The **==** operator that appear in many examples throughout the book compares variables by their values. When two variables are said to be *equal* in that sense, their values are equal.
- **Value identity:** In the sense of owning separate values, **speed** and **newSpeed** have separate identities. Even when their values are equal, they are different variables.

```
int speed = 123;
int newSpeed = speed;
assert(speed == newSpeed);
speed = 200;
assert(speed != newSpeed);
```

46.1.3 Address-of operator, &

We have been using the **&** operator so far with **readf()**. The **&** operator tells **readf()** where to put the input data.

The addresses of variables can be used for other purposes as well. The following code simply prints the addresses of two variables:

```
int speed = 123;
int newSpeed = speed;

writeln("speed : ", speed, " address: ", &speed);
writeln("newSpeed: ", newSpeed, " address: ", &newSpeed);
```

speed and **newSpeed** have the same value but their addresses are different:

```
speed : 123 address: 7FFF4B39C738
newSpeed: 123 address: 7FFF4B39C73C
```

Note: It is normal for the addresses to have different values every time the program is run. Variables live at parts of memory that happen to be available during that particular execution of the program.

Addresses are normally printed in hexadecimal format.

Additionally, the fact that the two addresses are 4 apart indicates that those two integers are placed next to each other in memory. (Note that the value of hexadecimal C is 12, so the difference between 8 and 12 is 4.)

46.2 Reference variables

Before getting to reference types let's first define reference variables.

Terminology: We have been using the phrase *to provide access to* so far in several contexts throughout the book. For example, slices and associative arrays do not own any elements but provide access to elements that are owned by the D runtime. Another phrase that is identical in meaning is *being a reference of* as in "slices are references of zero or more elements", which is sometimes used even shorter as *to reference* as in "this slice references two elements". Finally, the act of accessing a value through a reference is *dereferencing*.

Reference variables are variables that act like aliases of other variables. Although they look and are used like variables, they do not have values themselves. Modifications made on a reference variable change the value of the actual variable.

We have already used reference variables so far in two contexts:

- **ref in foreach loops:** The **ref** keyword makes the loop variable the *actual* element that corresponds to that iteration. When the **ref** keyword is not used, the loop variable is a *copy* of the actual element.

This can be demonstrated by the **&** operator as well. If their addresses are the same, two variables would be referencing the same value (or the *same element* in this case):

```
int[] slice = [ 0, 1, 2, 3, 4 ];
foreach (i, ref element; slice) {
    assert(&element == &slice[i]);
}
```

Although they are separate variables, the fact that the addresses of **element** and **slice[i]** are the same proves that they have the same value identity.

In other words, **element** and **slice[i]** are references of the same value. Changing either of those affects the actual value. The following memory layout indicates a snapshot of the iteration when **i** is 3:

slice[0]	slice[1]	slice[2]	slice[3]	slice[4]
→	→	→	(element)	
0	1	2	3	4

- **ref and out function parameters:** Function parameters that are specified as **ref** or **out** are aliases of the actual variable the function is called with.

The following example demonstrates this case by passing the same variable to separate **ref** and **out** parameters of a function. Again, the **&** operator indicates that both parameters have the same value identity:

```
import std.stdio;

void main()
{
    int originalVariable;
    writeln("address of originalVariable: ", &originalVariable);
    foo(originalVariable, originalVariable);
}

void foo(ref int refParameter, out int outParameter)
{
    writeln("address of refParameter : ", &refParameter);
    writeln("address of outParameter : ", &outParameter);
    assert(&refParameter == &outParameter);
}
```

Although they are defined as separate parameters, `refParameter` and `outParameter` are aliases of `originalVariable`:

```
address of originalVariable: 7FFF24172958
address of refParameter    : 7FFF24172958
address of outParameter    : 7FFF24172958
```

46.3 Reference types

Variables of reference types have individual identities but they do not have individual values. They *provide access to* existing variables.

We have already seen this concept with slices. Slices do not own elements, they provide access to existing elements:

```
void main()
{
    // Although it is named as 'array' here, this variable is
    // a slice as well. It provides access to all of the
    // initial elements:
    int[] array = [ 0, 1, 2, 3, 4 ];

    // A slice that provides access to elements other than the
    // first and the last:
    int[] slice = array[1 .. $ - 1];

    // At this point slice[0] and array[1] provide access to
    // the same value:
    assert(&slice[0] == &array[1]);

    // Changing slice[0] changes array[1]:
    slice[0] = 42;
    assert(array[1] == 42);
}
```

Contrary to reference variables, reference types are not simply aliases. To see this distinction, let's define another slice as a copy of one of the existing slices:

```
int[] slice2 = slice;
```

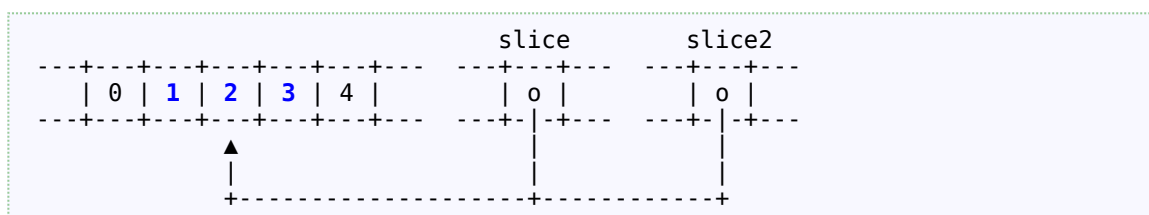
The two slices have their own addresses. In other words, they have separate identities:

```
assert(&slice != &slice2);
```

The following list is a summary of the differences between reference variables and reference types:

- Reference variables do not have identities, they are aliases of existing variables.
- Variables of reference types have identities but they do not own values; rather, they provide access to existing values.

The way `slice` and `slice2` live in memory can be illustrated as in the following figure:



The three elements that they both reference are highlighted in blue.

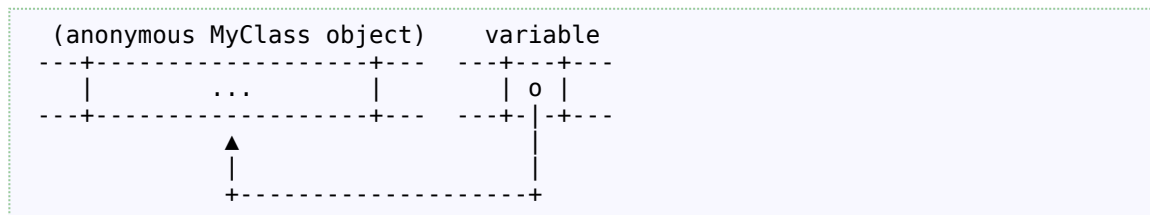
One of the differences between C++ and D is that classes are reference types in D. Although we will cover classes in later chapters in detail, the following is a short example to demonstrate this fact:

```
class MyClass
{
    int member;
}
```

Class objects are constructed by the **new** keyword:

```
auto variable = new MyClass;
```

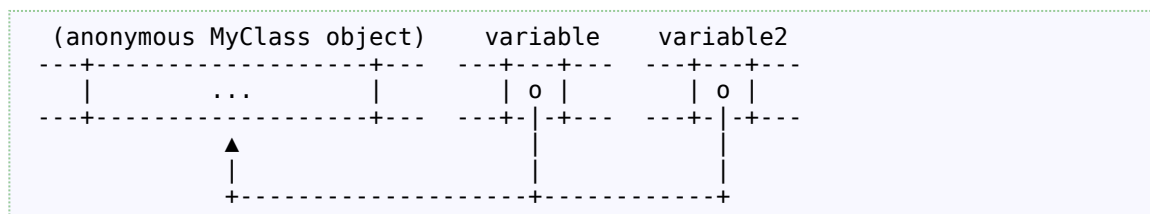
variable is a reference to an anonymous **MyClass** object that has been constructed by **new**:



Just like with slices, when **variable** is copied, the copy becomes another reference to the same object. The copy has its own address:

```
auto variable = new MyClass;
auto variable2 = variable;
assert(variable == variable2);
assert(&variable != &variable2);
```

They are equal from the point of view of referencing the same object, but they are separate variables:



This can also be shown by modifying the member of the object:

```
auto variable = new MyClass;
variable.member = 1;

auto variable2 = variable; // They share the same object
variable2.member = 2;

assert(variable.member == 2); // The object that variable
                             // provides access to has
                             // changed.
```

Another reference type is associative arrays. When assigned, they too provide access to the same set of elements:

```
string[int] byName =
[
```

```

    1   : "one",
    10  : "ten",
    100 : "hundred",
];

// The two associative arrays will be sharing the same
// set of elements
string[int] byName2 = byName;

// The mapping added through the second ...
byName2[4] = "four";

// ... is visible through the first.
assert(byName[4] == "four");

```

46.3.1 The difference in the assignment operation

With value types and reference variables, the assignment operation changes *the actual value*:

```

void main()
{
    int number = 8;

    halve(number); // The actual value changes
    assert(number == 4);
}

void halve(ref int dividend)
{
    dividend /= 2;
}

```

On the other hand, with reference types, the assignment operation changes *which value is being accessed*. For example, the assignment of the `slice3` variable below does not change the value of any element; rather, it changes what elements `slice3` is now a reference of:

```

int[] slice1 = [ 10, 11, 12, 13, 14 ];
int[] slice2 = [ 20, 21, 22 ];

int[] slice3 = slice1[1 .. 3]; // Access to the elements
                               // of slice1 with indexes 1
                               // and 2

slice3[0] = 777;
assert(slice1 == [ 10, 777, 12, 13, 14 ]);

// This assignment does not modify the elements that
// slice3 is providing access to, it makes slice3 provide
// access to other elements.
slice3 = slice2[$ - 1 .. $]; // Access to the last element

slice3[0] = 888;
assert(slice2 == [ 20, 21, 888 ]);

```

Let's demonstrate the same effect this time with two objects of the `MyClass` type:

```

auto variable1 = new MyClass;
variable1.member = 1;

auto variable2 = new MyClass;
variable2.member = 2;

auto aCopy = variable1;
aCopy.member = 3;

aCopy = variable2;

```

```
aCopy.member = 4;
assert(variable1.member == 3);
assert(variable2.member == 4);
```

The **aCopy** variable above first references the same object as **variable1**, and then the same object as **variable2**. As a consequence, the **.member** that is modified through **aCopy** is first **variable1**'s and then **variable2**'s.

46.3.2 Variables of reference types may not be referencing any object

With a reference variable, there is always an actual variable that it is an alias of; it can not start its life without a variable. On the other hand, variables of reference types can start their lives without referencing any object.

For example, a **MyClass** variable can be defined without an actual object having been created by **new**:

```
MyClass variable;
```

Such variables have the special value of **null**. We will cover **null** and the **is** keyword in the next chapter.

46.4 Fixed-length arrays are value types, slices are reference types

D's arrays and slices diverge when it comes to value type versus reference type.

As we have already seen above, slices are reference types. On the other hand, fixed-length arrays are value types. They own their elements and behave as individual values:

```
int[3] array1 = [ 10, 20, 30 ];
auto array2 = array1; // array2's elements are different
                    // from array1's
array2[0] = 11;

// First array is not affected:
assert(array1[0] == 10);
```

array1 is a fixed-length array because its length is specified when it has been defined. Since **auto** makes the compiler infer the type of **array2**, it is a fixed-length array as well. The values of **array2**'s elements are copied from the values of the elements of **array1**. Each array has its own elements. Modifying an element through one does not affect the other.

46.5 Experiment

The following program is an experiment of applying the **==** operator to different types. It applies the operator to both variables of a certain type and to the addresses of those variables. The program produces the following output:

Type of variable	a == b	&a == &b
variables with equal values (value type)	true	false
variables with different values (value type)	false	false
foreach with 'ref' variable	true	true
foreach without 'ref' variable	true	false
function with 'out' parameter	true	true
function with 'ref' parameter	true	true
function with 'in' parameter	true	false
slices providing access to same elements	true	false
slices providing access to different elements	false	false
MyClass variables to same object (reference type)	true	false
MyClass variables to different objects (reference type)	false	false

The table above has been generated by the following program:

```
import std.stdio;
import std.conv;
import std.array;

int moduleVariable = 9;

class MyClass
{
    int member;
}

void printHeader()
{
    immutable dchar[] header =
        "                Type of variable"
        "                a == b  &a == &b";

    writeln();
    writeln(header);
    writeln(replicate("=", header.length));
}

void printInfo(const dchar[] label,
              bool valueEquality,
              bool addressEquality)
{
    writefln("%55s%9s%9s",
            label,
            to!string(valueEquality),
            to!string(addressEquality));
}

void main()
{
    printHeader();

    int number1 = 12;
    int number2 = 12;
    printInfo("variables with equal values (value type)",
            number1 == number2,
            &number1 == &number2);

    int number3 = 3;
    printInfo("variables with different values (value type)",
            number1 == number3,
            &number1 == &number3);

    int[] slice = [ 4 ];
    foreach (i, ref element; slice) {
        printInfo("foreach with 'ref' variable",
                element == slice[i],
                &element == &slice[i]);
    }

    foreach (i, element; slice) {
        printInfo("foreach without 'ref' variable",
                element == slice[i],
```

```

        &element == &slice[i]);
    }

    outParameter(moduleVariable);
    refParameter(moduleVariable);
    inParameter(moduleVariable);

    int[] longSlice = [ 5, 6, 7 ];
    int[] slice1 = longSlice;
    int[] slice2 = slice1;
    printInfo("slices providing access to same elements",
        slice1 == slice2,
        &slice1 == &slice2);

    int[] slice3 = slice1[0 .. $ - 1];
    printInfo("slices providing access to different elements",
        slice1 == slice3,
        &slice1 == &slice3);

    auto variable1 = new MyClass;
    auto variable2 = variable1;
    printInfo(
        "MyClass variables to same object (reference type)",
        variable1 == variable1,
        &variable1 == &variable2);

    auto variable3 = new MyClass;
    printInfo(
        "MyClass variables to different objects (reference type)",
        variable1 == variable3,
        &variable1 == &variable3);
}

void outParameter(out int parameter)
{
    printInfo("function with 'out' parameter",
        parameter == moduleVariable,
        &parameter == &moduleVariable);
}

void refParameter(ref int parameter)
{
    printInfo("function with 'ref' parameter",
        parameter == moduleVariable,
        &parameter == &moduleVariable);
}

void inParameter(in int parameter)
{
    printInfo("function with 'in' parameter",
        parameter == moduleVariable,
        &parameter == &moduleVariable);
}

```

Notes:

- The program makes use of a module variable when comparing different types of function parameters. Module variables are defined at module level, outside of all of the functions. They are globally accessible to all of the code in the module.
- The **replicate()** function of the **std.array** module takes a range (the "=" above) and repeats it specified number of times.

46.6 Summary

- Variables of value types have their own values and addresses.

- Reference variables do not have their own values nor addresses. They are aliases of existing variables.
- Variables of reference types have their own addresses but the values that they reference do not belong to them.
- With reference types, assignment does not change value, it changes which value is being accessed.
- Variables of reference types may be **null**.

47 The null Value and the is Operator

In the previous chapter we have seen that a variable of a reference type need not be referencing an object:

```
MyClass referencesAnObject = new MyClass;
MyClass variable; // does not reference an object
```

Being a reference type, **variable** above does have an identity but it does not reference any object yet. Such an object can be pictured as having a place in the memory as in the following picture:

```
variable
---+-----+---
  | null |
---+-----+---
```

A reference that does not reference any value is **null**. We will expand on this below.

Such a variable is in an almost useless state. Since there is no **MyClass** object that it references, it cannot be used in a context where an actual **MyClass** object is needed:

```
import std.stdio;

class MyClass
{
    int member;
}

void use(MyClass variable)
{
    writeln(variable.member); // ← BUG
}

void main()
{
    MyClass variable;
    use(variable);
}
```

As there is no object that is referenced by the parameter that **use()** receives, attempting to access a member of a non-existing object results in a program crash:

```
$ ./deneme
Segmentation fault
```

"Segmentation fault" is an indication that the program has been terminated by the operating system because of attempting to access an illegal memory address.

47.1 The null value

The special value **null** can be printed just like any other value.

Note: Printing **null** directly does not compile due to a limitation of the compiler that has been used when writing this chapter (dmd 2.059). That is why its value is first being casted to **int** below. (The **cast** operator will be explained in the next chapter.)

```
writeln(cast(int)null);
```

Like in many other languages, the integer representation of **null** is 0:

```
0
```

A **null** variable can be used only in two contexts:

1. Assigning an object to it

```
variable = new MyClass; // now references an object
```

The assignment above makes **variable** provide access to the newly constructed object. From that point on, **variable** can be used for any valid operation of the **MyClass** type.

2. Determining whether it is **null**

However, because the **==** operator needs actual objects to compare, the expression below cannot be compiled:

```
if (variable == null) // ← compilation ERROR
```

For that reason, whether a variable is **null** must be determined by the **is** operator.

47.2 The is operator

This operator answers the question "does have the null value?":

```
if (variable is null) {
    // Does not reference any object
}
```

is can be used with other types of variables as well. In the following use, it compares the **values** of two integers:

```
if (speed is newSpeed) {
    // Their values are equal
} else {
    // Their values are different
}
```

When used with slices, it determines whether the two slices reference the same set of elements:

```
if (slice is slice2) {
    // They provide access to the same elements
}
```

47.3 The !is operator

!is is the opposite of **is**. It produces **true** when the values are different:

```
if (speed !is newSpeed) {
    // Their values are different
}
```

47.4 Assigning the null value

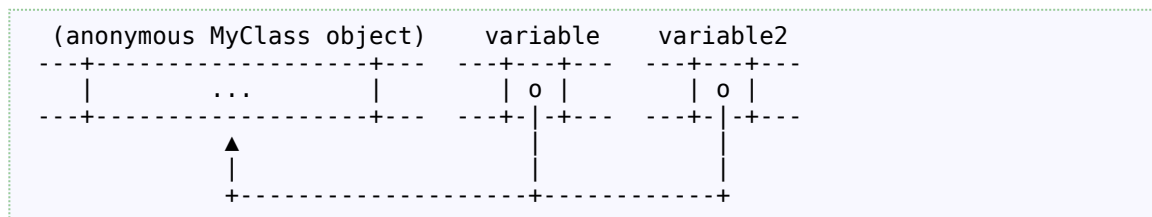
Assigning the `null` value to a variable of a reference type makes that variable stop referencing its current object.

If that assignment happens to be terminating the very last reference to the actual object, then the actual object becomes a candidate for finalization by the garbage collector. After all, not being referenced by any variable means that the object is not being used in the program at all.

Let's look at the example from the previous chapter where two variables have been referencing the same object:

```
auto variable = new MyClass;
auto variable2 = variable;
```

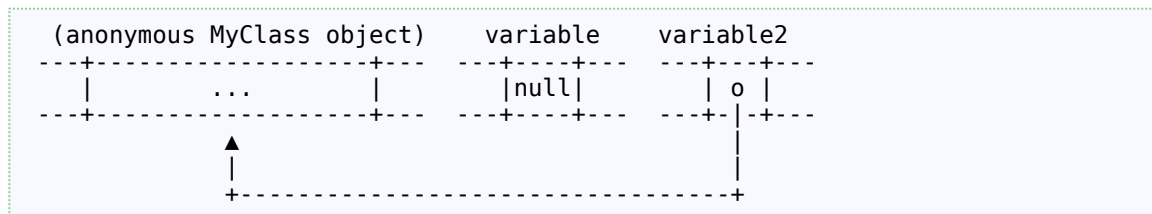
The following is a representation of the state of the memory after executing that code:



Assigning the `null` value to one of these variables breaks its relationship with the object:

```
variable = null;
```

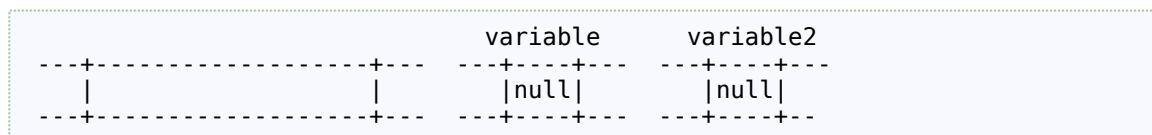
At this point there is only `variable2` that references the `MyClass` object:



Assigning `null` to the last reference would make the `MyClass` object unreachable:

```
variable2 = null;
```

Such unreachable objects are finalized by the garbage collector sometime in the future. From the point of view of the program, the object does not exist:



We had discussed ways of emptying an associative array in the exercises section of the [Associative Arrays chapter](#). We now know a fourth method: Assigning `null` to an associative array variable will break the relationship of that variable with the elements:

```
string[int] names;
// ...
names = null;    // Not providing access to any element
```

Similar to the **MyClass** examples, if **names** has been the last reference to the elements of the associative array, those elements would be finalized by the garbage collector.

Slices can be assigned **null** as well:

```
int[] slice = otherSlice[ 10 .. 20 ];  
// ...  
slice = null;    // Not providing access to any element
```

47.5 Summary

- **null** is the value indicating that a variable does not provide access to any value
- References that have the **null** value can only be used in two operations: assigning a value to them and determining whether they are **null** or not
- Since the **==** operator may have to access an actual object, determining whether a variable is **null** must be performed by the **is** operator
- **!is** is the opposite of **is**
- Assigning **null** to a variable makes that variable provide access to nothing
- Objects that are not referenced by any variable are finalized by the garbage collector

48 Type Conversions

Variables must be compatible with the expressions that they take parts in. As it has probably been obvious by the programs that we have seen so far, D is a *statically typed language*, meaning that the compatibility of types are validated at compile time.

All of the expressions that we have written so far always had compatible types because otherwise the code would be rejected by the compiler. The following is an example of a code that has incompatible types:

```
char[] slice;
writeln(slice + 5); // ← compilation ERROR
```

The compiler rejects the code due to the incompatible types `char[]` and `int` for the addition operation:

```
Error: incompatible types for ((slice) + (5)): 'char[]' and 'int'
```

Type incompatibility does not mean that the types are different; different types can indeed be used in expressions safely. For example, an `int` variable can safely be used in place of a `double` value:

```
double sum = 1.25;
int increment = 3;
sum += increment;
```

Even though `sum` and `increment` are of different types, the code above is valid because incrementing a `double` variable by an `int` value is legal.

48.1 Automatic type conversions

Automatic type conversions are also called *implicit type conversions*.

Although `double` and `int` are compatible types in the expression above, the addition operation must still be evaluated as a specific type at the microprocessor level. As you would remember from the [Floating Point Types chapter](#), the 64-bit type `double` is *wider* (or *larger*) than the 32-bit type `int`. Additionally, any value that can fit an `int` can fit a `double`.

When the compiler encounters an expression that involves mismatched types, it first converts the parts of the expressions to a common type and then evaluates the overall expression. The automatic conversions that are performed by the compiler are in the direction that avoids data loss. For example, `double` can hold any value that `int` can hold but the opposite is not true. The `+=` operation above can work because any `int` value can safely be converted to `double`.

The value that has been generated automatically as a result of a conversion is always an anonymous (and often temporary) variable. The original value does not change. For example, the automatic conversion during `+=` above does not change the type of `increment`; it is always an `int`. Rather, a temporary value of type `double` is constructed by the value of `increment`. The conversion that takes place in the background is equivalent to the following code:

```
{
    double an_anonymous_double_value = increment;
    sum += an_anonymous_double_value;
}
```

The compiler converts the **int** value to a temporary **double** value and uses that value in the operation. In this example, the temporary variable lives only during the **+=** operation.

Automatic conversions are not limited to arithmetic operations. There are other cases where types are converted to other types automatically. As long as the conversions are valid, the compiler takes advantage of type conversions to be able to use values in expressions. For example, a **byte** value can be passed for an **int** parameter:

```
void func(int number)
{
    // ...
}

void main()
{
    byte smallValue = 7;
    func(smallValue);    // automatic type conversion
}
```

In the code above, first a temporary **int** value is constructed and the function is called with that value.

48.1.1 Integer promotions

Values of types that are on the left-hand side of the following table never take part in arithmetic expressions as their actual types. Each type is first promoted to the type that is on the right-hand side of the table.

From	To
bool	int
byte	int
ubyte	int
short	int
ushort	int
char	int
wchar	int
dchar	uint

Integer promotions are applied to **enum** values as well.

The reasons for integer promotions are both historical (where the rules come from C) and the fact that the natural arithmetic type for the microprocessor is **int**. For example, although the following two variables are both **ubyte**, the addition operation is performed only after both of the values are individually promoted to **int**:

```
ubyte a = 1;
ubyte b = 2;
writeln(typeof(a + b).stringof);    // the addition is not in ubyte
```

The output:

```
int
```

Note that the types of the variables **a** and **b** do not change; only their values are temporarily promoted to **int** for the duration of the addition operation.

48.1.2 Arithmetic conversions

There are other conversion rules that are applied for arithmetic operations. In general, automatic arithmetic conversions are applied in the safe direction: from the *narrow* type to the *wider* type. Although this rule is easy to remember and is correct in most cases, automatic conversion rules are very complicated and in the case of signed-to-unsigned conversions, carry some risk of bugs.

The arithmetic conversion rules are the following:

1. If one of the values is **real**, then the other value is converted to **real**
2. Else, if one of the values is **double**, then the other value is converted to **double**
3. Else, if one of the values is **float**, then the other value is converted to **float**
4. Else, first *integer promotions* are applied according to the table above, and then the following rules are followed:
 1. If both types are the same, then no more steps needed
 2. If both types are signed or both types are unsigned, then the narrower value is converted to the wider type
 3. If the signed type is wider than the unsigned type, then the unsigned value is converted to the signed type
 4. Otherwise the signed type is converted to the unsigned type

Unfortunately, the last rule above can cause subtle bugs:

```
int    a = 0;
int    b = 1;
size_t c = 0;
writeln(a - b + c); // Surprising result!
```

Surprisingly, the output is not -1, but **size_t.max**:

```
18446744073709551615
```

Although one would expect $(0 - 1 + 0)$ to be calculated as -1, according to the rules above, the type of the entire expression is **size_t**, not **int**; and since **size_t** cannot hold negative values, the result underflows and becomes **size_t.max**.

48.1.3 const conversions

As we have seen earlier in the [Function Parameters chapter](#), reference types can automatically be converted to the **const** of the same type. Conversion to **const** is safe because the width of the type does not change and **const** is a promise to not modify the variable:

```
dchar[] parenthesized(const dchar[] text)
{
    return "{" ~ text ~ "}";
}

void main()
{
    dchar[] greeting;
    greeting ~= "hello world";
    parenthesized(greeting);
}
```

The mutable **greeting** above is automatically converted to a **const dchar[]** as it is passed to **parenthesized()**.

As we have also seen earlier, the opposite conversion is not automatic. A **const** reference is not automatically converted to a mutable reference:

```
dchar[] parenthesized(const dchar[] text)
{
    dchar[] argument = text; // ← compilation ERROR
    // ...
}
```

Note that this topic is only about references; since variables of value types are copied, it is not possible to affect the original through the copy anyway:

```
const int totalCorners = 4;
int theCopy = totalCorners; // compiles (value type)
```

The conversion from **const** to mutable above is legal because the copy is not a reference to the original.

48.1.4 immutable conversions

Because **immutable** specifies that a variable can never change, neither conversion from **immutable** nor to **immutable** are automatic:

```
string a = "hello"; // immutable characters
char[] b = a; // ← compilation ERROR
string c = b; // ← compilation ERROR
```

As with **const** conversions above, this topic is also only about reference types. Since variables of value types are copied anyway, conversions to and from **immutable** are valid:

```
immutable a = 10;
int b = a; // compiles (value type)
```

48.1.5 enum conversions

As we have seen in the [enum chapter](#), **enum** is for defining *named constants*:

```
enum Suit { spades, hearts, diamonds, clubs }
```

Remember that since no values are specified explicitly above, the values of the **enum** members start with zero and are automatically incremented by one. Accordingly, the value of **Suit.clubs** is 3.

enum values are automatically converted to integral types. For example, the value of **Suit.hearts** is taken to be 1 in the following calculation and the result becomes 11:

```
int result = 10 + Suit.hearts;
assert(result == 11);
```

The opposite conversion is not automatic: Integer values are not automatically converted to corresponding **enum** values. For example, the **suit** variable below might be expected to become **Suit.diamonds**, but the code cannot be compiled:

```
Suit suit = 2; // ← compilation ERROR
```

As we will see below, conversions from integers to **enum** values are still possible but they must be explicit.

48.1.6 bool conversions

false and **true** are automatically converted to 0 and 1, respectively:

```
int a = false;
assert(a == 0);

int b = true;
assert(b == 1);
```

The opposite conversion is also automatic but only for two special values. The literal values 0 and 1 are converted automatically to **false** and **true**, respectively:

```
bool a = 0;
assert(!a);    // false

bool b = 1;
assert(b);     // true
```

Other literal values cannot be converted to **bool** automatically:

```
bool b = 2;    // ← compilation ERROR
```

48.1.7 Automatic conversions to bool in conditional statements

Some statements make use of logical expressions: **if**, **while**, etc. For the logical expressions of such statements, not only **bool** but most other types can be used. The value zero is automatically converted to **false** and the non-zero values are automatically converted to **true**.

```
int i;
// ...

if (i) {    // ← int value is being used as a logical expression
    // ... 'i' is not zero
} else {
    // ... 'i' is zero
}
```

Similarly, **null** references are automatically converted to **false** and non-**null** references are automatically converted to **true**. This makes it easy to ensure that a reference is non-**null** before actually using it:

```
int[] a;
// ...

if (a) {    // ← automatic bool conversion
    // ... not null; 'a' can be used ...
} else {
    // ... null; 'a' cannot be used ...
}
```

48.2 Explicit type conversions

As we have seen above, there are cases where automatic conversions are not available:

- Conversions from wider types to narrower types
- Conversions from **const** to mutable

- **immutable** conversions
- Conversions from integers to **enum** values
- etc.

If such a conversion is known to be safe, the programmer can explicitly ask for a type conversion by one of three methods:

- By calling the **std.conv.to** function
- By calling the **std.exception.assertUnique** function
- By the **cast** operator

48.2.1 to() for most conversions

The **to()** function, which we have already used mostly to convert values to **string**, can actually be used for many other types. Its complete syntax is the following:

```
to!(DestinationType)(value)
```

Being a template, **to()** can take advantage of the shortcut template parameter notation: When the destination type consists only of a single token (generally, *a single word*), it can be called without the first pair of parentheses:

```
to!DestinationType(value)
```

The following program is trying to convert a **double** value to **short** and a **string** value to **int**:

```
void main()
{
    double d = -1.75;

    short s = d;    // ← compilation ERROR
    int i = "42";  // ← compilation ERROR
}
```

Since not every **double** value can be represented as a **short** and not every **string** can be represented as an **int**, those conversions are not automatic. When it is known by the programmer that the conversions are in fact safe or that the potential consequences are acceptable, then the types can be converted by **to()**:

```
import std.conv;

void main()
{
    double d = -1.75;

    short s = to!short(d);
    assert(s == -1);

    int i = to!int("42");
    assert(i == 42);
}
```

Note that because **short** cannot carry fractional values, the converted value is -1.

to() is safe: It throws an exception when a conversion is not possible.

48.2.2 assertUnique() for fast immutable conversions

to() can perform **immutable** conversions as well:

```
int[] slice = [ 10, 20, 30 ];
auto immutableSlice = to!(immutable int[])(slice);
```

In order to guarantee that the elements of `immutableSlice` will never change, it cannot share the same elements with `slice`. For that reason, `to()` creates an additional slice with `immutable` elements above. Otherwise, modifications to the elements of `slice` would cause the elements of `immutableSlice` change as well. This behavior is the same with the `.idup` property of arrays.

We can see that the elements of `immutableSlice` are indeed copies of the elements of `slice` by looking at the addresses of their first elements:

```
assert(&(slice[0]) != &(immutableSlice[0]));
```

Sometimes this copy is unnecessary and may slow the speed of the program noticeably in certain cases. As an example of this, let's look at the following function that takes an `immutable` slice:

```
void calculate(immutable int[] coordinates)
{
    // ...
}

void main()
{
    int[] numbers;
    numbers ~= 10;
    // ... various other modifications ...
    numbers[0] = 42;

    calculate(numbers);    // ← compilation ERROR
}
```

The program above cannot be compiled because the caller is not passing an `immutable` argument to `calculate()`. As we have seen above, an `immutable` slice can be created by `to()`:

```
import std.conv;
// ...
auto immutableNumbers = to!(immutable int[])(numbers);
calculate(immutableNumbers);    // ← now compiles
```

However, if `numbers` is needed only to produce this argument and will never be used after the function is called, copying its elements to `immutableNumbers` would be unnecessary.

`assumeUnique()` makes the elements of a slice `immutable` without copying:

```
import std.exception;
// ...
auto immutableNumbers = assumeUnique(numbers);
calculate(immutableNumbers);
assert(numbers is null);    // the original slice becomes null
```

`assumeUnique()` returns a new slice that provides `immutable` access to the existing elements. It also makes the original slice `null` to prevent the elements from accidentally being modified through it.

48.2.3 The cast operator

Both `to()` and `assumeUnique()` make use of the conversion operator `cast`, which is available to the programmer as well.

The `cast` operator takes the destination type in parentheses:

```
cast(DestinationType)value
```

`cast` is powerful even for conversions that `to()` cannot safely perform. `to()` does not perform the following conversions:

```
Suit suit = to!Suit(2);    // ← compilation ERROR
bool b = to!bool(2);     // ← compilation ERROR
```

Sometimes only the programmer can know whether an integer value corresponds to a valid `enum` value or that it makes sense to treat an integer value as a `bool`. The `cast` operator can be used when the conversion is known to be correct according to the program's logic:

```
Suit suit = cast(Suit)2;
assert(suit == Suit.diamonds);

bool b = cast(bool)2;
assert(b);
```

48.3 Summary

- Automatic type conversions are mostly in the safe direction: From the narrower type to the wider type and from mutable to `const`.
- However, conversions to unsigned types may have surprising effects because unsigned types cannot have negative values.
- `enum` types can automatically be converted to integer values but the opposite conversion is not automatic.
- `false` and `true` are automatically converted to 0 and 1 respectively. Similarly, zero values are automatically converted to `false` and non-zero values are automatically converted to `true`.
- `null` references are automatically converted to `false` and non-`null` references are automatically converted to `true`.
- `to()` covers most of the explicit conversions.
- `assumeUnique()` converts to `immutable` without copying.
- The `cast` operator is the most powerful conversion tool.

49 Structs

As has been mentioned several times earlier in the book, fundamental types are not suitable to represent higher-level concepts. For example, although a value of type `int` is suitable to represent the hour of day, two `int` variables would be more suitable together to represent a point in time: one for the hour and the other for the minute.

Structs are the feature that allow defining new types by combining already existing other types. The new type is defined by the `struct` keyword. Most of the content of this chapter is directly applicable to classes as well. Especially the concept of *combining existing types to define a new type* is exactly the same for them.

To understand how useful structs are, let's take a look at the `addDuration()` function that we had defined earlier in the [assert and enforce chapter](#). The following definition is from the exercise solution of that chapter:

```
void addDuration(in int startHour, in int startMinute,
                in int durationHour, in int durationMinute,
                out int resultHour, out int resultMinute)
{
    resultHour = startHour + durationHour;
    resultMinute = startMinute + durationMinute;

    if (resultMinute > 59) {
        ++resultHour;
    }
}
```

Note: I will ignore the `in`, `out`, and `unittest` blocks in this chapter to keep the code samples short.

Although the function above clearly takes six parameters, when the three pairs of parameters are considered, it is conceptually taking only three bits of information for the starting time, the duration, and the result.

49.1 Definition

The `struct` keyword defines a new type by combining variables that are related in some way:

```
struct TimeOfDay
{
    int hour;
    int minute;
}
```

The code above defines a new type named `TimeOfDay`, which consists of two variables named `hour` and `minute`. That definition allows the new `TimeOfDay` type to be used in the program just like any other type. The following code demonstrates how similar its use is to an `int`'s:

```
int number;           // a variable
number = otherNumber; // taking the value of otherNumber

TimeOfDay time;      // an object
time = otherTime;    // taking the value of otherTime
```

The syntax of `struct` definition is the following:

```
struct TypeName
{
```

```

} // ... member variables and functions ...
}

```

We will see member functions in later chapters.

The variables that a struct combines are called its *members*. According to this definition, **TimeOfDay** has two members: **hour** and **minute**.

49.1.1 struct defines a type, not a variable

There is an important distinction here: Especially after the [Name Space](#) and [Lifetimes and Fundamental Operations](#) chapters, the curly brackets of **struct** definitions may give the wrong impression that the struct members start and end their lives inside that scope. This is not true.

Member definitions are not variable definitions:

```

struct TimeOfDay
{
    int hour; // ← not a variable definition!
    int minute; // ← not a variable definition!
}

```

The definition of a **struct** determines the types and the names of the members that the objects of that **struct** will have:

```

struct TimeOfDay
{
    int hour; // ← will become a part of a variable
    int minute; // ← will become a part of a variable
}

```

Those member variables will be constructed as parts of **TimeOfDay** objects that take part in the program:

```

TimeOfDay bedTime; // This object contains its own hour
                  // and minute member variables.

TimeOfDay wakeUpTime; // This object contains its own hour
                    // and minute member variables as
                    // well. The member variables of
                    // this object are not related to
                    // the member variables of the
                    // previous object.

```

49.1.2 Coding convenience

Being able to combine the concepts of hour and minute together as a new type is a great convenience. For example, the function above can be rewritten in a more meaningful way by taking three **TimeOfDay** parameters instead of the existing six **int** parameters:

```

void addDuration(in TimeOfDay start,
                in TimeOfDay duration,
                out TimeOfDay result)
{
    // ...
}

```

Note: It is not normal to add two variables that represent two points in time. For example, it is meaningless to add the lunch time 12:00 to the breakfast time 7:30. It would make more sense to define another type, appropriately called **Duration**, and to add objects of that type to **TimeOfDay** objects. Despite this design error, I will continue using only **TimeOfDay** objects in this chapter and introduce **Duration** in a later chapter.

As you remember, functions return up-to a single value. That has precisely been the reason why the earlier definition of **addDuration()** was taking two **out** parameters: It could not return the hour and minute information as a single value.

Structs remove this limitation as well: Since multiple values can be combined as a single **struct** type, functions can return an object of such a **struct**, effectively returning multiple values at once. **addDuration()** can now be defined as returning its result:

```
TimeOfDay addDuration(in TimeOfDay start,
                    in TimeOfDay duration)
{
    // ...
}
```

As a consequence, **addDuration()** now becomes a function that produces a value, as opposed to being a function that has side effects. As you would remember from the [Functions chapter](#), producing results is preferred over having side effects.

Structs can be members of other structs. For example, the following **struct** has two **TimeOfDay** members:

```
struct Meeting
{
    string    topic;
    size_t    attendanceCount;
    TimeOfDay start;
    TimeOfDay end;
}
```

Meeting can in turn be a member of another **struct**. Assuming that there is also the **Meal** struct:

```
struct DailyPlan
{
    Meeting projectMeeting;
    Meal    lunch;
    Meeting budgetMeeting;
}
```

49.2 Accessing the members

Struct members are used like any other variable. The only difference is that the actual struct variable and a *dot* must be specified before the name of the member:

```
start.hour = 10;
```

The line above assigns the value 10 to the **hour** member of the **start** object.

Let's rewrite the **addDuration()** function with what we have seen so far:

```
TimeOfDay addDuration(in TimeOfDay start,
                    in TimeOfDay duration)
{
```

```

    TimeOfDay result;

    result.minute = start.minute + duration.minute;
    result.hour = start.hour + duration.hour;
    result.hour += result.minute / 60;

    result.minute %= 60;
    result.hour %= 24;

    return result;
}

```

Notice that the names of the variables are now much shorter in this version of the function: **start**, **duration**, and **result**. Additionally, instead of using complex names like **startHour**, it is possible to access struct members through their respective struct variables as in **start.hour**.

Here is a code that uses the new **addDuration()** function. Given the start time and the duration, the following code calculates when a class period at a school would end:

```

void main()
{
    TimeOfDay periodStart;
    periodStart.hour = 8;
    periodStart.minute = 30;

    TimeOfDay periodDuration;
    periodDuration.hour = 1;
    periodDuration.minute = 15;

    immutable periodEnd = addDuration(periodStart,
                                      periodDuration);

    writeln("Period end: %s:%s",
           periodEnd.hour, periodEnd.minute);
}

```

The output:

```
Period end: 9:45
```

The **main()** above has been written only by what we have seen so far. We will make this code even shorter and cleaner soon.

49.3 Construction

The first three lines of **main()** is about constructing the **periodStart** object and the next three lines are about constructing the **periodDuration** object. In each three lines of code first an object is being defined and then its hour and minute values are being set.

In order for a variable to be used in a safe way, that variable must first be constructed in a consistent state. Because construction is so common, there is a special construction syntax for struct objects:

```

TimeOfDay periodStart = TimeOfDay(8, 30);
TimeOfDay periodDuration = TimeOfDay(1, 15);

```

The values are automatically assigned to the members in the order that they are specified: Because **hour** is defined first in the **struct**, the value 8 is assigned to **periodStart.hour** and 30 is assigned to **periodStart.minute**.

49.3.1 Constructing objects as immutable

Being able to construct the object by specifying the values of its members at once makes it possible to define objects as **immutable**:

```
immutable periodStart = TimeOfDay(8, 30);
immutable periodDuration = TimeOfDay(1, 15);
```

Otherwise it would not be possible to mark an object first as **immutable** and then modify its members:

```
immutable TimeOfDay periodStart;
periodStart.hour = 8; // ← compilation ERROR
periodStart.minute = 30; // ← compilation ERROR
```

49.3.2 Trailing members need not be specified

There may be less values specified than the number of members. In that case, the remaining members are initialized by the **.init** values of their respective types.

The following program constructs **Test** objects each time with one less constructor parameter. The **assert** checks indicate that the unspecified members are initialized automatically by their **.init** values. (The reason for needing to call **isnan()** is explained after the program):

```
import std.math;

struct Test
{
    char    c;
    int     i;
    double  d;
}

void main()
{
    // The initial values of all of the members are specified
    auto t1 = Test('a', 1, 2.3);
    assert(t1.c == 'a');
    assert(t1.i == 1);
    assert(t1.d == 2.3);

    // Last one is missing
    auto t2 = Test('a', 1);
    assert(t2.c == 'a');
    assert(t2.i == 1);
    assert(isnan(t2.d));

    // Last two are missing
    auto t3 = Test('a');
    assert(t3.c == 'a');
    assert(t3.i == int.init);
    assert(isnan(t3.d));

    // No initial value specified
    auto t4 = Test();
    assert(t4.c == char.init);
    assert(t4.i == int.init);
    assert(isnan(t4.d));

    // The same as above
    Test t5;
    assert(t5.c == char.init);
    assert(t5.i == int.init);
}
```

```
    assert(isnan(t5.d));
}
```

As you would remember from the [Floating Point Types chapter](#), the initial value of **double** is **double.nan**. Since the **.nan** value is *unordered*, it is meaningless to use it in equality comparisons. That is why calling **std.math.isnan** is the correct way of determining whether a value equals to **.nan**.

49.3.3 Specifying default values for members

It is important that member variables are automatically initialized with known initial values. This prevents the program from continuing with indeterminate values. However, the **.init** value of their respective types may not be suitable for every type. For example, **char.init** is not even a valid value.

The initial values of the members of a struct can be determined when the struct is defined. This is useful for example to initialize floating point members by **0.0**, instead of the mostly-unusable **.nan**.

The default values are specified by the assignment syntax as the members are defined:

```
struct Test
{
    char    c = 'A';
    int     i = 11;
    double  d = 0.25;
}
```

Please note that the syntax above is not really assignment. The code above merely determines the default values that will be used when objects of that struct are actually constructed later in the program.

For example, the following **Test** object is being constructed without any specific values:

```
Test t; // no value is specified for the members
writeln("%s,%s,%s", t.c, t.i, t.d);
```

All of the members are initialized by their default values:

```
A,11,0.25
```

49.3.4 Constructing by the { } syntax

Struct objects can also be constructed by the following syntax:

```
TimeOfDay periodStart = { 8, 30 };
```

Similar to the earlier syntax, the specified values are assigned to the members in the order that they are specified; and the trailing members get their default values.

This syntax is inherited from the C programming language:

```
auto periodStart = TimeOfDay(8, 30); // ← recommended
TimeOfDay periodEnd = { 9, 30 };    // ← as in C
```

49.4 Copying and assignment

Structs are value types. As has been described in the [Value Types and Reference Types chapter](#), this means that every **struct** object has its own value. Objects get their own values when constructed, and their values change when they are assigned new values.

```

auto yourLunchTime = TimeOfDay(12, 0);
auto myLunchTime = yourLunchTime;

// Only my lunch time becomes 12:05:
myLunchTime.minute += 5;

// ... your lunch time is still the same:
assert(yourLunchTime.minute == 0);

```

During a copy, all of the members of the source object are automatically copied to the corresponding members of the destination object. Similarly, assignment involves assigning each member of the source to the corresponding member of the destination.

Struct members that are of reference types need careful attention.

49.4.1 Careful with members that are of reference types!

As you remember, copying or assigning variables of reference types does not change any value, it changes what object is being referenced. As a result, copying or assigning causes one more reference to the right-hand side object. The relevance of this for struct members is that, the members of two separate struct objects would start providing access to the same value.

To see an example of this, let's have a look at a struct where one of the members is a reference type. This struct is used for keeping the student number and the grades of a student:

```

struct Student
{
    int number;
    int[] grades;
}

```

The following code constructs a second **Student** object by copying an existing one:

```

// Constructing the first object:
auto student1 = Student(1, [ 70, 90, 85 ]);

// Constructing the second student as a copy of the first
// one and then changing its number:
auto student2 = student1;
student2.number = 2;

// WARNING: The grades are now being shared by the two objects!

// Changing the grades of the first student ...
student1.grades[0] += 5;

// ... affects the second student as well:
writeln(student2.grades[0]);

```

When **student2** is constructed, its members get the values of the members of **student1**. Since **int** is a value type, the second object gets its own **number** value.

The two **Student** objects also have individual **grades** members as well. However, since slices are reference types, the actual elements that the two slices share are the same. Consequently, a change made through one of the slices is seen through the other slice.

The output of the code indicates that the grade of the second student has been increased as well:

75

For that reason, a better approach might be to construct the second object by the copies of the grades of the first one:

```
// The second Student is being constructed by the copies
// of the grades of the first one:
auto student2 = Student(2, student1.grades.dup);

// Changing the grades of the first student ...
student1.grades[0] += 5;

// ... does not affect the grades of the second student:
writeln(student2.grades[0]);
```

Since the grades have been copied by **.dup**, this time the grades of the second student is not affected:

70

Note: It is possible to have even the reference members copied automatically. We will see how this is done later when covering struct member functions.

49.5 Struct literals

Similar to being able to use integer literal values like 10 in expressions without needing to define a variable, struct objects can be used as literals as well.

Struct literals are constructed by the object construction syntax.

```
TimeOfDay(8, 30) // ← struct literal value
```

Let's first rewrite the **main()** function above with what we have learned since its last version. The variables are constructed by the construction syntax and are **immutable** this time:

```
void main()
{
    immutable periodStart = TimeOfDay(8, 30);
    immutable periodDuration = TimeOfDay(1, 15);

    immutable periodEnd = addDuration(periodStart,
                                     periodDuration);

    writefln("Period end: %s:%s",
            periodEnd.hour, periodEnd.minute);
}
```

Note that **periodStart** and **periodDuration** need not be defined as named variables in the code above. Those are in fact temporary variables in this simple program, which are used only for calculating the **periodEnd** variable. They could be passed to **addDuration()** as literal values:

```

void main()
{
    immutable periodEnd = addDuration(TimeOfDay(8, 30),
                                     TimeOfDay(1, 15));

    writeln("Period end: %s:%s",
           periodEnd.hour, periodEnd.minute);
}

```

49.6 static members

Although objects mostly need to have individual copies of the struct members, sometimes it makes sense for the objects of a particular struct type to share some variables. This may be needed e.g. when a general information about that struct type need to be maintained.

As an example, let's imagine a type that assigns a separate identifier for every object that is constructed of that type:

```

struct Point
{
    // The identifier of each object
    size_t id;

    int line;
    int column;
}

```

In order to be able to assign different ids to each object, a separate variable is needed to keep the next available id. It would be incremented every time a new object is created.

Assume that `nextId` is to be defined elsewhere and to be available in the following function:

```

Point makePoint(int line, int column)
{
    size_t id = nextId;
    ++nextId;

    return Point(id, line, column);
}

```

A decision must be made regarding where the common `nextId` variable is to be defined.

static members are useful in such cases. Such common information is defined as a **static** member of the struct. Contrary to the regular members, there is a single variable of each **static** member throughout the program. That single variable is shared by all of the objects of that struct:

```

import std.stdio;

struct Point
{
    // The identifier of each object
    size_t id;

    int line;
    int column;

    // The id of the next object to construct
    static size_t nextId;
}

Point makePoint(int line, int column)
{
    size_t id = Point.nextId;
}

```

```

    ++Point.nextId;
    return Point(id, line, column);
}

void main()
{
    auto top = makePoint(7, 0);
    auto middle = makePoint(8, 0);
    auto bottom = makePoint(9, 0);

    writeln(top.id);
    writeln(middle.id);
    writeln(bottom.id);
}

```

As `nextId` is incremented at each object construction, each object gets a unique id:

```

0
1
2

```

Since `static` members are owned by the entire type, there need not be an object to access them. As we have seen above, such objects can be accessed by the name of the type, as well as by the name of any object of that struct:

```

++Point.nextId;
++bottom.nextId;    // would be the same as above

```

49.7 Exercises

1. Design a struct named `Card` to represent a playing card. This struct can have two members for the suit and the value. It may make sense to use an `enum` to represent the suit, or you can simply use the characters ♠, ♥, ♦, and ♣.

An `int` or a `dchar` value can be used for the card value. If you decide to use an `int`, the values 1, 11, 12, and 13 may represent the cards that do not have numbers (ace, jack, queen, and king).

There are other design choices to make. For example, the card values can be represented by an `enum` type as well.

The way objects of this struct will be constructed will depend on the choice of the types of its members. For example, if both members are `dchar`, then `Card` objects can be constructed like this:

```

auto card = Card('♣', '2');

```

2. Define a function named `printCard()`, which takes a `Card` object as a parameter and simply prints it:

```

void printCard(in Card card)
{
    // ... please define the function body ...
}

void main()
{

```

```

    auto card = Card(/* ... */);
    printCard(card);
}

```

For example, the function can print the 2 of clubs as:

```
♣2
```

The implementation of that function may depend on the choice of the types of the members.

3. Define a function named **newDeck()** and have it return 52 cards of a deck as a **Card** slice:

```

Card[] newDeck()
out (result)
{
    assert(result.length == 52);
}
body
{
    // ... please define the function body ...
}

```

It should be possible to call **newDeck()** as in the following code:

```

void main()
{
    Card[] deck = newDeck();

    foreach (card; deck) {
        printCard(card);
        write(' ');
    }

    writeln();
}

```

The output should be similar to the following with 52 distinct cards:

```

♠2 ♠3 ♠4 ♠5 ♠6 ♠7 ♠8 ♠9 ♠0 ♠J ♠Q ♠K ♠A ♥2 ♥3 ♥4
♥5 ♥6 ♥7 ♥8 ♥9 ♥0 ♥J ♥Q ♥K ♥A ♦2 ♦3 ♦4 ♦5 ♦6 ♦7
♦8 ♦9 ♦0 ♦J ♦Q ♦K ♦A ♣2 ♣3 ♣4 ♣5 ♣6 ♣7 ♣8 ♣9 ♣0
♣J ♣Q ♣K ♣A

```

4. Write a function that shuffles the deck. One way is to pick two cards at random by **std.random.uniform**, to swap those two cards, and to repeat this process a sufficient number of times. The function should take the number of repetition as a parameter:

```

void shuffle(Card[] deck, in int repetition)
{
    // ... please define the function body ...
}

```

Here is how it should be used:

```

void main()
{
    Card[] deck = newDeck();
    shuffle(deck, 1);

    foreach (card; deck) {

```

```

        printCard(card);
        write(' ');
    }

    writeln();
}

```

The function should swap cards **repetition** number of times. For example, when called by 1, the output should be similar to the following:

```

♠2 ♠3 ♠4 ♠5 ♠6 ♠7 ♠8 ♠9 ♠0 ♠J ♠Q ♠K ♠A ♥2 ♥3 ♥4
♥5 ♥6 ♥7 ♥8 ♣4 ♥0 ♥J ♥Q ♥K ♥A ♦2 ♦3 ♦4 ♦5 ♦6 ♦7
♦8 ♦9 ♦0 ♦J ♦Q ♦K ♦A ♠2 ♠3 ♥9 ♠5 ♠6 ♠7 ♠8 ♠9 ♠0
♣J ♣Q ♣K ♣A

```

A higher value for **repetition** should result in a more shuffled deck:

```

shuffled(deck, 100);

```

The output:

```

♠4 ♠7 ♦9 ♦6 ♥2 ♠6 ♠6 ♦A ♠5 ♦8 ♦3 ♥Q ♦J ♠K ♠8 ♠4
♥J ♠Q ♠Q ♠9 ♦0 ♥A ♠A ♥9 ♠7 ♥3 ♦K ♦2 ♥0 ♠J ♦7 ♥7
♠8 ♥4 ♠J ♦4 ♠0 ♥6 ♦5 ♥5 ♥K ♠3 ♦Q ♠2 ♠5 ♠2 ♥8 ♠A
♠K ♠9 ♠0 ♠3

```

Note: A much better way of shuffling the deck is explained in the solutions.

[... the solutions](#)

50 Variable Number of Parameters

This chapter covers two D features that bring flexibility on parameters when calling functions:

- Default parameter values
- Variadic functions

50.1 Default parameter values

A convenience with function parameters is the ability to specify default values for them. This is similar to the default initial values of struct members.

Some of the parameters of some functions are called mostly by the same values. To see an example of this, let's consider a function that prints the elements of an associative array of type `string[string]`. Let's assume that the function takes the separator characters as parameters as well:

```
void printAA(in char[] title,
            in string[string] aa,
            in char[] keySeparator,
            in char[] elementSeparator)
{
    writeln("-- ", title, " --");
    auto keys = aa.keys.sort;
    foreach (i, key; keys) {
        // No separator before the first element
        if (i != 0) {
            write(elementSeparator);
        }
        write(key, keySeparator, aa[key]);
    }
    writeln();
}
```

That function is being called below by ":" as the key separator and ", " as the element separator:

```
string[string] dictionary = [
    "blue":"mavi", "red":"kırmızı", "gray":"gri" ];
printAA("Color Dictionary", dictionary, ":", ", ");
```

The output:

```
-- Color Dictionary --
blue:mavi, gray:gri, red:kırmızı
```

If the separators are almost always going to be those two, they can be defined with default values:

```
void printAA(in char[] title,
            in string[string] aa,
            in char[] keySeparator = ":",
            in char[] elementSeparator = ", ")
{
```

```
} // ...
```

Parameters with default values need not be specified when the function is called:

```
printAA("Color Dictionary",
        dictionary); /* ← No separator specified. Both
                    *   parameters will get their
                    *   default values. */
```

The parameter values can still be specified when needed, and not necessarily all of them:

```
printAA("Color Dictionary", dictionary, "=");
```

The output:

```
-- Color Dictionary --
blue=mavi, gray=gri, red=kırmızı
```

The following call specifies both of the parameters:

```
printAA("Color Dictionary", dictionary, "=", "\n");
```

The output:

```
-- Color Dictionary --
blue=mavi
gray=gri
red=kırmızı
```

Default values can only be defined for the parameters that are at the end of the parameter list.

50.2 Variadic functions

Despite appearances, default parameter values do not change the number of parameters that a function receives. For example, even though some parameters may be assigned their default values, `printAA()` always takes four parameters and uses them according to its implementation.

The variadic functions feature on the other hand allows calling functions with unspecified number of arguments. We have already been taking advantage of this feature with functions like `writeln()`. `writeln()` can be called with any number of arguments:

```
writeln(
    "hello", 7, "world", 9.8 /*, and any number of other
                          * arguments as needed */);
```

There are four ways of defining variadic functions in D:

- The feature that works only for functions that are marked as **extern "C"**. This feature defines the hidden `_argptr` variable that is used for accessing the parameters. This book does not cover this feature partly because it is unsafe.
- The feature that works with regular D functions, which also uses the hidden `_argptr` variable, as well as the `_arguments` variable, the latter being of type `TypeInfo[]`. This

book does not cover this feature as well both because it relies on *pointers*, which have not been covered yet, and because this feature can be used in unsafe ways as well.

- A safe feature with the limitation that the unspecified number of parameters must all be of the same type. This is the feature that is covered in this section.
- Unspecified number of template parameters. This feature will be explained later in the templates chapters.

The parameters of variadic functions are passed to the function as a slice. Variadic functions are defined with a single parameter of a specific type of slice followed immediately by the `...` characters:

```
double sum(in double[] numbers ...)
{
    double result = 0.0;

    foreach (number; numbers) {
        result += number;
    }

    return result;
}
```

That definition makes `sum()` a variadic function, meaning that it is able to receive any number of arguments as long as they are **double** or any other type that can implicitly be convertible to **double**:

```
writeln(sum(1.1, 2.2, 3.3));
```

The single slice parameter and the `...` characters represent all of the arguments. For example, the slice would have five elements if the function were called with five **double** values.

In fact, the variable number of parameters can also be passed as a single slice:

```
writeln(sum([ 1.1, 2.2, 3.3 ])); // same as above
```

Variadic functions can also have required parameters, which must be defined first in the parameter list. For example, the following function prints an unspecified number of parameters within parentheses. Although the function leaves the number of the elements flexible, it requires that the parentheses are always specified:

```
char[] parenthesize(
    in char[] opening, // ← The first two parameters must be
    in char[] closing, // specified when the function is called.
    in char[][] words ...) // ← Need not be specified
{
    char[] result;

    foreach (word; words) {
        result ~= opening;
        result ~= word;
        result ~= closing;
    }

    return result;
}
```

The first two parameters are mandatory:

```
parenthesize("{}"); // ← compilation ERROR
```

As long as the mandatory parameters are specified, the rest are flexible:

```
writeln(parenthesize("{", "}", "apple", "pear", "banana"));
```

The output:

```
{apple}{pear}{banana}
```

50.3 Exercise

- Assume that the following **enum** is already defined:

```
enum Operation { add, subtract, multiply, divide }
```

Also assume that there is a **struct** that represents the calculation of an operation and its two operands:

```
struct Calculation
{
    Operation op;
    double first;
    double second;
}
```

For example, the object **Calculation(Operation.divide, 7.7, 8.8)** would represent the division of 7.7 by 8.8.

Design a function that receives an unspecified number of these **struct** objects, calculates the result of each **Calculation**, and then returns all of the results as a slice of type **double[]**.

For example, it should be possible to call the function as in the following code:

```
void main()
{
    writeln(calculate(Calculation(Operation.add, 1.1, 2.2),
                    Calculation(Operation.subtract, 3.3, 4.4),
                    Calculation(Operation.multiply, 5.5, 6.6),
                    Calculation(Operation.divide, 7.7, 8.8)));
}
```

The output of the code should be similar to the following:

```
[3.3, -1.1, 36.3, 0.875]
```

... the solution

51 Function Overloading

Defining more than one function having the same name is *function overloading*. In order to be able to differentiate these functions, their parameters must be different.

The following code has multiple overloads of the `info()` function, each taking a different type of parameter:

```
import std.stdio;

void info(in double number)
{
    writeln("Floating point: ", number);
}

void info(in int number)
{
    writeln("Integer      : ", number);
}

void info(in char[] str)
{
    writeln("String        : ", str);
}

void main()
{
    info(1.2);
    info(3);
    info("hello");
}
```

Although all of the functions are named `info()`, the compiler picks the one that matches the argument that is used when making the call. For example, because the literal `1.2` is of type `double`, the `info()` function that takes a `double` gets called for it.

The choice of which function to call is made at compile time, which may not always be easy or clear. For example, because `int` can implicitly be converted to both `double` and `real`, the compiler cannot decide which of the functions to call in the following program:

```
real sevenTimes(in real value)
{
    return 7 * value;
}

double sevenTimes(in double value)
{
    return 7 * value;
}

void main()
{
    int value = 5;
    auto result = sevenTimes(value);    // ← compilation ERROR
}
```

Note: It is usually unnecessary to write separate functions when the function bodies are exactly the same. We will see later in the [Templates chapter](#) how a single definition can be used for multiple types.

However, if there is another function overload that takes a `long` parameter, then the ambiguity would be resolved because `long` is a *better match* for `int` than `double` or `real`:

```

long sevenTimes(in long value)
{
    return 7 * value;
}

// ...

auto result = sevenTimes(value);    // now compiles

```

51.1 Overload resolution

The compiler picks the overload that is the *best match* for the arguments. This is called overload resolution.

Although overload resolution is simple and intuitive in most cases, it is sometimes complicated. The following are the rules of overload resolution. They are being presented in a simplified way in this book.

There are four states of match, listed from the worst to the best:

- mismatch
- match through automatic type conversion
- match through **const** qualification
- exact match

The compiler considers all of the overloads of a function during overload resolution. It first determines the match state of every parameter for every overload. For each overload, the least match state among the parameters is taken to be the match state of that overload.

After all of the match states of the overloads are determined, then the overload with the best match is selected. If there are more than one overload that has the best match, then more complicated resolution rules are applied. I will not get into more details of these rules in this book. If your program is in a situation where it depends on complicated overload resolution rules, it may be an indication that it is time to change the design of the program. Another option is to take advantage of other features of D like templates. An even simpler approach would be to abandon function overloading altogether by naming functions differently for each type e.g. like `sevenTimes_real()` and `sevenTimes_double()`.

51.2 Function overloading for structs

Function overloading is useful with structs and classes as well. Additionally, overload resolution ambiguities are much less frequent with user-defined types. Let's overload the `info()` function above for some of the types that we have defined in the [Structs chapter](#):

```

struct TimeOfDay
{
    int hour;
    int minute;
}

void info(in TimeOfDay time)
{
    writef("%02s:%02s", time.hour, time.minute);
}

```

That overload enables `TimeOfDay` objects to be used with `info()`. As a result, variables of user-defined types can be printed in exactly the same way as fundamental types:

```
auto breakfastTime = TimeOfDay(7, 0);
info(breakfastTime);
```

The **TimeOfDay** objects would be matched with that overload of **info()**:

```
07:00
```

The following is an overload of **info()** for the **Meeting** type:

```
struct Meeting
{
    string    topic;
    size_t    attendanceCount;
    TimeOfDay start;
    TimeOfDay end;
}

void info(in Meeting meeting)
{
    info(meeting.start);
    write('-');
    info(meeting.end);

    writef(" \"%s\" meeting with %s attendees",
           meeting.topic,
           meeting.attendanceCount);
}
```

Note that this overload makes use of the already-defined overload for **TimeOfDay**. **Meeting** objects can now be printed in exactly the same way as fundamental types as well:

```
auto bikeRideMeeting = Meeting("Bike Ride", 3,
                               TimeOfDay(9, 0),
                               TimeOfDay(9, 10));

info(bikeRideMeeting);
```

The output:

```
09:00-09:10 "Bike Ride" meeting with 3 attendees
```

51.3 Limitations

Although the **info()** function overloads above are a great convenience, this method has some limitations:

- **info()** always prints to **stdout**. It would be more useful if it could print to any **File**. One way of achieving this is to pass the output stream as a parameter as well e.g. for the **TimeOfDay** type:

```
void info(File file, in TimeOfDay time)
{
    file.writef("%02s:%02s", time.hour, time.minute);
}
```

That would enable printing **TimeOfDay** objects to any file, including **stdout**:

```
info(stdout, breakfastTime);

auto file = File("a_file", "w");
info(file, breakfastTime);
```

Note: The special objects `stdin`, `stdout`, and `stderr` are of type `File`.

- More importantly, `info()` does not solve the more general problem of producing the string representation of variables. For example, it does not help with passing objects of user-defined types to `writeln()`:

```
writeln(breakfastTime); // Not useful: prints in generic format
```

The code above prints the object in a generic format that includes the name of the type and the values of its members, not in a way that would be useful in the program:

```
TimeOfDay(7, 0)
```

It would be much more useful if there were a function that converted `TimeOfDay` objects to `string` in their special format as in `"12:34"`. We will see how to define `string` representations of struct objects in the next chapter.

51.4 Exercise

- Overload the `info()` function for the following structs as well:

```
struct Meal
{
    TimeOfDay time;
    string    address;
}

struct DailyPlan
{
    Meeting amMeeting;
    Meal    lunch;
    Meeting pmMeeting;
}
```

Since `Meal` has only the start time, add an hour and a half to determine its end time. You can use the `addDuration()` function that we have defined earlier in the structs chapter:

```
TimeOfDay addDuration(in TimeOfDay start,
                    in TimeOfDay duration)
{
    TimeOfDay result;

    result.minute = start.minute + duration.minute;
    result.hour   = start.hour   + duration.hour;
    result.hour += result.minute / 60;

    result.minute %= 60;
    result.hour   %= 24;

    return result;
}
```

Once the end times of `Meal` objects are calculated by `addDuration()`, `DailyPlan` objects should be printed as in the following output:

```
10:30-11:45 "Bike Ride" meeting with 4 attendees
12:30-14:00 Meal, Address: İstanbul
15:30-17:30 "Budget" meeting with 8 attendees
```


... the solution

52 Member Functions

Although this chapter focuses only on structs, most of the information in this chapter is applicable to classes as well.

In this chapter we will cover member functions of structs and define the special `toString()` member function that is used for representing objects in the `string` format.

When a struct or class is defined, usually a number of functions are also defined alongside with it. We have seen examples of such functions in the earlier chapters: `addDuration()` and an overload of `info()` have been written specifically to be used with the `TimeOfDay` type. In a sense, these two functions define the *interface* of `TimeOfDay`.

The first parameter of both `addDuration()` and `info()` has been the `TimeOfDay` object that each function would be operating on. Additionally, just like all of the other functions that we have seen so far, both of the functions have been defined at the *module level*, outside of any other scope.

The concept of a set of functions determining the interface of a struct is very common. For that reason, functions that are closely related to a type can be defined within the body of that type.

52.1 Defining member functions

Functions that are defined within the curly brackets of a `struct` are called *member functions*:

```
struct SomeStruct
{
    void member_function(/* the parameters of the function */)
    {
        // ... the definition of the function ...
    }
    // ... the other members of the struct ...
}
```

Member functions are accessed the same way as member variables, separated from the name of the object by a dot:

```
object.member_function(arguments);
```

We have used member functions before when specifying `stdin` and `stdout` explicitly during input and output operations:

```
stdin.readf(" %s", &number);
stdout.writeln(number);
```

The `readf()` and `writeln()` above are member function calls, operating on the objects `stdin` and `stdout`, respectively.

Let's define `info()` as a member function. Its previous definition has been the following:

```
void info(in TimeOfDay time)
{
    writef("%02s:%02s", time.hour, time.minute);
}
```

Making `info()` a member function is not as simple as moving its definition inside the struct. The function must be modified in two ways:

```
struct TimeOfDay
{
    int hour;
    int minute;

    void info()    // (1)
    {
        printf("%02s:%02s", hour, minute);    // (2)
    }
}
```

1. The member function does not take the object explicitly as a parameter.
2. For that reason, it refers to the member variables simply as `hour` and `minute`.

The reason is that member functions are always called on an existing object. The object is implicitly available to the member function:

```
auto time = TimeOfDay(10, 30);
time.info();
```

The `info()` member function is being called on the `time` object above. The members `hour` and `minute` that are referred to within the function definition correspond to the members of the `time` object, specifically `time.hour` and `time.minute`.

The member function call above is almost the equivalent of the following regular function call:

```
time.info();    // member function
info(time);    // regular function (the previous definition)
```

Whenever a member function is called on an object, the members of the object are implicitly accessible by the function:

```
auto morning = TimeOfDay(10, 0);
auto evening = TimeOfDay(22, 0);

morning.info();
write('-');
evening.info();
writeln();
```

When called on `morning`, the `hour` and `minute` that are used inside the member function refer to `morning.hour` and `morning.minute`. Similarly, when called on `evening`, they refer to `evening.hour` and `evening.minute`:

```
10:00-22:00
```

52.1.1 toString() for string representations

We have discussed the limitations of the `info()` function in the previous chapter. There is at least one more inconvenience with it: Although it prints the time in human-readable format, printing the '-' character and terminating the line still needs to be done explicitly by the programmer.

However, it would be more convenient if `TimeOfDay` objects could be used as easy as fundamental types as in the following code:

```
writefln("%s-%s", morning, evening);
```

In addition to reducing four lines of code to one, it would also allow printing objects to any stream:

```
auto file = File("time_information", "w");
file.writefln("%s-%s", morning, evening);
```

The `toString()` member function of user-defined types are treated specially: It is called automatically to produce the **string** representations of objects. `toString()` must return the **string** representation of the object.

Without getting into more detail, let's first see how the `toString()` function is defined:

```
import std.stdio;

struct TimeOfDay
{
    int hour;
    int minute;

    string toString()
    {
        return "todo";
    }
}

void main()
{
    auto morning = TimeOfDay(10, 0);
    auto evening = TimeOfDay(22, 0);

    writefln("%s-%s", morning, evening);
}
```

`toString()` does not produce anything meaningful yet, but the output shows that it has been called by `writefln()` twice for the two object:

```
todo-todo
```

Also note that `info()` is not needed anymore. `toString()` is replacing its functionality.

The simplest implementation of `toString()` would be to call `format()` of the `std.string` module. `format()` works in the same way as the formatted output functions like `writeln()`. The only difference is that instead of printing variables, it returns the formatted result in **string** format.

`toString()` can simply return the result of `format()` directly:

```
import std.string;
// ...
struct TimeOfDay
{
    // ...
    string toString()
    {
        return format("%02s:%02s", hour, minute);
    }
}
```

Note that `toString()` returns the representation of only *this* object. The rest of the output is handled by `writeln()`: It calls the `toString()` member function for the two objects separately, prints the '-' character in between, and finally terminates the line:

```
10:00-22:00
```

52.1.2 Example: `increment()` member function

Let's define a member function that adds a duration to `TimeOfDay` objects.

Before doing that, let's first correct a design flaw that we have been living with. We have seen in the [Structs chapter](#) that adding two `TimeOfDay` objects in `addDuration()` is not a meaningful operation:

```
TimeOfDay addDuration(in TimeOfDay start,
                    in TimeOfDay duration)    // meaningless
{
    // ...
}
```

What is natural to add to a point in time is *duration*. For example, adding the travel duration to the departure time would result in the arrival time.

On the other hand, subtracting two points in time is a natural operation, in which case the result would be a *duration*.

The following program defines a `Duration` struct with minute-precision, and an `addDuration()` function that uses it:

```
struct Duration
{
    int minute;
}

TimeOfDay addDuration(in TimeOfDay start,
                    in Duration duration)
{
    // Begin with a copy of start
    TimeOfDay result = start;

    // Add the duration to it
    result.minute += duration.minute;

    // Take care of overflows
    result.hour += result.minute / 60;
    result.minute %= 60;
    result.hour %= 24;

    return result;
}

unittest
{
    // A trivial test
    assert(addDuration(TimeOfDay(10, 30), Duration(10))
           == TimeOfDay(10, 40));

    // A time at midnight
    assert(addDuration(TimeOfDay(23, 9), Duration(51))
           == TimeOfDay(0, 0));

    // A time in the next day
    assert(addDuration(TimeOfDay(17, 45), Duration(8 * 60))
```

```

    == TimeOfDay(1, 45));
}

```

Let's redefine a similar function this time as a member function. `addDuration()` has been producing a new object as its result. Let's define an `increment()` member function that will directly modify *this* object instead:

```

struct Duration
{
    int minute;
}

struct TimeOfDay
{
    int hour;
    int minute;

    string toString()
    {
        return format("%02s:%02s", hour, minute);
    }

    void increment(in Duration duration)
    {
        minute += duration.minute;

        hour += minute / 60;
        minute %= 60;
        hour %= 24;
    }

    unittest
    {
        auto time = TimeOfDay(10, 30);

        // A trivial test
        time.increment(Duration(10));
        assert(time == TimeOfDay(10, 40));

        // 15 hours later must be in the next day
        time.increment(Duration(15 * 60));
        assert(time == TimeOfDay(1, 40));

        // 22 hours 20 minutes later must be midnight
        time.increment(Duration(22 * 60 + 20));
        assert(time == TimeOfDay(0, 0));
    }
}

```

`increment()` increments the value of the object by the specified amount of duration. In a later chapter we will see how the *operator overloading* feature of D will make it possible to add a duration by the `+=` operator syntax:

```

time += Duration(10);    // to be explained in a later chapter

```

Also note that `unittest` blocks can be written inside `struct` definitions as well, mostly for testing member functions. It is still possible to move such `unittest` blocks outside of the body of the struct:

```

struct TimeOfDay
{
    // ... struct definition ...
}

unittest

```

```
{
    // ... struct tests ...
}
```

52.2 Exercises

1. Add a **decrement()** member function to **TimeOfDay**, which should reduce the time by the specified amount of duration. Similar to **increment()**, it should *underflow* to the previous day when there is not enough time in the current day. For example, subtracting 10 minutes from 00:05 should result in 23:55.

In other words, implement **decrement()** to pass the following unit tests:

```
struct TimeOfDay
{
    // ...

    void decrement(in Duration duration)
    {
        // ... please implement this function ...
    }

    unittest
    {
        auto time = TimeOfDay(10, 30);

        // A trivial test
        time.decrement(Duration(12));
        assert(time == TimeOfDay(10, 18));

        // 3 days and 11 hours earlier
        time.decrement(Duration(3 * 24 * 60 + 11 * 60));
        assert(time == TimeOfDay(23, 18));

        // 23 hours and 18 minutes earlier must be midnight
        time.decrement(Duration(23 * 60 + 18));
        assert(time == TimeOfDay(0, 0));

        // 1 minute earlier
        time.decrement(Duration(1));
        assert(time == TimeOfDay(23, 59));
    }
}
```

2. Convert **Meeting**, **Meal**, and **DailyPlan** overloads of **info()** to **toString()** member functions as well. (See [the exercise solutions of the Function Overloading chapter](#) for their **info()** overloads.)

You will notice that in addition to making their respective structs more convenient, the implementations of the **toString()** member functions will all consist of single lines.

[... the solutions](#)

53 const ref Parameters and const Member Functions

This chapter is about how parameters and member functions are marked as **const** so that they can be used with **immutable** variables as well. As we have already covered **const** parameters in earlier chapters, some information in this chapter will be a review of some of the features that you already know.

Although the examples in this chapter use only structs, **const** member functions apply to classes as well.

53.1 immutable objects

We have already seen that it is not possible to modify **immutable** variables:

```
immutable readingTime = TimeOfDay(15, 0);
```

readingTime cannot be modified:

```
readingTime = TimeOfDay(16, 0);    // ← compilation ERROR
readingTime.minute += 10;         // ← compilation ERROR
```

The compiler does not allow modifying **immutable** objects in any way.

53.2 ref parameters that are not const

We have seen this concept earlier in the [Function Parameters chapter](#). Parameters that are marked as **ref** can freely be modified by the function. For that reason, even if the function does not actually modify the parameter, the compiler does not allow passing **immutable** objects as that parameter:

```
/* Although not being modified by the function, 'duration'
 * is not marked as 'const' */
int totalSeconds(ref Duration duration)
{
    return 60 * duration.minute;
}
// ...
immutable warmUpTime = Duration(3);
totalSeconds(warmUpTime);    // ← compilation ERROR
```

The compiler does not allow passing the **immutable warmUpTime** to **totalSeconds** because that function does not guarantee that the parameter will not be modified.

53.3 const ref parameters

const ref means that the parameter is not modified by the function:

```
int totalSeconds(const ref Duration duration)
{
    return 60 * duration.minute;
}
// ...
immutable warmUpTime = Duration(3);
totalSeconds(warmUpTime);    // ← now compiles
```


Such functions can receive **immutable** objects as parameters because the immutability of the object is enforced by the compiler:

```
int totalSeconds(const ref Duration duration)
{
    duration.minute = 7;    // ← compilation ERROR
    // ...
}
```

53.4 Non-const member functions

As we have seen with the `TimeOfDay.increment` member function, objects can be modified through member functions as well. `increment()` modifies the members of the object that it is called on:

```
struct TimeOfDay
{
    // ...
    void increment(in Duration duration)
    {
        minute += duration.minute;

        hour += minute / 60;
        minute %= 60;
        hour %= 24;
    }
    // ...
}
// ...
auto start = TimeOfDay(5, 30);
start.increment(Duration(30));    // 'start' gets modified
```

53.5 const member functions

Some member functions do not make any modifications to the object that they are called on. An example of such a function is `toString()`:

```
struct TimeOfDay
{
    // ...
    string toString()
    {
        return format("%02s:%02s", hour, minute);
    }
    // ...
}
```

Since the whole purpose of `toString()` is to represent the object in string format anyway, it should not modify the object.

The fact that a member function does not modify the object is declared by the **const** keyword after the parameter list:

```
struct TimeOfDay
{
    // ...
    string toString() const
    {
        return format("%02s:%02s", hour, minute);
    }
}
```

That **const** guarantees that the object itself is not going to be modified by the member function. As a consequence, **toString()** member function is allowed to be called even on **immutable** objects. Otherwise, the struct's **toString()** would not be called:

```
struct TimeOfDay
{
// ...
// Inferior design: Not marked as 'const'
string toString()
{
    return format("%02s:%02s", hour, minute);
}
}
// ...
immutable start = TimeOfDay(5, 30);
writeln(start); // TimeOfDay.toString() is not called!
```

The output is not the expected **05:30**, indicating that a generic function gets called instead of **TimeOfDay.toString()**:

```
immutable(TimeOfDay)(5, 30)
```

Further, calling **toString()** on an **immutable** object explicitly would cause a compilation error:

```
auto s = start.toString(); // ← compilation ERROR
```

Accordingly, the **toString()** functions that we have defined in the previous chapter have all been designed incorrectly; they should have been marked as **const**.

Note: The **const** keyword can be specified before the definition of the function as well:

```
// The same as above
const string toString()
{
    return format("%02s:%02s", hour, minute);
}
```

Since this version may give the incorrect impression that the **const** is a part of the return type, I recommend that you specify it after the parameter list.

53.6 How to use

- To give the guarantee that a parameter is not modified by the function, mark that parameter as **in**, **const**, or **const ref**.
- Mark member functions that do not modify the object as **const**:

```
struct TimeOfDay
{
// ...
string toString() const
{
    return format("%02s:%02s", hour, minute);
}
}
```

This would make the struct (or class) more useful by removing an unnecessary limitation. The examples in the rest of the book will observe this guideline.

54 Constructor and Other Special Functions

Although this chapter focuses only on structs, the topics that are covered here apply mostly to classes as well. The differences will be explained in later chapters.

Four member functions of structs are special because they define the fundamental operations of that type:

- Constructor: `this()`
- Destructor: `~this()`
- Postblit: `this(this)`
- Assignment operator: `opAssign()`

Although these fundamental operations are handled automatically for structs, hence need not be defined by the programmer, they can be overridden to make the `struct` behave in special ways.

54.1 Constructor

The responsibility of the constructor is to prepare an object for use by assigning appropriate values to its members.

We have already used constructors in previous chapters. When the name of a type is used like a function, it is actually the constructor that gets called. We can see this on the right-hand side of the following line:

```
auto busArrival = TimeOfDay(8, 30);
```

Similarly, a *class* object is being constructed on the right hand side of the following line:

```
auto variable = new SomeClass();
```

The arguments that are specified within parentheses correspond to the constructor parameters. For example, the values 8 and 30 above are passed to the `TimeOfDay` constructor as its parameters.

54.1.1 Constructor syntax

Different from other functions, constructors do not have return values. The name of the constructor is always `this`:

```
struct SomeStruct
{
    // ...

    this(/* constructor parameters */)
    {
        // ... operations that prepare the object for use ...
    }
}
```

The constructor parameters include information that is needed to make a useful and consistent object.

54.1.2 Compiler-generated automatic constructor

All of the structs that we have seen so far have been taking advantage of a constructor that has been generated automatically by the compiler. The automatic constructor assigns the parameter values to the members in the order that they are specified.

As you will remember from [the Structs chapter](#), the initial values for the trailing members need not be specified. The members that are not specified get initialized by the `.init` value of their respective types. The `.init` values of a member could be provided during the definition of that member after the `=` operator:

```
struct Test
{
    int member = 42;
}
```

Also considering the *default parameter values* feature from [the Variable Number of Parameters chapter](#), we can imagine that the automatic constructor for the following `struct` would be the equivalent of the following `this()`:

```
struct Test
{
    char    c;
    int     i;
    double  d;

    /* The equivalent of the compiler-generated automatic
     * constructor (Note: This is only for demonstration; the
     * following constructor would not actually be called
     * when default-constructing the object as Test().) */
    this(in char    c_parameter = char.init,
         in int     i_parameter = int.init,
         in double  d_parameter = double.init)
    {
        c = c_parameter;
        i = i_parameter;
        d = d_parameter;
    }
}
```

For most structs, the compiler-generated constructor is sufficient: Usually, providing appropriate values for each member is all that is needed for objects to be constructed.

54.1.3 Accessing the members by `this`.

To avoid mixing the parameters with the members, the parameter names above had `_parameter` appended to their names. There would be compilation errors without doing that:

```
struct Test
{
    char    c;
    int     i;
    double  d;

    this(in char    c = char.init,
         in int     i = int.init,
         in double  d = double.init)
    {
        // An attempt to assign an 'in' parameter to itself!
        c = c;    // ← compilation ERROR
        i = i;
        d = d;
    }
}
```

The reason is; `c` alone would mean the parameter, not the member, and as the parameters above are defined as `in`, they cannot be modified:

Error: variable `deneme.Test.this.c` cannot modify `const`

A solution is to prepend the member names with `this..` Inside member functions, `this` means "this object", making `this.c` mean "the `c` member of this object":

```

this(in char  c = char.init,
     in int   i = int.init,
     in double d = double.init)
{
    this.c = c;
    this.i = i;
    this.d = d;
}

```

Now `c` alone means the parameter and `this.c` means the member, and the code compiles and works as expected: The member `c` gets initialized by the value of the parameter `c`.

54.1.4 User-defined constructors

I have described the behavior of the compiler-generated constructor. Since that constructor is suitable for most cases, there is no need to define a constructor by hand.

Still, there are cases where constructing an object involves more complicated operations than assigning values to each member in order. As an example, let's consider **Duration** from the earlier chapters:

```

struct Duration
{
    int minute;
}

```

The compiler-generated constructor is sufficient for this single-member struct:

```
time.decrement(Duration(12));
```

Since that constructor takes the duration in minutes, the programmers would sometimes need to make calculations:

```

// 23 hours and 18 minutes earlier
time.decrement(Duration(23 * 60 + 18));

// 22 hours and 20 minutes later
time.increment(Duration(22 * 60 + 20));

```

To eliminate the need for these calculations, we can design a **Duration** constructor that takes two parameters and makes the calculation automatically:

```

struct Duration
{
    int minute;

    this(int hour, int minute)
    {
        this.minute = hour * 60 + minute;
    }
}

```

Since hour and minute are now separate parameters, the users simply provide their values without needing to make the calculation themselves:

```
// 23 hours and 18 minutes earlier
time.decrement(Duration(23, 18));

// 22 hours and 20 minutes later
time.increment(Duration(22, 20));
```

54.1.5 User-defined constructor disables compiler-generated constructor

A constructor that is defined by the programmer makes some uses of the compiler-generated constructor invalid: Objects cannot be constructed by *default parameter values* anymore. For example, trying to construct **Duration** by a single parameter is a compilation error:

```
time.decrement(Duration(12)); // ← compilation ERROR
```

Calling the constructor with a single parameter does not match the programmer's constructor and the compiler-generated constructor is disabled.

One solution is to *overload* the constructor by providing another constructor that takes just one parameter:

```
struct Duration
{
    int minute;

    this(int hour, int minute)
    {
        this.minute = hour * 60 + minute;
    }

    this(int minute)
    {
        this.minute = minute;
    }
}
```

A user-defined constructor disables constructing objects by the `{ }` syntax as well:

```
Duration duration = { 5 }; // ← compilation ERROR
```

Initializing without providing any parameter is still valid:

```
auto d = Duration(); // compiles
```

The reason is, in D, the `.init` value of every type must be known at compile time. The value of `d` above is equal to the initial value of **Duration**:

```
assert(d == Duration.init);
```

54.1.6 static opCall instead of the default constructor

Because the initial value of every type must be known at compile time, it is impossible to define the default constructor explicitly.

Let's consider the following constructor that tries to print some information every time an object of that type is constructed:

```

struct Test
{
    this()    // ← compilation ERROR
    {
        writeln("A Test object is being constructed.");
    }
}

```

The compiler output:

```

Error: constructor deneme.Deneme.this default constructor for
structs only allowed with @disable and no body

```

Note: We will see in later chapters that it is possible to define the default constructor for classes.

As a workaround, a no-parameter **static opCall()** can be used for constructing objects without providing any parameters. Note that this has no effect on the **.init** value of the type.

For this to work, **static opCall()** must construct and return an object of that struct type:

```

import std.stdio;

struct Test
{
    static Test opCall()
    {
        writeln("A Test object is being constructed.");
        Test test;
        return test;
    }
}

void main()
{
    auto test = Test();
}

```

The **Test()** call in **main()** executes **static opCall()**:

```

A Test object is being constructed.

```

Note that it is not possible to type **Test()** inside **static opCall()**. That syntax would execute **static opCall()** as well and cause an infinite recursion:

```

static Test opCall()
{
    writeln("A Test object is being constructed.");
    return Test();    // ← Calls 'static opCall()' again
}

```

The output:

```

A Test object is being constructed.
A Test object is being constructed.
A Test object is being constructed.
... ← repeats the same message

```

54.1.7 Calling other constructors

Constructors can call other constructors to avoid code duplication. Although **Duration** is too simple to demonstrate how useful this feature is, the following single-parameter constructor takes advantage of the two-parameter constructor:

```

this(int hour, int minute)
{
    this.minute = hour * 60 + minute;
}

this(int minute)
{
    this(0, minute);    // calls the other constructor
}

```

The constructor that only takes the minute value calls the other constructor by passing 0 as the value of hour.

Warning: *There is a design flaw in the **Duration** constructors above because the intention is not clear when the objects are constructed by a single parameter:*

```

// 10 hours or 10 minutes?
auto travelDuration = Duration(10);

```

Although it is possible to determine by reading the documentation or the code of the struct that the parameter actually means "10 minutes," it is an inconsistency as the first parameter of the two-parameter constructor is *hours*.

Such design mistakes are causes of bugs and must be avoided.

54.1.8 Immutability of constructor parameters

In the [Immutability chapter](#) we have seen that it is not easy to decide whether parameters of reference types should be defined as **const** or **immutable**. Although the same considerations apply for constructor parameters as well, **immutable** is usually a better choice for constructor parameters.

The reason is, it is common to assign the parameters to members to be used at a later time. When a parameter is not **immutable**, there is no guarantee that the original variable will not change by the time the member gets used.

Let's consider a constructor that takes a file name as a parameter. The file name will be used later on when writing student grades. According to the guidelines in the [Immutability chapter](#), to be more useful, let's assume that the constructor parameter is defined as **const char[]**:

```

import std.stdio;

struct Student
{
    const char[] fileName;
    int[] grades;

    this(const char[] fileName)
    {
        this.fileName = fileName;
    }

    void save()
    {
        auto file = File(fileName.idup, "w");
    }
}

```



```

        file.writeln("The grades of the student:");
        file.writeln(grades);
    }
    // ...
}

void main()
{
    char[] fileName;
    fileName ~= "student_grades";

    auto student = Student(fileName);

    // ...

    /* Assume the fileName variable is modified later on
     * perhaps unintentionally (all of its characters are
     * being set to 'A' here): */
    fileName[] = 'A';

    // ...

    /* The grades would be written to the wrong file: */
    student.save();
}

```

The program above saves the grades of the student under a file name that consists of A characters, not to "student_grades". For that reason, sometimes it is more suitable to define constructor parameters and members of reference types as **immutable**. We know that this is easy for strings by using aliases like **string**. The following code shows the parts of the struct that would need to be modified:

```

struct Student
{
    string fileName;
    // ...
    this(string fileName)
    {
        // ...
    }
    // ...
}

```

Now the users of the struct must provide **immutable** strings and as a result the confusion about the name of the file would be prevented.

54.1.9 Type conversions through single-parameter constructors

Single-parameter constructors can be thought of as providing a sort of type conversion: They produce an object of the particular struct type starting from a constructor parameter. For example, the following constructor produces a **Student** object from a **string**:

```

struct Student
{
    string name;

    this(string name)
    {
        this.name = name;
    }
}

```

to() and **cast** observe this behavior as a *conversion* as well. To see examples of this, let's consider the following **salute()** function. Sending a **string** parameter when it expects a **Student** would naturally cause a compilation error:

```
void salute(Student student)
{
    writeln("Hello ", student.name);
}
// ...
salute("Jane");    // ← compilation ERROR
```

On the other hand, all of the following lines ensure that a **Student** object is constructed before calling the function:

```
import std.conv;
// ...
salute(Student("Jane"));
salute(to!Student("Jean"));
salute(cast(Student)"Jim");
```

to and **cast** take advantage of the single-parameter constructor by constructing a temporary **Student** object and calling **salute()** with that object.

54.2 Destructor

The destructor includes the operations that must be executed when the lifetime of an object ends.

The compiler-generated automatic destructor executes the destructors of all of the members in order. For that reason, as it is with the constructor, there is no need to define a destructor for most structs.

However, sometimes some special operations may need to be executed when an object's lifetime ends. For example, an operating system resource that the object owns may need to be returned to the system; a member function of another object may need to be called; a server running somewhere on the network may need to be notified that a connection to it is about to be terminated; etc.

The name of the destructor is **~this** and just like constructors, it has no return type.

54.2.1 Destructor is executed automatically

The destructor is executed as soon as the lifetime of the struct ends. As you would remember from the [Lifetimes and Fundamental Operations chapter](#), the lifetime of an object ends when leaving the scope that it is defined in.

The following are times when the lifetime of a struct ends:

- When leaving the scope of the object either normally or due to a thrown exception:

```
if (aCondition) {
    auto duration = Duration(7);
    // ...
} // ← The destructor is executed for 'duration' at this point
```

- Anonymous objects are destroyed at the end of the whole expression that they are constructed in:

```
time.increment(Duration(5)); // ← The Duration(5) object
                             // gets destroyed at the end
                             // of the whole expression.
```

- All of the struct members of a struct object get destroyed when the outer object is destroyed.

54.2.2 Destructor example

Let's design a type for generating simple XML documents. XML elements are defined by angle brackets. They contain data and other XML elements. XML elements can have attributes as well; we will ignore them here.

Our aim will be to ensure that an element that has been *opened* by a `<name>` tag will always be *closed* by a matching `</name>` tag:

```
<class1> ← opening the outer XML element
  <grade> ← opening the inner XML element
    57 ← the data
  </grade> ← closing the inner element
</class1> ← closing the outer element
```

A struct that can produce the output above can be designed by two members that store the tag for the XML element and the indentation to use when printing it:

```
struct XmlElement
{
  string name;
  string indentation;
}
```

If the responsibilities of opening and closing the XML element are given to the constructor and the destructor, respectively, the desired output can be produced by managing the lifetimes of `XmlElement` objects. For example, the constructor can print `<tag>` and the destructor can print `</tag>`.

The following definition of the constructor produces the opening tag:

```
this(in string name, in int level)
{
  this.name = name;
  this.indentation = indentationString(level);

  writeln(indentation, '<', name, '>');
}
```

`indentationString()` is the following function:

```
import std.array;
// ...
string indentationString(in int level)
{
  return replicate(" ", level * 2);
}
```

The function calls `replicate()` from the `std.array` module, which makes and returns a new string made up of the specified value repeated the specified number of times.

The destructor can be defined similar to the constructor to produce the closing tag:

```

~this()
{
    writeln(indentation, "</", name, '>');
}

```

Here is a test code to demonstrate the effects of the automatic constructor and destructor calls:

```

void main()
{
    immutable classes = XmlElement("classes", 0);

    foreach (classId; 0 .. 2) {
        immutable classTag = "class" ~ to!string(classId);
        immutable classElement = XmlElement(classTag, 1);

        foreach (i; 0 .. 3) {
            immutable gradeElement = XmlElement("grade", 2);
            immutable randomGrade = uniform(50, 101);

            writeln(indentationString(3), randomGrade);
        }
    }
}

```

Note that the **XmlElement** objects are created in three separate scopes in the program above. The opening and closing tags of the XML elements in the output are produced solely by the constructor and the destructor of **XmlElement**.

I have indicated the matching opening and closing tags of the outer, the middle, and the inner scopes by different colors:

```

<classes>
  <class0>
    <grade>
      72
    </grade>
    <grade>
      97
    </grade>
    <grade>
      90
    </grade>
  </class0>
  <class1>
    <grade>
      77
    </grade>
    <grade>
      87
    </grade>
    <grade>
      56
    </grade>
  </class1>
</classes>

```

The `<classes>` element is produced by the `classesElement` variable. Because that variable is constructed first in `main()`, the output contains the output of its construction first. Since it is also the variable that is destroyed last, upon leaving `main()`, the output contains the output of the destructor call for its destruction last.

54.3 Postblit

Copying is constructing a new object from an existing one. Copying involves two steps:

1. Copying the members of the existing object to the new object. This step is called *blit*, short for *block transfer*.
2. Making further adjustments to the new object. This step is called *postblit*.

The first step is handled automatically by the compiler: It copies the members of the existing object to the members of the new object:

```
auto returnTripDuration = tripDuration; // copying
```

Do not confuse copying with *assignment*. The `auto` keyword above is an indication that a new object is being defined. The actual type name could have been spelled out instead of `auto`.

For an operation to be assignment, the object on the left-hand side must be an existing object. For example, assuming that `returnTripDuration` has already been defined:

```
returnTripDuration = tripDuration; // assignment (see below)
```

Sometimes it is necessary to make adjustments to the members of the new object after the automatic blit. These operations are defined in the `postblit` function of the struct.

Since it is about object construction, the name of the `postblit` is `this` as well. To separate it from the other constructors, its parameter list contains the keyword `this`:

```
this(this)
{
    // ...
}
```

We have defined a `Student` type in the [Structs chapter](#), which had a problem about copying objects of that type:

```
struct Student
{
    int number;
    int[] grades;
}
```

Being a slice, the `grades` member of that `struct` is a reference type. The consequence of copying a `Student` object is that the `grades` members of both the original and the copy provide access to the same actual array elements of type `int`. As a result, the effect of modifying a grade through one of these objects is seen through the other object as well:

```
auto student1 = Student(1, [ 70, 90, 85 ]);

auto student2 = student1; // copying
student2.number = 2;

student1.grades[0] += 5; // this changes the grade of the
```

```

assert(student2.grades[0] == 75); // second student as well:

```

To avoid such a confusion, the elements of the **grades** member of the second object must be separate and belong only to that object. Such *adjustments* are done in the postblit:

```

struct Student
{
    int number;
    int[] grades;

    this(this)
    {
        grades = grades.dup;
    }
}

```

Remember that all of the members have already been copied automatically before **this(this)** started executing. The single line in the postblit above makes a copy of the *elements* of the original array and assigns a slice of it back to **grades**. As a result, the new object gets its own copy of the grades.

Making modifications through the first object does not affect the second object anymore:

```

student1.grades[0] += 5;
assert(student2.grades[0] == 70);

```

54.4 Assignment operator

Assignment is giving a new value to an existing object:

```

returnTripDuration = tripDuration; // assignment

```

Assignment is the most involved operation among the special functions because it is a combination of two operations that we have seen above:

- Destroying the left-hand side object
- Constructing the left-hand side object from the right-hand side object

However, applying the two steps above in that order is risky because the original object would be destroyed before ensuring that the copying would succeed. An exception that is thrown during the copying would leave the left-hand side object in an inconsistent (fully-destroyed or half-destroyed) state.

For that reason, the compiler-generated assignment operator must act safely by applying the following steps:

1. Copy the left-hand side object to a temporary object
2. Copy the right-hand side object to the left-hand side object
3. Destroy the temporary object

Although the compiler-generated assignment operator is suitable in most cases, it can be defined by the programmer. When you do that, consider potential exceptions and write the assignment operator in a way that works even at the presence of thrown exceptions.

The syntax of the assignment operator is the following:

- The name of the function is **opAssign**.

- The type of the parameter is the same as the **struct** type. This parameter is often named as **rhs**, short for *right-hand side*.
- The return type is the same as the **struct** type.
- The function is exited by **return this**.

As an example, let's consider a simple **Duration** struct where the assignment operator prints a message:

```
struct Duration
{
    int minute;

    Duration opAssign(Duration rhs)
    {
        writefln("minute is being changed from %s to %s",
                this.minute, rhs.minute);

        this.minute = rhs.minute;

        return this;
    }
}
// ...
auto duration = Duration(100);
duration = Duration(200);           // assignment
```

The output:

```
minute is being changed from 100 to 200
```

54.4.1 Assigning from other types

Sometimes it is convenient to assign values of types that are different from the type of the **struct**. For example, instead of requiring a **Duration** object on the right-hand side, it may be useful to assign from an integer:

```
duration = 300;
```

This is possible by defining another assignment operator that takes an **int** parameter:

```
struct Duration
{
    int minute;

    Duration opAssign(Duration rhs)
    {
        writefln("minute is being changed from %s to %s",
                this.minute, rhs.minute);

        this.minute = rhs.minute;

        return this;
    }

    Duration opAssign(int minute)
    {
        writefln("minute is being replaced by an int");

        this.minute = minute;

        return this;
    }
}
// ...
```

```
duration = Duration(200);  
duration = 300;
```

The output:

```
minute is being changed from 100 to 200  
minute is being replaced by an int
```

Note: Although convenient, assigning different types to each other may cause confusions and errors.

54.5 Summary

- Constructor (**this**) is for preparing objects for use. The compiler-generated default constructor is sufficient in most cases.
- The behavior of the default constructor may not be changed in structs; **static opCall** can be used instead.
- Single-parameter constructors can be used during type conversions by **to** and **cast**.
- Destructor (**~this**) is for the operations that must be executed when the lifetimes of objects end.
- Postblit (**this(this)**) is for adjustments to the object after the automatic member copies.
- Assignment operator (**opAssign**) is for changing values of existing objects.

55 Operator Overloading

The topics covered in this chapter apply mostly for classes as well. The biggest difference is that the behavior of assignment operation `opAssign()` cannot be overloaded for classes.

Operator overloading enables defining how user-defined types behave when used with operators. In this context, the term *overload* means providing the definition of an operator for a specific type.

We have seen how to define structs and their member functions in previous chapters. As an example, we have defined the `increment()` member function to be able to add `Duration` objects to `TimeOfDay` objects. Here are the two structs from previous chapters, with only the parts that are relevant to this chapter:

```
struct Duration
{
    int minute;
}

struct TimeOfDay
{
    int hour;
    int minute;

    void increment(in Duration duration)
    {
        minute += duration.minute;

        hour += minute / 60;
        minute %= 60;
        hour %= 24;
    }
}

void main()
{
    auto lunchTime = TimeOfDay(12, 0);
    lunchTime.increment(Duration(10));
}
```

A benefit of member functions is being able to define operations of a type alongside the member variables of that type.

Despite their advantages, member functions can be seen as being limited compared to operations on fundamental types. After all, fundamental types can readily be used with operators:

```
int weight = 50;
weight += 10; // by an operator
```

According to what we have seen so far, similar operations can only be achieved by member functions for user-defined types:

```
auto lunchTime = TimeOfDay(12, 0);
lunchTime.increment(Duration(10)); // by a member function
```

Operator overloading enables using structs and classes with operators as well. For example, assuming that the `+=` operator is defined for `TimeOfDay`, the operation above can be written in exactly the same way as with fundamental types:

```
lunchTime += Duration(10); // by an operator even for the struct
```

Before getting to the details of operator overloading, let's first see how the line above would be enabled for `TimeOfDay`. What is needed is to redefine the `increment()` member function under the special name `opOpAssign(string op)` and also to specify that this definition is for the `+` character. As it will be explained below, this definition actually corresponds to the `+=` operator.

The definition of this member function does not look like the ones that we have seen so far. That is because `opOpAssign` is actually a *function template*. Since we will see templates in much later chapters, I will have to ask you to accept the operator overloading syntax as-is for now:

```
struct TimeOfDay
{
// ...
  ref TimeOfDay opOpAssign(string op)(in Duration duration) // (1)
  if (op == "+") // (2)
  {
    minute += duration.minute;

    hour += minute / 60;
    minute %= 60;
    hour %= 24;

    return this;
  }
}
```

The template definition consists of two parts:

1. `opOpAssign(string op)`: This part must be written as-is and should be accepted as the *name* of the function. We will see below that there are other member functions in addition to `opOpAssign`.
2. `if (op == "+")`: `opOpAssign` is used for more than one operator overload. `+` specifies that this is the operator overload that corresponds to the `+` character. This syntax is a *template constraint*, which will also be covered in later chapters.

Also note that this time the return type is different from the return type of the `increment()` member function: It is not `void` anymore. We will discuss the return types of operators later below.

Behind the scenes, the compiler replaces the uses of the `+=` operator with calls to the `opOpAssign! "+"` member function:

```
lunchTime += Duration(10);

// The following line is the equivalent of the previous one
lunchTime.opOpAssign! "+"(Duration(10));
```

The `! "+"` part that is after `opOpAssign` specifies that this call is for the definition of the operator for the `+` character. We will cover this template syntax in later chapters as well.

Note that the operator definition that corresponds to `+=` is defined by `+`, not by `+=`. The `Assign` in the name of `opOpAssign()` already implies that this name is for an assignment operator.

55.1 Do not change conventional behaviors

Being able to define the behaviors of operators brings a responsibility: The programmer must observe expectations. As an extreme example, the previous operator could have been defined

to decrement the time value instead of incrementing it. However, people who read the code would still expect the value to be incremented by the `+=` operator.

To some extent, the return types of operators can also be chosen freely. Still, general expectations must be observed for the return types as well.

Keep in mind that operators that behave unnaturally would cause confusions and bugs.

55.2 Overloadable unary operators

An operator that takes a single operand is called a unary operator:

```
++weight;
```

`++` is a unary operator because it works on a single variable.

Unary operators are defined by member functions named `opUnary`. `opUnary` does not take any parameters because it uses only the object that the operator is being executed on.

The overloadable unary operators and the corresponding operator strings are the following:

Operator	Description	Operator String
-object	negative value of	" - "
+object	the same value as (or a copy of)	" + "
~object	bitwise negation	" ~ "
*object	access to what it points to	" * "
++object	increment	" ++ "
--object	decrement	" -- "

For example, the `++` operator for `Duration` can be defined like this:

```
struct Duration
{
    int minute;

    ref Duration opUnary(string op)()
    {
        if (op == "++")
        {
            ++minute;
        }
    }
}
```

Note that the return type of the operator is marked as `ref` here as well. This will be explained later below.

`Duration` objects can now be incremented by `++`:

```
auto duration = Duration(20);
++duration;
```

The post-increment and post-decrement operators cannot be overloaded. The `object++` and `object--` uses are handled by the compiler automatically by saving the previous value of the object. For example, the compiler applies the equivalent of the following code for post-increment:

```
/* The previous value is copied by the compiler
 * automatically: */
Duration __previousValue__ = duration;
```

```
/* The ++ operator is called: */  
++duration;  
  
/* Then __previousValue__ is used as the value of the  
* post-increment operation. */
```

55.3 Overloadable binary operators

An operator that takes two operands is called a binary operator:

```
totalWeight = boxWeight + chocolateWeight;
```

The line above has two separate binary operators: the + operator, which adds the values of the two operands that are on its two sides, and the = operator that assigns the value of its right-hand operand to its left-hand operand.

The rightmost column below describes the category of each operator. The ones marked as "=" assign to the left-hand side object.

Operator	Description	Function name	Function name for right-hand side	Category
+	add	opBinary	opBinaryRight	arithmetic
-	subtract	opBinary	opBinaryRight	arithmetic
*	multiply	opBinary	opBinaryRight	arithmetic
/	divide	opBinary	opBinaryRight	arithmetic
%	remainder of	opBinary	opBinaryRight	arithmetic
^^	to the power of	opBinary	opBinaryRight	arithmetic
&	bitwise <i>and</i>	opBinary	opBinaryRight	bitwise
	bitwise <i>or</i>	opBinary	opBinaryRight	bitwise
^	bitwise <i>xor</i>	opBinary	opBinaryRight	bitwise
<<	left-shift	opBinary	opBinaryRight	bitwise
>>	right-shift	opBinary	opBinaryRight	bitwise
>>>	logical right-shift	opBinary	opBinaryRight	bitwise
~	concatenate	opBinary	opBinaryRight	
in	whether contained in	opBinary	opBinaryRight	
==	whether equal to	opEquals	-	logical
!=	whether not equal to	opEquals	-	logical
<	whether before	opCmp	-	sorting
<=	whether not after	opCmp	-	sorting
>	whether after	opCmp	-	sorting
>=	whether not before	opCmp	-	sorting
=	assign	opAssign	-	=
+=	increment	opOpAssign	-	=
-=	decrement	opOpAssign	-	=
*=	multiply and assign	opOpAssign	-	=
/=	divide and assign	opOpAssign	-	=
%=	assign the remainder of	opOpAssign	-	=
^^=	assign the power of	opOpAssign	-	=
&=	assign the result of &	opOpAssign	-	=
=	assign the result of	opOpAssign	-	=
^=	assign the result of ^	opOpAssign	-	=
<<=	assign the result of <<	opOpAssign	-	=
>>=	assign the result of >>	opOpAssign	-	=
>>>=	assign the result of >>>	opOpAssign	-	=
~=	append	opOpAssign	-	=

opBinaryRight is for when the object can appear on the right-hand side of the operator. Let's assume a binary operator that we shall call *op* appears in the program:

```
x op y
```

In order to determine what member function to call, the compiler considers the following two options:

```
x.opBinary!"op"(y); // the definition for x being on the left
y.opBinaryRight!"op"(x); // the definition for y being on the right
```

The compiler picks the option that is a better match than the other.

In most cases it is not necessary to define `opBinaryRight`, except for the `in` operator: It usually makes more sense to define `opBinaryRight` for `in`.

The parameter name `rhs` that appears in the following definitions is short for *right-hand side*. It denotes the operand that appears on the right-hand side of the operator:

```
x op y
```

For the expression above, the `rhs` parameter would represent the variable `y`.

55.4 Other operators that can be overloaded

In addition to the unary and binary operators above, the following operators can also be overloaded:

Description	Function Name	Sample Usage
function call	<code>opCall</code>	<code>object(42)</code>
element access	<code>opIndex</code>	<code>collection[i]</code>
assignment to element	<code>opIndexAssign</code>	<code>collection[i] = 7</code>
unary operation on element	<code>opIndexUnary</code>	<code>++collection[i]</code>
operation with assignment on element	<code>opIndexOpAssign</code>	<code>collection[i] *= 2</code>
slice to all elements	<code>opSlice</code>	<code>collection[]</code>
slice to some elements	<code>opSlice</code>	<code>collection[i, j]</code>
unary operation on all elements	<code>opSliceUnary</code>	<code>++collection[]</code>
unary operation on some elements	<code>opSliceUnary</code>	<code>++collection[i, j]</code>
assignment to all elements	<code>opSliceAssign</code>	<code>collection[] = 42</code>
assignment to some elements	<code>opSliceAssign</code>	<code>collection[i, j] = 7</code>
operation with assignment on all elements	<code>opSliceOpAssign</code>	<code>collection[] *= 2</code>
operation with assignment on some elements	<code>opSliceOpAssign</code>	<code>collection[i, j] *= 2</code>
dispatch for non-existent function	<code>opDispatch</code>	<code>object.nonExistent()</code>

These operators will be explained below under their own sections.

55.5 Return types of operators

When overloading an operator, it is advisable to observe the return type of that operator on fundamental types. This would help with making sense of code and reducing confusions.

None of the operators on fundamental types return `void`. This fact should be obvious for some operators. For example, the result of adding two `int` values as `a + b` is `int`:

```
int a = 1;
int b = 2;
int c = a + b; // c gets initialized by the return value
               // of the + operator
```

The return values of some other operators may not be so obvious. For example, even operators like `++i` have values:

```
int i = 1;
writeln(++i); // prints 2
```

The `++` operator not only increments `i`, it also produces the new value of `i`. Further, the value that is produced by `++` is not just the value of `i`, rather *the variable `i` itself*. We can demonstrate this fact by assigning to the result of the `++` operator:

```
int i = 1;
++i = 42; // assigning to the result of ++
writeln(i);
```

The output:

```
42
```

As can be seen by the code above, `++i` returns the variable itself and as a result the assignment changes `i`.

I recommend that you observe the following guidelines when overloading operators for your own types:

- **Operators that modify the object**

With the exception of `opAssign`, it is recommended that the operators that modify the object return the object itself. This guideline has been observed above with the `TimeOfDay.opAssign! "+"` and `Duration.opUnary! "++"`.

The following two steps achieve returning the object itself:

1. The return type is the type of the struct, marked by the `ref` keyword to mean *reference*.
2. The function is exited by `return this` to mean *return this object*.

The operators that modify the object are `opUnary! "++"`, `opUnary! "--"`, and all of the `opOpAssign` overloads.

- **Logical operators**

`opEquals` that represents both `==` and `!=` must return `bool`. Although the `in` operator normally returns *the contained object*, it can simply return `bool` as well.

- **Sort operators**

`opCmp` that represents `<`, `<=`, `>`, and `>=` must return `int`.

- **Operators that make a new object**

Some operators must make and return a new object:

- Unary operators `-`, `+`, and `~`; and the binary operator `~`.
- Arithmetic operators `+`, `-`, `*`, `/`, `%`, and `^^`.
- Bitwise operators `&`, `|`, `^`, `<<`, `>>`, and `>>>`.
- As has been seen in the previous chapter, `opAssign` returns a copy of this object by `return this`.

Note: As an optimization, sometimes it makes more sense for `opAssign` to return `const ref` for large structs. I will not apply this optimization in this book.

As an example of an operator that makes a new object, let's define the `opBinary! "+"` overload for `Duration`. This operator should add two `Duration` objects to make and return a new one:

```
struct Duration
{
    int minute;
```

```

Duration opBinary(string op)(in Duration rhs) const
    if (op == "+")
    {
        return Duration(minute + rhs.minute); // new object
    }
}

```

That definition enables adding **Duration** objects by the **+** operator:

```

auto travelDuration = Duration(10);
auto returnDuration = Duration(11);
Duration totalDuration;
// ...
totalDuration = travelDuration + returnDuration;

```

The compiler replaces that expression with the following member function call on the **travelDuration** object:

```

// the equivalent of the expression above
totalDuration = travelDuration.opBinary!"+"(returnDuration);

```

- **Unconstrained operators**

The return types of some of the operators depend entirely on the design of the user-defined type: The unary operator *****, **opCall**, **opCast**, **opDispatch**, **opIndex** and other indexing operators, and **opSlice** and other slicing operators.

55.6 opEquals() for equality comparisons

This member function defines the behaviors of the **==** and the **!=** operators.

The return type of **opEquals** is **bool**.

According to language rules, the parameter of **opEquals** for structs must be defined as **const ref**.

As an important rule, **opEquals** and **opCmp** must work consistently. For two objects that **opEquals** returns **true**, **opCmp** must return zero.

Once **opEquals()** is defined for equality, the compiler uses its opposite for inequality:

```

x == y;
x.opEquals(y); // the equivalent of the previous expression

x != y;
!(x.opEquals(y)); // the equivalent of the previous expression

```

Normally, it is not necessary to define **opEquals()** for structs. The compiler generates it for structs automatically. The automatically-generated **opEquals** compares all of the members individually.

Sometimes the equality of two objects must be defined differently from this automatic behavior. For example, some of the members may not be significant in this comparison, or the equality may depend on a more complex logic.

Just as an example, let's define **opEquals()** in a way that disregards the minute information altogether:

```

struct TimeOfDay
{

```



```

int hour;
int minute;

bool opEquals(const ref TimeOfDay rhs) const
{
    return hour == rhs.hour;
}
// ...
assert(TimeOfDay(20, 10) == TimeOfDay(20, 59));

```

Since the equality comparison considers the values of only the **hour** members, 20:10 and 20:59 end up being equal. (This is just an example; it should be clear that such an equality comparison would cause confusions.)

55.7 opCmp() for sorting

Sort operators determine the sort orders of objects. All of the ordering operators **<**, **<=**, **>**, and **>=** are covered by the **opCmp()** member function.

According to language rules, the parameter of **opCmp** for structs must be defined as **const ref**.

As an important rule, **opEquals** and **opCmp** must work consistently. For two objects that **opEquals** returns **true**, **opCmp** must return zero.

Let's assume that one of these four operators is used as in the following code:

```
if (x op y) { // op is one of <, <=, >, or >=
```

The compiler converts that expression to the following logical expression and uses the result of the new logical expression:

```
if (x.opCmp(y) op 0) {
```

Let's consider the **<=** operator:

```
if (x <= y) {
```

The compiler generates the following code behind the scenes:

```
if (x.opCmp(y) <= 0) {
```

For the user-defined **opCmp()** to work correctly, this member function must return a result according to the following rules:

- A *negative value* if the left-hand object is considered to be before the right-hand object
- A *positive value* if the left-hand object is considered to be after the right-hand object
- *Zero* if the objects are considered to have the same sort order

To be able to support those values, the return type of **opCmp()** must be **int**, not **bool**.

The following is a way of ordering **TimeOfDay** objects by first comparing the values of the **hour** members, and then comparing the values of the **minute** members (only if the **hour** members are equal):

```
int opCmp(const ref TimeOfDay rhs) const
{
```

```

    return (hour == rhs.hour
           ? minute - rhs.minute
           : hour - rhs.hour);
}

```

That definition returns the difference between the **minute** values when the **hour** members are the same, and the difference between the **hour** members otherwise. The return value would be a *negative value* when the *left-hand* object comes before in chronological order, a *positive value* if the *right-hand* object is before, and *zero* when they represent exactly the same time of day.

Once `opCmp()` is defined, the type can be used in algorithms that compare objects of that type. The `.sort` property of arrays is such an algorithm. The following program constructs 10 objects by random values and sorts them by `.sort`. As `.sort` works on the elements, it is the `opCmp()` operator that gets called behind the scenes to determine the sort orders of the elements:

```

import std.random;
import std.stdio;
import std.string;

struct TimeOfDay
{
    int hour;
    int minute;

    int opCmp(const ref TimeOfDay rhs) const
    {
        return (hour == rhs.hour
               ? minute - rhs.minute
               : hour - rhs.hour);
    }

    string toString() const
    {
        return format("%02s:%02s", hour, minute);
    }
}

void main()
{
    TimeOfDay[] times;

    foreach (i; 0 .. 10) {
        times ~= TimeOfDay(uniform(0, 24), uniform(0, 60));
    }

    times.sort;

    writeln(times);
}

```

As expected, the elements are sorted from the earliest time to the latest time:

```
[03:40, 04:10, 09:06, 10:03, 10:09, 11:04, 13:42, 16:40, 18:03, 21:08]
```

55.8 opCall() to call objects as functions

The parentheses around the parameter list when calling functions are operators as well. We have already seen how `static opCall()` makes it possible to use the name of the *type* as a function. `static opCall()` allows creating objects with default values at runtime.

Not-static `opCall()` on the other hand allows using the *objects* of user-defined types as functions:

```
Foo foo;
foo();
```

The object `foo` above is being called like a function.

As an example, let's consider a **struct** that represents a linear equation. This **struct** will be used for calculating the y values of the following linear equation for specific x values:

$$y = ax + b$$

The following `opCall()` simply calculates and returns the value of y according to that equation:

```
struct LinearEquation
{
    double a;
    double b;

    double opCall(double x) const
    {
        return a * x + b;
    }
}
```

With that definition, each object of **LinearEquation** represents a linear equation for specific a and b values. Such an object can be used as a function that calculates the y values:

```
LinearEquation equation = { 1.2, 3.4 };
double y = equation(5.6); // the object is being used like a function
```

Note: Defining `opCall()` for a **struct** disables the compiler-generated automatic constructor. That is why the `{ }` syntax is used above instead of the recommended `LinearEquation(1.2, 3.4)`. When the latter syntax is desired, a **static** `opCall()` that takes two **double** parameters must also be defined.

`equation` above represents the $y = 1.2x + 3.4$ linear equation. Using that object as a function executes the `opCall()` member function.

This feature can be useful to define and store the a and b values in an object once and to use that object multiple times later on. The following code uses such an object in a loop:

```
LinearEquation equation = { 0.01, 0.4 };

for (double x = 0.0; x <= 1.0; x += 0.125) {
    writefln("%f: %f", x, equation(x));
}
```

That object represents the $y = 0.01x + 0.4$ equation. It is being used for calculating the results for x values in the range from 0.0 to 1.0.

55.9 Indexing operators

`opIndex()`, `opIndexAssign()`, `opIndexUnary()`, and `opIndexOpAssign()` make it possible to use indexing operators for user-defined types similar to arrays as in `object[index]`. Different from arrays, there can be more than one index in square brackets. The exact

behavior of these operators depend on the user-defined type. We will see a sample program that includes these operators below.

opIndex is for element access. The indexes within brackets become the parameters of the operator function:

```
grade = grades[3, 1];           // access the element at 3,1
grade = grades.opIndex(3, 1);  // the equivalent of the above
```

opIndexUnary is similar to **opUnary**. Its difference is that the operation is applied to the *element* that corresponds to the specified index(es):

```
++grades[4, 0];                // increment the element at 4,0
grades.opIndexUnary!("++")(4, 0); // the equivalent of the above
```

opIndexAssign is for assignment. Its first parameter is the value that is to be assigned, and the rest of the parameters are the index(es):

```
grades[1, 1] = 95;              // assign 95 to element 1,1
grades.opIndexAssign(95, 1, 1); // the equivalent of the above
```

opIndexOpAssign is similar to **opOpAssign**. Its difference is that the result of the operation is assigned back to the specified element:

```
grades[2, 1] += 42;             // add to the element at 2,1
grades.opIndexOpAssign!("+")(42, 2, 1); // the equivalent of the above
```

55.9.1 Indexing operators example

The following struct includes examples of each of these operators. In this design, the first index specifies the student and the second index specifies the grade of that student. For example, `[1, 2]` specifies the grade id 2 of student id 1.

```
import std.stdio;

struct StudentGrades
{
    // Two grades for five students
    int[2][5] grades;

    // Returns the specified grade
    int opIndex(size_t studentId, size_t gradeId) const
    {
        return grades[studentId][gradeId];
    }

    // Increments the specified grade
    ref int opIndexUnary(string op)(size_t studentId,
                                    size_t gradeId)
    {
        if (op == "++")
        {
            return ++grades[studentId][gradeId];
        }
    }

    // Assigns to the specified grade of the specified student
    int opIndexAssign(int grade,
                      size_t studentId,
                      size_t gradeId)
    {
        return grades[studentId][gradeId] = grade;
    }
}
```

```

// Adds to the specified grade
ref int opIndexOpAssign(string op)(int amount,
                                   size_t studentId,
                                   size_t gradeId)
{
    if (op == "+")
    {
        return grades[studentId][gradeId] += amount;
    }
}

void main()
{
    StudentGrades grades;

    writeln("Grade for index 3,1: ",
            grades[3, 1]);           // access
    ++grades[4, 0];                 // increment
    grades[1, 1] = 95;              // assignment
    grades[2, 1] += 42;             // add-assignment

    writeln(grades);
}

```

The output:

```

Grade for index 3,1: 0
StudentGrades([[0, 0], [0, 95], [0, 42], [0, 0], [1, 0]])

```

Note that the return types of `opIndexUnary` and `opIndexOpAssign` are marked as `ref` according to the return type guidelines above. Additionally, the operators do not do anything special for the return values; they simply return the results of the `++` and `+=` operators that they use internally.

55.10 Slicing operators

`opSlice`, `opSliceUnary`, `opSliceAssign`, and `opSliceOpAssign` are similar to the indexing operators. Their difference is that they allow *slicing* the objects of user-defined types.

All of these operators have two distinct uses: The use where the square brackets are empty and the use where the square brackets contain a number range. Similar to slices, these should be defined to mean *all of the elements* and *some of the elements*, respectively.

Slicing operators are the most complex operators because they involve the two distinct concepts *container* and *range*. We will see these concepts in more detail in later chapters. These operators will make more sense only after understanding ranges.

For now, I will present an example that consists only of stub types and stub functions. Similar to earlier examples, this example uses only the `+++` and `++` characters. Normally, many other operator characters must also be defined.

```

/* The type of the elements contained in the container. */
struct Element
{
    // ...
}

/* Similar to slices, represents the concept of "element
 * range". */
struct Range
{
    // ...
}

```

```

/* Similar to arrays and associative arrays, brings elements
 * together. */
struct Container
{
    // For the object[] syntax
    Range opSlice()
    {
        Range allElements;
        // ... must provide access to all elements ...
        return allElements;
    }

    // For the object[i .. j] syntax
    Range opSlice(size_t i, size_t j)
    {
        Range specifiedElements;
        // ... must provide access to the specified elements ...
        return specifiedElements;
    }

    // For the ++object[] syntax
    Range opSliceUnary(string op)()
    {
        if (op == "+")
        {
            Range allElements;
            // ... must increment the values of all elements ...
            return allElements;
        }
    }

    // For the ++object[i .. j] syntax
    Range opSliceUnary(string op)(size_t i, size_t j)
    {
        if (op == "+")
        {
            Range specifiedElements;
            // ... must increment the values of the specified elements ...
            return specifiedElements;
        }
    }

    // For the (object[] = value) syntax
    Range opSliceAssign(Element value)
    {
        Range allElements;
        // ... must assign to all elements ...
        return allElements;
    }

    // For the (object[i .. j] = value) syntax
    Range opSliceAssign(Element value, size_t i, size_t j)
    {
        Range specifiedElements;
        // ... must assign to the specified elements ...
        return specifiedElements;
    }

    // For the (object[] += value) syntax
    Range opSliceOpAssign(string op)(Element value)
    {
        if (op == "+")
        {
            Range allElements;
            // ... must add to all elements ...
            return allElements;
        }
    }

    // For the (object[i .. j] += value) syntax
    Range opSliceOpAssign(string op)(Element value,
                                     size_t i, size_t j)
    {
        if (op == "+")
        {
            Range specifiedElements;
            // ... must add to the specified elements ...

```

```

        return specifiedElements;
    }

    // ... other operator overloads as needed ...
}

void main()
{
    Container c;
    Range r;
    Element e;

    /* The following lines are examples of what the previous
     * operator overloads make possible. The equivalents that
     * the compiler uses behind the scenes are shown as code
     * comments. */

    r = c[];           // c.opSlice();
    r = c[1 .. 5];     // c.opSlice(1, 5);

    ++c[];            // c.opSliceUnary!"++"();
    ++c[2 .. 9];      // c.opSliceUnary!"++"(2, 9);

    c[] = e;          // c.opSliceAssign(e);
    c[4 .. 6] = e;    // c.opSliceAssign(e, 4, 6);

    c[] += e;         // c.opSliceOpAssign!"+"(e);
    c[5 .. 8] += e;   // c.opSliceOpAssign!"+"(e, 5, 8);
}

```

55.11 opCast for type conversions

opCast defines the explicit type conversions. It can be overloaded separately for each target type. As you would remember from the earlier chapters, explicit type conversions are performed by the **to** function and the **cast** operator.

opCast is a template as well, but it has a different format: The target type is specified by the **(T : target_type)** syntax:

```

target_type opCast(T : target_type)()
{
    // ...
}

```

This syntax will become clear later after the templates chapter as well.

Let's change the definition of **Duration** so that it now has two members: hours and minutes. The operator that converts objects of this type to **double** can be defined as in the following code:

```

import std.stdio;
import std.conv;

struct Duration
{
    int hour;
    int minute;

    double opCast(T : double)() const
    {
        return hour + (to!double(minute) / 60);
    }
}

void main()

```

```
{
    auto duration = Duration(2, 30);
    double d = to!double(duration); // can be cast(double)duration as well
    writeln(d);
}
```

The compiler replaces the type conversion call above with the following one:

```
double d = duration.opCast!double();
```

The **double** conversion function above produces 2.5 for two hours and thirty minutes:

```
2.5
```

55.12 Catch-all operator opDispatch

opDispatch gets called whenever a *missing* member of an object is accessed. All attempts to access non-existent members are dispatched to this function.

The name of the missing member becomes the template parameter value of **opDispatch**.

The following code demonstrates a simple definition:

```
import std.stdio;

struct Foo
{
    void opDispatch(string name, T)(T parameter)
    {
        writefln("Foo.opDispatch - name: %s, value: %s",
            name, parameter);
    }
}

void main()
{
    Foo foo;
    foo.aNonExistentFunction(42);
    foo.anotherNonExistentFunction(100);
}
```

There are no compiler errors for the calls to non-existent members. Instead, all of those calls are dispatched to **opDispatch**. The first template parameter is the name of the member. The parameter values that are used when calling the function appear as the parameters of **opDispatch**:

```
Foo.opDispatch - name: aNonExistentFunction, value: 42
Foo.opDispatch - name: anotherNonExistentFunction, value: 100
```

The **name** template parameter can be used inside the function to make decisions on how the call to that specific non-existent function should be handled:

```
switch (name) {
    // ...
}
```


55.13 Inclusion query by `opBinaryRight!"in"`

This operator allows defining the behavior of the `in` operator for user-defined types. `in` is commonly used with associative arrays to determine whether a value for a specific key exists in the array.

Different from other operators, this operator is normally overloaded for the case where the object appears on the right-hand side:

```
if (time in lunchBreak) {
```

The compiler would use `opBinaryRight` behind the scenes:

```
// the equivalent of the above:
if (lunchBreak.opBinaryRight!"in"(time)) {
```

55.13.1 Example of the `in` operator

The following program defines a `TimeSpan` type in addition to `Duration` and `TimeOfDay`. The `in` operator that is defined for `TimeSpan` determines whether a moment in time is within that time span.

To keep the code short, the following program defines only the necessary member functions.

Note how the `TimeOfDay` object is used seamlessly in the `for` loop. That loop is a demonstration of how useful operator overloading can be.

```
import std.stdio;
import std.string;

struct Duration
{
    int minute;
}

struct TimeOfDay
{
    int hour;
    int minute;

    ref TimeOfDay opOpAssign(string op)(in Duration duration)
    {
        if (op == "+")
        {
            minute += duration.minute;

            hour += minute / 60;
            minute %= 60;
            hour %= 24;

            return this;
        }
    }

    int opCmp(const ref TimeOfDay rhs) const
    {
        return (hour == rhs.hour
            ? minute - rhs.minute
            : hour - rhs.hour);
    }

    string toString() const
    {
        return format("%02s:%02s", hour, minute);
    }
}
```

```

struct TimeSpan
{
    TimeOfDay begin;
    TimeOfDay end; // end is considered to be outside of the span

    bool opBinaryRight(string op)(TimeOfDay time) const
    {
        if (op == "in")
        {
            return (time >= begin) && (time < end);
        }
    }
}

void main()
{
    auto lunchBreak = TimeSpan(TimeOfDay(12, 00),
                               TimeOfDay(13, 00));

    for (auto time = TimeOfDay(11, 30);
         time < TimeOfDay(13, 30);
         time += Duration(15)) {

        if (time in lunchBreak) {
            writeln(time, " is during the lunch break");
        } else {
            writeln(time, " is outside of the lunch break");
        }
    }
}

```

The output:

```

11:30 is outside of the lunch break
11:45 is outside of the lunch break
12:00 is during the lunch break
12:15 is during the lunch break
12:30 is during the lunch break
12:45 is during the lunch break
13:00 is outside of the lunch break
13:15 is outside of the lunch break

```

55.14 Defining more than one operator at the same time

To keep the code samples short, we have used only the `++`, `+`, and `+=` operators above. It is conceivable that when one operator is overloaded for a type, many others would also need to be overloaded. For example, the `--` and `-=` operators are also defined for the following

Duration:

```

struct Duration
{
    int minute;

    ref Duration opUnary(string op)()
    {
        if (op == "++")
        {
            ++minute;
            return this;
        }

        ref Duration opUnary(string op)()
        {
            if (op == "--")
            {

```

```

        --minute;
        return this;
    }

    ref Duration opOpAssign(string op)(in int amount)
    {
        if (op == "+")
        {
            minute += amount;
            return this;
        }
    }

    ref Duration opOpAssign(string op)(in int amount)
    {
        if (op == "-")
        {
            minute -= amount;
            return this;
        }
    }
}

unittest
{
    auto duration = Duration(10);

    ++duration;
    assert(duration.minute == 11);

    --duration;
    assert(duration.minute == 10);

    duration += 5;
    assert(duration.minute == 15);

    duration -= 3;
    assert(duration.minute == 12);
}

void main()
{}

```

The operator overloads above have code duplications. The only differences between the similar functions are highlighted by yellow. Such code duplications can be reduced and sometimes avoided altogether by *string mixins*. We will see the **mixin** keyword in a later chapter as well. I would like to show briefly how this keyword helps with operator overloading.

mixin inserts the specified string as source code right where the **mixin** statement appears in code. The following struct is the equivalent of the one above:

```

struct Duration
{
    int minute;

    ref Duration opUnary(string op)()
    {
        if ((op == "++") || (op == "--"))
        {
            mixin (op ~ "minute;");
            return this;
        }
    }

    ref Duration opOpAssign(string op)(in int amount)
    {
        if ((op == "+") || (op == "-"))
        {
            mixin ("minute " ~ op ~ "= amount;");
            return this;
        }
    }
}

```

If the **Duration** objects also need to be multiplied and divided by an amount, all that is needed is to add two more conditions to the template constraint:

```

struct Duration
{
// ...

    ref Duration opOpAssign(string op)(in int amount)
    if ((op == "+" || (op == "-" ||
        (op == "*" || (op == "/")))
    {
        mixin ("minute " ~ op ~ "= amount;");
        return this;
    }
}

unittest
{
    auto duration = Duration(12);

    duration *= 4;
    assert(duration.minute == 48);

    duration /= 2;
    assert(duration.minute == 24);
}

```

55.15 Exercise

- Define a fraction type that holds its numerator and denominator as members of type **long**. Such a type may be useful, because it does not lose value like **float**, **double**, and **real** do due to their precisions. For example, although the result of multiplying a **double** value of 1.0/3 by 3 is *not* 1.0, multiplying a **Fraction** object that represents the fraction 1/3 by 3 would be exactly 1:

```

struct Fraction
{
    long num; // numerator
    long den; // denominator

    /* As a convenience, the constructor uses the default
     * value of 1 for the denominator. */
    this(long num, long den = 1)
    {
        enforce(den != 0, "The denominator cannot be zero");

        this.num = num;
        this.den = den;

        /* Ensuring that the denominator is always positive
         * will simplify the definitions of some of the
         * operator functions. */
        if (this.den < 0) {
            this.num = -this.num;
            this.den = -this.den;
        }
    }

    /* ... you define the operator overloads ... */
}

```

Define operators as needed for this type to make it a convenient type as close to fundamental types as possible. Ensure that the definition of the type passes all of the following unit tests. The unit tests ensure the following behaviors:

- An exception must be thrown when constructing an object with zero denominator. (This is already taken care of by the **enforce** expression above.)
- Producing the negative of the value: For example, the negative of $1/3$ should be $-1/3$.
- Incrementing and decrementing the value by **++** and **--**.
- Support for four arithmetic operations: Both modifying the value of an object by **+=**, **-=**, ***=**, and **/=**; and producing the result of using two objects with the **+**, **-**, *****, and **/** operators. (Similar to the constructor, dividing by zero should be prevented.)

As a reminder, here are the formulas of arithmetic operations that involve two fractions a/b and c/d :

 - Addition: $a/b + c/d = (a*d + c*b)/(b*d)$
 - Subtraction: $a/b - c/d = (a*d - c*b)/(b*d)$
 - Multiplication: $a/b * c/d = (a*c)/(b*d)$
 - Division: $(a/b) / (c/d) = (a*d)/(b*c)$
- The actual (and necessarily lossful) value of the object can be converted to **double**.
- Sort order and equality comparisons are performed by the actual values of the fractions, not by the values of the numerators and denominators. For example, the fractions $1/3$ and $20/60$ must be considered to be equal.

```

unittest
{
    /* Must throw when denominator is zero. */
    assertThrown(Fraction(42, 0));

    /* Let's start with 1/3. */
    auto a = Fraction(1, 3);

    /* -1/3 */
    assert(-a == Fraction(-1, 3));

    /* 1/3 + 1 == 4/3 */
    ++a;
    assert(a == Fraction(4, 3));

    /* 4/3 - 1 == 1/3 */
    --a;
    assert(a == Fraction(1, 3));

    /* 1/3 + 2/3 == 3/3 */
    a += Fraction(2, 3);
    assert(a == Fraction(1));

    /* 3/3 - 2/3 == 1/3 */
    a -= Fraction(2, 3);
    assert(a == Fraction(1, 3));

    /* 1/3 * 8 == 8/3 */
    a *= Fraction(8);
    assert(a == Fraction(8, 3));

    /* 8/3 / 16/9 == 3/2 */
    a /= Fraction(16, 9);
    assert(a == Fraction(3, 2));

    /* Must produce the equivalent value in type 'double'.
     *
     * Note that although double cannot represent every value
     * precisely, 1.5 is an exception. That is why this test
     * is being applied at this point. */
    assert(to!double(a) == 1.5);
}

```

```

/* 1.5 + 2.5 == 4 */
assert(a + Fraction(5, 2) == Fraction(4, 1));

/* 1.5 - 0.75 == 0.75 */
assert(a - Fraction(3, 4) == Fraction(3, 4));

/* 1.5 * 10 == 15 */
assert(a * Fraction(10) == Fraction(15, 1));

/* 1.5 / 4 == 3/8 */
assert(a / Fraction(4) == Fraction(3, 8));

/* Must throw when dividing by zero. */
assertThrown(Fraction(42, 1) / Fraction(0));

/* The one with lower numerator is before. */
assert(Fraction(3, 5) < Fraction(4, 5));

/* The one with larger denominator is before. */
assert(Fraction(3, 9) < Fraction(3, 8));
assert(Fraction(1, 1_000) > Fraction(1, 10_000));

/* The one with lower value is before. */
assert(Fraction(10, 100) < Fraction(1, 2));

/* The one with negative value is before. */
assert(Fraction(-1, 2) < Fraction(0));
assert(Fraction(1, -2) < Fraction(0));

/* The ones with equal values must be both <= and >=. */
assert(Fraction(-1, -2) <= Fraction(1, 2));
assert(Fraction(1, 2) <= Fraction(-1, -2));
assert(Fraction(3, 7) <= Fraction(9, 21));
assert(Fraction(3, 7) >= Fraction(9, 21));

/* The ones with equal values must be equal. */
assert(Fraction(1, 3) == Fraction(20, 60));

/* The ones with equal values with sign must be equal. */
assert(Fraction(-1, 2) == Fraction(1, -2));
assert(Fraction(1, 2) == Fraction(-1, -2));
}

```

std.exception.assertThrown that appear in the tests above is for ensuring that an expression does indeed throw an exception. It works as the equivalent of the following code:

```

assertThrown(Fraction(42, 1) / Fraction(0));
// ...
/* The equivalent of the line above */
{
    auto isThrown = false;

    try {
        auto result = Fraction(42, 1) / Fraction(0);
    } catch (Exception exc) {
        isThrown = true;
    }

    assert(isThrown);
}

```

... the solution

56 Classes

Similar to structs, **class** is a feature for defining new types. Different from structs, classes provide the *object oriented programming* (OOP) paradigm in D. The major aspects of OOP are the following:

- **Encapsulation:** Controlling access to members (*Encapsulation is available for structs as well but it has not been mentioned until this chapter.*)
- **Inheritance:** Acquiring members of another type
- **Polymorphism:** Being able to use a more special type in place of a more general type

Encapsulation is achieved by *protection attributes*, which we will see in a later chapter. Inheritance is for acquiring *implementations* of other types. Polymorphism is for abstracting parts of programs from each other and is achieved by class *interfaces*.

This chapter will introduce classes at a high level, underlining the fact that they are reference types. Classes will be explained in more detail in later chapters.

56.1 Comparing with structs

In general, classes are very similar to structs. Most of the features that we have seen for structs in the following chapters apply to classes as well:

- [Structs](#)
- [Member Functions](#)
- [const ref Parameters and const Member Functions](#)
- [Constructor and Other Special Functions](#)
- [Operator Overloading](#)

However, there are important differences between classes and structs.

56.1.1 Classes are reference types

The biggest difference from structs is that structs are *value types* and classes are *reference types*. The other differences outlined below are mostly due to this fact.

56.1.2 Class objects may be null

As it has been mentioned briefly in [The null Value and the is Operator chapter](#), class variables can be **null**. In other words, class variables may not be providing access to any object. Class variables do not have values themselves; the actual class objects must be constructed by the **new** keyword.

As you would also remember, comparing a reference to **null** by the **==** or the **!=** operator is an error. Instead, the comparison must be done by the **is** or the **!is** operator, accordingly:

```
MyClass referencesAnObject = new MyClass;
assert(referencesAnObject !is null);

MyClass variable; // does not reference an object
assert(variable is null);
```

The reason is that, the **==** operator may need to consult the values of the members of the objects and that attempting to access the members through a potentially **null** variable would cause a memory access error. For that reason, class variables must always be compared by the **is** and **!is** operators.

56.1.3 Class variables versus class objects

Class variable and class object are separate concepts.

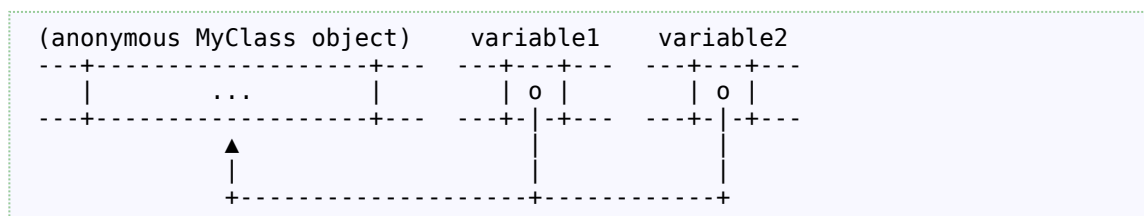
Class objects are constructed by the **new** keyword; they do not have names. The actual concept that a class type represents in a program is provided by a class object. For example, assuming that a **Student** class represents students by their names and grades, such information would be stored by the members of **Student** objects. Partly because they are anonymous, it is not possible to access class objects directly.

A class variable on the other hand is a language feature for accessing class objects. Although it may seem syntactically that operations are being performed on a class *variable*, the operations are actually dispatched to a class *object*.

Let's consider the following code that we have seen previously in the [Value Types and Reference Types](#) chapter:

```
auto variable1 = new MyClass;
auto variable2 = variable1;
```

The **new** keyword constructs an anonymous class object. **variable1** and **variable2** above merely provide access to that anonymous object:



56.1.4 Copying

Copying affects only the variables, not the object.

Because classes are reference types, defining a new class variable as a copy of another makes two variables that provide access to the same object. The actual object is not copied.

Since no object gets copied, the postblit function **this(this)** is not available for classes.

```
auto variable2 = variable1;
```

In the code above, **variable2** is being initialized by **variable1**. The two variables start providing access to the same object.

When the actual object needs to be copied, the class must have a member function for that purpose. To be compatible with arrays, this function may be named **dup()**. This function must create and return a new class object. Let's see this on a class that has various types of members:

```
class Foo
{
    S    o; // assume S is a struct type
    char[] s;
    int  i;

    // ...

    this(S o, const char[] s, int i)
    {
        this.o = o;
    }
}
```



```

        this.s = s.dup;
        this.i = i;
    }

    Foo dup() const
    {
        return new Foo(o, s, i);
    }
}

```

The `dup()` member function makes a new object by taking advantage of the constructor of `Foo` and returns the new object. Note that the constructor copies the `s` member explicitly by the `.dup` property of arrays. Being value types, `o` and `i` are copied automatically.

The following code makes use of `dup()` to create a new object:

```

auto var1 = new Foo(S(1.5), "hello", 42);
auto var2 = var1.dup();

```

As a result, the objects that are associated with `var1` and `var2` are different.

Similarly, an `immutable` copy of an object can be provided by a member function appropriately named `idup()`:

```

class Foo
{
// ...

    immutable(Foo) idup() const
    {
        return new immutable(Foo)(o, s, i);
    }
}

// ...

immutable(Foo) imm = var1.idup();

```

56.1.5 Assignment

Just like copying, assignment affects only the variables.

Assigning to a class variable disassociates that variable from its current object and associates it with a new object.

If there is no other class variable that still provides access to the object that has been disassociated from, then that object is going to be destroyed some time in the future by the garbage collector.

```

auto variable1 = new MyClass();
auto variable2 = new MyClass();
variable1 = variable2;

```

The assignment above makes `variable1` leave its object and start providing access to `variable2`'s object. Since there is no other variable for `variable1`'s original object, that object will be destroyed by the garbage collector.

The behavior of assignment cannot be changed for classes. In other words, `opAssign` cannot be overloaded for them.

56.1.6 Definition

Classes are defined by the **class** keyword instead of the **struct** keyword:

```
class ChessPiece
{
    // ...
}
```

56.1.7 Construction

As with structs, the name of the constructor is **this**. Unlike structs, class objects cannot be constructed by the { } syntax.

```
class ChessPiece
{
    dchar shape;

    this(dchar shape)
    {
        this.shape = shape;
    }
}
```

Unlike structs, there is no automatic object construction where the constructor parameters are assigned to members sequentially:

```
class ChessPiece
{
    dchar shape;
    size_t value;
}

void main()
{
    auto king = new ChessPiece('♔', 100); // ← compilation ERROR
}
```

Error: no constructor for ChessPiece

For that syntax to work, a constructor must be defined explicitly by the programmer.

56.1.8 Destruction

As with structs, the name of the destructor is **~this**:

```
~this()
{
    // ...
}
```

56.1.9 Member access

Same as structs, the members are accessed by the *dot* operator:

```
auto king = new ChessPiece('♔');
writeln(king.shape);
```

Although the syntax makes it look as if a member of the *variable* is being accessed, it is actually the member of the *object*. Class variables do not have members, the class objects do. The **king** variable does not have a **shape** member, the anonymous object does.

Note: *It is usually not proper to access members directly as in the code above. When that exact syntax is desired, properties should be preferred, which will be explained in a later chapter.*

56.1.10 Operator overloading

Other than the fact that **opAssign** cannot be overloaded for classes, operator overloading is the same as structs. For classes, the meaning of **opAssign** is always *associating a class variable with a class object*.

56.1.11 Member functions

Member functions are generally the same as structs.

A difference is that some member functions are automatically inherited from the **Object** class. We will see in the next chapter how the definition of **toString** is changed by the **override** keyword.

56.1.12 The **is** and **!is** operators

These operators operate on class variables.

is specifies whether two class variables provide access to the same class object. It returns **true** if the object is the same and **false** otherwise. **!is** is the opposite of **is**.

```
auto myKing = new ChessPiece('♔');
auto yourKing = new ChessPiece('♔');
assert(myKing !is yourKing);
```

Since the objects of **myKing** and **yourKing** variables are different, the **!is** operator returns **true**. Even though the two objects are constructed by the same character '♔', they are still two separate objects.

When the variables provide access to the same object, **is** returns **true**:

```
auto myKing2 = myKing;
assert(myKing2 is myKing);
```

Both of the variables above provide access to the same object.

56.2 Summary

- Classes and structs share common features but have big differences.
- Classes are reference types. The **new** keyword constructs an anonymous *class object* and returns a *class variable*.
- Class variables that are not associated with any object are **null**. Checking against **null** must be done by **is** or **!is**, not by **==** or **!=**.
- The act of copying associates an additional variable with an object. In order to copy class objects, the type must have a special function likely named **dup()**.
- Assignment associates a variable with an object. This behavior cannot be changed.

57 Inheritance

Inheritance is defining a more special type based on an existing more general type. The special type acquires the members of the general type and as a result, can take place of the general type.

Inheritance is available for classes, not structs. The class that inherits another class is called the *subclass*, and the class that gets inherited is called the *superclass*.

There are two types of inheritance in D. We will cover *implementation inheritance* in this chapter and leave *interface inheritance* to a later chapter.

When defining a subclass, the superclass is specified after a colon character:

```
class SubClass : SuperClass
{
    // ...
}
```

To see an example of this, let's assume that there is already the following class that represents a clock:

```
class Clock
{
    int hour;
    int minute;
    int second;

    void adjust(int hour, int minute, int second = 0)
    {
        this.hour = hour;
        this.minute = minute;
        this.second = second;
    }
}
```

Apparently, the members of that class do not need special values during construction; so there is no constructor. Instead, the members are set by the **adjust()** member function:

```
auto deskClock = new Clock;
deskClock.adjust(20, 30);
writeln(
    "%02s:%02s:%02s",
    deskClock.hour, deskClock.minute, deskClock.second);
```

Note: It would be more useful to produce the time string by a **toString()** function. It will be added later when explaining the **override** keyword below.

The output:

```
20:30:00
```

With only that much functionality, **Clock** could be a struct as well, and depending on the needs of the program, that could be sufficient.

However, being a class makes it possible to inherit from **Clock**.

To see an example of inheritance, let's consider an **AlarmClock** that not only includes all of the functionality of **Clock**, but also provides a way of setting the alarm. Let's first define this type without regard to **Clock**. If we did that, we would have to include the same three

members of **Clock** and the same **adjust()** function that adjusted them. **AlarmClock** would also have other members for its additional functionality:

```
class AlarmClock
{
    int hour;
    int minute;
    int second;
    int alarmHour;
    int alarmMinute;

    void adjust(int hour, int minute, int second = 0)
    {
        this.hour = hour;
        this.minute = minute;
        this.second = second;
    }

    void adjustAlarm(int hour, int minute)
    {
        alarmHour = hour;
        alarmMinute = minute;
    }
}
```

The members that appear exactly in **Clock** are highlighted. As can be seen, defining **Clock** and **AlarmClock** separately results in code duplication.

Inheritance is helpful in such cases. Inheriting **AlarmClock** from **Clock** simplifies the new class and reduces code duplication:

```
class AlarmClock : Clock
{
    int alarmHour;
    int alarmMinute;

    void adjustAlarm(int hour, int minute)
    {
        alarmHour = hour;
        alarmMinute = minute;
    }
}
```

The new definition of **AlarmClock** is the equivalent of the previous one. The highlighted part of the new definition corresponds to the highlighted parts of the old definition.

Because **AlarmClock** inherits the members of **Clock**, it can be used just like a **Clock**:

```
auto bedSideClock = new AlarmClock;
bedSideClock.adjust(20, 30);
bedSideClock.adjustAlarm(7, 0);
```

The members that are inherited from the superclass are accessed as if they are the members of the subclass:

```
writeln("%02s:%02s:%02s ♪%02s:%02s",
        bedSideClock.hour,
        bedSideClock.minute,
        bedSideClock.second,
        bedSideClock.alarmHour,
        bedSideClock.alarmMinute);
```

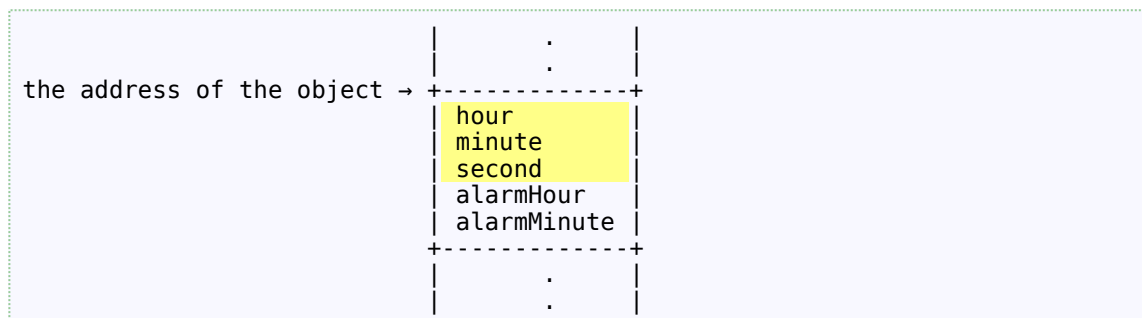
The output:

20:30:00 🎵07:00

Note: An `AlarmClock.toString` function would be more useful in this case. It will be defined later below.

The inheritance used in this example is *implementation inheritance*.

If we imagine the memory as a ribbon going from top to bottom, the placement of the members of `AlarmClock` in memory can be pictured as in the following illustration:



The illustration above is just to give an idea on how the members of the superclass and the subclass may be combined together. The actual layout of the members would depend on decisions that are made by the compiler during compilation.

57.1 Warning: Inherit only if "is a"

We have seen that implementation inheritance is about acquiring members. Consider this kind of inheritance only if the subtype can be thought of being a kind of the supertype as in the phrase "alarm clock *is a* clock."

"Is a" is not the only relationship between types; a more common relationship is the "has a" relationship. For example, let's assume that we want to add the concept of a `Battery` to the `Clock` class. It would not be appropriate to add `Battery` to `Clock` by inheritance because the statement "clock is a battery" is not true:

```
class Clock : Battery    ← Wrong design
{
    // ...
}
```

A clock is not a battery; it *has a* battery. When there is such a relationship of containment, the type that is contained must be defined as a member of the type that contains it:

```
class Clock
{
    Battery battery;    ← Correct design
    // ...
}
```

57.2 Interitance from at most one class

Class inheritance is limited by a single class.

For example, assuming that there is also a `SoundEmitter` class, and even though "alarm clock is a sound emitting object" is also true, it is not possible to inherit `AlarmClock` both from `Clock` and `SoundEmitter`:

```

class SoundEmitter
{
    // ...
}

class AlarmClock : Clock, SoundEmitter    // ← compilation ERROR
{
    // ...
}

```

On the other hand, there is no limit to the number of *interfaces* that a class can inherit from. We will see the **interface** keyword in a later chapter.

Additionally, there is no limit to how deep the inheritance hierarchy can go:

```

class MusicalInstrument
{
    // ...
}

class StringInstrument : MusicalInstrument
{
    // ...
}

class Violin : StringInstrument
{
    // ...
}

```

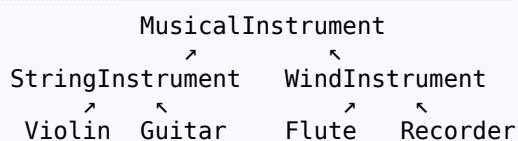
The inheritance hierarchy above defines a relationship from the more general to the more specific: musical instrument, string instrument, and violin.

57.3 Hierarchy charts

Types that are related by the "is a" relationship form a *class hierarchy*.

According to OOP conventions, class hierarchies are represented by superclasses being on the top and the subclasses being at the bottom. The inheritance relationships are indicated by arrows pointing from the subclasses to the superclasses.

For example, the following can be a hierarchy of musical instruments:



57.4 Accessing superclass members

The **super** keyword allows referring to members that are inherited from the superclass.

```

class AlarmClock : Clock
{
    // ...

    void foo()
    {
        super.minute = 10; // The inherited 'minute' member
        minute = 10;      // Same thing if there is no ambiguity
    }
}

```

```

    }
}

```

The **super** keyword is not always necessary; **minute** alone has the same meaning in the code above. The **super** keyword is needed when both the superclass and the subclass have members under the same names. We will see this below when we will need to write **super.reset()** and **super.toString()**.

57.5 Constructing superclass members

The other use of the **super** keyword is to call the constructor of the superclass. This is similar to calling the constructors of the current class: **this** when calling constructors of this class and **super** when calling constructors of the superclass.

It is not required to call the superclass constructor explicitly; when the superclass has a default constructor, the inherited members of the superclass are automatically default-constructed.

We have not defined constructors for the **Clock** and **AlarmClock** classes yet. For that reason, the members of both of those classes are initialized by the **.init** values of their respective types, which is 0 for **int**.

Let's assume that **Clock** has the following constructor:

```

class Clock
{
    this(int hour, int minute, int second)
    {
        this.hour = hour;
        this.minute = minute;
        this.second = second;
    }
    // ...
}

```

That constructor must be used when constructing **Clock** objects:

```

auto clock = new Clock(17, 15, 0);

```

Naturally, the programmers who use the **Clock** type directly would have to use that syntax. However, when constructing an **AlarmClock** object, they cannot construct its **Clock** part separately. Besides, the users of **AlarmClock** need not even know that it inherits from **Clock**.

A user of **AlarmClock** should simply construct an **AlarmClock** object and use it in the program without needing to pay attention to its **Clock** heritage:

```

auto bedSideClock = new AlarmClock(/* ... */);
// ... use as an AlarmClock ...

```

For that reason, constructing the superclass part is the responsibility of the subclass. The subclass calls the constructor of the superclass by the **super()** syntax:

```

class AlarmClock : Clock
{
    this(int hour, int minute, int second, // for Clock's members
         int alarmHour, int alarmMinute) // for AlarmClock's members
    {
        super(hour, minute, second);
    }
}

```



```

        this.alarmHour = alarmHour;
        this.alarmMinute = alarmMinute;
    }
    // ...
}

```

The constructor of **AlarmClock** takes arguments for both its own members and the members of its superclass. It then uses some of that information to construct its superclass part.

If the superclass already has a default constructor, it is not necessary to call **super()** explicitly.

57.6 Overriding the definitions of member functions

One of the benefits of inheritance is being able to redefine the member functions of the superclass in the subclass. This is called *overriding*: The existing definition of the member function of the superclass is overridden by the subclass by the **override** keyword.

Let's assume that **Clock** has a member function that is used for resetting its members all to zero:

```

class Clock
{
    void reset()
    {
        hour = 0;
        minute = 0;
        second = 0;
    }
    // ...
}

```

That function is inherited by **AlarmClock** and can be called on an **AlarmClock** object:

```

auto bedSideClock = new AlarmClock(20, 30, 0, 7, 0);
// ...
bedSideClock.reset();

```

However, decidedly ignorant of the members of **AlarmClock**, **Clock.reset** can only reset its own members. For that reason, to reset the members of the subclass as well, **reset()** must be overridden:

```

class AlarmClock : Clock
{
    override void reset()
    {
        super.reset();
        alarmHour = 0;
        alarmMinute = 0;
    }
    // ...
}

```

The subclass resets only its own members and dispatches the rest of the task to **Clock** by the **super.reset()** call. Note that writing just **reset()** would not work as it would mean the **reset()** function of **AlarmClock** itself. Calling **reset()** from within itself would cause an infinite recursion.

The reason that I have delayed the definition of `toString()` until this point is that it must be defined by the `override` keyword for classes. As we will see in the next chapter, every class is automatically inherited from a superclass called `Object` and `Object` already defines a `toString()` member function.

For that reason, the `toString()` member function for classes must be defined by the `override` keyword:

```
class Clock
{
    override string toString() const
    {
        return format("%02s:%02s:%02s", hour, minute, second);
    }
    // ...
}

class AlarmClock : Clock
{
    override string toString() const
    {
        return format("%s ♪%02s:%02s", super.toString(),
            alarmHour, alarmMinute);
    }
    // ...
}
```

Note that `AlarmClock` is again dispatching some of the task to `Clock` by the `super.toString()` call.

Those two overrides of `toString()` allow using `AlarmClock` objects as strings:

```
void main()
{
    auto deskClock = new AlarmClock(10, 15, 0, 6, 45);
    writeln(deskClock);
}
```

The output:

```
10:15:00 ♪06:45
```

57.7 Using the subclass in place of the superclass

Since the superclass is more *general* and the subclass is more *special*, objects of a subclass can be used in places where an object of superclass is required. This is called *polymorphism*.

The concepts of general and special types can be seen in statements like "this type is of that type": "alarm clock is a clock", "student is a person", "cat is an animal", etc. Accordingly, an alarm clock can be used where a clock is needed, a student can be used where a person is needed, and a cat can be used where an animal is needed.

When a subclass object is being used as a superclass object, it does not lose its own special type. This is similar to its examples in real life: Using an alarm clock simply as a clock does not change the fact that it is an alarm clock; it would still behave like an alarm clock.

Let's assume that a function takes a `Clock` object as parameter, which it resets at some point during its execution:

```
void use(Clock clock)
{
    // ...
    clock.reset();
    // ...
}
```

Polymorphism makes it possible to send an **AlarmClock** to such a function:

```
auto deskClock = new AlarmClock(10, 15, 0, 6, 45);
writeln("Before: ", deskClock);
use(deskClock);
writeln("After : ", deskClock);
```

This is in accordance with the relationship "alarm clock is a clock." As a result, the members of the **deskClock** object gets reset:

```
Before: 10:15:00 ↵06:45
After : 00:00:00 ↵00:00
```

The important observation here is that not only the members of **Clock** but also the members of **AlarmClock** have been reset.

Although **use()** calls **reset()** on a **Clock** object, since the actual object is an **AlarmClock**, the function that gets called is **AlarmClock.reset**. According to its definition above, **AlarmClock.reset** resets the members of both **Clock** and **AlarmClock**.

In other words, although **use()** uses the object as a **Clock**, the actual object may be an inherited type that behaves in its own special way.

Let's add another class to the **Clock** hierarchy. The **reset()** function of this type sets its members to random values:

```
import std.random;

class BrokenClock : Clock
{
    this()
    {
        super(0, 0, 0);
    }

    override void reset()
    {
        hour = uniform(0, 24);
        minute = uniform(0, 60);
        second = uniform(0, 60);
    }
}
```

When an object of **BrokenClock** is sent to **use()**, then the special **reset()** function of **BrokenClock** would be called. Again, although it is passed as a **Clock**, the actual object is still a **BrokenClock**:

```
auto shelfClock = new BrokenClock;
use(shelfClock);
writeln(shelfClock);
```

The output shows random time values as a result of resetting a **BrokenClock**:

```
22:46:37
```

57.8 Inheritance is transitive

Polymorphism is not limited to two classes. Subclasses of subclasses can also be used in place of any superclass in the hierarchy.

Let's consider the **MusicalInstrument** hierarchy:

```
class MusicalInstrument
{
    // ...
}

class StringInstrument : MusicalInstrument
{
    // ...
}

class Violin : StringInstrument
{
    // ...
}
```

The inheritances above build the following relationships: "string instrument is a musical instrument" and "violin is a string instrument." Therefore, it is also true that "violin is a musical instrument." Consequently, a **Violin** object can be used in place of a **MusicalInstrument**.

Assuming that all of the supporting code below have also been defined:

```
void playInTune(MusicalInstrument instrument,
               MusicalPiece piece)
{
    instrument.tune();
    instrument.play(piece);
}

// ...

auto myViolin = new Violin;
playInTune(myViolin, improvisation);
```

Although **playInTune()** expects a **MusicalInstrument**, it is being called with a **Violin** due to the relationship "violin is a musical instrument."

Inheritance can be as deep as needed.

57.9 Abstract member functions and abstract classes

Sometimes there are member functions that are natural to appear on a class interface even though that class cannot provide its definition. When there is no *concrete* definition of a member function, that function is an *abstract* member function. A class that has at least one abstract member function is an abstract class.

For example, the **ChessPiece** superclass in a hierarchy may have an **isValid()** member function that determines whether a given move is valid for that chess piece. Since validity of a move depends on the actual type of the chess piece, the **ChessPiece** general class cannot make this decision itself. The valid moves can only be known by the subclasses like **Pawn**, **King**, etc.

The **abstract** keyword specifies that the definition of the member function is not provided by this class and *must be* provided by a subclass:

```
class ChessPiece
{
    abstract bool isValid(in Square from, in Square to);
}
```

The `isValid()` function does not have a definition; it is a declaration.

It is not possible to construct objects of abstract classes:

```
auto piece = new ChessPiece; // ← compilation ERROR
```

Objects of an abstract class cannot be used because of the missing function definitions. Such a class would not know how to respond to a call like `piece.isValid(thisSquare, thatSquare)`.

On the other hand, it is possible to construct objects of subclasses as long as there are no function definitions that are still left undefined (i.e. that are still abstract):

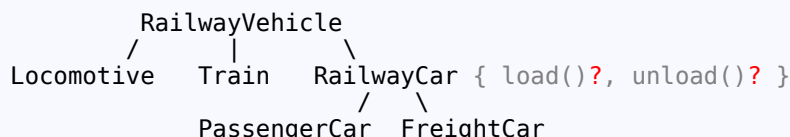
```
class Pawn : ChessPiece
{
    override bool isValid(in Square from, in Square to)
    {
        // ... the implementation of isValid for pawn ...
        return decision;
    }
}
```

Because it provides the definition for the abstract function, it is possible to construct objects of `Pawn`:

```
auto piece = new Pawn; // compiles
```

57.10 Example

Let's consider a class hierarchy that represents railway vehicles:



The functions that the `RailwayCar` will declare as **abstract** are indicated by red question marks.

Since my goal is only to present a class hierarchy and point out some of its design decisions, I will not fully implement these classes. Instead of doing actual work, they will simply print messages.

The most general class of the hierarchy above is `RailwayVehicle`. In this program, it will only know how to move itself:

```
class RailwayVehicle
{
    void advance(in size_t kilometers)
    {
        writefln("The vehicle is advancing %s kilometers",
                kilometers);
    }
}
```

A class that inherits from **RailwayVehicle** is **Locomotive**, which does not have any special members yet:

```
class Locomotive : RailwayVehicle
{}
```

We will add a special **makeSound()** member function to **Locomotive** later during one of the exercises.

RailwayCar is a **RailwayVehicle** as well. However, if the hierarchy supports different types of railway cars, then certain behaviors like loading and unloading must be done according to their exact types. For that reason, **RailwayCar** can only declare these two functions as abstract:

```
class RailwayCar : RailwayVehicle
{
    abstract void load();
    abstract void unload();
}
```

Loading and unloading a passenger car is as simple as opening the doors of the car, while loading and unloading a freight car may involve porters and winches. The following subclasses provide definitions for the abstract functions of **RailwayCar**:

```
class PassengerCar : RailwayCar
{
    override void load()
    {
        writeln("The passengers are getting on");
    }

    override void unload()
    {
        writeln("The passengers are getting off");
    }
}

class FreightCar : RailwayCar
{
    override void load()
    {
        writeln("The crates are being loaded");
    }

    override void unload()
    {
        writeln("The crates are being unloaded");
    }
}
```

Being an abstract class does not preclude the use of **RailwayCar** in the program. Objects of **RailwayCar** can not be constructed but **RailwayCar** can be used as an interface. As the subclasses define the two relationships "passenger car is a railway car" and "freight car is a railway car", the objects of **PassengerCar** and **FreightCar** can be used in places of **RailwayCar**. This will be seen in the **Train** class below.

The class that represents a train can consist of a locomotive and an array of railwaycars:

```
class Train : RailwayVehicle
{
    Locomotive locomotive;
    RailwayCar[] cars;
```

```
} // ...
```

I would like to repeat an important point: Although both **Locomotive** and **RailwayCar** inherit from **RailwayVehicle**, it would not be correct to inherit **Train** from either of them. Inheritance models the "is a" relationship and a train is neither a locomotive nor a passenger car. A train consists of them.

If we require that every train must have a locomotive, the **Train** constructor must ensure that it takes a valid **Locomotive** object. Similarly, if the railway cars are optional, they can be added by a member function:

```
import std.exception;
// ...

class Train : RailwayVehicle
{
    // ...

    this(Locomotive locomotive)
    {
        enforce(locomotive !is null,
                "Locomotive cannot be null");
        this.locomotive = locomotive;
    }

    void addCar(RailwayCar[] cars...)
    {
        this.cars ~= cars;
    }

    // ...
}
```

Note that **addCar()** can validate the **RailwayCar** objects as well. I am ignoring that validation here.

We can imagine that the departures and arrivals of trains should also be supported:

```
class Train : RailwayVehicle
{
    // ...

    void departStation(string station)
    {
        foreach (car; cars) {
            car.load();
        }

        writeln("Departing from the %s station", station);
    }

    void arriveStation(string station)
    {
        writeln("Arriving at the %s station", station);

        foreach (car; cars) {
            car.unload();
        }
    }
}
```

The following **main()** is making use of the **RailwayVehicle** hierarchy:

```

import std.stdio;
// ...
void main()
{
    auto locomotive = new Locomotive;
    auto train = new Train(locomotive);

    train.addCar(new PassengerCar, new FreightCar);

    train.departStation("Ankara");
    train.advance(500);
    train.arriveStation("Haydarpaşa");
}

```

The **Train** class is being used by functions that are provided by two separate interfaces:

1. When the **advance()** function is called, the **Train** object is being used as a **RailwayVehicle** because that function is declared by **RailwayVehicle**.
2. When the **departStation()** and **arriveStation()** functions are called, **train** is being used as a **Train** because those functions are declared by **Train**.

The blue arrows indicate that the **load()** and **unload()** functions work according to the actual type of **RailwayCar**:

```

The passengers are getting on      ←
The crates are being loaded        ←
Departing from the Ankara station
The vehicle is advancing 500 kilometers
Arriving at the Haydarpaşa station
The passengers are getting off     ←
The crates are being unloaded      ←

```

57.11 Summary

- Inheritance is for the "is a" relationship.
- Each class can inherit from up to one **class**.
- **super** has two uses: Calling the constructor of the superclass and accessing the members of the superclass.
- **override** is for redefining member functions of the superclass specially for this class.
- **abstract** requires that a member function must be overridden.

57.12 Exercises

1. Let's modify **RailwayVehicle**. In addition to reporting the distance that it advances, let's have it also make sounds. To keep the output short, let's print the sounds per 100 kilometers:

```

class RailwayVehicle
{
    void advance(in size_t kilometers)
    {
        writeln("The vehicle is advancing %s kilometers",
            kilometers);

        foreach (i; 0 .. kilometers / 100) {
            writeln(" %s", makeSound());
        }
    }
}

```



```
} // ...
```

However, `makeSound()` cannot be defined by `RailwayVehicle` because vehicles may have different sounds:

- "choo choo" for `Locomotive`
- "clack clack" for `RailwayCar`

Note: Leave `Train.makeSound` to the next exercise.

Because it must be overridden, `makeSound()` must be declared as **abstract** by the superclass:

```
class RailwayVehicle
{
    // ...

    abstract string makeSound();
}
```

Implement `makeSound()` for the subclasses and try the code with the following `main()`:

```
void main()
{
    auto railwayCar1 = new PassengerCar;
    railwayCar1.advance(100);

    auto railwayCar2 = new FreightCar;
    railwayCar2.advance(200);

    auto locomotive = new Locomotive;
    locomotive.advance(300);
}
```

Make the program produce the following output:

```
The vehicle is advancing 100 kilometers
clack clack
The vehicle is advancing 200 kilometers
clack clack
clack clack
The vehicle is advancing 300 kilometers
choo choo
choo choo
choo choo
```

Note that there is no requirement that the sounds of `PassengerCar` and `FreightCar` be different. They can share the same implementation from `RailwayCar`.

3. Think about how `makeSound()` can be implemented for `Train`. One idea is that `Train.makeSound` may return a **string** that consists of the sounds of the members of `Train`.

[... the solutions](#)

58 Object

Classes that do not explicitly inherit any class automatically inherit the **Object** class.

By that definition, the topmost class in any class hierarchy inherits **Object**:

```
// ": Object" is not written; it is automatic
class MusicalInstrument : Object-
{
    // ...
}

// Inherits Object indirectly
class StringInstrument : MusicalInstrument
{
    // ...
}
```

Since the topmost class inherits **Object**, every class indirectly inherits **Object** as well. In that sense, every class "is an" **Object**.

Every class inherits the following member functions of **Object**:

- **toString**: The **string** representation of the object.
- **opEquals**: Equality comparison with another object.
- **opCmp**: Sort order comparison with another object.
- **toHash**: Associative array hash value.

The last three of these functions emphasize the values of objects. They also make a class eligible for being the key type of associative arrays.

Because these functions are inherited, their redefinitions for the subclasses require the **override** keyword.

Note: **Object** defines other members as well. This chapter will include only these four functions.

58.1 toString

Same with structs, **toString** enables using objects as strings:

```
auto clock = new Clock(20, 30, 0);
writeln(clock);           // Calls clock.toString()
```

The inherited **toString()** is usually not useful; it produces just the name of the type:

```
deneme.Clock
```

The part before the name of the type is the name of the module. The output above indicates that **Clock** has been defined in the **deneme** module.

As we have seen in the previous chapter, this function is almost always overridden to produce a more meaningful **string** representation:

```
import std.string;

class Clock
{
    override string toString() const
    {
```

```

        return format("%02s:%02s:%02s", hour, minute, second);
    }
    // ...
}
class AlarmClock : Clock
{
    override string toString() const
    {
        return format("%s 🎵%02s:%02s", super.toString(),
            alarmHour, alarmMinute);
    }
    // ...
}
// ...

auto bedSideClock = new AlarmClock(20, 30, 0, 7, 0);
writeln(bedSideClock);

```

The output:

```
20:30:00 🎵07:00
```

58.2 opEquals

As we have seen in the [Operator Overloading chapter](#), this member function is about the behavior of the `==` operator (and the `!=` operator indirectly). The return value of the operator must be **true** if the objects are considered to be equal and **false** otherwise.

Warning: The definition of this function must be consistent with `opCmp()`; for two objects that `opEquals()` returns **true**, `opCmp()` must return zero.

Contrary to structs, the compiler does not call `a.opEquals(b)` right away when it sees the expression `a == b`. When two class objects are compared by the `==` operator, a four-step algorithm is executed:

```

bool opEquals(Object a, Object b) {
    if (a is b) return true;           // (1)
    if (a is null || b is null) return false; // (2)
    if (typeid(a) == typeid(b)) return a.opEquals(b); // (3)
    return a.opEquals(b) && b.opEquals(a); // (4)
}

```

1. If the two variables provide access to the same object (or they are both **null**), then they are equal.
2. Following from the previous check, if only one is **null** then they are not equal.
3. If both of the objects are of the same type, then `a.opEquals(b)` is called to determine the equality.
4. Otherwise, for the two objects to be considered equal, `opEquals` must have been defined for both of their types and `a.opEquals(b)` and `b.opEquals(a)` must agree that the objects are equal.

Accordingly, if `opEquals()` is not provided for a class type, then the values of the objects are not considered; rather, equality is determined by checking whether the two class variables provide access to the same object:

```

auto variable0 = new Clock(6, 7, 8);
auto variable1 = new Clock(6, 7, 8);

assert(variable0 != variable1); // They are not equal
                                // because the objects are
                                // different

```

Even though the two objects are constructed by the same arguments above, the variables are not equal because they are not associated with the same object.

On the other hand, because the following two variables provide access to the same object, they are *equal*:

```

auto partner0 = new Clock(9, 10, 11);
auto partner1 = partner0;

assert(partner0 == partner1); // They are equal because
                               // the object is the same

```

Sometimes it makes more sense to compare objects by their values instead of their identities. For example, it is conceivable that **variable0** and **variable1** above compare equal because their values are the same.

Different from structs, the type of the parameter of **opEquals** for classes is **Object**:

```

class Clock
{
    override bool opEquals(Object o) const
    {
        // ...
    }

    // ...
}

```

As you will see below, the parameter is almost never used directly. For that reason, it should be acceptable to name it simply as **o**. Most of the time the first thing to do with that parameter is to use it in a type conversion.

The parameter of **opEquals** is the object that appears on the right-hand side of the **==** operator:

```

variable0 == variable1; // o represents variable1

```

Since the purpose of **opEquals()** is to compare two objects of this class type, the first thing to do is to convert **o** to a variable of the same type of this class. Since it would not be appropriate to modify the right-hand side object in an equality comparison, it is also proper to convert the type as **const**:

```

override bool opEquals(Object o) const
{
    auto rhs = cast(const Clock)o;

    // ...
}

```

As you would remember, **rhs** is a common abbreviation for *right-hand side*. Also, **std.conv.to** can be used for the conversion as well:

```
import std.conv;
// ...
auto rhs = to!(const Clock)o;
```

If the original object on the right-hand side can be converted to **Clock**, then **rhs** becomes a non-**null** class variable. Otherwise, **rhs** is set to **null**, indicating that the objects are not of the same type.

According to the design of a program, it may make sense to compare objects of two incompatible types. I will assume here that for the comparison to be valid, **rhs** must not be **null**; so, the first logical expression in the following **return** statement checks that it is not **null**. Otherwise, it would be an error to try to access the members of **rhs**:

```
class Clock
{
    int hour;
    int minute;
    int second;

    override bool opEquals(Object o) const
    {
        auto rhs = cast(const Clock)o;

        return (rhs &&
                (hour == rhs.hour) &&
                (minute == rhs.minute) &&
                (second == rhs.second));
    }
    // ...
}
```

With that definition, **Clock** objects can now be compared by their values:

```
auto variable0 = new Clock(6, 7, 8);
auto variable1 = new Clock(6, 7, 8);

assert(variable0 == variable1); // Now they are equal
                                // because their values
                                // are equal
```

When defining **opEquals** it is important to remember the members of the superclass. For example, when comparing objects of **AlarmClock** it would make sense to also consider the inherited members:

```
class AlarmClock : Clock
{
    int alarmHour;
    int alarmMinute;

    override bool opEquals(Object o) const
    {
        auto rhs = cast(const AlarmClock)o;

        return (rhs &&
                (alarmHour == rhs.alarmHour) &&
                (alarmMinute == rhs.alarmMinute) &&
                super.opEquals(o));
    }
    // ...
}
```

That expression could be written as `super == o` as well. However, that would initiate the four-step algorithm again and as a result, the code might be a little slower.

58.3 opCmp

This operator is used when sorting class objects. `opCmp` is the function that gets called behind the scenes for the `<`, `<=`, `>`, and `>=`.

This operator must return a negative value when the left-hand object is before, a positive value when the left-hand object is after, and zero when both objects have the same sorting order.

Warning: The definition of this function must be consistent with `opEquals()`; for two objects that `opEquals()` returns `true`, `opCmp()` must return zero.

Unlike `toString` and `opEquals`, there is no default implementation of this function in `Object`. If the implementation is not available, comparing objects for sort order causes an exception to be thrown:

```
auto variable0 = new Clock(6, 7, 8);
auto variable1 = new Clock(6, 7, 8);

assert(variable0 <= variable1);    // ← Causes exception
```

object.Exception: need opCmp for class deneme.Clock

It is up to the design of the program what happens when the left-hand and right-hand objects are of different types. One way is to take advantage of the order of types that is maintained by the compiler automatically. This is achieved by calling the `opCmp` function on the `typeid` values of the two types:

```
class Clock
{
    int hour;
    int minute;
    int second;

    override int opCmp(Object o) const
    {
        /* Taking advantage of the automatically-maintained
         * order of the types. */
        if (typeid(this) != typeid(o)) {
            return typeid(this).opCmp(typeid(o));
        }

        auto rhs = cast(const Clock)o;
        /* No need to check whether rhs is null, because it is
         * known at this line that it has the same type as o. */

        if (hour != rhs.hour) {
            return hour - rhs.hour;
        } else if (minute != rhs.minute) {
            return minute - rhs.minute;
        } else {
            return second - rhs.second;
        }
    }

    // ...
}
```

The definition above first checks whether the types of the two objects are the same. If not, it uses the ordering of the types themselves. Otherwise, it compares the objects by the values of their **hour**, **minute**, and **second** members.

A chain of ternary operators may also be used:

```

override int opCmp(Object o) const
{
    if (typeid(this) != typeid(o)) {
        return typeid(this).opCmp(typeid(o));
    }

    auto rhs = cast(const Clock)o;

    return (hour != rhs.hour
        ? hour - rhs.hour
        : (minute != rhs.minute
            ? minute - rhs.minute
            : second - rhs.second));
}

```

If important, the comparison of the members of the superclass must also be considered. The following **AlarmClock.opCmp** is calling **Clock.opCmp** first:

```

class AlarmClock : Clock
{
    override int opCmp(Object o) const
    {
        auto rhs = cast(const AlarmClock)o;

        const int superResult = super.opCmp(o);

        if (superResult != 0) {
            return superResult;
        } else if (alarmHour != rhs.alarmHour) {
            return alarmHour - rhs.alarmHour;
        } else {
            return alarmMinute - rhs.alarmMinute;
        }
    }
    // ...
}

```

Above, if the superclass comparison returns a non-zero value then that result is used because the sort order of the objects is already determined by that value.

AlarmClock objects can now be compared for their sort orders:

```

auto ac0 = new AlarmClock(8, 0, 0, 6, 30);
auto ac1 = new AlarmClock(8, 0, 0, 6, 31);

assert(ac0 < ac1);

```

opCmp is used by other language features and libraries as well. For example, the **.sort** property of arrays takes advantage of **opCmp** when sorting the elements.

58.3.1 opCmp for string members

When some of the members are strings, they can be compared explicitly to return a negative, positive, or zero value:

```

import std.exception;

class Student
{
    string name;

    override int opCmp(Object o) const
    {
        auto rhs = cast(Student)o;
        enforce(rhs);

        if (name < rhs.name) {
            return -1;
        } else if (name > rhs.name) {
            return 1;
        } else {
            return 0;
        }
    }

    // ...
}

```

Instead, the existing `std.algorithm.cmp` function can be used, which happens to be faster as well:

```

import std.algorithm;

class Student
{
    string name;

    override int opCmp(Object o) const
    {
        auto rhs = cast(Student)o;
        enforce(rhs);

        return cmp(name, rhs.name);
    }

    // ...
}

```

Also note that `Student` does not support comparing incompatible types by enforcing that the conversion from `Object` to `Student` is possible.

58.4 toHash

This function allows objects of a class type to be used as associative array *keys*. It does not affect the cases where the type is used as associative array *values*.

Warning: Defining only this function is not sufficient. In order for the class type to be used as associative array keys, consistent definitions of `opEquals` and `opCmp` must also be defined.

58.4.1 Hash table indexes

Associative arrays are a hash table implementation. Hash table is a very fast data structure when it comes to searching elements in the table. (*Note: Like most other things in software, this speed comes at a cost: Hash tables must keep elements in an unordered way, and they may be taking up space that is more than exactly necessary.*)

The high speed of hash tables comes from the fact that they first produce integer values for keys. These integers are called *hash values*. The hash values are then used for indexing into an internal array that is maintained by the table.

A benefit of this method is the fact that any type that can be converted to an integer value can be used as the key type of associative arrays. **toHash** is the function that returns the hash table index value for an object.

Although perhaps not very meaningful, even **Clock** objects can be used as associative array key values:

```
string[Clock] timeTags;
timeTags[new Clock(12, 0, 0)] = "Noon";
```

The default definition of **toHash** that is inherited from **Clock** produces different hash values for different objects. This is similar to how the default behavior of **opEquals** considers different objects as being not equal.

The code above compiles and runs even when there is no special definition of **toHash** for **Clock**. However, its default behavior is almost never what is desired. To see that default behavior, let's try to access an element by an object that is different from the one that has been used when inserting the element. Although the new **Clock** object below has the same value as the **Clock** object that has been used when inserting into the associative array above, the value cannot be found:

```
if (new Clock(12, 0, 0) in timeTags) {
    writeln("Exists");
} else {
    writeln("Missing");
}
```

According to the **in** operator, there is no element in the table that corresponds to the value **Clock(12, 0, 0)**:

Missing

The reason for this surprising behavior is that the key object that has been used when inserting the element is not the same as the key object that has been used when accessing the element.

58.4.2 Selecting members for toHash

Although the hash value is calculated from the members of an object, not every member is suitable for this task.

The candidates are the members that distinguish objects from each other. For example, the members **name** and **lastName** of a **Student** class would be suitable if those members can be used for identifying objects of that type.

On the other hand, a **grades** array of a **Student** class would not be suitable both because many objects may have the same array and also it is likely that the **grades** array may change over time.

58.4.3 Calculating hash values

The choice of hash values has a direct effect on the performance of associative arrays. Furthermore, a hash calculation that is effective on one type of data may not be as effective on another type of data. As *hash algorithms* are beyond the scope of this book, I will give just one guideline here: In general, it is better to produce different hash values for objects that are considered to have different values. However, it is not an error if two objects with different values produce the same index value; it is merely undesirable for performance reasons.

It is conceivable that all of the members of **Clock** are significant to distinguish its objects from each other. For that reason, the hash values can be calculated from the values of its three members. *The number of seconds since midnight* would be effective hash values for objects that represent different points in time:

```
class Clock
{
    int hour;
    int minute;
    int second;

    override size_t toHash() const
    {
        /* Because there are 3600 seconds in an hour and 60
         * seconds in a minute: */
        return (3600 * hour) + (60 * minute) + second;
    }

    // ...
}
```

Whenever **Clock** is used as the key type of associative arrays, that special definition of **toHash** would be used. As a result, even though the two key objects of **Clock(12, 0, 0)** above are distinct, they would now produce the same hash value.

The new output:

Exists

Similar to the other member functions, the superclass may need to be considered as well. For example, **AlarmClock.toHash** can take advantage of **Clock.toHash** during its index calculation:

```
class AlarmClock : Clock
{
    int alarmHour;
    int alarmMinute;

    override size_t toHash() const
    {
        return super.toHash() + alarmHour + alarmMinute;
    }

    // ...
}
```

Note: Take the calculation above just as an example. In general, adding integer values is not an effective way of generating hash values.

There are existing efficient algorithms for calculating hash values for variables of floating point, array, and struct types. These algorithms are available to the programmer as well.

What needs to be done is to call `getHash()` on the `typeid` of each member. The syntax of this method is the same for floating point, array, and struct types.

For example, hash values of a `Student` type can be calculated from its `name` member as in the following code:

```
class Student
{
    string name;

    override size_t toHash() const
    {
        return typeid(name).getHash(&name);
    }

    // ...
}
```

58.4.4 Hash values for structs

Since structs are value types, hash values for their objects are calculated automatically by an efficient algorithm. That algorithm takes all of the members of the object into consideration.

When there is a specific reason like needing to exclude certain members from the hash calculation, `toHash()` can be overridden for structs as well.

58.5 Exercises

1. Start with the following class, which represents colored points:

```
enum Color { blue, green, red }

class Point
{
    int x;
    int y;
    Color color;

    this(int x, int y, Color color)
    {
        this.x = x;
        this.y = y;
        this.color = color;
    }
}
```

Implement `opEquals` for this class in a way that ignores colors. When implemented in that way, the following `assert` check should pass:

```
// Different colors
auto bluePoint = new Point(1, 2, Color.blue);
auto greenPoint = new Point(1, 2, Color.green);

// They are still equal
assert(bluePoint == greenPoint);
```

2. Implement `opCmp` by considering first `x` then `y`. The following `assert` checks should pass:

```
auto redPoint1 = new Point(-1, 10, Color.red);
auto redPoint2 = new Point(-2, 10, Color.red);
auto redPoint3 = new Point(-2, 7, Color.red);

assert(redPoint1 < bluePoint);
```

```

assert(redPoint3 < redPoint2);

/* Even though blue is before green in enum Color,
 * because color is being ignored, bluePoint must not be
 * before greenPoint. */
assert(!(bluePoint < greenPoint));

```

Like the **Student** class above, you can implement **opCmp** by excluding incompatible types by the help of **enforce**.

3. Consider the following class that combines three **Point** objects in an array:

```

class TriangularArea
{
    Point[3] points;

    this(Point one, Point two, Point three)
    {
        points = [ one, two, three ];
    }
}

```

Implement **toHash** for that class. Again, the following **assert** checks should pass:

```

/* area1 and area2 are constructed by distinct points that
 * happen to have the same values. (Remember that
 * bluePoint and greenPoint should be considered equal.) */
auto area1 = new TriangularArea(bluePoint, greenPoint, redPoint1);
auto area2 = new TriangularArea(greenPoint, bluePoint, redPoint1);

// The areas should be equal
assert(area1 == area2);

// An associative array
double[TriangularArea] areas;

// A value is being entered by area1
areas[area1] = 1.25;

// The value is being accessed by area2
assert(area2 in areas);
assert(areas[area2] == 1.25);

```

Remember that **opEquals** and **opCmp** must also be defined when **toHash** is defined.

... the solutions

59 Interfaces

The **interface** keyword is for defining interfaces in class hierarchies. **interface** is very similar to **class** with the following restrictions:

- The member functions that it declares (but not implements) are abstract even without the **abstract** keyword.
- The member functions that it implements must be **static** or **final**. (**static** and **final** member functions are explained below.)
- Its member variables must be **static**.
- Interfaces can inherit only interfaces.

Despite these restrictions, there is no limit on the number of **interfaces** that a class can inherit from. (In contrast, a class can inherit from up to one **class**.)

59.1 Definition

Interfaces are defined by the **interface** keyword, the same way as classes:

```
interface SoundEmitter
{
    // ...
}
```

An **interface** is for declaring member functions that are implicitly abstract:

```
interface SoundEmitter
{
    string emitSound();    // Declared (not implemented)
}
```

Classes that inherit from that interface would have to provide the implementations of the abstract functions of the interface.

59.2 Inheriting from an interface

The **interface** inheritance syntax is the same as **class** inheritance:

```
class Violin : SoundEmitter
{
    string emitSound()
    {
        return "♪♪♪";
    }
}

class Bell : SoundEmitter
{
    string emitSound()
    {
        return "ding";
    }
}
```

Interfaces support polymorphism: Functions that take interface parameters can use those parameters without needing to know the actual types of objects. For example, the following function that takes a parameter of **SoundEmitter** calls **emitSound()** on that parameter without needing to know the actual type of the object:

```
void useObjectWithSound(SoundEmitter object)
{
    // ... some operations ...
    writeln(object.emitSound());
    // ... more operations ...
}
```

Just like with classes, that function can be called with any type of object that inherits from the **SoundEmitter** interface:

```
useObjectWithSound(new Violin);
useObjectWithSound(new Bell);
```

The special **emitSound()** function for each object would get called and the outputs of **Violin.emitSound** and **Bell.emitSound** would be printed:

```
♪♪♪
ding
```

59.3 Inheriting from more than one interface

A class can be inherited from up to one **class**. There is no limit on the number of **interfaces** to inherit from.

Let's consider the following interface that represents communication devices:

```
interface CommunicationDevice
{
    void talk(string message);
    string listen();
}
```

If a **Phone** class needs to be used both as a sound emitter and a communication device, it can inherit both of those interfaces:

```
class Phone : SoundEmitter, CommunicationDevice
{
    // ...
}
```

That definition represents both of these relationships: "phone is a sound emitter" and "phone is a communication device."

In order to construct objects of this class, **Phone** must implement the abstract functions of both of the interfaces:

```
class Phone : SoundEmitter, CommunicationDevice
{
    string emitSound()           // for SoundEmitter
    {
        return "rrring";
    }

    void talk(string message)    // for CommunicationDevice
    {
        // ... put the message on the line ...
    }

    string listen()             // for CommunicationDevice
    {
```

```

    string soundOnTheLine;
    // ... get the message from the line ...
    return soundOnTheLine;
}
}

```

A class can inherit from any number of interfaces as it makes sense according to the design of the program.

59.4 Inheriting from interface and class

Classes can still inherit from up to one **class** as well:

```

class Clock
{
    // ... clock implementation ...
}

class AlarmClock : Clock, SoundEmitter
{
    string emitSound()
    {
        return "beep";
    }
}

```

AlarmClock inherits the members of **Clock**. Additionally, it also provides the **emitSound()** function that the **SoundEmitter** interface requires.

59.5 Inheriting interface from interface

An interface that is inherited from another interface effectively increases the number of functions that the subclasses must implement:

```

interface MusicalInstrument : SoundEmitter
{
    void adjustTuning();
}

```

According to the definition above, in order to be a **MusicalInstrument**, both the **emitSound()** function that **SoundEmitter** requires and the **adjustTuning()** function that **MusicalInstrument** requires must be implemented.

For example, if **Violin** inherits from **MusicalInstrument** instead of **SoundEmitter**, it must now also implement **adjustTuning()**:

```

class Violin : MusicalInstrument
{
    string emitSound() // for SoundEmitter
    {
        return "♪♪♪";
    }

    void adjustTuning() // for MusicalInstrument
    {
        // ... special tuning of the violin ...
    }
}

```

59.6 static member functions

I have delayed explaining **static** member functions until this chapter to keep the earlier chapters shorter. **static** member functions are available for structs, classes, and interfaces.

Regular member functions are always called on an object. The member variables that are referenced inside the member function are the members of a particular object:

```
struct Foo
{
    int i;

    void modify(int value)
    {
        i = value;
    }
}

void main()
{
    auto object0 = Foo();
    auto object1 = Foo();

    object0.modify(10);    // object0.i changes
    object1.modify(10);    // object1.i changes
}
```

The members can also be referenced by **this**:

```
void modify(int value)
{
    this.i = value;    // equivalent of the previous one
}
```

A **static** member function does not operate on an object; there is no object that the **this** keyword would refer to, so **this** is not valid inside a **static** function. For that reason, none of the regular member variables are available inside **static** member functions:

```
struct Foo
{
    int i;

    static void commonFunction(int value)
    {
        i = value;    // ← compilation ERROR
        this.i = value;    // ← compilation ERROR
    }
}
```

static member functions can use only the **static** member variables.

Let's redesign the **Point** struct that we have seen earlier in [the Structs chapter](#), this time with a **static** member function. In the following code, every **Point** object gets a unique id, which is determined by a **static** member function:

```
import std.stdio;

struct Point
{
    size_t id;    // Object id
    int line;
    int column;

    // The id to be used for the next object
}
```



```

static size_t nextId;

this (int line, int column)
{
    this.line = line;
    this.column = column;
    this.id = makeNewId();
}

static size_t makeNewId()
{
    immutable newId = nextId;
    ++nextId;
    return newId;
}

void main()
{
    auto top = Point(7, 0);
    auto middle = Point(8, 0);
    auto bottom = Point(9, 0);

    writeln(top.id);
    writeln(middle.id);
    writeln(bottom.id);
}

```

The **static** `makeNewId()` function can use the common variable `nextId`. As a result, every object gets a unique id:

```

0
1
2

```

Although the example above contains a **struct**, **static** member functions are available for classes and interfaces as well.

59.7 final member functions

I have delayed explaining **final** member functions until this chapter to keep the earlier chapters shorter. **final** member functions are available only for classes and interfaces; they are not relevant to structs because structs do not support inheritance.

final specifies that a member function cannot be redefined by a subclass. In a sense, the implementation that this **class** or **interface** provides is the *final* implementation of that function. An example of a case where this feature is useful is where the general steps of an algorithm are defined by an interface and the finer details are left to subclasses.

Let's see an example of this with a **Game** interface. The general steps of playing a game is being determined by the `play()` function of the following **interface**:

```

interface Game
{
    final void play()
    {
        string name = gameName();
        writefln("Starting %s", name);

        introducePlayers();
        prepare();
        begin();
    }
}

```

```

        end();

        writefln("Ending %s", name);
    }

    string gameName();
    void introducePlayers();
    void prepare();
    void begin();
    void end();
}

```

It is not possible for subclasses to modify the definition of the **play()** member function. The subclasses can (and must) provide the definitions of the five abstract member functions that are declared by the interface. By doing so, the subclasses complete the missing steps of the algorithm:

```

import std.stdio;
import std.string;
import std.random;
import std.conv;

class DiceSummingGame : Game
{
    string player;
    size_t count;
    size_t sum;

    string gameName()
    {
        return "Dice Summing Game";
    }

    void introducePlayers()
    {
        write("What is your name? ");
        player =.chomp(readln());
    }

    void prepare()
    {
        write("How many times to throw the dice? ");
        readf(" %s", &count);
        sum = 0;
    }

    void begin()
    {
        foreach (i; 0 .. count) {
            immutable dice = uniform(1, 7);
            writefln("%s: %s", i, dice);
            sum += dice;
        }
    }

    void end()
    {
        writefln("Player: %s, Dice sum: %s, Average: %s",
            player, sum, to!double(sum) / count);
    }
}

void useGame(Game game)
{
    game.play();
}

void main()

```

```
{
    useGame(new DiceSummingGame());
}
```

Although the example above contains an **interface**, **final** member functions are available for classes as well.

59.8 How to use

interface is a commonly used feature. There is one or more **interface** at the top of almost every class hierarchy. A kind of hierarchy that is commonly encountered in programs involves a single **interface** and a number of classes that implement that interface:

```

MusicalInstrument
  (interface)
 /   |   \   \
Violin Guitar Flute ...

```

Although there are more complicated hierarchies in practice, the simple hierarchy above solves many problems.

It is also common to move common implementation details of class hierarchies to intermediate classes. The subclasses inherit from these intermediate classes. The **StringInstrument** and **WindInstrument** classes below can contain the common members of their respective subclasses:

```

MusicalInstrument
  (interface)
 /       \
StringInstrument  WindInstrument
 /  |  \   /  |  \
Violin Viola ... Flute Clarinet ...

```

The subclasses would implement their respective special definitions of member functions.

59.9 Abstraction

Interfaces help make parts of programs independent from each other. This is called *abstraction*. For example, a program that deals with musical instruments can be written primarily by using the **MusicalInstrument** interface, without ever specifying the actual types of the musical instruments.

A **Musician** class can contain a **MusicalInstrument** without ever knowing the actual type of the instrument:

```
class Musician
{
    MusicalInstrument instrument;
    // ...
}
```

Different types of musical instruments can be combined in a collection without regard to the actual types of these instruments:

```
MusicalInstrument[] orchestraInstruments;
```

Most of the functions of the program can be written only by using this interface:

```

bool needsTuning(MusicalInstrument instrument)
{
    bool result;
    // ...
    return result;
}

void playInTune(MusicalInstrument instrument)
{
    if (needsTuning(instrument)){
        instrument.adjustTuning();
    }

    writeln(instrument.emitSound());
}

```

Abstracting away parts of a program from each other allows making changes in one part of the program without needing to modify the other parts. When implementations of certain parts of the program are *behind* a particular interface, the code that uses only that interface does not get affected.

59.10 Example

The following program defines the **SoundEmitter**, **MusicalInstrument**, and **CommunicationDevice** interfaces:

```

import std.stdio;

interface SoundEmitter
{
    string emitSound();
}

/* This class needs to implement only emitSound(). */
class Bell : SoundEmitter
{
    string emitSound()
    {
        return "ding";
    }
}

interface MusicalInstrument : SoundEmitter
{
    void adjustTuning();
}

/* This class needs to implement both emitSound() and
 * adjustTuning(). */
class Violin : MusicalInstrument
{
    string emitSound()
    {
        return "♪♪♪";
    }

    void adjustTuning()
    {
        // ... tuning of the violin ...
    }
}

interface CommunicationDevice
{
    void talk(string message);
    string listen();
}

```

```

}

/* This class needs to implement emitSound(), talk(), and
 * listen(). */
class Phone : SoundEmitter, CommunicationDevice
{
    string emitSound()
    {
        return "rrring";
    }

    void talk(string message)
    {
        // ... put the message on the line ...
    }

    string listen()
    {
        string soundOnTheLine;
        // ... get the message from the line ...
        return soundOnTheLine;
    }
}

class Clock
{
    // ... the implementation of Clock ...
}

/* This class needs to implement only emitSound(). */
class AlarmClock : Clock, SoundEmitter
{
    string emitSound()
    {
        return "beep";
    }

    // ... the implementation of AlarmClock ...
}

void main()
{
    SoundEmitter[] devices;

    devices ~= new Bell;
    devices ~= new Violin;
    devices ~= new Phone;
    devices ~= new AlarmClock;

    foreach (device; devices) {
        writeln(device.emitSound());
    }
}

```

Because **devices** is a **SoundEmitter** slice, it can contain objects of any type that inherits from **SoundEmitter** (i.e. types that have an "is a" relationship with **SoundEmitter**). As a result, the output of the program consists of different sounds that are emitted by the different types of objects:

```

ding
]]]]
rrring
beep

```

59.11 Summary

- **interface** defines an interface, which is similar to classes that consist only of abstract functions. **interface** can contain only **static** member variables and **static** or **final** member functions.
- In order to be able to construct objects of a class, that class must include implementations for all of the member functions of all of the interfaces that it inherits from.
- **interface** does not have a restriction of **class**: It is possible to inherit from unlimited number of **interfaces**.
- A common hierarchy consists of a single **interface** and a number of subclasses that implement that interface.

60 destroy and scoped

We have seen the lifetimes of objects in the [Lifetimes and Fundamental Operations](#) chapter.

In later chapters, we have seen that the objects are prepared for use in the constructor, which is called `this()`; and the final operations of objects are applied in the destructor, which is called `~this()`.

For structs and other value types, the destructor is executed at the time when the lifetime of an object ends. For classes and other reference types, it is executed by the garbage collector some time in the future. The important distinction is that the destructor of a class object is not executed when its lifetime ends.

System resources are commonly returned back to the system in destructors. For example, `std::stdio::File` returns the file resource back to the operating system in its destructor. As it is not certain when the destructor of a class object will be called, the system resources that it holds may not be returned until too late when other objects cannot get a hold of the same resource.

60.1 An example of calling destructors late

Let's define a class to see the effects of executing class destructors late. The following constructor increments a `static` counter, and the destructor decrements it. As you remember, there is only one of each `static` member, which is shared by all of the objects of a type. Such a counter would indicate the number of objects that are yet to be destroyed.

```
class LifetimeObserved
{
    int[] array;           // ← Belongs to each object
    static size_t counter; // ← Shared by all objects

    this()
    {
        /*
         * We are using a relatively large array to make each
         * object consume a large amount of memory. Hopefully
         * this will make the garbage collector call object
         * destructors more frequently to open up space for
         * more objects.
         */
        array.length = 30_000;

        /*
         * Increment the counter for this object that is being
         * constructed.
         */
        ++counter;
    }

    ~this()
    {
        /*
         * Decrement the counter for this object that is being
         * destroyed.
         */
        --counter;
    }
}
```

The following program constructs objects of that class inside a loop:

```
import std.stdio;

void main()
{
    foreach (i; 0 .. 20) {
        auto variable = new LifetimeObserved; // ← start
        write(LifetimeObserved.counter, ' ');
    } // ← end

    writeln();
}
```

The lifetime of each **LifetimeObserved** object is in fact very short: Its life starts when it is constructed by the **new** keyword and ends at the closing curly bracket of the **foreach** loop. Each object then becomes the responsibility of the garbage collector. The **start** and **end** comments indicate the start and end of the lifetimes.

Even though there is up to one object alive at a given time, the value of the counter indicates that the destructor is not executed when the lifetime ends:

```
1 2 3 4 5 6 7 8 2 3 4 5 6 7 8 2 3 4 5 6
```

According to that output, the memory sweep algorithm of the garbage collector has delayed executing the destructor for up to 8 objects. (*Note: The output may be different depending on the garbage collection algorithm, available memory, and other factors.*)

60.2 destroy() to execute the destructor

destroy() executes the destructor for an object:

```
void main()
{
    foreach (i; 0 .. 20) {
        auto variable = new LifetimeObserved;
        write(LifetimeObserved.counter, ' ');
        destroy(variable);
    }

    writeln();
}
```

Like before, the value of **LifetimeObserved.counter** is incremented by the constructor as a result of **new**, and becomes 1. This time, right after it gets printed, **destroy()** executes the destructor for the object and the value of the counter is decremented again down to zero. For that reason, this time its value is always 1:

```
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Once destroyed, the object should be considered to be in an invalid state and must not be used anymore:

```
destroy(variable);
// ...
// Warning: Using a potentially invalid object
writeln(variable.array);
```

Although **destroy()** is primarily for reference types, it can also be called on **struct** objects to destroy them before the end of their normal lifetimes.

60.3 When to use

As has been seen in the previous example, `destroy()` is used when resources need to be released at a specific time without relying on the garbage collector.

60.4 Example

We had designed an `XmlElement` struct in the [Constructor and Other Special Functions chapter](#). That struct was being used for printing XML elements in the format `<tag>value</tag>`. Printing the closing tag has been the responsibility of the destructor:

```
struct XmlElement
{
    // ...

    ~this()
    {
        writeln(indentation, "</", name, '>');
    }
}
```

The following output was produced by a program that used that struct. This time, I am replacing the word "class" with "course" to avoid confusing it with the `class` keyword:

```
<courses>
  <course0>
    <grade>
      72
    </grade> ← The closing tags appear on correct lines
    <grade>
      97
    </grade> ←
    <grade>
      90
    </grade> ←
  </course0> ←
  <course1>
    <grade>
      77
    </grade> ←
    <grade>
      87
    </grade> ←
    <grade>
      56
    </grade> ←
  </course1> ←
</courses> ←
```

The previous output happens to be correct because `XmlElement` is a `struct`. The desired output is achieved simply by placing the objects in appropriate scopes:

```
void main()
{
    immutable courses = XmlElement("courses", 0);
}
```

```

foreach (courseId; 0 .. 2) {
    immutable courseTag = "course" ~ toString(courseId);
    immutable courseElement = XmlElement(courseTag, 1);

    foreach (i; 0 .. 3) {
        immutable gradeElement = XmlElement("grade", 2);
        immutable randomGrade = uniform(50, 101);

        writeln(indentationString(3), randomGrade);

    } // ← gradeElement is destroyed
} // ← courseElement is destroyed
} // ← courses is destroyed

```

The destructor prints the closing tags as the objects gets destroyed.

To see how classes behave differently, let's convert **XmlElement** to a class:

```

import std.stdio;
import std.array;
import std.random;
import std.conv;

string indentationString(in int level)
{
    return replicate(" ", level * 2);
}

class XmlElement
{
    string name;
    string indentation;

    this(in string name, in int level)
    {
        this.name = name;
        this.indentation = indentationString(level);

        writeln(indentation, '<', name, '>');
    }

    ~this()
    {
        writeln(indentation, "</", name, ">");
    }
}

void main()
{
    immutable courses =
        new immutable(XmlElement)("courses", 0);

    foreach (courseId; 0 .. 2) {
        immutable courseTag = "course" ~ toString(courseId);
        immutable courseElement =
            new immutable(XmlElement)(courseTag, 1);

        foreach (i; 0 .. 3) {
            immutable gradeElement =
                new immutable(XmlElement)("grade", 2);
            immutable randomGrade = uniform(50, 101);

            writeln(indentationString(3), randomGrade);
        }
    }
}

```

As the responsibility of calling the destructors are now left to the garbage collector, the program does not produce the desired output:

```

<courses>
  <course0>
    <grade>
      57
    <grade>
      98
    <grade>
      87
  <course1>
    <grade>
      84
    <grade>
      60
    <grade>
      99
  </grade> ← The closing tags appear at the end
</grade> ←
</grade> ←
</course1> ←
</grade> ←
</grade> ←
</grade> ←
</course0> ←
</courses> ←

```

The destructor is still executed for every object but this time at the end when the program is exiting. (Note: The garbage collector does not guarantee that the destructor will be called for every object. In reality, it is possible that there are no closing tags printed at all.)

destroy() ensures that the destructor is called at desired points in the program:

```

void main()
{
  immutable courses =
    new immutable(XmlElement)("courses", 0);

  foreach (courseId; 0 .. 2) {
    immutable courseTag = "course" ~ toString(courseId);
    immutable courseElement =
      new immutable(XmlElement)(courseTag, 1);

    foreach (i; 0 .. 3) {
      immutable gradeElement =
        new immutable(XmlElement)("grade", 2);
      immutable randomGrade = uniform(50, 101);

      writeln(indentationString(3), randomGrade);
      destroy(gradeElement);
    }
    destroy(courseElement);
  }
  destroy(courses);
}

```

With those changes, the output of the code now matches the output of the code that use structs:

```

<courses>
  <course0>
    <grade>
      66
    </grade> ← The closing tags appear on correct lines
    <grade>
      75
    </grade> ←
    <grade>
      68
    </grade> ←
  </course0> ←
  <course1>
    <grade>
      73
    </grade> ←
    <grade>
      62
    </grade> ←
    <grade>
      100
    </grade> ←
  </course1> ←
</courses> ←

```

60.5 `scoped()` to call the destructor automatically

The program above has a weakness: The scopes may be exited before the `destroy()` lines are executed, commonly by thrown exceptions. If the `destroy()` lines must be executed even when exceptions are thrown, a solution is to take advantage of `scoped()` and other features that we have seen in the [Exceptions chapter](#).

Another solution is to construct class objects by `std::typecons::scoped` instead of by the `new` keyword. `scoped()` wraps the class object inside a `struct` and the destructor of that `struct` object destroys the class object when itself goes out of scope.

The effect of `scoped()` is to make class objects behave similar to struct objects regarding lifetimes.

With the following changes, the program produces the expected output as before:

```

import std.typecons;
// ...
void main()
{
  immutable courses =
    scoped!(immutable(XmlElement))("courses", 0);

  foreach (courseId; 0 .. 2) {
    immutable courseTag = "course" ~ toString(courseId);
    immutable courseElement =
      scoped!(immutable(XmlElement))(courseTag, 1);

    foreach (i; 0 .. 3) {

```

```

        immutable gradeElement =
            scoped!(immutable(XmlElement))("grade", 2);
        immutable randomGrade = uniform(50, 101);
        writeln(indentationString(3), randomGrade);
    }
}

```

Note that there are no **destroy()** lines anymore.

Also note that the **immutable** keywords made this program look more complicated than necessary. The code would be simpler with mutable objects:

```

auto courses = scoped!XmlElement("courses", 0);

```

60.6 Summary

- **destroy()** is for executing the destructor of a class object explicitly.
- Objects that are constructed by **scoped()** are destroyed upon leaving their respective scopes.

61 Modules and Libraries

The building blocks of D programs (and libraries) are modules.

D modules are based on a simple concept: Every source file is a module. Accordingly, the single files that we have been writing our programs in have all been individual modules.

By default, the name of a module is the same as its filename without the `.d` extension. When explicitly specified, the name of the module is defined by the `module` keyword, which must appear as the first non-comment line in the source file.

For example, assuming that the name of a source file is "cat.d", the name of the module would be specified by the `module` keyword:

```
module cat;

class Cat
{
    // ...
}
```

The `module` line is optional. When not specified, it is the same as the file name without the `.d` extension.

61.1 File and module names

D supports Unicode in source code and module names. However, the Unicode support of file systems vary. For example, although most Linux file systems support Unicode, the file names in Windows file systems may not distinguish between lower and upper case letters. Additionally, most file systems limit the characters that can be used in file and directory names.

For portability reasons, I recommend that you use only lower case ASCII letters in file names. For example, "resume.d" would be a suitable file name for a class named **Résumé**.

Accordingly, the name of the module would consist of ASCII letters as well:

```
module resume; // Module name consisting of ASCII letters

class Résumé // Program code consisting of Unicode characters
{
    // ...
}
```

61.2 Packages

A combination of related modules are called a *package*. D packages are a simple concept as well: The source files that are inside the same directory are considered to belong to the same package. The name of the directory becomes the name of the package, which must also be specified as the first parts of module names.

For example, if "cat.d" and "dog.d" are inside the directory "animal", then specifying the directory name along with the module name makes them be a part of the same package:

```
module animal.cat;

class Cat
{
```

```
} // ...
```

Similarly, for the **dog** module:

```
module animal.dog;
class Dog
{
    // ...
}
```

Since package names correspond to directory names, the package names of modules that are deeper than one directory level must reflect that hierarchy. For example, if the "animal" directory included a "vertebrate" directory, the name of a module inside that directory would include **vertebrate** as well:

```
module animal.vertebrate.cat;
```

The directory hierarchies can be arbitrarily complex depending on the needs of the program. Relatively small programs usually have all of their source files in a single directory.

61.3 Using modules in programs

The **import** keyword, which we have been using in almost every program so far, is for introducing a module to the current module:

```
import std.stdio;
```

The module name may contain the package name as well. For example, the **std.** part above indicates that **stdio** is a module that is a part of the **std** package.

The **animal.cat** and **animal.dog** modules would be imported similarly. Let's assume that the following code is inside a file named "deneme.d":

```
module deneme; // the name of this module
import animal.cat; // a module that it uses
import animal.dog; // another module that it uses

void main()
{
    auto cat = new Cat();
    auto dog = new Dog();
}
```

Note: As described below, for the program to be built correctly, those module files must also be provided to the linker.

61.4 Locations of modules

The compiler finds the module files by converting the package and module names directly to directory and file names.

For example, the previous two modules would be located as "animal/cat.d" and "animal/dog.d", respectively (or "animal\cat.d" and "animal\dog.d", depending on the file system). Considering the main source file as well, the program above consists of three files.

61.5 Long and short module names

The names that are used in the program may be spelled out with the module and package names:

```
auto cat0 = Cat();
auto cat1 = animal.cat.Cat(); // same as above
```

The long names are normally not needed but sometimes there are name conflicts. For example, when referring to a name that appears in more than one module, the compiler cannot decide which one is meant.

The following program is spelling out the long names to distinguish between two separate **Jaguar** structs that are defined in two separate modules: **animal** and **car**:

```
import animal.jaguar;
import car.jaguar;

// ...

auto conflicted = Jaguar(); // ← compilation ERROR

auto myAnimal = animal.jaguar.Jaguar(); // ← compiles
auto myCar = car.jaguar.Jaguar(); // ← compiles
```

61.6 Adding module definitions to the program

The **import** keyword is not sufficient to make modules become parts of the program. It simply makes available the features of a module inside the current module. That much is needed only to *compile* the code.

It is not possible to build the previous program only by the main source file, "deneme.d":

```
$ dmd deneme.d -w
deneme.o: In function `_Dmain':
deneme.d: undefined reference to `_D6animal3cat3Cat7__ClassZ'
deneme.d: undefined reference to `_D6animal3dog3Dog7__ClassZ'
collect2: ld returned 1 exit status
--- errorlevel 1
```

Those error messages are generated by the *linker*. Although they are not user-friendly messages, they indicate that some definitions that are needed by the program are missing.

The actual build of the program is the responsibility of the linker, which gets called automatically by the compiler behind the scenes. The compiler passes the modules that it has just compiled to the linker, and the linker combines those modules (and libraries) to produce the executable program.

For that reason, all of the modules that make up the program must be provided to the linker. For the program above to be built, "animal/cat.d" and "animal/dog.d" must also be specified on the compilation line:

```
$ dmd deneme.d animal/cat.d animal/dog.d -w
```

Instead of having to mention the modules individually every time on the command line, they can be combined as libraries.

61.7 Libraries

A collection of compiled modules is called a library. Libraries are not programs themselves; they do not have the `main()` function. Libraries contain compiled definitions of functions, structs, classes, and other features of modules, which are to be linked later by the linker to produce the program.

dmd's `-lib` command line option is for making libraries. The following command makes a library that contains the "cat.d" and the "dog.d" modules. The name of the library is specified by the `-of` switch:

```
$ dmd animal/cat.d animal/dog.d -lib -ofanimal -w
```

The actual name of the library file depends on the platform. For example, the extension of library files is `.a` under Linux systems: `animal.a`.

Once that library is built, it is not necessary to specify the "animal/cat.d" and "animal/dog.d" modules individually anymore. The library file is sufficient:

```
$ dmd deneme.d animal.a -w
```

The command above replaces the following one:

```
$ dmd deneme.d animal/cat.d animal/dog.d -w
```

As an exception, the D standard library Phobos need not be specified on the command line. That library is automatically included behind the scenes. Otherwise, it could be specified similar to the following line:

```
$ dmd deneme.d animal.a /usr/lib64/libphobos2.a -w
```

Note: *The name and location of the Phobos library may be different on different systems.*

62 Encapsulation and Protection Attributes

All of the structs and classes that we have defined so far have all been accessible from the *outside*.

Let's consider the following struct:

```
enum Gender { female, male }

struct Student
{
    string name;
    Gender gender;
}
```

The members of that struct is freely accessible to the rest of the program:

```
auto student = Student("Tim", Gender.male);
writeln("%s is a %s student.", student.name, student.gender);
```

Such freedom is a convenience in programs. For example, the previous line has been useful to produce the following output:

```
Tim is a male student.
```

However, this freedom is also a liability. As an example, let's assume that perhaps by mistake, the name of a student object gets modified in the program:

```
student.name = "Anna";
```

That assignment may put the object in an invalid state:

```
Anna is a male student.
```

As another example, let's consider a **School** class. Let's assume that this class has two member variables that store the numbers of the male and female students separately:

```
class School
{
    Student[] students;
    size_t femaleCount;
    size_t maleCount;

    void add(in Student student)
    {
        students ~= student;

        final switch (student.gender) {
            case Gender.female:
                ++femaleCount;
                break;

            case Gender.male:
                ++maleCount;
                break;
        }
    }

    override string toString() const
    {
        return format("%s female, %s male; total %s students",
```

```

        femaleCount, maleCount, students.length);
    }
}

```

The `add()` member function adds students while ensuring that the counts are always correct:

```

auto school = new School;
school.add(Student("Lindsey", Gender.female));
school.add(Student("Mark", Gender.male));
writeln(school);

```

The program produces the following consistent output:

```
1 female, 1 male; total 2 students
```

However, being able to access the members of `School` freely would not guarantee that this consistency would always be maintained. Let's consider adding a new element to the `students` member, this time directly:

```
school.students ~= Student("Nancy", Gender.female);
```

Because the new student has been added to the array directly, without going through the `add()` member function, the `School` object is now in an inconsistent state:

```
1 female, 1 male; total 3 students
```

62.1 Encapsulation

Encapsulation is a programming concept of restricting access to members to avoid problems similar to the one above.

Another benefit of encapsulation is to eliminate the need to know the implementation details of types. In a sense, encapsulation allows presenting a type as a black box that is used only through its interface.

Additionally, preventing users from accessing the members directly allows changing the members of a class freely in the future. As long as the functions that define the interface of a class is kept the same, its implementation can be changed freely.

Encapsulation is not for restricting access to sensitive data like a credit card number or a password, and it cannot be used for that purpose. Encapsulation is a development tool: It allows using and coding types easily and safely.

62.2 Protection attributes

Protection attributes limit access to members of structs, classes, and modules. There are two ways of specifying protection attributes:

- At struct or class level to specify the protection of every struct or class member individually.
- At module level to specify the protection of every feature of a module individually: class, struct, function, enum, etc.

Protection attributes can be specified by the following keywords. The default attribute is **public**.

- **public**: Specifies accessibility by any part of the program without any restriction.
An example of this is **stdout**. Merely importing **std.stdio** makes **stdout** available to every module that imported it.
- **private**: Specifies restricted accessibility.
private class members and module members can only be accessed by the module that defines that member.

Additionally, **private** member functions cannot be overridden by subclasses.

- **package**: Specifies package-level accessibility.
A feature that is marked as **package** can be accessed by all of the code that is a part of the same package. The **package** attribute involves only the inner-most package.

For example, a **package** definition that is inside the **animal.vertebrate.cat** module can be accessed by any other module of the **vertebrate** package.

Similar to the **private** attribute, **package** member functions cannot be overridden by subclasses.
- **protected**: Specifies accessibility by derived classes.
This attribute extends the **private** attribute: A **protected** member can be accessed not only by the module that defines it, but also by the classes that inherit from the class that defines that **protected** member.

Additionally, the **export** attribute specifies accessibility from the outside of the program.

62.3 Definition

Protection attributes can be specified in three ways.

When written in front of a single definition, it specifies the protection attribute of that definition only. This is similar to the Java programming language:

```
private int foo;

private void bar()
{
    // ...
}
```

When specified by a colon, it specifies the protection attributes of all of the following definitions until the next specification of a protection attribute. This is similar to the C++ programming language:

```
private:
    // ...
    // ... all of the definitions here are private ...
    // ...

protected:
    // ...
    // ... all of the definitions here are protected ...
    // ...
```

When specified for a block, the protection attribute is for all of the definitions that are inside that block:

```
private {
    // ...
}
```

```

    // ... all of the definitions here are private ...
    // ...
}

```

62.4 Module imports are private by default

A module that is imported by **import** is private to the module that imports it. It would not be visible to other modules that import it indirectly. For example, if a **school** module imports **std.stdio**, modules that import **school** cannot automatically use the **std.stdio** module.

Let's assume that the **school** module starts by the following lines:

```

module school.school;

import std.stdio;    // imported for this module's own use...

// ...

```

The following program cannot be compiled because **writeln** is not visible to it:

```

import school.school;

void main()
{
    writeln("hello");    // ← compilation ERROR
}

```

std.stdio must be imported by that module as well:

```

import school.school;
import std.stdio;

void main()
{
    writeln("hello");    // now compiles
}

```

Sometimes it is desired that a module presents other modules indirectly. For example, it would make sense for a **school** module to automatically import a **student** module for its users. This is achieved by marking the **import** as **public**:

```

module school.school;

public import school.student;

// ...

```

With that definition, modules that import **school** can use the definitions that are inside the **student** module without needing to import it:

```

import school.school;

void main()
{
    auto student = Student("Tim", Gender.male);

    // ...
}

```

Although the program above imports only the **school** module, the **student.Student** struct is also available to it.

62.5 When to use encapsulation

Encapsulation avoids problems similar to the one we have seen in the introduction section of this chapter. It is an invaluable tool to ensure that objects are always in consistent states. Encapsulation helps preserve struct and class *invariants* by protecting members from direct modifications by the users of the type.

Encapsulation provides freedom of implementation by abstracting implementations away from user code. Otherwise, if users had direct access to for example `School.students`, it would be hard to modify the design of the class by changing that array e.g. to an associative array, because this would affect all user code that has been accessing that member.

Encapsulation is one of the most powerful benefits of object oriented programming.

62.6 Example

Let's define the `Student` struct and the `School` class by taking advantage of encapsulation and let's use them in a small test program.

This example program will consist of three files. As you remember from the previous chapter, being parts of the `school` package, two of these files will be under the "school" directory:

- "school/student.d": The `student` module that defines the `Student` struct
- "school/school.d": The `school` module that defines the `School` class
- "deneme.d": A small test program

Here is the "school/student.d" file:

```
module school.student;

import std.string;
import std.conv;

enum Gender { female, male }

struct Student
{
    package string name;
    package Gender gender;

    string toString() const
    {
        return format("%s is a %s student.",
                      name, to!string(gender));
    }
}
```

The members of this struct are marked as `package` to enable access only to modules of the same package. We will soon see that `School` will be accessing these members directly. (Note that even this should be considered as violating the principle of encapsulation. Still, let's stick with the `package` attribute in this example program.)

The following is the "school/school.d" module that makes use of the previous one:

```
module school.school;

public import school.student; // 1

import std.string;

class School
{
```

```

private: // 2
    Student[] students;
    size_t femaleCount;
    size_t maleCount;

public: // 3
    void add(in Student student)
    {
        students ~= student;

        final switch (student.gender) { // 4a
        case Gender.female:
            ++femaleCount;
            break;

        case Gender.male:
            ++maleCount;
            break;
        }
    }

    override string toString() const
    {
        string result = format(
            "%s female, %s male; total %s students",
            femaleCount, maleCount, students.length);

        foreach (i, student; students) {
            result ~= (i == 0) ? ": " : ", ";
            result ~= student.name; // 4b
        }

        return result;
    }
}

```

1. **school.student** is being imported publicly so that the users of **school.school** will not need to import that module explicitly. In a sense, the **student** module is made available by the **school** module.
2. All of the member variables of **School** are marked as private. This is important to help protect the consistency of the member variables of this class.
3. For this class to be useful, it must present some member functions. **add()** and **toString()** are made available to the users of this class.
4. As the two member variables of **Student** have been marked as **package**, being a part of the same package, **School** can access those variables.

Finally, the following is a test program that uses those types:

```

import std.stdio;
import school.school;

void main()
{
    auto student = Student("Tim", Gender.male);
    writeln(student);

    auto school = new School;

    school.add(Student("Lindsey", Gender.female));
    school.add(Student("Mark", Gender.male));
    school.add(Student("Nancy", Gender.female));

    writeln(school);
}

```

This program can use **Student** and **School** only through their public interfaces. It cannot access the member variables of those types. As a result, the objects would always be consistent:

```
Tim is a male student.  
2 female, 1 male; total 3 students: Lindsey, Mark, Nancy
```

Note that the program interacts with **School** only by its **add()** and **toString()** functions. As long as the interfaces of these functions are kept the same, changes in their implementations would not affect the program above.

63 Skipping chapters...

There are the following chapters of the book at this point but they haven't been translated from Turkish yet:

- [Nitelikler](#) @property
- [Yapı ve Sınıflarda Sözleşmeli Programlama](#) invariant

Please click [Next →] above or below for the next available chapter in English.

64 Templates

Templates are the feature that allows describing the code as a pattern, for the compiler to generate program code automatically. Parts of the source code may be left to the compiler to be filled in until that part is actually used in the program. The compiler fills in what is left for later.

Templates are very useful especially in libraries because they enable writing generic algorithms and data structures, instead of tying them to specific types.

Compared to the template supports in other languages, D's templates are very powerful and extensive. I will not get into all of the details of templates in this chapter. I will cover only function, struct, and class templates and only *type* template parameters. For a complete reference on D templates, see [Philippe Sigaud's D Templates: A Tutorial](#).

To see the benefits of templates let's start with a function that prints values in parentheses:

```
void printInParens(int value)
{
    writeln("(%s)", value);
}
```

Because the parameter is specified as **int**, that function can only be used with values of type **int**, or values that can automatically be converted to **int**. For example, the compiler would not allow calling it with a floating point type.

Let's assume that the requirements of a program changes and that other types need to be printed in parentheses as well. One of the solutions for this would be to take advantage of function overloading and provide overloads of the function for all of the types that the function is used with:

```
// The function that already exists
void printInParens(int value)
{
    writeln("(%s)", value);
}

// Overloading the function for 'double'
void printInParens(double value)
{
    writeln("(%s)", value);
}
```

This solution does not scale well because this time the function cannot be used with e.g. **real** or any user-defined type. Although it is possible to overload the function for other types, the cost of doing this would be prohibitive.

An important observation here is that regardless of the type of the parameter, the contents of the overloads would all be *generically* the same: a single **writeln()** expression.

Such duplication is common in algorithms and data structures. For example, the binary search algorithm is independent of the type of the elements: it is about the specific steps and operations of the search. Similarly, the linked list data structure is independent of the type of the elements: linked list is merely about *how* the elements are stored in the container, regardless of their types.

Templates are useful in such situations: once a piece of code is described as a template, the compiler generates overloads of the same code automatically according to the actual uses of that code in the program.

As I have mentioned above, in this chapter I will cover only function, struct, and class templates, and *type* template parameters.

64.1 Function templates

Defining a function as a template is leaving one or more of the types that it uses as unspecified, to be deduced later by the compiler.

The types that are being left unspecified are defined within the template parameter list, which comes between the name of the function and the function parameter list. For that reason, function templates have two parameter lists: the template parameter list and the function parameter list:

```
void printInParens(T)(T value)
{
    writefln("%s", value);
}
```

The **T** within the template parameter list above means that **T** can be any type. Although **T** is an arbitrary name, it is an acronym for "type" and is very common in templates.

Since **T** represents any type, the templated definition of **printInParens()** above is sufficient to use it with almost every type, including the user-defined ones:

```
import std.stdio;

void printInParens(T)(T value)
{
    writefln("%s", value);
}

void main()
{
    printInParens(42);           // with int
    printInParens(1.2);        // with double

    auto myValue = MyStruct();
    printInParens(myValue);    // with MyStruct
}

struct MyStruct
{
    string toString() const
    {
        return "hello";
    }
}
```

The compiler considers all of the uses of **printInParens()** in the program and generates code to support all those uses. The program is then compiled as if the function has been overloaded explicitly for **int**, **double**, and **MyStruct**:

```

/*
 * Note: These functions are not part of the source
 *       code. They are the equivalents of the functions that
 *       the compiler would automatically generate.
 */

void printInParens(int value)
{
    writefln("(%s)", value);
}

void printInParens(double value)
{
    writefln("(%s)", value);
}

void printInParens(MyStruct value)
{
    writefln("(%s)", value);
}

```

The output of the program contains different lines from the different *instantiations* by the compiler of the function template:

```

(42)
(1.2)
(hello)

```

64.2 More than one template parameter

Let's change the function template to take the characters to use for the parentheses as parameters:

```

void printInParens(T)(T value, char opening, char closing)
{
    writeln(opening, value, closing);
}

```

Now we can call the same function with different sets of parentheses:

```
printInParens(42, '<', '>');
```

Although being able to specify the parentheses makes the function more usable, specifying the type of the parentheses as **char** makes it less flexible because it is not possible to call the function with characters of type **wchar** or **dchar**:

```
printInParens(42, '→', '←'); // ← compilation ERROR
```

```

Error: template deneme.printInParens(T) cannot deduce
template function from argument types !()(int,wchar,wchar)

```

One solution would be to specify the type of the parentheses as **dchar** but this would still be insufficient as this time the function could not be called e.g. with **string** or user-defined types.

Another solution is to leave the type of the parentheses to the compiler as well. Defining an additional template parameter instead of the specific **char** is sufficient:

```
void printInParens(T, ParenType)(T value,
                               ParenType opening,
                               ParenType closing)
{
    writeln(opening, value, closing);
}
```

The meaning of the new template parameter is similar to **T**'s: **ParenType** can be any type.

It is now possible to use many different types of parentheses. The following are with **wchar** and **string**:

```
printInParens(42, '→', '←');
printInParens(1.2, "-=", "=");
```

```
→42←
-=1.2=-
```

The flexibility of **printInParens()** has been increased, as it now works correctly for any combination of **T** and **ParenType** as long as those types are printable with **writeln()**.

64.3 Type deduction

The compiler's deciding on what type to use for a template parameter is called *type deduction*.

Continuing from the last example above, the compiler decides on the following types according to the two uses of the function template:

- **int** and **wchar** when 42 is printed
- **double** and **string** when 1.2 is printed

The compiler can deduce types only from the types of the parameter values that are passed to function templates. Although the compiler can usually deduce the types without any ambiguity, there are times when the types must be specified explicitly by the programmer.

64.4 Explicit type specification

Sometimes it is not possible for the compiler to deduce the template parameters. One occasion that this can happen is when the types do not appear in the function parameter list. When template parameters are not related to function parameters, the compiler cannot deduce the template parameter types.

To see an example of this, let's design a function that asks a question to the user, reads a value as a response, and returns that response. Additionally, let's make this a function template so that it can be used to read any type of response:

```
T getResponse(T)(string question)
{
    writef("%s (%s): ", question, T.stringof);

    T response;
    readf(" %s", &response);

    return response;
}
```

That function template would be very useful in programs to read different types of values from the input. For example, to read some user information, we may imagine calling it like the following:

```
getResponse("What is your age?");
```

Unfortunately, that call does not give the compiler any clue as to what the template parameter **T** should be. What is known is that the question is passed to the function as a **string**, but the type of the return value cannot be deduced:

```
Error: template deneme.getResponse(T) cannot deduce template
function from argument types !()(string)
```

In such cases, the template parameters must be specified explicitly by the programmer. Template parameters are specified in parentheses after an exclamation mark:

```
getResponse!(int)("What is your age?");
```

The code above can now be accepted by the compiler and the function template is compiled as **T** referring to **int** within the definition of the template.

When there is only one template parameter specified, the parentheses around it are optional:

```
getResponse!int("What is your age?"); // same as above
```

You may recognize that syntax from **to!string**, which we have been using in earlier programs. **to()** is a function template, which takes the target type of the conversion as a template parameter. Since it has only one template parameter that needs to be specified, it is commonly written as **to!string** instead of **to!(string)**.

64.5 Template instantiation

Automatic generation of code for different sets of template parameter types is called an *instantiation* of that template for that specific set of types. For example, **to!string** and **to!int** are two distinct instantiations of the **to** function template.

As I will mention again in a separate section below, distinct instantiations of templates produce distinct and incompatible types.

64.6 Template specializations

Although the **getResponse()** function template can in theory be used for any template type, the code that the compiler generates may not be suitable for every type. Let's assume that we have the following type to represent points on a surface:

```
struct Point
{
    int x;
    int y;
}
```

Although the instantiation of **getResponse()** for the **Point** type itself would be fine, the generated **readf()** call for **Point** cannot be compiled. This is because the standard library function **readf()** does not know how to read a **Point** object. The two lines that actually read

the response would look like the following in the **Point** instantiation of the **getResponse()** function template:

```
Point response;
readf(" %s", &response); // ← compilation ERROR
```

One way of reading a **Point** object would be to read the values of the **x** and **y** members separately and then to *construct* a **Point** object with those values.

Providing a special definition of a template for a specific type is called *template specialization*. The specialization is specified with the type name after a **:** character in the template parameter list. A **Point** specialization of the **getResponse()** function template can be defined as the following:

```
// The general definition of the function template
T getResponse(T)(string question)
{
    // ...
}

// The specialization of the function template for Point
T getResponse(T : Point)(string question)
{
    writeln("%s (Point)", question);

    auto x = getResponse!int(" x");
    auto y = getResponse!int(" y");

    return Point(x, y);
}
```

Note that the specialization takes advantage of the default definition of **getResponse()** to read two **int** values to be used as the values of the **x** and **y** members.

Instead of instantiating the template itself, now the compiler uses the specialization above whenever **getResponse()** is called for the **Point** type:

```
auto center = getResponse!Point("Where is the center?");
```

Assuming that the user enters 11 and 22:

```
Where is the center? (Point)
x (int): 11
y (int): 22
```

The **getResponse!int()** calls are directed to the general definition of the template and the **getResponse!Point()** calls are directed to the **Point** specialization of it. For that reason, both the general and special definitions of the template must be present for the code to be compiled and work correctly.

As another example, let's consider using the same template with **string**. As you remember from the [Strings chapter](#), **readf()** would read all of the characters from the input as part of a single **string** until the end of input. For that reason, the default definition of **getResponse()** would not be useful when reading **string** responses:

```
// Reads the entire input, not only the name!
auto name = getResponse!string("What is your name?");
```

We can provide a template specialization for **string** as well. The following specialization reads the entire *line* instead:

```
T getResponse(T : string)(string question)
{
    writef("%s (string): ", question);

    // Read and ignore whitespace characters which have
    // presumably been left-over from the previous user input
    string response;
    do {
        response =.chomp(readln());
    } while (response.length == 0);

    return response;
}
```

64.7 Struct and class templates

The **Point** struct may be seen as having a limitation: as its two members are defined specifically as **int**, it cannot represent fractional coordinate values. This limitation can be removed if the **Point** struct is defined as a template.

Let's first define a member function that returns the distance to another **Point** object:

```
import std.math;

// ...

struct Point
{
    int x;
    int y;

    int distanceTo(in Point that) const
    {
        immutable real xDistance = x - that.x;
        immutable real yDistance = y - that.y;

        immutable distance = sqrt((xDistance * xDistance) +
                                   (yDistance * yDistance));

        return cast(int)distance;
    }
}
```

Point is suitable when the required precision is relatively low: It can calculate the distance between two points at kilometer precision, e.g. between the center and branch offices of an organization:

```
auto center = getResponse!Point("Where is the center?");
auto branch = getResponse!Point("Where is the branch?");

writeln("Distance: ", center.distanceTo(branch));
```

Unfortunately, **Point** is inadequate at higher precisions than **int** can provide.

structs and classes can be defined as templates too by specifying a template parameter list after their names. For example, **Point** can be defined as a struct template by providing a template parameter and replacing the **ints** by that parameter:

```
struct Point(T)
{
```



```

T x;
T y;

T distanceTo(in Point that) const
{
    immutable real xDistance = x - that.x;
    immutable real yDistance = y - that.y;

    immutable distance = sqrt((xDistance * xDistance) +
                               (yDistance * yDistance));

    return cast(T)distance;
}

```

Since structs and classes are not functions, they cannot be called with parameters. This makes it impossible for the compiler to deduce their template parameters. The template parameter list must always be specified for struct and class templates:

```

auto center = Point!int(0, 0);
auto branch = Point!int(100, 100);

writeln("Distance: ", center.distanceTo(branch));

```

The definitions above make the compiler generate code for the `int` instantiation of the `Point` template, which is the equivalent of its earlier non-template definition. Naturally, now it can be used with any type. For example, when more precision is needed, with `double`:

```

auto point1 = Point!double(1.2, 3.4);
auto point2 = Point!double(5.6, 7.8);

writeln(point1.distanceTo(point2));

```

Although the template itself has been defined independently of any specific type, its single definition makes it possible to represent points of various precisions.

Simply converting `Point` to a template would cause compilation errors in code that has already been written according to its non-template definition. For example, now the `Point` specialization of `getResponse()` cannot be compiled:

```

T getResponse(T : Point)(string question)    // ← compilation ERROR
{
    writeln("%s (Point)", question);

    auto x = getResponse!int(" x");
    auto y = getResponse!int(" y");

    return Point(x, y);
}

```

The reason for the compilation error is that `Point` itself is not a type anymore: `Point` is now a *struct template*. Only instantiations of that template would be considered as types. The following changes are required to correctly specialize `getResponse()` for any instantiation of `Point`:

```

Point!T getResponse(T : Point!T)(string question)    // 2, 1
{
    writeln("%s (Point!%s)", question, T.stringof); // 5

    auto x = getResponse!T(" x");                  // 3a
    auto y = getResponse!T(" y");                  // 3b
}

```

```

    return Point!T(x, y); // 4
}

```

1. In order for this template specialization to support all instantiations of **Point**, the template parameter list must mention **Point!T**. This simply means that the **getResponse()** specialization is for **Point!T**, regardless of **T**. This specialization would match **Point!int**, **Point!double**, etc.
2. Similarly, to return the correct type as the response, the return type must be specified as **Point!T** as well.
3. Since the types of **x** and **y** members of **Point!T** are now **T**, as opposed to **int**, the members must be read by calling **getResponse!T()**, not **getResponse!int()**, as the latter would be correct only for **Point!int**.
4. Similar to items 1 and 2, the type of the return value is **Point!T**.
5. To print the name of the type accurately for every type, as in **Point!int**, **Point!double**, etc., **T.stringof** is used.

64.8 Default template parameters

Sometimes it is cumbersome to provide template parameter types every time a template is used, especially when that type is almost always one particular type. For example, **getResponse()** may almost always be called for the **int** type in the program, and only in a few places for the **double** type.

It is possible to specify default types for template parameters, which are assumed when the types are not explicitly provided. Default parameter types are specified after the = character:

```

T getResponse(T = int)(string question)
{
    // ...
}
// ...
auto age = getResponse("What is your age?");

```

As no type has been specified when calling **getResponse()** above, **T** becomes the default type **int** and the call ends up being the equivalent of **getResponse!int()**.

Default template parameters can be specified for struct and class templates as well, but in their case the template parameter list must always be written even when empty:

```

struct Point(T = int)
{
    // ...
}
// ...
Point!() center;

```

Similar to default function parameter values as we have seen in the [Variable Number of Parameters](#) chapter, default template parameters can be specified for all of the template parameters or for the *last* ones:

```

void myTemplate(T0, T1 = int, T2 = char)()
{
    // ...
}

```

The last two template parameters of that function may be left unspecified but the first one is required:

```
myTemplate!string();
```

In that usage, the second and third parameters are **int** and **char**, respectively.

64.9 Every different template instantiation is a distinct type

Every instantiation of a template for a given set of types is considered to be a distinct type. For example, **Point!int** and **Point!double** are separate types:

```
Point!int point3 = Point!double(0.25, 0.75); // ← compilation ERROR
```

The two different types cannot be assigned:

```
Error: cannot implicitly convert expression (Point(0.25,0.75))
of type Point!(double) to Point!(int)
```

64.10 A compile time feature

Templates are entirely a compile time feature. The instances of templates are generated by the compiler at compile time. They cannot be generated at runtime.

64.11 Class template example: stack data structure

Struct and class templates are commonly used in the implementations of data structures. Let's design a stack container that will be able to contain any type.

Stack is one of the simplest data structures. It represents a container where elements are placed conceptually on top of each other as would be in a stack of papers. New elements go on top, and only the topmost element is accessed. When an element is removed, it is always the topmost one.

If we also define a property that returns the total number of elements in the stack, all of the operations of this data structure would be the following:

- Add element (**push()**)
- Remove element (**pop()**)
- Access the topmost element (**.top**)
- Report the number of elements (**.length**)

An array can be used to store the elements such that the last element of the array would be representing the topmost element of the stack. Finally, it can be defined as a class template to be able to contain elements of any type:

```
class Stack(T)
{
    T[] elements;
public:
    void push(T element)
    {
        elements ~= element;
    }
}
```

```

void pop()
{
    --elements.length;
}

@property T top() const
{
    return elements[$ - 1];
}

@property size_t length() const
{
    return elements.length;
}
}

```

As a design decision, `push()` and `pop()` are defined as regular member functions, and `.top` and `.length` are defined as properties because they can be seen as providing simple informations about the stack collection.

Here is a `unittest` block for this class that uses its `int` instantiation:

```

unittest
{
    auto stack = new Stack!int;

    // The newly added element must appear on top
    stack.push(42);
    assert(stack.top == 42);
    assert(stack.length == 1);

    // .top and .length should not affect the elements
    assert(stack.top == 42);
    assert(stack.length == 1);

    // The newly added element must appear on top
    stack.push(100);
    assert(stack.top == 100);
    assert(stack.length == 2);

    // Removing the last element must expose the previous one
    stack.pop();
    assert(stack.top == 42);
    assert(stack.length == 1);

    // The stack must become empty when the last element is
    // removed
    stack.pop();
    assert(stack.length == 0);
}

```

To take advantage of this class template, let's try using it this time with a user-defined type. As an example, here is a modified version of `Point`:

```

struct Point(T)
{
    T x;
    T y;

    string toString() const
    {
        return format("%s,%s", x, y);
    }
}

```

A `Stack` that contains elements of type `Point!double` can be defined like the following:

```
auto points = new Stack!(Point!double);
```

Here is a test program that first adds ten elements to this stack and then removes them one by one:

```
import std.string;
import std.stdio;
import std.random;

struct Point(T)
{
    T x;
    T y;

    string toString() const
    {
        return format("%s,%s", x, y);
    }
}

// Returns a random value between -0.50 and 0.50.
double random_double()
out (result)
{
    assert((result >= -0.50) && (result < 0.50));
}
body
{
    return (cast(double)uniform(0, 100) - 50) / 100;
}

// Returns a Stack that contains 'count' number of random
// Point!double elements.
Stack!(Point!double) randomPoints(int count)
in
{
    assert(count >= 0);
}
out (result)
{
    assert(result.length == count);
}
body
{
    auto points = new Stack!(Point!double);

    foreach (i; 0 .. count) {
        immutable point = Point!double(random_double(),
                                         random_double());
        writeln("adding :", point);
        points.push(point);
    }

    return points;
}

void main()
{
    auto stackedPoints = randomPoints(10);

    while (stackedPoints.length) {
        writeln("removing: ", stackedPoints.top);
        stackedPoints.pop();
    }
}
```

As the output of the program shows, the elements are removed in the reverse order as they have been added:

```

adding : (-0.02,-0.01)
adding : (0.17,-0.5)
adding : (0.12,0.23)
adding : (-0.05,-0.47)
adding : (-0.19,-0.11)
adding : (0.42,-0.32)
adding : (0.48,-0.49)
adding : (0.35,0.38)
adding : (-0.2,-0.32)
adding : (0.34,0.27)
removing: (0.34,0.27)
removing: (-0.2,-0.32)
removing: (0.35,0.38)
removing: (0.48,-0.49)
removing: (0.42,-0.32)
removing: (-0.19,-0.11)
removing: (-0.05,-0.47)
removing: (0.12,0.23)
removing: (0.17,-0.5)
removing: (-0.02,-0.01)

```

64.12 Function template example: binary search algorithm

Binary search is the fastest algorithm to search for an element among the elements of an already sorted array. It is a very simple algorithm: The element in the middle is considered; if that element is the one that has been sought, then the search is over. If not, then the algorithm is repeated on the elements that are either on the left-hand side or on the right-hand side of the middle element, depending on whether the sought element is greater or less than the middle element.

Algorithms that repeat themselves on a smaller range of the initial elements are recursive. Let's write the binary search algorithm recursively by calling itself from within its implementation.

Before converting it to a template, let's first write this function to support only arrays of `int`. We can easily convert it to a template later, by adding a template parameter list and replacing appropriate `ints` in its definition by `Ts`. Here is a binary search algorithm that works on arrays of `int`:

```

/*
 * This function returns the index of the value if it exists
 * in the array, size_t.max otherwise.
 */
size_t binarySearch(const int[] values, in int value)
{
    // The value is not in the array if the array is empty.
    if (values.length == 0) {
        return size_t.max;
    }

    immutable midPoint = values.length / 2;

    if (value == values[midPoint]) {
        // Found.
        return midPoint;
    } else if (value < values[midPoint]) {

```

```

    // The value can only be in the left-hand side; let's
    // search in a slice that represents that half.
    return binarySearch(values[0 .. midPoint], value);

} else {
    // The value can only be in the right-hand side; let's
    // search in the right-hand side.
    auto index =
        binarySearch(values[midPoint + 1 .. $], value);

    if (index != size_t.max) {
        // Adjust the index; it is 0-based in the
        // right-hand side slice.
        index += midPoint + 1;
    }

    return index;
}

assert(false, "We should have never gotten to this line");
}

```

The function above implements this simple algorithm in four steps:

- If the array is empty, return `size_t.max` to indicate that the value has not been found.
- If the element at the mid-point is equal to the sought value, then return the index of that element.
- If the value is less than the element at the mid-point, then repeat the same algorithm on the left-hand side.
- Else, repeat the same algorithm on the right-hand side.

Here is a unittest block that tests the function:

```

unittest
{
    auto array = [ 1, 2, 3, 5 ];
    assert(binarySearch(array, 0) == size_t.max);
    assert(binarySearch(array, 1) == 0);
    assert(binarySearch(array, 4) == size_t.max);
    assert(binarySearch(array, 5) == 3);
    assert(binarySearch(array, 6) == size_t.max);
}

```

Now that the function has been implemented and tested for `int`, we can convert it to a template. `int` appears only in two places in the definition of the template:

```

size_t binarySearch(const int[] values, in int value)
{
    // ... int does not appear here ...
}

```

The `ints` that appear in the parameter list are the types of the elements and the value. Specifying those as template parameters is sufficient to make this algorithm a template and to be usable with other types as well:

```

size_t binarySearch(T)(const T[] values, in T value)
{
    // ...
}

```

That function template can be used with any type that matches the operations that are applied to that type in the template. In `binarySearch()`, the elements are used only with comparison operators `==` and `<`:

```

if (value == values[midPoint]) {
    // ...
} else if (value < values[midPoint]) {
    // ...
}

```

For that reason, **Point** is not ready to be used with **binarySearch()** yet:

```

struct Point(T)
{
    T x;
    T y;

    string toString() const
    {
        return format("%s,%s", x, y);
    }
}

void main()
{
    Point!int[] points;

    foreach (i; 0 .. 15) {
        points ~= Point!int(i, i);
    }

    assert(binarySearch(points, Point!int(10, 10)) == 10);
}

```

The program above would cause a compilation error:

```

Error: need member function opCmp() for struct
const(Point!(int)) to compare

```

According to the error message, **opCmp()** needs to be defined for **Point**, which has been covered in the [Operator Overloading chapter](#):

```

struct Point(T)
{
    // ...

    int opCmp(const ref Point that) const
    {
        return (x == that.x
            ? y - that.y
            : x - that.x);
    }
}

```

64.13 Summary

We will see the other features of templates in a later chapter. For a complete reference on D templates, see [Philippe Sigaud's D Templates: A Tutorial](#). The following are what we have covered in this chapter:

- Templates define the code as a pattern, for the compiler to generate instances of it according to the actual uses in the program.
- Templates are a compile-time feature.
- Specifying template parameter lists is sufficient to make function, struct, and class definitions templates.


```

void functionTemplate(T)(T functionParameter)
{
    // ...
}

class ClassTemplate(T)
{
    // ...
}

```

- Template parameters can be specified explicitly after an exclamation mark; the parentheses are not necessary when there is only one parameter to specify.

```

auto object1 = new ClassTemplate!(double);
auto object2 = new ClassTemplate!double;    // same thing

```

- Every different instantiation of a template is a different type.

```

assert(typeid(ClassTemplate!int) !=
       typeid(ClassTemplate!uint));

```

- Template parameters can only be deduced for function templates.

```

functionTemplate(42);    // functionTemplate!int() is deduced

```

- Templates can be specialized for the type that is specified after the : character.

```

class ClassTemplate(T : dchar)
{
    // ...
}

```

- Default template parameter types can be specified after the = character.

```

void functionTemplate(T = long)(T functionParameter)
{
    // ...
}

```

65 Skipping chapters...

There are the following chapters of the book at this point but they haven't been translated from Turkish yet:

- [alias ve alias this](#) `alias, alias this`
- [Göstergeler](#) `* &`
- [Bit İşlemleri](#) `~ & | ^ >> >>> <<`
- [Koşullu Derleme](#) `debug, version, static if, static assert, is ifadesi`
- [İşlev Göstergeleri ve Kapamalar](#) `function, delegate`
- [Yapı ve Sınıflarda foreach](#) `opApply empty popFront front (ve daha başka)`
- [Birlikler](#) `union`
- [Etiketler ve goto](#) `goto`
- [Ayrıntılı Şablonlar](#) `template`
- [Diğer İşlev Olanakları](#) `pure nothrow @safe (ve daha başka)`
- [Katmalar](#) `mixin`

Please click [Next →] above or below for the next available chapter in English.

66 Ranges

Ranges are an abstraction of element access. This abstraction enables the use of great number of algorithms over great number of container types. Ranges emphasize how container elements are accessed, as opposed to how the containers are implemented.

Ranges are a very simple concept that is based on whether a type defines certain sets of member functions. We have already seen this concept in the [foreach with Structs and Classes](#) chapter: any type that provides the member functions `empty`, `front`, and `popFront()` can be used with the `foreach` loop. The set of those three member functions is the requirement of the range type `InputRange`.

I will start introducing ranges with `InputRange`, the simplest of all the range types. The other ranges require more member functions over `InputRange`.

Before going further, I would like to provide the definitions of containers and algorithms.

Container (data structure): Container is a very useful concept that appears in almost every program. Variables are put together for a purpose and are used together as elements of a container. D's containers are its core features arrays and associative arrays, and special container types that are defined in the `std.container` module. Every container is implemented as a specific data structure. For example, associative arrays are a *hash table* implementation.

Every data structure stores its elements and provides access to them in ways that are special to that data structure. For example, in the array data structure the elements are stored side by side and accessed by an element index; in the linked list data structure the elements are stored in nodes and are accessed by going through those nodes one by one; in a sorted binary tree data structure, the nodes provide access to the preceding and successive elements through separate branches; etc.

In this chapter, I will use the terms *container* and *data structure* interchangeably.

Algorithm (function): Processing of data structures for specific purposes in specific ways is called an *algorithm*. For example, *linear search* is an algorithm that searches by iterating over a container from the beginning to the end; *binary search* is an algorithm that searches for an element by eliminating half of the candidates at every step; etc.

In this chapter, I will use the terms *algorithm* and *function* interchangeably.

For most of the samples below, I will use `int` as the element type and `int[]` as the container type. In reality, ranges are more powerful when used with templated containers and algorithms. In fact, most of the containers and algorithms that ranges tie together are all templates. I will leave examples of templated ranges to the next chapter.

66.1 History

A very successful library that abstracts algorithms and data structures from each other is the Standard Template Library (STL), which also appears as a part of the C++ standard library. STL provides this abstraction with the *iterator* concept, which is implemented by C++'s templates.

Although they are a very useful abstraction, iterators do have some weaknesses. Ranges are designed by Andrei Alexandrescu partly to overcome these weaknesses. D's standard library Phobos is the only library that uses ranges that are the subject of this chapter.

Andrei Alexandrescu introduces ranges in the seminal paper [On Iteration](#) and demonstrates how they are superior to iterators.

66.2 Ranges are a part of D

Ranges are an integral part of D. D's slices happen to be implementations of the most powerful range **RandomAccessRange**, and there are many range features in Phobos. Although implementing new range types or algorithms are not common in most programs, it is important to understand how ranges are used in Phobos.

Many Phobos algorithms return temporary range objects. For example **filter()**, which chooses elements that are greater than 10 in the following code, actually returns a range object, not an array:

```
import std.stdio;
import std.algorithm;

void main()
{
    int[] values = [ 1, 20, 7, 11 ];
    writeln(filter!(value => value > 10)(values));
}
```

writeln uses that range object lazily and accesses the elements as it needs them:

```
[20, 11]
```

That output may suggest that **filter()** returns an **int[]** but this is not the case. We can see this from the fact that the following assignment produces a compilation error:

```
int[] chosen = filter!(value => value > 10)(values); // ← compilation ERROR
```

The error message contains the type of the range object:

```
Error: cannot implicitly convert expression (filter(values))
of type Result to int[]
```

Note: The type may be different in the version of Phobos that you are using.

It is possible to convert that temporary object to an actual array, as we will see later in the chapter.

66.3 Traditional implementations of algorithms

In traditional implementations of algorithms, the algorithms know how the data structures that they operate on are implemented. For example, the following function that prints the elements of a linked list must know that the nodes of the linked list have members named **element** and **next**:

```
struct Node
{
    int element;
    Node * next;
}

void print(const(Node) * list)
{
```

```

for ( ; list; list = list.next) {
    write(' ', list.element);
}

```

Similarly, a function that prints the elements of an array must know that arrays have a **length** property and their elements are accessed by the **[]** operator:

```

void print(const int[] array)
{
    for (int i = 0; i != array.length; ++i) {
        write(' ', array[i]);
    }
}

```

Note: We know that **foreach** is more useful when iterating over arrays. As a demonstration of how traditional algorithms are tied to data structures, let's assume a case where the use of **for** is really needed.

Having algorithms tied to data structures makes it necessary to write them specially for each type. For example, the functions `find()`, `sort()`, `swap()`, etc. must be written separately for array, linked list, associative array, binary tree, heap, etc. As a result, N algorithms that support M data structures must be written $N \times M$ times. (Note: In reality, the count is less than $N \times M$, because not every algorithm can be used with every data structure, e.g. associative arrays cannot be sorted.)

Conversely, because ranges abstract algorithms away from data structures, implementing just N algorithms and M data structures would be sufficient. A newly implemented data structure can work with all of the existing algorithms that support the type of range that the new data structure provides, and a newly implemented algorithm can work with all of the existing data structures that support the range type that the new algorithm requires.

66.4 Phobos ranges

The ranges in this chapter are different from number ranges that are written in the form **begin..end**. We had seen how number ranges are used with the **foreach** loop and with slices:

```

foreach (value; 3..7) {           // number range,
                                // NOT a Phobos range

int[] slice = array[5..10];     // number range,
                                // NOT a Phobos range

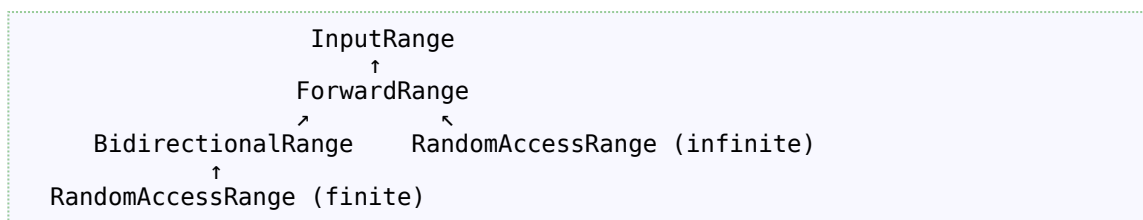
```

When I write *range*, I mean a Phobos range in this chapter.

Ranges form a *range hierarchy*. At the bottom of this hierarchy is the simplest range **InputRange**. The other ranges bring more requirements on top of the range that they are based on. The following are all of the ranges with their requirements, sorted from the simplest to the more capable:

- **InputRange**: requires the **empty**, **front** and **popFront()** member functions
- **ForwardRange**: additionally requires the **save** member function
- **BidirectionalRange**: additionally requires the **back** and **popBack()** member functions
- **RandomAccessRange**: additionally requires the **[]** operator (and another property depending on whether the range is finite or infinite)

This hierarchy can be shown as in the following graph. **RandomAccessRange** has finite and infinite versions:



The graph above is in the style of class hierarchies where the lowest level type is at the top.

Those ranges are about providing element access. There is one more range, which is about element *output*:

- **OutputRange**: requires support for the `put(range, element)` operation

These five range types are sufficient to abstract algorithms from data structures.

66.4.1 Iterating by shortening the range

Normally, iterating over the elements of a container does not change the container itself. For example, iterating over a slice with **foreach** or **for** does not affect the slice:

```

int[] slice = [ 10, 11, 12 ];

for (int i = 0; i != slice.length; ++i) {
    write(' ', slice[i]);
}

assert(slice.length == 3); // ← the length doesn't change
  
```

Another way of iteration requires a different way of thinking: iteration can be achieved by shortening the range from the beginning. In this method, always the first element is used for element access and the first element is *popped* from the beginning in order to get to the next element:

```

for ( ; slice.length; slice = slice[1..$]) {
    write(' ', slice[0]); // ← always the first element
}
  
```

Iteration is achieved by removing the first element by the `slice = slice[1..$]` expression. The slice above is completely consumed by going through the following stages:

```

[ 10, 11, 12 ]
  [ 11, 12 ]
    [ 12 ]
      [ ]
  
```

The iteration concept of Phobos ranges is based on this new thinking of shortening the range from the beginning. (**BidirectionalRange** and finite **RandomAccessRange** types can be shortened from the end as well.)

Please note that the code above is just to demonstrate this type of iteration; it should not be considered normal to iterate as in that example.

Since losing elements just to iterate over a range would not be desired in most cases, a surrogate range may be consumed instead. The following code uses a separate slice to preserve the elements of the original one:

```

int[] slice = [ 10, 11, 12 ];
int[] surrogate = slice;

for ( ; surrogate.length; surrogate = surrogate[1..$]) {
    write(' ', surrogate[0]);
}

assert(surrogate.length == 0); // ← surrogate is consumed
assert(slice.length == 3);    // ← slice remains the same

```

This is the method employed by most of the Phobos range functions: they return special range objects to be consumed in order to preserve the original containers.

66.5 InputRange

This type of range models the type of iteration where elements are accessed in sequence as we have seen in the `print()` functions above. This iteration is always in the forward direction; there is no way of going back to the elements that have already been iterated over. Despite this *limitation*, many algorithms are based merely on sequential access. **InputRange** covers the standard input streams of programs as well, where elements are removed from the stream as they are read.

For completeness, here are the three functions that **InputRange** requires:

- **empty**: specifies whether the range is empty; it must return **true** when the range is considered to be empty, and **false** otherwise
- **front**: provides access to the element at the beginning of the range
- **popFront()**: shortens the range from the beginning by removing the first element

Note: I write **empty** and **front** without parentheses, as they can be seen as properties of the range; and **popFront()** with parentheses as it is a function with side effects.

Here is how `print()` can be implemented by using these range functions:

```

void print(T)(T range)
{
    for ( ; !range.empty; range.popFront()) {
        write(' ', range.front);
    }

    writeln();
}

```

Please also note that `print()` is now a function template to avoid limiting the range type arbitrarily. `print()` can now work with any type that provides the three **InputRange** functions.

66.5.1 InputRange example

Let's redesign the **School** type that we have seen before, this time as an **InputRange**. We can imagine **School** as a **Student** container so when designed as a range, it can be seen as a range of **Students**.

In order to keep the example short, let's disregard some important design aspects. Let's

- implement only the members that are related to this section
- design all types as structs
- ignore specifiers and qualifiers like **private**, **public**, and **const**
- not take advantage of contract programming and unit testing

```

import std.string;

struct Student
{
    string name;
    int number;

    string toString()
    {
        return format("%s(%s)", name, number);
    }
}

struct School
{
    Student[] students;
}

void main()
{
    auto school = School( [ Student("Ebru", 1),
                           Student("Derya", 2) ,
                           Student("Damla", 3) ] );
}

```

To make **School** be accepted as an **InputRange**, we must define the three **InputRange** member functions.

For **empty** to return **true** when the range is empty, we can use the length of the **students** array. When the length of that array is 0, the range is considered empty:

```

struct School
{
    // ...

    @property bool empty()
    {
        return students.length == 0;
    }
}

```

empty is defined as **@property** to be able to write it without parentheses as in **school.empty**.

For **front** to return the first element of the range, we can return the first element of the array:

```

struct School
{
    // ...

    @property ref Student front()
    {
        return students[0];
    }
}

```

Note: I have used the **ref** keyword to be able to provide access to the actual element instead of a copy of it. Otherwise the elements would be copied because **Student** is a struct.

For **popFront()** to shorten the range from the beginning, we can shorten the **students** array from the beginning:

```

struct School
{
    // ...

```



```

void popFront()
{
    students = students[1 .. $];
}
}

```

Note: As I have mentioned above, it is not normal to lose the original elements from the container just to iterate over them. We will address this issue below by introducing a special range type.

These three functions are sufficient to make **School** to be used as an **InputRange**. As an example, let's add the following line at the end of **main()** above to have our new **print()** function template to use **school** as a range:

```
print(school);
```

print() uses that object as an **InputRange** and prints its elements to the output:

```
Ebru(1) Derya(2) Damla(3)
```

We have achieved our goal of defining a user type as an **InputRange**; we have sent it to an algorithm that operates on **InputRange** types. **School** is actually ready to be used with algorithms of Phobos or any other library that work with **InputRange** types. We will see examples of this below.

66.5.2 The `std.array` module to use slices as ranges

Merely importing the **std.array** module makes the most common container type conform to the most capable range type: slices can seamlessly be used as **RandomAccessRange** objects.

std.array defines **empty**, **front**, **popFront()** and other range functions specially for slices. As a result, slices are ready to be used with any range function, for example with **print()**:

```

import std.array;
// ...
print([ 1, 2, 3, 4 ]);

```

It is not necessary to import **std.array** if the **std.range** module has already been imported.

Since it is not possible to remove elements from fixed-length arrays, **popFront()** cannot be defined for them. For that reason, fixed-length arrays cannot be used as ranges themselves:

```

void print(T)(T range)
{
    for ( ; !range.empty; range.popFront()) { // ← compilation ERROR
        write(' ', range.front);
    }
    writeln();
}

void main()
{
    int[4] array = [ 1, 2, 3, 4 ];
    print(array);
}

```

It would be better if the compilation error appeared on the line where `print()` is called. This is possible by adding a template constraint to `print()`. The following template constraint takes advantage of `isInputRange`, which we will see in the next chapter. By the help of the template constraint, now the compilation error is for the line where `print()` is called, not for a line where `print()` is defined:

```
void print(T)(T range)
    if (isInputRange!T)    // template constraint
{
    // ...
}
// ...
print(array);    // ← compilation ERROR
```

It is still possible to access the elements of a fixed-length array by the range functions. What needs to be done is to use a whole slice of the array, not the array itself:

```
print(array[]);    // now compiles
```

Although slices can be used as ranges, not every range type can be used as an array. When necessary, all of the elements can be copied one by one into an array. `std.array.array` is a helper function to simplify this task; `array()` iterates over `InputRange` ranges, copies the elements, and returns a new array:

```
import std.array;
// ...
auto copiesOfStudents = array(school);
writeln(copiesOfStudents);
```

The output:

```
[Ebru(1), Derya(2), Damla(3)]
```

66.5.3 Special convenience for strings

Being character arrays by definition, strings can also be used as ranges just by importing `std.array`. The two exceptions are, `char` and `wchar` strings cannot be used as `RandomAccessRange`.

`std.array` provides a special functionality with all types of strings: iterating over strings becomes iterating over Unicode code points, not over UTF code units. As a result, strings appear as ranges of Unicode characters.

The following strings contain ç and é, which cannot be represented by a single `char` and the *gothic letter ahsa* (?), which cannot be represented by a single `wchar` (note that these characters may not be displayed correctly in the environment that you are reading this chapter):

```
import std.array;
// ...
print("abcçdeé?"c);
print("abcçdeé?"w);
print("abcçdeé?"d);
```

The output of the program is what we would normally expect from a *range of letters*:

```
a b c ç d e é ?
a b c ç d e é ?
a b c ç d e é ?
```

As you can see, that output does not match what we have seen in the [Characters](#) and [Strings](#) chapters. Since the elements of `string` and `wstring` are `char` and `wchar` respectively, one might expect to see UTF code units in the previous output.

As a reminder, let's consider the following function to see how strings used as ranges are different from strings used as arrays:

```
void printElements(T)(T str)
{
    for (int i = 0; i != str.length; ++i) {
        write(' ', str[i]);
    }

    writeln();
}

// ...

printElements("abcçdeé?"c);
printElements("abcçdeé?"w);
printElements("abcçdeé?"d);
```

Indexing provides direct access to the elements of the arrays:

```
a b c     d e            
a b c ç d e é        
a b c ç d e é ?
```

66.5.4 Ranges without actual elements

The elements of the `School` objects did actually exist; `School.front` returned a reference to an existing `Student` object.

One of the powers of ranges is the flexibility of not actually owning elements. `front` need not return an actual element of an actual container. The returned *element* can be calculated each time when `popFront()` is called, and can be used as the value that is returned by `front`.

We have already seen a range without actual elements above: Since `char` and `wchar` cannot represent Unicode characters, the Unicode characters that appear as range elements cannot be actual elements of any `char` or `wchar` array. In the case of strings, `front` returns a `dchar` that is *constructed* from the corresponding UTF code units of arrays:

```
import std.array;

void main()
{
    dchar letter = "é".front; // The dchar that is returned by
                            // front is constructed from the
                            // two chars that represent é
}
```

Although the element type of the array is `char`, the return type of `front` above is `dchar` and that `dchar` is not an element of the array, but is made up of the elements of the array.

Similarly, some ranges do not own any elements but are used for providing access to elements of other ranges. This is a solution to the problem of losing elements while iterating

over **School** objects above. In order to preserve the elements of the actual **School** objects, a special **InputRange** can be used.

To see how this is done, let's define a new struct named **StudentRange** and move all of the range member functions from **School** to this new struct. Note that **School** itself is not a range anymore:

```
struct School
{
    Student[] students;
}

struct StudentRange
{
    Student[] students;

    this(School school)
    {
        this.students = school.students;
    }

    @property bool empty()
    {
        return students.length == 0;
    }

    @property ref Student front()
    {
        return students[0];
    }

    void popFront()
    {
        students = students[1 .. $];
    }
}
```

The new range starts with a private slice that provides access to the students of **School** and consumes this private slice in **popFront()**. As a result, the actual slice in **School** is preserved:

```
auto school = School( [ Student("Ebru", 1),
                       Student("Derya", 2) ,
                       Student("Damla", 3) ] );

print(StudentRange(school));

// The actual array is now preserved:
assert(school.students.length == 3);
```

Note: Since all its work is dispatched to its member slice, **StudentRange** may not be seen as a good example of a range. In fact, assuming that **students** is an accessible member of **School**, the user code could have created a slice of **School.students** directly and could have used that slice as a range.

66.5.5 Infinite ranges

Another benefit of ranges without actual elements is the ability to create infinite ranges.

Making an infinite range is as simple as having **empty** always return **false**. Since it is constant, **empty** need not even be a function and is defined as an **enum** value:

```
enum empty = false; // ← infinite range
```

Another option is to use an immutable **static** member:

```
static immutable empty = false; // same as above
```

As an example of this, let's design a range that represents the Fibonacci series. Despite having only two **int** members, the following range can be used as the infinite Fibonacci series:

```
struct FibonacciSeries
{
    int first = 0;
    int second = 1;

    enum empty = false; // ← infinite range

    @property int front() const
    {
        return first;
    }

    void popFront()
    {
        int third = first + second;
        first = second;
        second = third;
    }
}
```

Note: Although it is infinite, because the members are of type **int**, the elements of this Fibonacci series would be wrong beyond **int.max**.

Since **empty** is always **false** for **FibonacciSeries** objects, the **for** loop in **print()** never terminates for them:

```
print(FibonacciSeries()); // never terminates
```

An infinite range is useful when the range need not be consumed completely right away. We will see how to use only some of the elements of a **FibonacciSeries** below.

66.5.6 Functions that return ranges

Earlier, we have created a **StudentRange** object by explicitly writing **StudentRange(school)**.

In most cases, a convenience function that returns the object of such a range is used instead. For example, a function with the whole purpose of returning a **StudentRange** would simplify the code:

```
StudentRange studentsOf(ref School school)
{
    return StudentRange(school);
}

// ...

print(studentsOf(school));
```

This is a convenience over having to remember and spell out the names of range types explicitly, which can get quite complicated in practice.

We can see an example of this with the simple **std.range.take** function. **take()** is a function that provides access to a specified number of elements of a range, from the beginning. In

reality, this functionality is not achieved by the `take()` function itself, but by a special range object that it returns. This fact need not be explicit when using `take()`:

```
import std.range;

// ...

auto school = School( [ Student("Ebru", 1),
                        Student("Derya", 2),
                        Student("Damla", 3) ] );

print(take(studentsOf(school), 2));
```

`take()` returns a temporary range object above, which provides access to the first 2 elements of `school`. In turn, `print()` uses that object and produces the following output:

```
Ebru(1) Derya(2)
```

The operations above still don't make any changes to `school`; it still has 3 elements:

```
print(take(studentsOf(school), 2));
assert(school.students.length == 3);
```

The specific types of the range objects that are returned by functions like `take()` are not important. These types may sometimes be exposed in error messages, or we can print them ourselves with the help of `typeof` and `stringof`:

```
writeln(typeof(take(studentsOf(school), 2)).stringof);
```

According to the output, `take()` returns a template named `Take`:

```
Take! (StudentRange)
```

66.5.7 std.range and std.algorithm modules

A great benefit of defining our types as ranges is being able to use them not only with our own functions, but with Phobos and other libraries as well.

`std.range` includes a large number of range functions, structs, and classes. `std.algorithm` includes many algorithms that are commonly found also in the standard libraries of other languages.

To see an example of how our types can be used with standard modules, let's use `School` with the `std.algorithm.swapFront` algorithm. `swapFront()` swaps the first elements of two `InputRange` ranges. (More correctly, first elements of two ranges that *have swappable elements*. Arrays satisfy this condition.)

```
import std.algorithm;

// ...

auto turkishSchool = School( [ Student("Ebru", 1),
                               Student("Derya", 2),
                               Student("Damla", 3) ] );

auto americanSchool = School( [ Student("Mary", 10),
                                Student("Jane", 20) ] );

swapFront(studentsOf(turkishSchool),
          studentsOf(americanSchool));
```

```
print(studentsOf(turkishSchool));
print(studentsOf(americanSchool));
```

The first elements of the two schools are swapped:

```
Mary(10) Derya(2) Damla(3)
Ebru(1) Jane(20)
```

As another example, let's now look at the `std.algorithm.filter` algorithm. `filter()` returns a special range that filters out elements that do not satisfy a specific condition (a *predicate*). The operation of filtering out the elements only affects accessing the elements; the original range is preserved.

Predicates are expressions that must evaluate to **true** for the elements that are considered to satisfy a condition, and **false** for the elements that do not. There are a number of ways of specifying the predicate that `filter()` should use. As we have seen in earlier examples, one way is to use a lambda expression. The parameter `a` below represents each student:

```
filter!(a => a.number % 2)(studentsOf(school))
```

The predicate above selects the elements of the range `studentsOf(school)` that have odd numbers.

Like `take()`, `filter()` too returns a special range object. That range object in turn can be passed to other range functions. For example, it can be passed to `print()`:

```
print(filter!(a => a.number % 2)(studentsOf(school)));
```

That expression can be explained from right to left: *start with the range `studentsOf(school)`, construct a range object that will filter out the elements of that initial range, and pass the new range object to `print()`.*

The output consists of students with odd numbers:

```
Ebru(1) Damla(3)
```

As long as it returns **true** for the elements that satisfy the condition, the predicate can also be specified as a function:

```
import std.array;
// ...
bool startsWithD(Student student)
{
    return student.name.front == 'D';
}
print(filter!startsWithD(studentsOf(school)));
```

The predicate function above returns **true** for students having names starting with the letter D, and **false** for the others.

Note: Using `student.name[0]` would have meant the first UTF-8 code unit, not the first letter. As I have mentioned above, `front` uses `name` as a range and always returns the first Unicode character.

This time the students whose names start with D are selected and printed:

```
Derya(2) Damla(3)
```

66.5.8 Laziness

Another benefit of functions' returning range objects is that, those objects can be used lazily. This may be essential for execution speed and memory consumption. Additionally, the fact that infinite ranges can exist at all is possible due to lazy ranges.

Lazy ranges produce their elements one at a time and only when needed. We see an example of this with the **FibonacciSeries** range: The elements are calculated by **popFront()** only as they are needed. If **FibonacciSeries** were an eager range and tried to produce all of the elements up front, it could never end or find room for the elements that it produced.

Another problem of eager ranges is the fact that they would have to spend time and space for elements that may have never going to be needed.

Like most of the algorithms in Phobos, **take()** and **filter()** benefit from laziness. For example, we can pass **FibonacciSeries** to **take()** and have it generate finite number of elements:

```
print(take(FibonacciSeries(), 10));
```

Although **FibonacciSeries** is infinite, the output contains only the first 10 numbers:

```
0 1 1 2 3 5 8 13 21 34
```

66.6 ForwardRange

InputRange models a range where elements are taken out of the range as they are iterated over.

Some ranges are capable of saving their states, as well as operating as an **InputRange**. For example, **FibonacciSeries** objects can save their states because these objects can freely be copied and the two copies continue their lives independently from each other.

ForwardRange provides the **save** member function, which is expected to return a copy of the range. The copy that **save** returns must operate independently from the range object that it was copied from: iterating over one copy must not affect the other copy.

Importing **std.array** automatically makes slices become **ForwardRange** ranges.

In order to implement **save** for **FibonacciSeries**, we can simply return a copy of the object:

```
struct FibonacciSeries
{
    // ...
    @property FibonacciSeries save() const
    {
        return this;
    }
}
```

The returned copy is a separate range that would continue from the point where it was copied from.

We can demonstrate that the copied object is independent from the actual range with the following program. The algorithm `std.range.popFrontN()` in the following code removes a specified number of elements from the specified range:

```
import std.range;

// ...

void report(T)(const dchar[] title, const ref T range)
{
    writeln("%40s: %s", title, take(range, 5));
}

void main()
{
    auto range = FibonacciSeries();
    report("Original range", range);

    popFrontN(range, 2);
    report("After removing two elements", range);

    auto theCopy = range.save;
    report("The copy", theCopy);

    popFrontN(range, 3);
    report("After removing three more elements", range);
    report("The copy", theCopy);
}
```

The output of the program shows that removing elements from the range does not affect its saved copy:

```
Original range: [0, 1, 1, 2, 3]
After removing two elements: [1, 2, 3, 5, 8]
The copy: [1, 2, 3, 5, 8]
After removing three more elements: [5, 8, 13, 21, 34]
The copy: [1, 2, 3, 5, 8]
```

Also note that the range is passed directly to `writeln` in `report()`. Like our `print()` function, the output functions of the `stdio` module can take `InputRange` objects. I will use `stdio`'s output functions from now on.

An algorithm that works with `ForwardRange` is `std.range.cycle.cycle()` iterates over the elements of a range repeatedly from the beginning to the end. In order to be able to start over from the beginning it must be able to save a copy of the initial state of the range, so it requires a `ForwardRange`.

Since `FibonacciSeries` is now a `ForwardRange`, we can try `cycle()` with a `FibonacciSeries` object; but in order to avoid having `cycle()` iterate over an infinite range, and as a result never finding the end of it, we must first make a finite range by passing `FibonacciSeries` through `take()`:

```
writeln(take(cycle(take(FibonacciSeries(), 5)), 20));
```

In order to make the resultant range finite as well, the range that is returned by `cycle` is also passed through `take()`. The output consists of *the first twenty elements of cycling through the first five elements of `FibonacciSeries`*:

```
[0, 1, 1, 2, 3, 0, 1, 1, 2, 3, 0, 1, 1, 2, 3, 0, 1, 1, 2, 3]
```

We could have defined intermediate variables to make the code more readable. The following is an equivalent of the single line of code above:

```

auto series          = FibonacciSeries();
auto firstPart      = take(series, 5);
auto cycledThrough  = cycle(firstPart);
auto firstPartOfCycledThrough = take(cycledThrough, 20);

writeln(firstPartOfCycledThrough);

```

I would like to point out the importance of laziness again: The first four lines above merely construct range objects that will eventually produce the elements. The numbers that are part of the result are calculated by `FibonacciSeries.popFront()` as needed.

Note: Although we have started with `FibonacciSeries` as `ForwardRange`, we have actually passed the result of `take(FibonacciSeries())` to `cycle()`. `take()` is adaptive: the range that it returns is a `ForwardRange` if its parameter is a `ForwardRange`. We will see how this is accomplished with `isForwardRange` in the next chapter.

66.7 BidirectionalRange

`BidirectionalRange` provides two member functions over the member functions of `ForwardRange`. `back` is similar to `front`: it provides access to the last element of the range. `popBack()` is similar to `popFront()`: it removes the last element from the range.

Importing `std.array` automatically makes slices become `BidirectionalRange` ranges.

A good `BidirectionalRange` example is the `std.range.retro` function. `retro()` takes a `BidirectionalRange` and ties its `front` to `back`, and `popFront()` to `popBack()`. As a result, the original range is iterated over in reverse order:

```
writeln(retro([ 1, 2, 3 ]));
```

The output:

```
[3, 2, 1]
```

Let's define a range that behaves similarly to the special range that `retro()` returns. Although the following range has limited functionality, it shows how powerful ranges are:

```

struct Reversed
{
    int[] range;

    this(int[] range)
    {
        this.range = range;
    }

    @property bool empty() const
    {
        return range.empty;
    }

    @property int front() const
    {
        return range.back; // ← reverse
    }

    @property int back() const
    {

```

```

    return range.front; // ← reverse
}
void popFront()
{
    range.popBack();    // ← reverse
}
void popBack()
{
    range.popFront();  // ← reverse
}
}
void main()
{
    writeln(Reversed([ 1, 2, 3]));
}

```

The output is the same as `retro()`:

```
[3, 2, 1]
```

66.8 RandomAccessRange

RandomAccessRange represents ranges that allow accessing elements by the `[]` operator. As we have seen in the [Operator Overloading chapter](#), `[]` operator is defined by the `opIndex()` member function.

Importing `std.array` module makes slices become **RandomAccessRange** ranges only if possible. For example, since UTF-8 and UTF-16 encodings do not allow accessing Unicode characters by an index, `char` and `wchar` arrays cannot be used as **RandomAccessRange** ranges of Unicode characters. On the other hand, since the codes of the UTF-32 encoding correspond one-to-one to Unicode character codes, `dchar` arrays can be used as **RandomAccessRange** ranges of Unicode characters.

It is natural that every type would define the `opIndex()` member function according to its functionality. However, computer science has an expectation on its algorithmic complexity: random access must be *at constant time*. Constant time access means that the time spent when accessing an element is independent of the number of elements in the container. Therefore, no matter how large the range is, element access should not depend on the length of the range.

In order to be considered a **RandomAccessRange**, *one* of the following conditions must also be satisfied:

- to be an infinite **ForwardRange**

or

- to be a **BidirectionalRange** that also provides the `length` property

Depending on the condition that is satisfied, the range is either infinite or finite.

66.8.1 Infinite RandomAccessRange

The following are all of the requirements of a **RandomAccessRange** that is based on an *infinite ForwardRange*:

- `empty`, `front` and `popFront()` that **InputRange** requires

- **save** that **ForwardRange** requires
- **opIndex()** that **RandomAccessRange** requires
- the value of **empty** to be known at compile time as **false**

We were able to define **FibonacciSeries** as a **ForwardRange**. However, **opIndex()** cannot be implemented to operate at constant time for **FibonacciSeries** because accessing an element requires accessing all of the previous elements first.

As an example where **opIndex()** can operate at constant time, let's define an infinite range that consists of squares of integers. Although this range is infinite, accessing any one of its elements can happen at constant time:

```
class SquaresRange
{
    int first;

    this(int first = 0)
    {
        this.first = first;
    }

    enum empty = false;

    @property int front() const
    {
        return opIndex(0);
    }

    void popFront()
    {
        ++first;
    }

    @property SquaresRange save() const
    {
        return new SquaresRange(first);
    }

    int opIndex(size_t index) const
    {
        /* This function operates at constant time */
        immutable integerValue = first + cast(int)index;
        return integerValue * integerValue;
    }
}
```

Although no space has been allocated for the elements of this range, the elements can be accessed by the **[]** operator:

```
auto squares = new SquaresRange();
writeln(squares[5]);
writeln(squares[10]);
```

The output contains the elements at indexes 5 and 10:

```
25
100
```

The element with index 0 should always represent the first element of the range. We can take advantage of **popFrontN()** when testing whether this really is the case:

```
popFrontN(squares, 5);
writeln(squares[0]);
```

The first 5 elements of the range are 0, 1, 4, 9 and 16; the squares of 0, 1, 2, 3 and 4. After removing those, the square of the next value becomes the first element of the range:

```
25
```

Being a **RandomAccessRange** (the most functional range), **SquaresRange** can also be used as other types of ranges. For example as an **InputRange** when passing to **filter()**:

```
bool are_lastTwoDigitsSame(int value)
{
    /* Must have at least two digits */
    if (value < 10) {
        return false;
    }

    /* Last two digits must be divisible by 11 */
    immutable lastTwoDigits = value % 100;
    return (lastTwoDigits % 11) == 0;
}

writeln(filter!are_lastTwoDigitsSame(take(squares, 50)));
```

The output consists of elements among the first 50, where last two digits are the same:

```
[100, 144, 400, 900, 1444, 1600]
```

66.8.2 Finite RandomAccessRange

The following are all of the requirements of a **RandomAccessRange** that is based on a *finite BidirectionalRange*:

- **empty**, **front** and **popFront()** that **InputRange** requires
- **save** that **ForwardRange** requires
- **back** and **popBack()** that **BidirectionalRange** requires
- **opIndex()** that **RandomAccessRange** requires
- **length**, which provides the length of the range

As an example of a finite **RandomAccessRange**, let's define a range that works similarly to **std.range.chain.chain()** presents the elements of a number of separate ranges as if they are elements of a single larger range. Although **chain()** works with any type of element and any type of range, to keep the example short, let's implement a range that works only with **int** slices.

Let's name this range **Together** and expect the following behavior from it:

```
auto range = Together([ 1, 2, 3 ], [ 101, 102, 103]);
writeln(range[4]);
```

When constructed with the two separate arrays above, **range** should present all of those elements as a single range. For example, although neither array has an element at index 4, the element 102 should be the element that corresponds to index 4 of the collective range:

```
102
```

As expected, printing the entire range should contain all of the elements:

```
writeln(range);
```

The output:

```
[1, 2, 3, 101, 102, 103]
```

Together will operate lazily: the elements will not be copied to a new larger array; they will be accessed from the original slices.

We can take advantage of *variadic functions* that was introduced in the [Variable Number of Parameters](#) chapter when initializing the range by any number of original slice:

```
struct Together
{
    int[][] slices;

    this(int[][] slices ...)
    {
        this.slices = slices;

        clearFront();
        clearBack();
    }

    // ...
}
```

The `clearFront()` and `clearBack()` calls that the constructor makes are to remove empty slices from the beginning and the end of the original slices. Such empty slices do not change the behavior of **Together** and removing them up front will simplify the implementation:

```
struct Together
{
    // ...

    private void clearFront()
    {
        while (!slices.empty && slices.front.empty) {
            slices.popFront();
        }
    }

    private void clearBack()
    {
        while (!slices.empty && slices.back.empty) {
            slices.popBack();
        }
    }
}
```

We will call those functions later from `popFront()` and `popBack()` as well.

Since `clearFront()` and `clearBack()` remove all of the empty slices from the beginning and the end; still having a slice would mean that the collective range is not yet empty. In other words, the range should be considered empty only if there is no slice left:

```
struct Together
{
    // ...

    @property bool empty()
    {
        return slices.empty;
    }
}
```

```

    }
}

```

The first element of the first slice is the first element of this **Together** range:

```

struct Together
{
// ...
    @property int front()
    {
        return slices.front.front;
    }
}

```

Removing the first element of the first slice removes the first element of this range as well. Since this operation may leave the first slice empty, we must call **clearFront()** to remove that empty slice and the ones that are after that one:

```

struct Together
{
// ...
    void popFront()
    {
        slices.front.popFront();
        clearFront();
    }
}

```

A copy of this range can be constructed from a copy of the **slices** member:

```

struct Together
{
// ...
    @property Together save()
    {
        return Together(slices.dup);
    }
}

```

*Please note that **.dup** copies only the slices, not the elements that they contain.*

The operations at the end of the range are similar to the ones at the beginning:

```

struct Together
{
// ...
    @property int back()
    {
        return slices.back.back;
    }

    void popBack()
    {
        slices.back.popBack();
        clearBack();
    }
}

```

The length of the range can be calculated as the sum of the length of the slices:

```

struct Together
{
// ...

    @property size_t length()
    {
        size_t totalLength = 0;

        foreach (slice; slices) {
            totalLength += slice.length;
        }

        return totalLength;
    }
}

```

Alternatively, the length may be calculated with less code by taking advantage of `std.algorithm.reduce`. `reduce()` takes an operation as its template parameter and applies that operation to all of the elements of a range:

```

import std.algorithm;

// ...

@property size_t length()
{
    return reduce!((a, b) => a + b.length)(size_t.init, slices);
}

```

The `a` in the template parameter represents the current result (*the sum* in this case) and `b` represents the current element. The first function parameter is the initial value of the result (`size_t.init` is 0) and the second function parameter is the range that contains the elements.

Note: Further, instead of calculating the length every time when `length` is called, it may be measurably faster to maintain a member variable perhaps named `length_`, which always equals the correct length of the collective range. That member may be calculated once in the constructor and adjusted accordingly as elements are removed by `popFront()` and `popBack()`.

One way of returning the element that corresponds to a specific index is to look at every slice to determine whether the element would be among the elements of that slice:

```

struct Together
{
// ...

    int opIndex(size_t index)
    {
        /* Save the index for the error message */
        immutable originalIndex = index;

        foreach (slice; slices) {
            if (slice.length > index) {
                return slice[index];
            } else {
                index -= slice.length;
            }
        }

        throw new Exception(
            format("Invalid index: %s (length: %s)",
                originalIndex, this.length));
    }
}

```



```
}
}
```

Note: This `opIndex()` does not satisfy the constant time requirement that has been mentioned above. For this implementation to be acceptably fast, the `slices` member must not be too long.

This new range is now ready to be used with any number of `int` slices. With the help of `take()` and `array()`, we can even include the range types that we have defined earlier in this chapter:

```
auto range = Together(array(take(FibonacciSeries(), 10)),
                      [ 777, 888 ],
                      array(take(new SquaresRange(), 5)));

writeln(range.save);
```

The elements of the three slices are accessed as if they are elements of a single large array:

```
[0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 777, 888, 0, 1, 4, 9, 16]
```

We can pass this range to other range algorithms. For example to `retro()`, which requires a `BidirectionalRange`:

```
writeln(retro(range.save));
```

The output:

```
[16, 9, 4, 1, 0, 888, 777, 34, 21, 13, 8, 5, 3, 2, 1, 1, 0]
```

Of course you should use the more functional `std.range.chain` instead of `Together` in your programs.

66.9 OutputRange

All of the range types that we have seen so far are about element access. `OutputRange` represents streamed element output, similar to sending characters to `stdout`.

I have mentioned earlier that `OutputRange` requires support for the `put(range, element)` operation. `put()` is a function defined in the `std.range` module. It determines the capabilities of the range and the element at compile time and uses the most appropriate method to use to *output* the elements.

`put()` considers the following cases in the order that they are listed below, and applies the method for the first matching case. `R` represents the type of the range; `range`, a range object; `E`, the type of the element; and `e` an element of the range:

Case Considered	Method Applied
R has a member function named put and put can take an E as a parameter	<code>range.put(e);</code>
R has a member function named put and put can take an E[] as a parameter	<code>range.put([e]);</code>
R is an InputRange and e can be assigned to range.front	<code>range.front = e; range.popFront();</code>
E is an InputRange and can be copied to R	<code>for (; !e.empty; e.popFront()) put(range, e.front);</code>
R can take E as a parameter (e.g. R could be a delegate)	<code>range(e);</code>
R can take E[] as a parameter (e.g. R could be a delegate)	<code>range([e]);</code>

Let's define a range that matches the first case: The range will have a member function named **put()**, which takes a parameter that matches the type of the output range.

This output range will be used for outputting elements to multiple files, including **stdout**. When elements are outputted with **put()**, they will all be written to all of those files. As an additional functionality, let's add the ability to specify a delimiter to be written after each element.

```

struct MultiFile
{
    string delimiter;
    File[] files;

    this(string delimiter, string[] fileNames ...)
    {
        this.delimiter = delimiter;

        /* stdout is always included */
        this.files ~= stdout;

        /* A File object for each file name */
        foreach (fileName; fileNames) {
            this.files ~= File(fileName, "w");
        }
    }

    void put(T)(T element)
    {
        foreach (file; files) {
            file.write(element, delimiter);
        }
    }
}

```

In order to be used as an output range of any type of elements, **put()** is also templated on the element type.

An algorithm in Phobos that uses **OutputRange** is **std.algorithm.copy.copy()** is a very simple algorithm, which copies the elements of an **InputRange** to an **OutputRange**.

```
import std.algorithm;

// ...

auto output = MultiFile("\n", "output_0", "output_1");
copy([ 1, 2, 3], output);
copy([ "red", "blue", "green" ], output);
```

That code outputs the elements of the input ranges both to **stdout** and to files named "output_0" and "output_1":

```
1
2
3
red
blue
green
```

66.9.1 Using slices as OutputRange

The **std.range** module makes slices **OutputRange** objects as well. (By contrast, **std.array** makes them only input ranges.) Unfortunately, using slices as **OutputRange** objects has a confusing effect: slices lose an element for each **put()** operation on them; and that element is the element that has just been outputted!

The reason for this behavior is a consequence of slices' not having a **put()** member function. As a result, the third case of the previous table is matched for slices and the following method is applied:

```
range.front = e;
range.popFront();
```

As the code above is executed for each **put()**, the front element of the slice is assigned to the value of the *outputted* element, to be subsequently removed from the slice with **popFront()**:

```
import std.stdio;
import std.range;

void main()
{
    int[] slice = [ 1, 2, 3 ];
    put(slice, 100);
    writeln(slice);
}
```

As a result, although the slice is used as an **OutputRange**, it surprisingly *loses* elements:

```
[2, 3]
```

To avoid this, a separate slice must be used as an **OutputRange** instead:

```
import std.stdio;
import std.range;

void main()
{
    int[] slice = [ 1, 2, 3 ];
    int[] slice2 = slice;
```

```

    put(slice2, 100);

    writeln(slice2);
    writeln(slice);
}

```

This time the second slice is consumed and the original slice has the expected elements:

```

[2, 3]
[100, 2, 3] ← expected result

```

Another important fact is that the length of the slice does not grow when used as an **OutputRange**. It is the programmer's responsibility to ensure that there is enough room in the slice:

```

int[] slice = [ 1, 2, 3 ];
int[] slice2 = slice;

foreach (i; 0 .. 4) { // ← no room for 4 elements
    put(slice2, i * 100);
}

```

When the slice becomes completely empty because of the indirect **popFront()** calls, the program terminates with an exception:

```

core.exception.AssertError@...: Attempting to fetch the front
of an empty array of int

```

std.array.appender allows using slices as an **OutputRange** where the elements are appended. The **put()** function of the special range object that **appender()** returns actually appends the elements to the original slice:

```

import std.array;

// ...

auto a = appender([ 1, 2, 3 ]);

foreach (i; 0 .. 4) {
    a.put(i * 100);
}

```

In the code above, **appender** is called with an array and returns a special range object. That range object is in turn used as an **OutputRange** by calling its **put()** member function. The resultant elements are accessed by its **.data** property:

```

writeln(a.data);

```

The output:

```

[1, 2, 3, 0, 100, 200, 300]

```

66.10 Range templates

Although we have used mostly **int** ranges in this chapter, ranges and range algorithms are much more useful when defined as templates.

The `std.range` module includes many range templates. We will see these templates in the next chapter.

66.11 Summary

- Ranges abstract data structures from algorithms and allow them to be used with algorithms seamlessly.
- Ranges are a D concept and are the basis for many features of Phobos.
- Many Phobos algorithms return lazy range objects to accomplish their special tasks.
- When used as **InputRange** objects, the elements of strings are Unicode characters.
- **InputRange** requires **empty**, **front** and **popFront()**.
- **ForwardRange** additionally requires **save**.
- **BidirectionalRange** additionally requires **back** and **popBack()**.
- Infinite **RandomAccessRange** requires **opIndex()** over **ForwardRange**.
- Finite **RandomAccessRange** requires **opIndex()** and **length** over **BidirectionalRange**.
- **std.array.appender** returns an **OutputRange** that appends to slices.
- Slices are ranges of finite **RandomAccessRange**
- Fixed-length arrays are not ranges.

67 Skipping chapters...

There are the following chapters of the book at this point but they haven't been translated from Turkish yet:

- [Başka Aralık Olanakları](#) `isInputRange ElementType hasLength inputRangeObject` (ve daha başka)
- [Çokuzlular](#) `tuple Tuple TypeTuple`

Please click [Next →] above or below for the next available chapter in English.

68 Parallelism

Most modern microprocessors consist of more than one *core*, each of which can operate as an individual processing unit. They can execute different parts of different programs at the same time. The features of the `std.parallelism` module make it possible for programs to take advantage of all of the cores in order to run faster.

This chapter covers the following range algorithms. These algorithms should be used only when the operations that are to be executed *in parallel* are truly independent from each other. *In parallel* means that operations are executed on multiple cores at the same time:

- **parallel**: Executes the iterations of a **foreach** loop in parallel.
- **task**: Creates tasks that are executed in parallel.
- **asyncBuf**: Iterates the elements of an **InputRange** semi-eagerly in parallel.
- **map**: Calls functions with the elements of an **InputRange** semi-eagerly in parallel.
- **amap**: Calls functions with the elements of a **RandomAccessRange** fully-eagerly in parallel.
- **reduce**: Makes calculations over the elements of a **RandomAccessRange** in parallel.

In the programs that we have written so far we have been assuming that the expressions of a program are executed in a certain order, at least in general line-by-line:

```
++i;
++j;
```

In the code above, we expect that the value of **i** is incremented before the value of **j** is incremented. Although that is semantically correct, it is rarely the case in reality: microprocessors and compilers use optimization techniques to have some variables reside in microprocessor's registers, which are independent from each other. When that is the case, the microprocessor would execute operations like the increments above in parallel.

Although these optimizations are effective, they cannot be applied automatically to layers higher than the very low-level operations. Only the programmer can determine that certain high-level operations are independent and that they can be executed in parallel.

In a loop, the elements of a range are normally processed one after the other, operations of each element following the operations of previous elements:

```
auto students =
    [ Student(1), Student(2), Student(3), Student(4) ];

foreach (student; students) {
    student.aSlowOperation();
}
```

Normally, a program would be executed on one of the cores of the microprocessor, which has been assigned by the operating system to execute the program. As the **foreach** loop normally operates on elements one after the other, **aSlowOperation()** would be called for each student sequentially. However, in many cases it is not necessary for the operations of preceding students to be completed before starting the operations of successive students. If the operations on the **Student** objects were truly independent, it would be wasteful to ignore the other microprocessor cores, which might potentially be waiting idle on the system.

To simulate long-lasting operations, the following examples call **Thread.sleep()** from the **core.thread** module. **Thread.sleep()** suspends the operations for the specified amount of time. One way of specifying the duration is the **dur** template. Its template parameter is the unit of time: "**msecs**" for milliseconds and "**seconds**" for seconds. **Thread.sleep** is admittedly an artificial method to use in the following examples because it takes time without

ever busy any core. Despite being an unrealistic tool, it is still useful in this chapter to demonstrate the power of parallelism.

```
import std.stdio;
import core.thread;

struct Student
{
    int number;

    void aSlowOperation()
    {
        writeln("The work on student %s has begun", number);

        // Wait 1 second to simulate a long-lasting operation
        Thread.sleep(dur!"seconds"(1));

        writeln("The work on student %s has ended", number);
    }
}

void main()
{
    auto students =
        [ Student(1), Student(2), Student(3), Student(4) ];

    foreach (student; students) {
        student.aSlowOperation();
    }
}
```

The execution time of the program can be measured in a console by **time**:

```
$ time ./deneme
The work on student 1 has begun
The work on student 1 has ended
The work on student 2 has begun
The work on student 2 has ended
The work on student 3 has begun
The work on student 3 has ended
The work on student 4 has begun
The work on student 4 has ended

real    0m4.005s   ← 4 seconds total
user    0m0.004s
sys     0m0.000s
```

Since the students are iterated over in sequence and since the work of each student takes 1 second, the total execution time comes out to be 4 seconds. However, if these operations were executed in an environment that had 4 cores, they could be operated on at the same time and the total time would be reduced to about 1 second.

Before seeing how this is done, let's first determine the number of cores that are available on the system by **std.parallelism.totalCPUs**:

```
import std.stdio;
import std.parallelism;

void main()
{
```



```
    writeln("There are %s cores on this system.", totalCPUs);
}
```

The output of the program in the environment that this chapter has been written is the following:

```
There are 4 cores on this system.
```

68.1 taskPool.parallel()

This function can also be called simply as **parallel()**.

parallel() executes the iterations of a **foreach** loop in parallel. Merely importing the **std.parallelism** module and replacing **students** with **parallel(students)** in the program above is sufficient to take advantage of all of the cores of the system:

```
import std.parallelism;
// ...
foreach (student; parallel(students)) {
```

We have seen earlier in the [foreach for structs and classes chapter](#) that the expressions that are in **foreach** blocks are passed to **opApply()** member functions as delegates. **parallel()** returns a range object that knows how to distribute the execution of the **delegate** to a separate core for each element.

As a result, passing the **Student** range through **parallel()** makes the program above finish in 1 second on a system that has 4 cores:

```
$ time ./deneme
The work on student 2 has begun
The work on student 1 has begun
The work on student 4 has begun
The work on student 3 has begun
The work on student 1 has ended
The work on student 2 has ended
The work on student 4 has ended
The work on student 3 has ended

real    0m1.005s   ← now only 1 second
user    0m0.004s
sys     0m0.004s
```

Note: The execution time of the program may be different on other systems but it is expected to be roughly "4 seconds divided by the number of cores".

A flow of execution through certain parts of a program is called a *thread of execution* or a *thread*. Programs can consist of multiple threads that are being actively executed at the same time. The operating system starts and executes each thread on a core and then suspends it to execute other threads. The execution of each thread may involve many cycles of starting and suspending.

All of the threads of all of the programs that are active at a given time are executed on the very cores of the microprocessor. The operating system decides when and under what condition to start and suspend each thread. That is the reason why the messages that are printed by **aSlowOperation()** are in mixed order in the output above. This undeterministic

order of thread executions may not matter if the operations of the **Student** objects are truly independent from each other.

It is the responsibility of the programmer to call **parallel()** only when the operations in the **foreach** loop are independent for each iteration. For example, if it were important that the messages appear in a certain order in the output, calling **parallel()** should be considered an error in the program above. The programming model that supports threads that depend on other threads is called *concurrency*. Concurrency is the topic of the next chapter.

By the time parallel **foreach** ends, all of the operations inside the loop have been completed for all of the elements. The program can safely continue after the **foreach** loop.

68.1.1 Work unit size

The second parameter of **parallel()** has an overloaded meaning and is ignored in some cases:

```
/* ... */ = parallel(range, work_unit_size = 100);
```

- When iterating over **RandomAccessRange** ranges:
The distribution of threads to cores has some minimal cost. This cost may sometimes be significant especially when the operations of the loop are completed in a very short time. In such cases, it may be faster to have each thread execute more than one iteration of the loop. The work unit size determines the number of elements that each thread should execute at each of its iterations:

```
foreach (student; parallel(students, 2)) {
```

The default value of work unit size is 100 and is suitable for most cases.

- When iterating over non-**RandomAccessRange** ranges:
parallel() does not start parallel executions until *work unit size* number of elements of a non-**RandomAccessRange** have been executed serially first. Due to the relatively high value of 100, **parallel()** may give the wrong impression that it is not effective when tried on short non-**RandomAccessRange** ranges.
- When iterating over the result ranges of **asyncBuf()** or parallel **map()** (both are explained later in this chapter):
When **parallel()** works on the results of **asyncBuf()** or **map()**, it ignores the work unit size parameter. Instead, **parallel()** reuses the internal buffer of the result range.

68.2 Task

Operations that are executed in parallel with other operations of a program are called *tasks*. Tasks are represented by the type **std.parallelism.Task**.

In fact, **parallel()** constructs a new **Task** object for every iteration of the **foreach** loop and starts that task automatically. **parallel()** then waits for all of the tasks to be completed before finally exiting the loop. **parallel()** is very convenient as it *constructs, starts, and waits for* the tasks automatically. Its limitation is that **parallel()** can only be used in **foreach** loops.

When tasks do not correspond to or cannot be represented by elements of a range, these three steps can be handled explicitly by the programmer. **task()** constructs,

`executeInNewThread()` starts, and `yieldForce()` waits for a task object. These three functions are explained further in the comments of the following program.

The `anOperation()` function is started twice in the following program. It prints the first letter of `id` to indicate which task it is working for.

Note: Normally, the characters that are printed to output streams like `stdout` do not appear on the output right away. They are instead stored in an output buffer until a line of output is completed. Since `write` does not output a new-line character, in order to observe the parallel execution of the following program, `stdout.flush()` is called to send the contents of the buffer to `stdout` even before reaching the end of a line.

```
import std.stdio;
import std.parallelism;
import std.array;
import core.thread;

/* Prints the first letter of 'id' every half a second. It
 * arbitrarily returns the value 1 to simulate functions that
 * do calculations. This result will be used later in main. */
int anOperation(string id, int duration)
{
    writeln("%s will take %s seconds", id, duration);

    foreach (i; 0 .. (duration * 2)) {
        Thread.sleep(dur!"msecs"(500)); /* half a second */
        write(id.front);
        stdout.flush();
    }

    return 1;
}

void main()
{
    /* Construct a task object that will execute
     * anOperation(). The function parameters that are
     * specified here are passed to the task function as its
     * function parameters. */
    auto theTask = task!anOperation("theTask", 5);

    /* Start the task object */
    theTask.executeInNewThread();

    /* As 'theTask' continues executing, 'anOperation()' is
     * being called again, this time directly in main. */
    immutable result = anOperation("main's call", 3);

    /* At this point we are sure that the operation that has
     * been started directly from within main has been
     * completed, because it has been started by a regular
     * function call, not as a task. */

    /* On the other hand, it is not certain at this point
     * whether 'theTask' has completed its operations
     * yet. yieldForce() waits for the task to complete its
     * operations; it returns only when the task has been
     * completed. Its return value is the return value of
     * the task function, i.e. anOperation(). */
    immutable taskResult = theTask.yieldForce();

    writeln();
    writeln("All finished; the result is %s.",
        result + taskResult);
}
```

The output of the program should be similar to the following. The fact that the **m** and **t** letters are printed in mixed order indicates that the operations are executed in parallel:

```
main's call will take 3 seconds
theTask will take 5 seconds
mtmttmmttmmttttt
All finished; the result is 2.
```

The task function has been specified as a template parameter of `task()`. Another overload of `task()` takes the function as its first function parameter:

```
void someFunction(int value)
{
    /* ... */
}

auto theTask = task(&someFunction, 42);
```

A lambda function or an object of a type that defines the `opCall` member can also be used as the task function. The following example starts a task that executes a lambda:

```
auto theTask = task((int value) {
    /* ... */
}, 42);
```

68.2.1 Exceptions

As tasks are executed on separate threads, the exceptions that they throw cannot be caught by the thread that started them. For that reason, the exceptions that are thrown are automatically caught by the tasks themselves, to be rethrown later when `Task` member functions like `yieldForce()` are called. This makes it possible for the main thread to catch exceptions that are thrown by a task.

```
import std.stdio;
import std.parallelism;
import core.thread;

void mayThrow()
{
    writeln("mayThrow() is started");
    Thread.sleep(dur!"seconds"(1));
    writeln("mayThrow() is throwing an exception");
    throw new Exception("Error message");
}

void main()
{
    auto theTask = task!mayThrow();
    theTask.executeInNewThread();

    writeln("main is continuing");
    Thread.sleep(dur!"seconds"(3));

    writeln("main is waiting for the task");
    theTask.yieldForce();
}
```

The output of the program shows that the uncaught exception that has been thrown by the task does not terminate the entire program right away (it terminates only the task):

```

main is continuing
mayThrow() is started
mayThrow() is throwing an exception           ← thrown
main is waiting for the task
object.Exception@deneme.d(10): Error message   ← terminated

```

yieldForce() can be called in a **try-catch** block to catch the exceptions that are thrown by the task. Note that this is different from single threads: In single-threaded programs like the samples that we have been writing until this chapter, **try-catch** wraps the code that may throw. In parallelism, it wraps **yieldForce()**:

```

try {
    theTask.yieldForce();
} catch (Exception exc) {
    writefln("Detected an error in the task: '%s'", exc.msg);
}

```

This time the exception is caught by the main thread instead of terminating the program:

```

main is continuing
mayThrow() is started
mayThrow() is throwing an exception           ← thrown
main is waiting for the task
Detected an error in the task: 'Error message' ← caught

```

68.2.2 Member functions of Task

- **done**: Specifies whether the task has been completed; rethrows the exception if the task has been terminated with an exception.

```

if (theTask.done) {
    writeln("Yes, the task has been completed");
} else {
    writeln("No, the task is still going on");
}

```

- **executeInNewThread()**: Starts the task in a new thread.
- **executeInNewThread(int priority)**: Starts the task in a new thread with the specified priority. (Priority is an operating system concept that determines execution priorities of threads.)

There are three functions to wait for the completion of a task:

- **yieldForce()**: Starts the task if it has not been started yet; if it has already been completed, returns its return value; if it is still running, waits for its completion without making the microprocessor busy; if an exception has been thrown, rethrows that exception.
- **spinForce()**: Works similarly to **yieldForce()**, except that it makes the microprocessor busy while waiting, in order to catch the completion as early as possible.
- **workForce()**: Works similarly to **yieldForce()**, except that it starts a new task in the current thread while waiting for the task to be completed.

In most cases **yieldForce()** is the most suitable function to call when waiting for a task to complete; it suspends the thread that calls **yieldForce()** until the task is completed.

Although `spinForce()` makes the microprocessor busy while waiting, it is suitable when the task is expected to be completed in a very short time. `workForce()` can be called when starting other tasks is preferred over suspending the current thread.

Please see the online documentation of Phobos for the other member functions of `Task`.

68.3 `taskPool.asyncBuf()`

Similar to the way `parallel()` iterates `foreach` loops in parallel, `asyncBuf()` iterates `InputRange` ranges in parallel. It stores the elements in a buffer as they are produced by the range, and serves the elements from that buffer to its user.

In order to avoid making a potentially fully-lazy input range a fully-eager range, it iterates the elements in *waves*. Once it prepares certain number of elements in parallel, it waits until those elements are consumed by `popFront()` before producing the elements of the next wave.

`asyncBuf()` takes a range and an optional *buffer size* that determines how many elements to be made available during each wave:

```
auto elements = taskPool.asyncBuf(range, buffer_size);
```

To see the effects of `asyncBuf()`, let's use a range that takes half a second to iterate and half a second to process each element. This range simply produces integers up to the specified limit:

```
import std.stdio;
import core.thread;

struct Range
{
    int limit;
    int i;

    bool empty() const @property
    {
        return i >= limit;
    }

    int front() const @property
    {
        return i;
    }

    void popFront()
    {
        writeln("Producing the element after %s", i);
        Thread.sleep(dur!"msecs"(500));
        ++i;
    }
}

void main()
{
    auto range = Range(10);

    foreach (element; range) {
        writeln("Using element %s", element);
        Thread.sleep(dur!"msecs"(500));
    }
}
```

The elements are produced and used lazily. Since it takes one second for each element, the whole range takes ten seconds to process in this program:

```
$ time ./deneme
Using element 0
Producing the element after 0
Using element 1
Producing the element after 1
Using element 2
...
Producing the element after 8
Using element 9
Producing the element after 9

real    0m10.007s   ← 10 seconds total
user    0m0.004s
sys     0m0.000s
```

According to that output, the elements are produced and used sequentially.

On the other hand, it may not be necessary to wait for preceding elements to be processed before starting to produce the successive elements. The program would take less time if other elements could be produced while the front element is in use:

```
import std.parallelism;
//...
foreach (element; taskPool.asyncBuf(range, 2)) {
```

In the call above, `asyncBuf()` makes two elements ready in its buffer. Elements are produced in parallel while they are being used:

```

$ time ./deneme
Producing the element after 0
Producing the element after 1
Using element 0
Producing the element after 2
Using element 1
Producing the element after 3
Using element 2
Producing the element after 4
Using element 3
Producing the element after 5
Using element 4
Producing the element after 6
Producing the element after 7
Using element 5
Using element 6
Producing the element after 8
Producing the element after 9
Using element 7
Using element 8
Using element 9

real    0m6.007s   ← now 6 seconds
user    0m0.000s
sys     0m0.004s

```

The default value of buffer size is 100. The buffer size that produces the best performance would be different under different situations.

`asyncBuf()` can be used outside of `foreach` loops, a trait that sets it apart from `parallel()`. For example, the following code uses the return value of `asyncBuf()` as an `InputRange` which operates semi-eagerly:

```

auto range = Range(10);
auto asyncRange = taskPool.asyncBuf(range, 2);
writeln(asyncRange.front);

```

68.4 taskPool.map()

It helps to explain `map()` from the `std.algorithm` module before explaining `taskPool.map()`. `std.algorithm.map` is an algorithm commonly found in many functional languages. It calls a function with the elements of a range one-by-one and returns a range that consists of the results of calling that function with each element. It is a lazy algorithm: It calls the function as needed.

The fact that `std.algorithm.map` operates lazily is very powerful in many programs. However, if the function needs to be called with every element anyway and the operations on each element are independent from each other, laziness may be unnecessarily slower than parallel execution. `taskPool.map()` and `taskPool.amap()` from the `std.parallelism` module take advantage of multiple cores and run faster in many cases.

Let's compare these three algorithms using the `Student` example. Let's assume that `Student` has a member function that returns the average grade of the student. To demonstrate how parallel algorithms are faster, let's again slow this function down with `Thread.sleep()`.

std.algorithm.map takes the function as its template parameter, and the range as its function parameter. It returns a range that consists of the results of applying that function to the elements of the range:

```
auto result_range = map!func(range);
```

The function may be specified by the `=>` syntax as a *lambda expression* as we have seen in earlier chapters. The following program uses **map()** to call the **averageGrade()** member function on each element:

```
import std.stdio;
import std.algorithm;
import core.thread;

struct Student
{
    int number;
    int[] grades;

    double averageGrade() @property
    {
        writeln("Started working on student %s",
            number);
        Thread.sleep(dur!"seconds"(1));

        immutable average =
            reduce!((a, b) => a + b)(0.0, grades) / grades.length;

        writeln("Finished working on student %s", number);
        return average;
    }
}

void main()
{
    Student[] students;

    foreach (i; 0 .. 10) {
        /* Two grades for each student */
        students ~= Student(i, [80 + i, 90 + i]);
    }

    auto results = map!(a => a.averageGrade)(students);

    foreach (result; results) {
        writeln(result);
    }
}
```

Note that the program uses **std.algorithm.reduce** when calculating the average grade. We have seen **reduce()** in an example in the [Ranges chapter](#) before. **reduce()** and its parallel counterpart **taskPool.reduce()** will be explained below.

The output of the program demonstrates that **map()** operates lazily; **averageGrade()** is called for each result as the **foreach** loop iterates:

```

$ time ./deneme
Started working on student 0
Finished working on student 0
85          ← calculated as foreach iterates
Started working on student 1
Finished working on student 1
86
...
Started working on student 9
Finished working on student 9
94

real    0m10.006s   ← 10 seconds total
user    0m0.000s
sys     0m0.004s

```

If `std.algorithm.map` were an eager algorithm, the messages about the starts and finishes of the operations would be printed altogether at the top.

`taskPool.map()` from the `std.parallelism` module works essentially the same as `std.algorithm.map`. The only difference is that it executes the function calls semi-eagerly and stores the results in a buffer to be served from as needed. The size of this buffer is determined by the second parameter. For example, the following code would make ready the results of the function calls for three elements at a time:

```

import std.parallelism;
// ...
double averageGrade(Student student)
{
    return student.averageGrade;
}
// ...
auto results = taskPool.map!averageGrade(students, 3);

```

Note: The free-standing `averageGrade()` function above is needed due to a limitation that involves using local delegates with member function templates like `TaskPool.map`:

```

auto results = taskPool.map!(a => a.averageGrade)(students, 3);    // ← compilation ERROR

```

This time the operations are executed in waves of three elements:

```

$ time ./deneme
Started working on student 1 ← in parallel
Started working on student 2 ← but in unpredictable order
Started working on student 0
Finished working on student 1
Finished working on student 2
Finished working on student 0
85
86
87
Started working on student 4
Started working on student 5
Started working on student 3
Finished working on student 4
Finished working on student 3
Finished working on student 5
88
89
90
Started working on student 7
Started working on student 8
Started working on student 6
Finished working on student 7
Finished working on student 6
Finished working on student 8
91
92
93
Started working on student 9
Finished working on student 9
94

real    0m4.007s    ← 4 seconds total
user    0m0.000s
sys     0m0.004s

```

The second parameter of `map()` has the same meaning as `asyncBuf()`: It determines the size of the buffer that `map()` uses to store the results in. The third parameter is the work unit size as in `parallel()`; the difference being its default value, which is `size_t.max`:

```

/* ... */ = taskPool.map!func(range,
                             buffer_size = 100
                             work_unit_size = size_t.max);

```

68.5 taskPool.amap()

Parallel `amap()` works the same as parallel `map()` with two differences:

- It is fully eager.
- It works with `RandomAccessRange` ranges.

```

auto results = taskPool.amap!averageGrade(students);

```

Since it is eager, all of the results are ready by the time `amap()` returns:

```
$ time ./deneme
Started working on student 1    ← all are executed up front
Started working on student 0
Started working on student 2
Started working on student 3
Finished working on student 1
Started working on student 4
Finished working on student 2
Finished working on student 3
Started working on student 6
Finished working on student 0
Started working on student 7
Started working on student 5
Finished working on student 4
Started working on student 8
Finished working on student 6
Started working on student 9
Finished working on student 7
Finished working on student 5
Finished working on student 8
Finished working on student 9
85
86
87
88
89
90
91
92
93
94

real    0m3.005s    ← 3 seconds total
user    0m0.000s
sys     0m0.004s
```

`amap()` works faster than `map()` at the expense of using an array that is large enough to store all of the results. It consumes more memory to gain speed.

The optional second parameter of `amap()` is the work unit size as well:

```
auto results = taskPool.amap!averageGrade(students, 2);
```

The results can also be stored in a `RandomAccessRange` that is passed to `amap()` as its third parameter:

```
double[] results;
results.length = students.length;
taskPool.amap!averageGrade(students, 2, results);
```

68.6 `taskPool.reduce()`

As with `map()`, it helps to explain `reduce()` from the `std.algorithm` module first.

We have seen `std.algorithm.reduce` in the [Ranges chapter](#) before. `reduce()` is another high-level algorithm commonly found in many functional languages. Just like `map()`, it takes one or more functions as template parameters. As its function parameters, it takes a value to be used as the initial value of the result, and a range. `reduce()` calls the functions with the current value of the result and each element of the range. When no initial value is specified, the first element of the range is used instead.

Assuming that it defines a variable named `result` in its implementation, the way that `reduce()` works can be described by the following steps:

1. Assigns the initial value to `result`
2. Executes the expression `result = func(result, element)` for every element
3. Returns the final value of `result`

For example, the sum of the squares of the elements of an array can be calculated as in the following program:

```
import std.stdio;
import std.algorithm;

void main()
{
    writeln(reduce!((a, b) => a + b * b)(0, [5, 10]));
}
```

When the function is specified by the `=>` syntax as in the program above, the first parameter (here `a`) represents the current value of the result (initialized by the parameter `0` above) and the second parameter (here `b`) represents the current element.

The program outputs the sum of 25 and 100, the squares of 5 and 10:

```
125
```

As obvious from its behavior, `reduce()` uses a loop in its implementation. Because that loop is normally executed on a single core, it may be unnecessarily slow when the function calls for each element are independent from each other. In such cases `taskPool.reduce()` from the `std.parallelism` module can be used for taking advantage of all of the cores.

To see an example of this let's use `reduce()` with a function that is slowed down again artificially:

```
import std.stdio;
import std.algorithm;
import core.thread;

int aCalculation(int result, int element)
{
    writefln("started - element: %s, result: %s",
            element, result);

    Thread.sleep(dur!"seconds"(1));
    result += element;

    writefln("finished - element: %s, result: %s",
            element, result);

    return result;
}
```

```

}
void main()
{
    writeln("Result: ", reduce!aCalculation(0, [1, 2, 3, 4]));
}

```

reduce() uses the elements in sequence to reach the final value of the result:

```

$ time ./deneme
started - element: 1, result: 0
finished - element: 1, result: 1
started - element: 2, result: 1
finished - element: 2, result: 3
started - element: 3, result: 3
finished - element: 3, result: 6
started - element: 4, result: 6
finished - element: 4, result: 10
Result: 10

real    0m4.003s   ← 4 seconds total
user    0m0.000s
sys     0m0.000s

```

As in the **parallel()** and **map()** examples, importing the **std.parallelism** module and calling **taskPool.reduce()** is sufficient to take advantage of all of the cores:

```

import std.parallelism;
// ...
writeln("Result: ", taskPool.reduce!aCalculation(0, [1, 2, 3, 4]));

```

However, there are important differences in the way **taskPool.reduce()** works.

Like the other parallel algorithms, **taskPool.reduce()** executes the functions in parallel by using elements in different tasks. Each task works on the elements that it is assigned to and calculates a **result** that corresponds to the elements of that task. Since **reduce()** is called with only a single initial value, every task must use that same initial value to initialize its own **result** (the parameter **0** above).

The final values of the results that each task produces are themselves used in the same **result** calculation one last time. These final calculations are executed sequentially, not in parallel. For that reason, **taskPool.reduce()** may execute slower in short examples as in this chapter as will be observed in the following output.

The fact that the same initial value is used for all of the tasks, effectively being used in the calculations multiple times, **taskPool.reduce()** may calculate a result that is different from what **std.algorithm.reduce()** calculates. For that reason, the initial value must be the *identity value* for the calculation that is being performed, e.g. the **0** in this example which does not have any effect in addition.

Additionally, as the results are used by the same functions one last time in the sequential calculations, the types of the parameters that the functions take must be compatible with the types of the values that the functions return.

taskPool.reduce() should be used only under these considerations.

```
import std.parallelism;
// ...
writeln("Result: ", taskPool.reduce!aCalculation(0, [1, 2, 3, 4]));
```

The output of the program indicates that first the calculations are performed in parallel, and then their results are calculated sequentially. The calculations that are performed sequentially are highlighted in green:

```
$ time ./deneme
started - element: 3, result: 0 ← first, the tasks in parallel
started - element: 2, result: 0
started - element: 1, result: 0
started - element: 4, result: 0
finished - element: 3, result: 3
finished - element: 1, result: 1
started - element: 1, result: 0 ← then, their results sequentially
finished - element: 4, result: 4
finished - element: 2, result: 2
finished - element: 1, result: 1
started - element: 2, result: 1
finished - element: 2, result: 3
started - element: 3, result: 3
finished - element: 3, result: 6
started - element: 4, result: 6
finished - element: 4, result: 10
Result: 10

real    0m5.006s ← parallel reduce is slower in this example
user    0m0.004s
sys     0m0.000s
```

Parallel `reduce()` is faster in many other calculations like the calculation of the math constant π by quadrature.

68.7 Multiple functions and tuple results

`std.algorithm.map()`, `taskPool.map()`, `taskPool.amap()`, and `taskPool.reduce()` can all take more than one function, in which case the results are returned as a **Tuple**. We have seen the **Tuple** type in the [Tuples chapter](#) before. The results of individual functions correspond to the elements of the tuple in the order that the functions are specified. For example, the result of the first function is the first member of the tuple.

The following program demonstrates multiple functions with `std.algorithm.map`. Note that the return types of the functions need not be the same, as seen in the `quarterOf()` and `tenTimes()` functions below. In that case, the types of the members of the tuples would be different as well:

```
import std.stdio;
import std.algorithm;
import std.conv;

double quarterOf(double value)
{
    return value / 4;
}
```

```

string tenTimes(double value)
{
    return to!string(value * 10);
}

void main()
{
    auto values = [10, 42, 100];
    auto results = map!(quarterOf, tenTimes)(values);

    writeln(" Quarters Ten Times");

    foreach (quarterResult, tenTimesResult; results) {
        writeln("%8.2f%8s", quarterResult, tenTimesResult);
    }
}

```

The output:

Quarters	Ten Times
2.50	100
10.50	420
25.00	1000

In the case of `taskPool.reduce()`, the initial values of the results must be specified as a tuple:

```
taskPool.reduce!(foo, bar)(tuple(0, 1), [1, 2, 3, 4]);
```

68.8 TaskPool

Behind the scenes, the parallel algorithms from the `std.parallelism` module all use task objects that are elements of a `TaskPool` container. Normally, all of the algorithms use the same container object named `taskPool`.

`taskPool` contains appropriate number of tasks depending on the environment that the program runs under. For that reason, usually there is no need to create any other `TaskPool` object. Even so, explicit `TaskPool` objects may be created and used as needed.

The `TaskPool` constructor takes the number of threads to use during the parallel operations that are later started through it. The default value of the number of threads is one less than the number of cores on the system. All of the features that we have seen in this chapter can be applied to a separate `TaskPool` object.

The following example calls `parallel()` on a local `TaskPool` object:

```

import std.stdio;
import std.parallelism;

void main()
{
    auto workers = new TaskPool(2);

    foreach (i; workers.parallel([1, 2, 3, 4])) {
        writeln("Working on %s", i);
    }

    workers.finish();
}

```


TaskPool.finish() tells the object to stop processing when all of its current tasks are completed.

68.9 Summary

- It is an error to execute operations in parallel unless those operations are independent from each other.
- **parallel()** executes the iterations of **foreach** loops in parallel.
- Tasks can explicitly be created, started, and waited for by **task()**, **executeInNewThread()**, and **yieldForce()**, respectively.
- The exceptions that are escaped from tasks can be caught later by most of the parallelism functions like **yieldForce()**.
- **asyncBuf()** iterates the elements of an **InputRange** semi-eagerly in parallel.
- **map()** calls functions with the elements of an **InputRange** semi-eagerly in parallel.
- **amap()** calls functions with the elements of a **RandomAccessRange** fully-eagerly in parallel.
- **reduce()** makes calculations over the elements of a **RandomAccessRange** in parallel.
- **map()**, **amap()**, and **reduce()** can take multiple functions and return the results as tuples.
- When needed, **TaskPool** objects other than **taskPool** can be used.

69 Message Passing Concurrency

Concurrency is similar to but different from the topic of the previous chapter, parallelism. As these two concepts both involve executing programs on threads, and as parallelism is based on concurrency, they are sometimes confused with each other.

The following are the differences between parallelism and concurrency:

- The main purpose of parallelism is to take advantage of microprocessor cores to improve the performance of programs. Concurrency on the other hand, is a concept that may be needed even on a single-core environment. Concurrency is about making a program run on more than one thread at a time. An example of a concurrent program would be a server program that is responding to requests of more than one client at the same time.
- In parallelism, tasks are independent from each other. In fact, it would be a bug if they did depend on results of other tasks that are running at the same time. In concurrency, it is normal for threads to depend on results of other threads.
- Although both programming models use operating system threads, in parallelism threads are encapsulated by the concept of task. Concurrency makes use of threads explicitly.
- Parallelism is easy to use, and as long as tasks are independent it is easy to produce programs that work correctly. Concurrency is easy only when it is based on *message passing*. It is very difficult to write correct concurrent programs if they are based on the traditional model of concurrency that involves lock-based data sharing.

D supports both models of concurrency: message passing and data sharing. We will cover message passing in this chapter and data sharing in the next chapter.

69.1 Concepts

Thread: Operating systems execute programs as work units called *threads*. D programs start executing with `main()` on a thread that has been assigned to that program by the operating system. All of the operations of the program are normally executed on that thread. The program is free to start other threads to be able to work on multiple tasks at the same time. In fact, tasks that have been covered in the previous chapter are based on threads that are started automatically by `std.parallelism`.

The operating system can pause threads at unpredictable times for unpredictable durations. As a result, even operations as simple as incrementing a variable may be paused mid operation:

```
++i;
```

The operation above involves three steps: Reading the value of the variable, incrementing the value, and assigning the new value back to the variable. The thread may be paused at any point between these steps to be continued after an unpredictable time.

Message: Data that is passed between threads are called messages. Messages may be composed of any type and any number of variables.

Thread identifier: Every thread has an id, which is used for specifying recipients of messages.

Owner: Any thread that starts another thread is called the owner of the new thread.

Worker: Any thread that is started by an owner is called a worker.

69.2 Starting threads

`spawn()` takes a function pointer as a parameter and starts a new thread from that function. Any operations that are carried out by that function, including other functions that it may call, would be executed on the new thread. As soon as a new thread is started, the owner and the worker start executing separately as if they were independent programs:

```
import std.stdio;
import std.concurrency;
import core.thread;

void worker()
{
    foreach (i; 0 .. 5) {
        Thread.sleep(dur!"msecs"(500));
        writeln(i, " (worker)");
    }
}

void main()
{
    spawn(&worker);

    foreach (i; 0 .. 5) {
        Thread.sleep(dur!"msecs"(300));
        writeln(i, " (main)");
    }

    writeln("main is done.");
}
```

The examples in this chapter call `Thread.sleep` to slow down threads to demonstrate that they run at the same time. The output of the program shows that the two threads, one that runs `main()` and the other that has been started by `spawn()`, execute independently at the same time:

```
0 (main)
0 (worker)
1 (main)
2 (main)
1 (worker)
3 (main)
2 (worker)
4 (main)
main is done.
3 (worker)
4 (worker)
```

The program automatically waits for all of the threads to finish executing. We can see this in the output above by the fact that `worker()` continues executing even after `main()` exits after printing "main is done."

The parameters that the thread function takes are passed to `spawn()` as its second and later arguments. The two worker threads in the following program print four numbers each. They take the starting number as the thread function parameter:

```
import std.stdio;
import std.concurrency;
import core.thread;
```

```

void worker(int firstNumber)
{
    foreach (i; 0 .. 4) {
        Thread.sleep(dur!"msecs"(500));
        writeln(firstNumber + i);
    }
}

void main()
{
    foreach (i; 1 .. 3) {
        spawn(&worker, i * 10);
    }
}

```

The output of one of the threads is indicated in blue:

```

10
20
11
21
12
22
13
23

```

The lines of the output may be different at different times depending on how the threads are paused and resumed by the operating system.

69.3 Thread identifiers

The `thisTid` variable that is available globally at the module level is always the id of the *current* thread.

```

import std.stdio;
import std.concurrency;

void printTid(string tag)
{
    writefln("%s: %s, address: %s", tag, thisTid, &thisTid);
}

void worker()
{
    printTid("Worker");
}

void main()
{
    spawn(&worker);
    printTid("Owner ");
}

```

The type of `thisTid` is `Tid`, but its value has no significance for the program. Further, both threads report it to be surprisingly at the same address:

```

Owner : Tid(std.concurrency.MessageBox), address: 809C360
Worker: Tid(std.concurrency.MessageBox), address: 809C360

```

Even so, the two **thisTid** variables above are independent from each other. This is because in D data is thread-local by default.

It is common for the owner to pass its **thisTid** to the worker to introduce itself:

```
import std.concurrency;

void worker(Tid myOwner)
{
    // ...
}

void main()
{
    spawn(&worker, thisTid);
}
```

The return value of **spawn()**, which I have been ignoring until this point is the id of the worker thread:

```
Tid myWorker = spawn(&worker, thisTid);
```

That is how the owner and the worker obtain each other's thread ids.

69.4 Message Passing

send() sends messages and **receiveOnly()** waits for a message of a particular type. (There is also **prioritySend()**, **receive()**, and **receiveTimeout()**, which will be explained later below.)

The owner in the following program sends its worker a message of type **int** and waits for a message from the worker of type **double**. The threads continue sending messages back and forth until the owner sends a negative **int**. This is the owner thread:

```
void main()
{
    Tid worker = spawn(&workerFunc, thisTid);

    foreach (value; 1 .. 5) {
        worker.send(value);
        double result = receiveOnly!double();
        writeln("sent: %s, received: %s", value, result);
    }

    /* Sending a negative value to the worker so that it
     * terminates. */
    worker.send(-1);
}
```

main() stores the return value of **spawn()** under the name **worker** and uses that variable when sending messages to the worker.

On the other side, the worker receives the message that it needs as an **int**, uses that value in a calculation, and sends the result as type **double** to its owner:

```
void workerFunc(Tid owner)
{
    int value = 0;

    while (value >= 0) {
        value = receiveOnly!int();
        double result = to!double(value) / 5;
    }
}
```

```

    }
    owner.send(result);
}

```

The main thread reports the messages that it sends and the messages that it receives:

```

sent: 1, received: 0.2
sent: 2, received: 0.4
sent: 3, received: 0.6
sent: 4, received: 0.8

```

It is possible to send more than one value as a part of the same message. The following message consists of three parts:

```
owner.send(thisTid, 42, 1.5);
```

Values that are passed as parts of a single message appear as a tuple on the receiver's side. In such cases the template parameters of `receiveOnly()` must match the types of the tuple members:

```

/* Wait for a message composed of Tid, int, and double. */
auto message = receiveOnly!(Tid, int, double)();

auto sender = message[0]; // of type Tid
auto integer = message[1]; // of type int
auto floating = message[2]; // of type double

```

If the types do not match, a `MessageMismatch` exception is thrown:

```

import std.concurrency;

void workerFunc(Tid owner)
{
    // Sending string
    owner.send("hello");
}

void main()
{
    spawn(&workerFunc, thisTid);

    // Expecting double
    auto message = receiveOnly!double();
}

```

The output:

```

std.concurrency.MessageMismatch@std/concurrency.d(235):
Unexpected message type

```

The exceptions that the worker may throw cannot be caught by the owner. One solution is to have the worker catch the exception to be sent as a message. We will see this below.

69.4.1 Example

Let's use what we have seen so far in a simulation program.

The following program simulates robots moving randomly and independently in a two dimensional space. The movement of each robot is handled by a separate thread, which takes four pieces of information when started:

- The id of its owner: This id will be used when sending messages about the robot's movements.
- The number of the robot: This information is sent back to the owner to identify the robot that the message is about.
- The origin: This is where the robot starts moving from.
- The duration between each step: This information is used for determining when the robot's next step will be.

Last three pieces of this information can be stored in the following **Job** struct:

```
struct Job
{
    uint robotId;
    Position origin;
    Duration restDuration;
}
```

The thread function that moves each robot continuously sends the number of the robot and its movement to the owner thread:

```
void robotMover(Tid owner, Job job)
{
    Position from = job.origin;

    while (true) {
        Thread.sleep(job.restDuration);

        Position to = randomNeighbor(from);
        Movement movement = Movement(from, to);
        from = to;

        owner.send(MovementMessage(job.robotId, movement));
    }
}
```

The owner simply waits for these messages in an infinite loop. It identifies the robots by the robot number that is sent as a part of the message and simply prints each movement:

```
while (true) {
    auto message = receiveOnly!MovementMessage();

    writefln("%s %s",
             robots[message.robotId], message.movement);
}
```

All of the messages in this simple program go from the worker to the owner. Message passing would normally be more complicated in many other types of programs.

Here is the complete program:

```
import std.stdio;
import std.random;
import std.string;
import std.concurrency;
import core.thread;

struct Position
{
    int line;
```

```

    int column;

    string toString()
    {
        return format("%s,%s", line, column);
    }
}

struct Movement
{
    Position from;
    Position to;

    string toString()
    {
        return ((from == to)
                ? format("%s (idle)", from)
                : format("%s -> %s", from, to));
    }
}

class Robot
{
    string image;
    Duration restDuration;

    this(string image, Duration restDuration)
    {
        this.image = image;
        this.restDuration = restDuration;
    }

    override string toString()
    {
        return format("%s(%s)", image, restDuration);
    }
}

/* Returns a random position around 0,0. */
Position randomPosition()
{
    return Position(uniform! "["(-10, 10),
                   uniform! "["(-10, 10));
}

/* Returns at most one step from the specified coordinate. */
int randomStep(int current)
{
    return current + uniform! "["(-1, 1);
}

/* Returns a neighbor of the specified Position. It may be one
 * of the neighbors at eight directions, or the specified
 * position itself. */
Position randomNeighbor(Position position)
{
    return Position(randomStep(position.line),
                   randomStep(position.column));
}

struct Job
{
    uint robotId;
    Position origin;
    Duration restDuration;
}

struct MovementMessage
{
    uint robotId;
    Movement movement;
}

```



```

}

void robotMover(Tid owner, Job job)
{
    Position from = job.origin;

    while (true) {
        Thread.sleep(job.restDuration);

        Position to = randomNeighbor(from);
        Movement movement = Movement(from, to);
        from = to;

        owner.send(MovementMessage(job.robotId, movement));
    }
}

void main()
{
    /* Robots with various restDurations. */
    Robot[] robots = [ new Robot("A", dur!"msecs"( 600)),
                       new Robot("B", dur!"msecs"(2000)),
                       new Robot("C", dur!"msecs"(5000)) ];

    /* Start a mover thread for each robot. */
    foreach (uint robotId, robot; robots) {
        spawn(&robotMover, thisTid, Job(robotId,
                                         randomPosition(),
                                         robot.restDuration));
    }

    /* Ready to collect information about the movements of the
     * robots. */
    while (true) {
        auto message = receiveOnly!MovementMessage();

        /* Print the movement of this robot. */
        writefln("%s %s",
                 robots[message.robotId], message.movement);
    }
}

```

The program prints each movement until terminated:

```

A(600 ms) 6,2 -> 7,3
A(600 ms) 7,3 -> 8,3
A(600 ms) 8,3 -> 7,3
B(2 secs) -7,-4 -> -6,-3
A(600 ms) 7,3 -> 6,2
A(600 ms) 6,2 -> 7,1
A(600 ms) 7,1 (idle)
B(2 secs) -6,-3 (idle)
A(600 ms) 7,1 -> 7,2
A(600 ms) 7,2 -> 7,3
C(5 secs) -4,-4 -> -3,-5
A(600 ms) 7,3 -> 6,4
...

```

The power of concurrency by message passing is demonstrated in this program. Movements of each robot are calculated independently by separate threads. It is the owner thread that *serializes* the printing of the movements by printing them in the order that it receives the messages.

69.5 Expecting different types of messages

`receiveOnly()` can expect only one type of message. `receive()` on the other hand can wait for more than one type of message. It dispatches messages to message handling delegates. When a message arrives, it is compared to the message type of each delegate. The delegate that matches the type of the particular message handles it.

For example, the following `receive()` call specifies two message handlers that would handle messages of types `int` and `string`, respectively:

```
void workerFunc()
{
    bool isDone = false;

    while (!isDone) {
        void intHandler(int message)
        {
            writeln("handling int message: ", message);

            if (message == -1) {
                writeln("exiting");
                isDone = true;
            }
        }

        void stringHandler(string message)
        {
            writeln("handling string message: ", message);
        }

        receive(&intHandler, &stringHandler);
    }
}
```

Messages of type `int` would match `intHandler()` and messages of type `string` would match `stringHandler()`. The worker thread above can be tested by the following program:

```
import std.stdio;
import std.concurrency;

// ...

void main()
{
    auto worker = spawn(&workerFunc);

    worker.send(10);
    worker.send(42);
    worker.send("hello");
    worker.send(-1); // ← to terminate the worker
}
```

The output of the program indicates that the messages are handled by matching functions on the receiver's side:

```
handling int message: 10
handling int message: 42
handling string message: hello
handling int message: -1
exiting
```

Lambda functions and objects of types that define the `opCall()` member function can also be passed to `receive()` as message handlers. The following worker handles messages by

lambda functions. The following program also defines a special type named **Exit** used for communicating to the thread that it is time for it to exit. Using such a specific type is more expressive than sending the arbitrary value of -1.

There are three anonymous functions below that are passed to **receive()** as message handlers. Their curly brackets are highlighted:

```
import std.stdio;
import std.concurrency;

struct Exit
{}

void workerFunc()
{
    bool isDone = false;

    while (!isDone) {
        receive(
            (int message) {
                writeln("int message: ", message);
            },
            (string message) {
                writeln("string message: ", message);
            },
            (Exit message) {
                writeln("exiting");
                isDone = true;
            }
        );
    }
}

void main()
{
    auto worker = spawn(&workerFunc);

    worker.send(10);
    worker.send(42);
    worker.send("hello");
    worker.send(Exit());
}
```

69.5.1 Receiving any type of message

std.variant.Variant is a type that can encapsulate any type of data. Messages that do not match the handlers that are specified earlier in the argument list always match a **Variant** handler:

```
import std.stdio;
import std.concurrency;

void workerFunc()
{
    receive(
        (int message) { /* ... */ },
        (double message) { /* ... */ },
        (Variant message) {
            writeln("Unexpected message: ", message);
        }
    );
}

struct SpecialMessage
{
```

```

    /* ... */
}

void main()
{
    auto worker = spawn(&workerFunc);
    worker.send(SpecialMessage());
}

```

The output:

```
Unexpected message: SpecialMessage()
```

The details of **Variant** are outside of the scope of this chapter.

69.6 Waiting for messages up to a certain time

It may not make sense to wait for messages beyond a certain time. The sender may have been busy temporarily or may have terminated with an exception. `receiveTimeout()` prevents blocking the receiving thread indefinitely.

The first parameter of `receiveTimeout()` determines how long the message should be waited for. Its return value is `true` if a message has been received within that time, `false` otherwise.

```

import std.stdio;
import std.concurrency;
import core.thread;

void workerFunc(Tid owner)
{
    Thread.sleep(dur!"seconds"(3));
    owner.send("hello");
}

void main()
{
    spawn(&workerFunc, thisTid);

    writeln("Waiting for a message");
    bool received = false;
    while (!received) {
        received = receiveTimeout(dur!"msecs"(600),
            (string message) {
                writeln("received: ", message);
            });

        if (!received) {
            writeln("... no message yet");

            /* ... other operations may be executed here ... */
        }
    }
}

```

The owner above waits for a message for up to 600 milliseconds. It can continue working on other things if a message does not arrive within that time:

```

Waiting for a message
... no message yet
... no message yet
... no message yet
... no message yet
received: hello

```

69.7 Exceptions during the execution of the worker

As we have seen in the previous chapter, the facilities of the `std.parallelism` module automatically catch exceptions that have been thrown during the execution of tasks and rethrow them in the context of the owner. This allows the owner to catch such exceptions:

```

try {
    theTask.yieldForce();
} catch (Exception exc) {
    writefln("Detected an error in the task: '%s'",
            exc.msg);
}

```

`std.concurrency` does not provide such a convenience for general exception types. The exceptions can be caught and sent explicitly by the worker. As we will see below, it is also possible to receive `OwnerTerminated` and `LinkTerminated` exceptions as messages.

The `calculate()` function below receives `string` messages, converts them to `double`, adds 0.5, and sends the result back as a message:

```

void calculate(Tid owner)
{
    while (true) {
        auto message = receiveOnly!string();
        owner.send(to!double(message) + 0.5);
    }
}

```

The `to!double()` call above would throw an exception if the string cannot be converted to a `double` value. Because such an exception would terminate the worker thread right away, the owner in the following program can receive a response only for the first message:

```

import std.stdio;
import std.concurrency;
import std.conv;

/* ... */

void main()
{
    Tid calculator = spawn(&calculate, thisTid);

    calculator.send("1.2");
    calculator.send("hello"); // ← incorrect input
    calculator.send("3.4");

    foreach (i; 0 .. 3) {
        auto message = receiveOnly!double();
        writefln("result %s: %s", i, message);
    }
}

```

The owner receives the response for "1.2" as 1.7 but because the worker has been terminated, the owner would be blocked waiting for a message that would never arrive:

result 0: 1.7

← waiting for a message that will never arrive

One thing that the worker can do is to catch the exception explicitly and to send it as a special error message. The following program sends the reason of the failure as a **CalculationFailure** message. Additionally, this program takes advantage of a special message type to signal to the worker when it is time to exit:

```
import std.stdio;
import std.concurrency;
import std.conv;

struct CalculationFailure
{
    string reason;
}

struct Exit
{}

void calculate(Tid owner)
{
    bool isDone = false;

    while (!isDone) {
        receive(
            (string message) {
                try {
                    owner.send(to!double(message) + 0.5);
                } catch (Exception exc) {
                    owner.send(CalculationFailure(exc.msg));
                }
            },
            (Exit message) {
                isDone = true;
            });
    }
}

void main()
{
    Tid calculator = spawn(&calculate, thisTid);

    calculator.send("1.2");
    calculator.send("hello"); // ← incorrect input
    calculator.send("3.4");
    calculator.send(Exit());

    foreach (i; 0 .. 3) {
        writef("result %s: ", i);

        receive(
            (double message) {
                writeln(message);
            },
            (CalculationFailure message) {
                writefln("ERROR! '%s'", message.reason);
            });
    }
}
```

This time the reason of the failure is printed by the owner:

```
result 0: 1.7
result 1: ERROR! 'no digits seen'
result 2: 3.9
```

Another method would be to send the actual exception object itself to the owner. The owner can use the exception object or simply rethrow it:

```
// ... at the worker ...
    try {
        // ...
    } catch (shared(Exception) exc) {
        owner.send(exc);
    },

// ... at the owner ...
receive(
    // ...

    (shared(Exception) exc) {
        throw exc;
    });
```

The reason why the **shared** specifiers are necessary is explained in the next chapter.

69.8 Detecting thread termination

Threads can detect that the receiver of a message has terminated.

69.8.1 OwnerTerminated exception

This exception is thrown when receiving a message from the owner if the owner has been terminated. The intermediate owner thread below simply exits after sending two messages to its worker. This causes an **OwnerTerminated** exception to be thrown at the worker thread:

```
import std.stdio;
import std.concurrency;

void main()
{
    spawn(&intermediaryFunc);
}

void intermediaryFunc()
{
    auto worker = spawn(&workerFunc);
    worker.send(1);
    worker.send(2);
} // ← Terminates after sending two messages

void workerFunc()
{
    while (true) {
        auto m = receiveOnly!int(); // ← An exception is
                                   //   thrown if the owner
                                   //   has terminated.

        writeln("Message: ", m);
    }
}
```

The output:

```

Message: 1
Message: 2
std.concurrency.OwnerTerminated@std/concurrency.d(248):
Owner terminated

```

The worker can catch that exception to exit gracefully:

```

void workerFunc()
{
    bool isDone = false;

    while (!isDone) {
        try {
            auto m = receiveOnly!int();
            writeln("Message: ", m);

        } catch (OwnerTerminated exc) {
            writeln("The owner has terminated.");
            isDone = true;
        }
    }
}

```

The output:

```

Message: 1
Message: 2
The owner has terminated.

```

We will see below that this exception can be received as a message as well.

69.8.2 LinkTerminated exception

`spawnLinked()` is used in the same way as `spawn()`. When a worker that has been started by `spawnLinked()` terminates, a `LinkTerminated` exception is thrown at the owner:

```

import std.stdio;
import std.concurrency;

void main()
{
    auto worker = spawnLinked(&workerFunc, thisTid);

    while (true) {
        auto m = receiveOnly!int(); // ← An exception is
                                   //   thrown if the worker
                                   //   has terminated.
        writeln("Message: ", m);
    }
}

void workerFunc(Tid owner)
{
    owner.send(10);
    owner.send(20);
} // ← Terminates after sending two messages

```

The worker above terminates after sending two messages. Since the worker has been started by `spawnLinked()`, the owner is notified of the worker's termination by a `LinkTerminated` exception:


```

Message: 10
Message: 20
std.concurrency.LinkTerminated@std/concurrency.d(263):
Link terminated

```

The owner can catch the exception to do something special like terminating gracefully:

```

bool isDone = false;
while (!isDone) {
    try {
        auto m = receiveOnly!int();
        writeln("Message: ", m);
    } catch (LinkTerminated exc) {
        writeln("The worker has terminated.");
        isDone = true;
    }
}

```

The output:

```

Message: 10
Message: 20
The worker has terminated.

```

This exception can be received as a message as well.

69.8.3 Receiving exceptions as messages

The **OwnerTerminated** and **LinkTerminated** exceptions can be received as messages as well. The following code demonstrates this for the **OwnerTerminated** exception:

```

bool isDone = false;
while (!isDone) {
    receive(
        (int message)
        {
            writeln("Message: ", message);
        },
        (OwnerTerminated exc)
        {
            writeln("The owner has terminated; exiting.");
            isDone = true;
        }
    );
}

```

69.9 Mailbox management

Every thread has a private mailbox that holds the messages that are sent to that thread. The number of messages in a mailbox may increase or decrease depending on how long it takes for the thread to receive and respond to each message. A continuously growing mailbox puts stress on the entire system and may point to a design flaw in the program. It may also mean that the thread may never get to the most recent messages.

setMaxMailboxSize() is used for limiting the number of messages that a mailbox can hold. Its three parameters specify the mailbox, the maximum number of messages that it can hold,

and what should happen when the mailbox is full, in that order. There are four choices for the last parameter:

- **OnCrowding.block**: The sender waits until there is room in the mailbox.
- **OnCrowding.ignore**: The message is discarded.
- **OnCrowding.throwException**: A **MailboxFull** exception is thrown when sending the message.
- A function pointer of type **bool function(Tid)**: The specified function is called.

Before seeing an example of **setMaxMailboxSize()**, let's first cause a mailbox to grow continuously. The worker in the following program sends messages back to back but the owner spends some time for each message:

```

/* WARNING: Your system may become unresponsive when this
 * program is running. */
import std.concurrency;
import core.thread;

void workerFunc(Tid owner)
{
    while (true) {
        owner.send(42);    // ← Produces messages continuously
    }
}

void main()
{
    spawn(&workerFunc, thisTid);

    while (true) {
        receive(
            (int message) {
                // Spends time for each message
                Thread.sleep(dur!"seconds"(1));
            });
    }
}

```

Because the consumer is slower than the producer, the memory that the program above uses would grow continuously. To prevent that, the owner may limit the size of its mailbox before starting the worker:

```

void main()
{
    setMaxMailboxSize(thisTid, 1000, OnCrowding.block);

    spawn(&workerFunc, thisTid);
    // ...
}

```

The **setMaxMailboxSize()** call above sets the main thread's mailbox size to 1000. **OnCrowding.block** causes the sender to wait until there is room in the mailbox.

The following example uses **OnCrowding.throwException**, which causes a **MailboxFull** exception to be thrown when sending a message to a mailbox that is full:

```

import std.concurrency;
import core.thread;

void workerFunc(Tid owner)
{
    while (true) {
        try {
            owner.send(42);
        }
    }
}

```

```

        } catch (MailboxFull exc) {
            /* Failed to send; will try again later. */
            Thread.sleep(dur!"msecs"(1));
        }
    }
}

void main()
{
    setMaxMailboxSize(thisTid, 1000, OnCrowding.throwException);

    spawn(&workerFunc, thisTid);

    while (true) {
        receive(
            (int message) {
                Thread.sleep(dur!"seconds"(1));
            });
    }
}

```

69.10 Priority messages

Messages can be sent with higher priority than regular messages by `prioritySend()`. These messages are handled before the other messages that are already in the mailbox:

```
prioritySend(owner, ImportantMessage(100));
```

If the receiver does not have a message handler that matches the type of the priority message, then a `PriorityMessageException` is thrown:

```
std.concurrency.PriorityMessageException@std/concurrency.d(280):
Priority message
```

69.11 Thread names

In the simple programs that we have used above, it was easy to pass the thread ids of owners and workers. Passing thread ids from thread to thread may be overly complicated in programs that use more than a couple of threads. To reduce this complexity, it is possible to assign names to threads, which are globally accessible from any thread.

The following three functions define an interface to an associative array that every thread has access to:

- `register()`: Associates a thread with a name.
- `locate()`: Returns the thread that is associated with the specified name. If there is no thread associated with that name, then `Tid.init` is returned.
- `unregister()`: Breaks the association between the specified name and the thread.

The following program starts two threads that find each other by their names. These threads continuously send messages to each other until instructed to terminate by an `Exit` message:

```

import std.stdio;
import std.concurrency;
import core.thread;

struct Exit
{}

```

```

void main()
{
    // A thread whose partner is named "second"
    auto first = spawn(&player, "second");
    register("first", first);
    scope(exit) unregister("first");

    // A thread whose partner is named "first"
    auto second = spawn(&player, "first");
    register("second", second);
    scope(exit) unregister("second");

    Thread.sleep(dur!"seconds"(2));

    prioritySend(first, Exit());
    prioritySend(second, Exit());
}

void player(string nameOfPartner)
{
    Tid partner;

    while (partner == Tid.init) {
        Thread.sleep(dur!"msecs"(1));
        partner = locate(nameOfPartner);
    }

    bool isDone = false;

    while (!isDone) {
        partner.send("hello " ~ nameOfPartner);
        receive(
            (string message) {
                writeln("Message: ", message);
                Thread.sleep(dur!"msecs"(500));
            },
            (Exit message) {
                writefln("%s, I am exiting.", nameOfPartner);
                isDone = true;
            });
    }
}

```

The output:

```

Message: hello second
Message: hello first
Message: hello second
Message: hello first
Message: hello first
Message: hello second
Message: hello first
Message: hello second
second, I am exiting.
first, I am exiting.

```

69.12 Summary

- When threads do not depend on other threads, prefer *parallelism*, which has been the topic of the previous chapter. Consider *concurrency* only when threads depend on operations of other threads.

- Because concurrency by data sharing is hard to implement correctly, prefer concurrency by message passing, which is the subject of this chapter.
- **spawn()** and **spawnLinked()** start threads.
- **thisTid** is the thread id of the current thread.
- **send()** and **prioritySend()** send messages.
- **receiveOnly()**, **receive()**, and **receiveTimeout()** wait for messages.
- **Variant** matches any type of message.
- **setMaxMailboxSize()** limits the size of mailboxes.
- **register()**, **unregister()**, and **locate()** allow referring to threads by name.
- Exceptions may be thrown during message passing: **MessageMismatch**, **OwnerTerminated**, **LinkTerminated**, **MailboxFull**, and **PriorityMessageException**.
- The owner cannot automatically catch exceptions that are thrown from the worker.

70 Skipping chapters...

There are the following chapters of the book at this point but they haven't been translated from Turkish yet:

- [Veri Paylaşarak Eş Zamanlı Programlama](#) `synchronized shared`
- [Tür Nitelikleri](#) `__traits isNumeric isNarrowString` (ve daha başka)
- [Bellek Yönetimi](#) `calloc emplace destroy static this static ~this .alignof`

Please click [Next →] above or below for the next available chapter in English.

71 Exercise Solutions

71.1 Assignment and Order of Evaluation

- The values of **a**, **b**, and **c** are printed on the right-hand side of each operation. The value that changes at every operation is printed in yellow:

```
at the beginning → a 1, b 2, c irrelevant
c = a           → a 1, b 2, c 1
a = b           → a 2, b 2, c 1
b = c           → a 2, b 1, c 1
```

At the end, the values of **a** and **b** have been swapped.

71.2 The Hello World Program

-

```
import std.stdio;

void main()
{
    writeln("Something else... :p");
}
```

-

```
import std.stdio;

void main()
{
    writeln("A line...");
    writeln("Another line...");
}
```

- The following program cannot be compiled because the semicolon at the end of the **writeln** line is missing:

```
import std.stdio;

void main()
{
    writeln("Hello world!")
}
```

71.3 writeln and write

- One method is to use another parameter in between:

```
writeln("Hello world!", " ", "Hello fish!");
```

- write** can take multiple parameters as well:

```
write("one", " two", " three");
```

71.4 Fundamental Types

1. We can use other types instead of `int`:

```
import std.stdio;

void main()
{
    writeln("Type           : ", short.stringof);
    writeln("Length in bytes: ", short.sizeof);
    writeln("Minimum value  : ", short.min);
    writeln("Maximum value  : ", short.max);

    writeln();

    writeln("Type           : ", ulong.stringof);
    writeln("Length in bytes: ", ulong.sizeof);
    writeln("Minimum value  : ", ulong.min);
    writeln("Maximum value  : ", ulong.max);
}
```

71.5 Variables

- 1.

```
import std.stdio;

void main()
{
    int amount = 20;
    double rate = 2.11;

    writeln("I have exchanged ", amount,
           " Euros at the rate of ", rate);
}
```

71.6 Standard Input and Output Streams

- 1.

```
import std.stdio;

void main()
{
    stdout.writeln(1, ",", 2);
}
```

71.7 Reading from the Standard Input

1. Since the exchange rate is normally not a whole value, we can use `double` to represent that variable. Remembering to use "`%s`":

```
import std.stdio;

void main()
{
    write("Please enter the exchange rate: ");
    double rate;
    readf(" %s", &rate);

    write("How many Euros do you have? ");
    int amount;
}
```



```

readf(" %s", &amount);

writeln("Exchanging ", amount,
        " Euros at the rate of ", rate);
}

```

- When the characters cannot be converted to the desired type, **stdin** gets in an unusable state. For example, entering "abc" when an **int** is expected would make **stdin** unusable.

71.8 Logical Expressions

- Because the compiler recognizes **10 < value** already as an expression, it expects a comma after it to accept it as a legal argument to **writeln**. Using parentheses around the whole expression would not work either, because this time a closing parenthesis would be expected after the same expression.
- Grouping the expression as **(10 < value) < 20** removes the compilation error, because in this case first **10 < value** is evaluated and then its result is used with **< 20**.
We know that the value of a logical expression like **10 < value** is either **false** or **true**. **false** and **true** take part in integer expressions as 0 and 1, respectively. (We will see automatic type conversions in a later chapter.) As a result, the whole expression is the equivalent of either **0 < 20** or **1 < 20**, both of which evaluate to **true**.
- The expression "greater than the lower value and less than the upper value" can be coded like the following:

```
writeln("Is between: ", (value > 10) && (value < 20));
```

- "There is a bicycle for everyone" can be coded as **personCount <= bicycleCount** or **bicycleCount >= personCount**. The rest of the logical expression can directly be translated to program code from the exercise:

```

writeln("We are going to the beach: ",
        ((distance < 10) && (bicycleCount >= personCount))
        ||
        ((personCount <= 5) && existsCar && existsLicense)
        );

```

Note the placement of the **||** operator to help with readability by separating the two main conditions.

71.9 The if Statement

- The statement **writeln("Washing the plate")** is written indented as if to be within the **else** scope. However, because the scope of that **else** is not written with curly brackets, only the **writeln("Eating pie")** statement is considered under that **else**.
Since whitespaces are not important in D programs, the *plate statement* is actually an independent statement within **main** and is executed unconditionally. It confuses the reader as well because of having been indented more than usual. If the *plate statement* must really be within the **else** scope, there must be curly brackets around that scope:

```

import std.stdio;

void main()

```

```

{
    bool existsLemonade = true;

    if (existsLemonade) {
        writeln("Drinking lemonade");
        writeln("Washing the cup");
    } else {
        writeln("Eating pie");
        writeln("Washing the plate");
    }
}

```

2. We can come up with more than one design for the conditions of this game. I will show two examples. In the first one, we apply the information directly from the exercise:

```

import std.stdio;

void main()
{
    write("What is the value of the die? ");
    int die;
    readf(" %s", &die);

    if (die == 1) {
        writeln("You won");
    } else if (die == 2) {
        writeln("You won");
    } else if (die == 3) {
        writeln("You won");
    } else if (die == 4) {
        writeln("I won");
    } else if (die == 5) {
        writeln("I won");
    } else if (die == 6) {
        writeln("I won");
    } else {
        writeln("ERROR: ", die, " is invalid");
    }
}

```

Unfortunately, that program has many repetitions. We can achieve the same result by other designs. Here is one:

```

import std.stdio;

void main()
{
    write("What is the value of the die? ");
    int die;
    readf(" %s", &die);

    if ((die == 1) || (die == 2) || (die == 3)) {
        writeln("You won");
    } else if ((die == 4) || (die == 5) || (die == 6)) {
        writeln("I won");
    } else {
        writeln("ERROR: ", die, " is invalid");
    }
}

```

- The previous designs cannot be used in this case. It is not practical to type 1000 different values in a program and expect them all be correct or readable. For that reason, it is better to determine whether the value of the die is *within a range*:

```
if ((die >= 1) && (die <= 500))
```

71.10 The while Loop

- Because the initial value of **number** is 0, the logical expression of the **while** loop is **false** since the very beginning, and this is preventing from entering the loop body. A solution is to use an initial value that will allow the **while** condition to be **true** at the beginning:

```
int number = 3;
```

- All of the variables in the following program are default initialized to 0. This allows entering both of the loops at least once:

```
import std.stdio;

void main()
{
    int secretNumber;

    while ((secretNumber < 1) || (secretNumber > 10)) {
        write("Please enter a number between 1 and 10: ");
        readf(" %s", &secretNumber);
    }

    int guess;

    while (guess != secretNumber) {
        write("Guess the secret number: ");
        readf(" %s", &guess);
    }

    writeln("That is correct!");
}
```

71.11 Integers and Arithmetic Operations

- We can use the / operator for the division and the % operator for the remainder:

```
import std.stdio;

void main()
{
    int first;
    write("Please enter the first number: ");
    readf(" %s", &first);

    int second;
    write("Please enter the second number: ");
    readf(" %s", &second);

    int quotient = first / second;
    int remainder = first % second;

    writeln(first, " = ",
            second, " * ", quotient, " + ", remainder);
}
```

- We can determine whether the remainder is 0 or not with an **if** statement:

```

import std.stdio;

void main()
{
    int first;
    write("Please enter the first number: ");
    readf(" %s", &first);

    int second;
    write("Please enter the second number: ");
    readf(" %s", &second);

    int quotient = first / second;
    int remainder = first % second;

    // We cannot call writeln up front before determining
    // whether the remainder is 0 or not. We must terminate
    // the line later with a writeln.
    write(first, " = ", second, " * ", quotient);

    // The remainder must be printed only if non-zero.
    if (remainder != 0) {
        write(" + ", remainder);
    }

    // We are now ready to terminate the line.
    writeln();
}

```

3.

```

import std.stdio;

void main()
{
    while (true) {
        write("0: Exit, 1: Add, 2: Subtract, 3: Multiply,",
            " 4: Divide - Please enter the operation: ");

        int operation;
        readf(" %s", &operation);

        // Let's first validate the operation
        if ((operation < 0) || (operation > 4)) {
            writeln("I don't know this operation");
            continue;
        }

        if (operation == 0){
            writeln("Goodbye!");
            break;
        }

        // If we are here, we know that we are dealing with
        // one of the four operations. Now is the time to read
        // two integers from the user:

        int first;
        int second;

        write(" First number: ");
        readf(" %s", &first);

        write("Second number: ");
        readf(" %s", &second);

        int result;

        if (operation == 1) {
            result = first + second;
        }
    }
}

```

```

    } else if (operation == 2) {
        result = first - second;
    } else if (operation == 3) {
        result = first * second;
    } else if (operation == 4) {
        result = first / second;
    } else {
        writeln(
            "There is an error! ",
            "This condition should have never occurred.");
        break;
    }
    writeln("      Result: ", result);
}

```

4.

```

import std.stdio;

void main()
{
    int value = 1;

    while (value <= 10) {
        if (value != 7) {
            writeln(value);
        }

        ++value;
    }
}

```

71.12 Floating Point Types

1. Replacing the three **ints** with three **doubles** is sufficient:

```

double first;
double second;

// ...

double result;

```

2. The following program demonstrates how much more complicated it would become if more than five variables were needed:

```

import std.stdio;

void main()
{
    double value_1;
    double value_2;
    double value_3;
    double value_4;
    double value_5;

    write("Value 1: ");
    readf(" %s", &value_1);
    write("Value 2: ");
    readf(" %s", &value_2);
}

```

```

write("Value 3: ");
readf(" %s", &value_3);
write("Value 4: ");
readf(" %s", &value_4);
write("Value 5: ");
readf(" %s", &value_5);

writeln("Twice the values:");
writeln(value_1 * 2);
writeln(value_2 * 2);
writeln(value_3 * 2);
writeln(value_4 * 2);
writeln(value_5 * 2);

writeln("One fifth the values:");
writeln(value_1 / 5);
writeln(value_2 / 5);
writeln(value_3 / 5);
writeln(value_4 / 5);
writeln(value_5 / 5);
}

```

71.13 Arrays

1.

```

import std.stdio;

void main()
{
    write("How many values will be entered? ");
    int count;
    readf(" %s", &count);

    double[] values;
    values.length = count;

    // The counter is commonly named as 'i'
    int i;
    while (i < count) {
        write("Value ", i, ": ");
        readf(" %s", &values[i]);
        ++i;
    }

    writeln("In sorted order:");
    values.sort;

    i = 0;
    while (i < count) {
        write(values[i], " ");
        ++i;
    }
    writeln();

    writeln("In reverse order:");
    values.reverse;

    i = 0;
    while (i < count) {
        write(values[i], " ");
        ++i;
    }
    writeln();
}

```

2. The explanations are included as code comments:

```

import std.stdio;

void main()
{
    // Using dynamic arrays because it is not known how many
    // values are going to be read from the input
    int[] odds;
    int[] evens;

    writeln("Please enter integers (-1 to terminate):");

    while (true) {

        // Reading the value
        int value;
        readf(" %s", &value);

        // The special value of -1 breaks the loop
        if (value == -1) {
            break;
        }

        // Adding to the corresponding array, depending on
        // whether the value is odd or even. It is an even
        // number if there is no remainder when divided by 2.
        if ((value % 2) == 0) {
            evens ~= value;
        } else {
            odds ~= value;
        }
    }

    // The odds and evens arrays are sorted separately
    odds.sort;
    evens.sort;

    // The two arrays are then appended to form a new array
    int[] result;
    result = odds ~ evens;

    writeln("First the odds then the evens, sorted:");

    // Printing the array elements in a loop
    int i;
    while (i < result.length) {
        write(result[i], " ");
        ++i;
    }

    writeln();
}

```

3. There are three mistakes (bugs) in this program. The first two are with the **while** loops: both of the loop conditions use the `<=` operator instead of the `<` operator. As a result, the program uses invalid indexes and attempts to access elements that are not parts of the arrays.

Since it is more beneficial for you to debug the third mistake yourself, I would like you to first run the program after fixing the previous two bugs. You will notice that the program will not print the results. Why?

The answer will be revealed if you select the area between the following horizontal lines with your mouse.

71.14 Slices and Other Array Features

- Iterating over elements by consuming a slice from the beginning is an interesting concept. This method is also the basis of Phobos ranges that we will see in a later chapter.

```
import std.stdio;

void main()
{
    double[] array = [ 1, 20, 2, 30, 7, 11 ];

    double[] slice = array;    // Start with a slice that
                               // provides access to all of
                               // the elements of the array

    while (slice.length) {    // As long as there is at least
                               // one element in that slice

        if (slice[0] > 10) {  // Always use the first element
            slice[0] /= 2;    // in the expressions
        }

        slice = slice[1 .. $]; // Shorten the slice from the
                               // beginning
    }

    writeln(array);          // The actual elements are
                               // changed
}
```

71.15 Strings

- Although some of the functions in Phobos modules will be easy to use with strings, library documentations are usually terse compared to tutorials. You may find especially the Phobos ranges confusing at this point. We will see Phobos ranges in a later chapter.
- Many other functions may be chained as well:

```
import std.stdio;
import std.string;

void main()
{
    write("First name: ");
    string first = capitalize(chomp(readln()));

    write("Last name: ");
    string last = capitalize(chomp(readln()));

    string fullName = first ~ " " ~ last;
    writeln(fullName);
}
```

- This program uses two indexes to make a slice:

```
import std.stdio;
import std.string;

void main()
{
    write("Please enter a line: ");
    string line = chomp(readln());

    sizediff_t first_e = indexOf(line, 'e');

    if (first_e == -1) {
```



```

        writeln("There is no letter e in this line.");
    } else {
        sizediff t last_e = lastIndexOf(line, 'e');
        writeln(line[first_e .. last_e + 1]);
    }
}

```

71.16 Redirecting Standard Input and Output Streams

1. Redirecting standard input and output of programs are commonly used especially on Unix-based operating system consoles. Some programs are designed to work well when piped to other programs.

For example, a file named **deneme.d** can be searched under a directory tree by piping **find** and **grep** as in the following line:

```
find | grep deneme.d
```

find prints the names of all of the files to its output. **grep** receives that output through its input and prints the lines that contain **deneme.d** to its own output.

71.17 Files

•

```

import std.stdio;
import std.string;

void main()
{
    write("Please enter the name of the file to read from: ");
    string inFileName =.chomp(readln());
    File inFile = File(inFileName, "r");

    string outFileName = inFileName ~ ".out";
    File outFile = File(outFileName, "w");

    while (!inFile.eof()) {
        string line =.chomp(inFile.readln());

        if (line.length != 0) {
            outFile.writeln(line);
        }
    }

    writeln(outFileName, " has been created.");
}

```

71.18 auto and typeof

1. We can use **typeof** to determine the type of the literal and **.stringof** to get the name of that type as **string**:

```

import std.stdio;

void main()
{
    writeln(typeof(1.2).stringof);
}

```

The output:

```
double
```

71.19 The for Loop

1.

```
import std.stdio;

void main()
{
    for (int line = 0; line != 9; ++line) {
        for (int column = 0; column != 9; ++column) {
            write(line, ',', column, ' ');
        }

        writeln();
    }
}
```

2. Triangle:

```
import std.stdio;

void main()
{
    for (int line = 0; line != 9; ++line) {
        int length = line + 1;

        for (int i = 0; i != length; ++i) {
            write('*');
        }

        writeln();
    }
}
```

Parallelogram:

```
import std.stdio;

void main()
{
    for (int line = 0; line != 9; ++line) {
        for (int i = 0; i != line; ++i) {
            write(' ');
        }

        writeln("*****");
    }
}
```

Can you produce the diamond pattern?

```
 *
***
*****
*****
*****
***
 *
```

71.20 The Ternary Operator ?:

1. Although it may make more sense to use an **if-else** statement in this exercise, the following program uses two **?:** operators:

```
import std.stdio;

void main()
{
    write("Please enter the net amount: ");

    int amount;
    readf(" %s", &amount);

    writeln("$",
            amount < 0 ? -amount : amount,
            amount < 0 ? " lost" : " gained");
}
```

The program prints "gained" even when the value is zero. Modify the program to print a message more appropriate for zero.

71.21 Literals

1. The problem here is that the value on the right-hand side is too large to fit in an **int**. According to the rules about integer literals, its type is **long**. For that reason it doesn't fit the type of the variable on the left-hand side. There are at least two solutions.

One solution is to leave the type of the variable to the compiler for example by the **auto** keyword:

```
auto amount = 10_000_000_000;
```

The type of **amount** would be deduced to be **long** from its initial value from the right-hand side.

Another solution is to make the type of the variable **long** as well:

```
long amount = 10_000_000_000;
```

2. We can take advantage of the special **'\r'** character that takes the printing to the beginning of the line.

```
import std.stdio;

void main()
{
    for (int number = 0; ; ++number) {
        write("\rNumber: ", number);
    }
}
```

The output of that program may be erratic due to its interactions with the output buffer. The following program flushes the output buffer and waits for 1 millisecond after each write:

```
import std.stdio;
import core.thread;

void main()
```

```

{
    for (int number = 0; ; ++number) {
        write("\rNumber: ", number);
        stdout.flush();
        Thread.sleep(dur!"msecs"(1));
    }
}

```

Flushing the output is normally not necessary as it is flushed automatically before getting to the next line e.g. by `writeln`, or before reading from `stdin`.

71.22 Formatted Output

1. We have already seen that this is trivial with format specifiers:

```

import std.stdio;

void main()
{
    writeln("(Enter 0 to exit the program.)");

    while (true) {
        write("Please enter a number: ");
        long number;
        readf(" %s", &number);

        if (number == 0) {
            break;
        }

        writefln("%1$d <=> %1$#x", number);
    }
}

```

2. Remembering that the % character must appear twice in the format string to be printed as itself:

```

import std.stdio;

void main()
{
    write("Please enter the percentage value: ");
    double percentage;
    readf(" %s", &percentage);

    writefln("%%.2f", percentage);
}

```

71.23 Formatted Input

1. Using a format string where the parts of the date are replaced with %s would be sufficient:

```

import std.stdio;

void main()
{
    int year;
    int month;
    int day;

    readf("%s.%s.%s", &year, &month, &day);
}

```

```
writeln("Month: ", month);
}
```

71.24 The do-while Loop

1. This program is not directly related to the **do-while** loop, as any problem that is solved by the **do-while** loop can also be solved by the other loop statements.

The program can guess the number that the user is thinking of by shortening the candidate range from top or bottom according to the user's answers. For example, if its first guess is 50 and the user's reply is that the secret number is greater, the program would then know that the number must be in the range [51,100]. If the program then guesses another number right in the middle of that range, this time the number would be known to be either in the range [51,75] or in the range [76,100].

When the size of the range is 1, the program would be sure that it must be the number that the user has guessed.

71.25 Associative Arrays

1.
 - The **.keys** property returns a slice (i.e. dynamic array) that includes all of the keys of the associative array. Iterating over this slice and removing the element for each key by calling **.remove** would result in an empty associative array:

```
import std.stdio;

void main()
{
    string[int] names =
    [
        1   : "one",
        10  : "ten",
        100 : "hundred",
    ];

    writeln("Initial length: ", names.length);

    int[] keys = names.keys;

    for (int i = 0; i != keys.length; ++i) {
        writeln("Removing the element for %s", keys[i]);
        names.remove(keys[i]);
    }

    writeln("Final length: ", names.length);
}
```

That solution may be slow especially for large arrays. The following methods would empty the array in a single step.

- Another solution is to assign an empty array:

```
string[int] emptyAA;
names = emptyAA;
```

- Since the initial value of an array is an empty array anyway, the following method would achieve the same result:

```
names = names.init;
```

- The goal is to store multiple grades per student. Since multiple grades can be stored in a dynamic array, an associative array that maps from **string** to **int[]** would work here. The grades can be appended to the dynamic arrays that are stored in the associative array:

```
import std.stdio;

void main()
{
    /* The key type of this associative array is string and
     * the value type is int[], i.e. an array of ints. The
     * associative array is being defined with an extra
     * space in between to help distinguish the value type: */
    int[] [string] grades;

    /* The array of ints that correspond to "emre" is being
     * used for appending the new grade to that array: */
    grades["emre"] ~= 90;
    grades["emre"] ~= 85;

    /* Printing the grades of "emre": */
    writeln(grades["emre"]);
}
```

The grades can also be assigned altogether as an array:

```
import std.stdio;

void main()
{
    int[][string] grades;

    grades["emre"] = [ 90, 85, 95 ];

    writeln(grades["emre"]);
}
```

71.26 The foreach Loop

- To have an associative array that works the opposite of **names**, the types of the key and the value must be swapped. The new associative array must be defined as of type **int[string]**.

Iterating over the keys and the values of the original associative array while using keys as values and values as keys would populate the **numbers** table:

```
import std.stdio;

void main()
{
    string[int] names = [ 1:"one", 7:"seven", 20:"twenty" ];
    int[string] values;

    foreach (key, value; names) {
        values[value] = key;
    }

    writeln(values["twenty"]);
}
```

71.27 switch and case

-

```

import std.stdio;
import std.string;

void main()
{
    string op;
    double first;
    double second;

    write("Please enter the operation: ");
    op =.chomp(readln());

    write("Please enter two values separated by a space: ");
    readf(" %s %s", &first, &second);

    double result;

    final switch (op) {
    case "add":
        result = first + second;
        break;

    case "subtract":
        result = first - second;
        break;

    case "multiply":
        result = first * second;
        break;

    case "divide":
        result = first / second;
        break;
    }

    writeln(result);
}

```

2. By taking advantage of distinct **case** values:

```

final switch (op) {
case "add", "+":
    result = first + second;
    break;

case "subtract", "-":
    result = first - second;
    break;

case "multiply", "*":
    result = first * second;
    break;

case "divide", "/":
    result = first / second;
    break;
}

```

3. Since the **default** section is needed to throw the exception from, it cannot be a **final switch** statement anymore. Here are the parts of the program that are modified:

```

// ...

switch (op) {
// ...

default:

```

```

        throw new Exception("Invalid operation: " ~ op);
    }
// ...

```

71.28 enum

1.

```

import std.stdio;
import std.conv;

enum Operation { exit, add, subtract, multiply, divide }

void main()
{
    // Print the supported operations
    write("Operations - ");
    for (Operation operation;
         operation <= Operation.max;
         ++operation) {

        writef("%d:%s ", operation, operation);
    }
    writeln();

    // Infinite loop until the user wants to exit
    while (true) {
        write("Operation? ");

        // The input must be read in the actual type (int) of
        // the enum
        int operationCode;
        readf(" %s", &operationCode);

        /*
         * We will start using enum values instead of magic
         * constants from this point on. So the operation code
         * that has been read in int must be converted to its
         * corresponding enum value.
         *
         * (Type conversions will be covered in more detail in
         * a later chapter.)
         */
        Operation operation = cast(Operation)operationCode;

        if ((operation < Operation.min) ||
            (operation > Operation.max)) {
            writeln("ERROR: Invalid operation");
            continue;
        }

        if (operation == Operation.exit) {
            writeln("Goodbye!");
            break;
        }

        double first;
        double second;
        double result;

        write(" First operand? ");
        readf(" %s", &first);

        write("Second operand? ");
        readf(" %s", &second);

        switch (operation) {

```



```

    case Operation.add:
        result = first + second;
        break;

    case Operation.subtract:
        result = first - second;
        break;

    case Operation.multiply:
        result = first * second;
        break;

    case Operation.divide:
        result = first / second;
        break;

    default:
        throw new Exception(
            "ERROR: This line should have never been reached.");
    }

    writeln("        Result: ", result);
}
}

```

71.29 Functions

1.

```

import std.stdio;

void printMenu(string[] items, int firstNumber)
{
    foreach (i, item; items) {
        writeln(' ', i + firstNumber, ' ', item);
    }
}

void main()
{
    string[] items =
        [ "Black", "Red", "Green", "Blue", "White" ];
    printMenu(items, 1);
}

```

2. Here are some ideas:

- Write a function named **drawHorizontalLine()** to draw horizontal lines.
- Write a function named **drawSquare()** to draw squares. This function could take advantage of **drawVerticalLine()** and **drawHorizontalLine()** when drawing the square.
- Improve the functions to also take the character that is used when "drawing". This would allow drawing each shape with a different character:

```

void putDot(Canvas canvas, int line, int column, dchar dot)
{
    canvas[line][column] = dot;
}

```

71.30 Function Parameters

1. Because the parameters of this function are the kind that gets copied from the arguments, what get swapped in the function are those copies.

To make the function swap the arguments, both of the parameters must be passed by reference:

```
void swap(ref int first, ref int second)
{
    immutable int temp = first;
    first = second;
    second = temp;
}
```

With that change, now the variables in `main()` would be swapped:

```
2 1
```

Although not related to the original problem, also note that `temp` is specified as `immutable` as it would not be changed in the function once having been initialized.

71.31 Program Environment

1.

```
import std.stdio;
import std.conv;

int main(string[] args)
{
    if (args.length != 4) {
        stderr.writeln(
            "ERROR! Usage: \n    ", args[0],
            " a_number operator another_number");
        return 1;
    }

    immutable first = to!double(args[1]);
    string op = args[2];
    immutable second = to!double(args[3]);

    switch (op) {
    case "+":
        writeln(first + second);
        break;

    case "-":
        writeln(first - second);
        break;

    case "x":
        writeln(first * second);
        break;

    case "/":
        writeln(first / second);
        break;

    default:
        throw new Exception("Invalid operator: " ~ op);
    }

    return 0;
}
```

2.

```
import std.stdio;
import std.process;
```

```
void main()
{
    write("Please enter the command line to execute: ");
    string commandLine = readln();

    writeln("Return value: ", system(commandLine));
}
```

71.32 assert and enforce

1. You will notice that the program terminates normally when you enter **06:09** and **1:2**. However, you may notice that the start time is not what has been entered by the user:

```
1 hours and 2 minutes after 09:06 is 10:08.
```

As you can see, although the time that has been entered as **06:09**, the output contains **09:06**. This error will be caught by the help of an **assert** in the next problem.

2. The **assert** failure after entering **06:09** and **15:2** takes us to the following line:

```
string timeToString(in int hour, in int minute)
{
    assert((hour >= 0) && (hour <= 23));
    // ...
}
```

For this **assert** check to fail, this function must have been called with invalid **hour** value.

The only two calls to **timeToString()** in the program do not appear to have any problems:

```
writeln("%s hours and %s minutes after %s is %s.",
        durationHour, durationMinute,
        timeToString(startHour, startMinute),
        timeToString(endHour, endMinute));
```

A little more investigation should reveal the actual cause of the bug: The hour and minute variables are swapped when reading the start time:

```
readTime("Start time", startMinute, startHour);
```

That programming error causes the time to be interpreted as **09:06** and incrementing it by duration **15:2** causes an invalid hour value.

An obvious correction is to pass the hour and minute variables in the right order:

```
readTime("Start time", startHour, startMinute);
```

The output:

```
Start time? (HH:MM) 06:09
Duration? (HH:MM) 15:2
15 hours and 2 minutes after 06:09 is 21:11.
```

3. It is again the same **assert** check:

```
assert((hour >= 0) && (hour <= 23));
```

The reason is that `addDuration()` can produce hour values that are greater than 23. Adding a *remainder* operation at the end would ensure one of the *output guarantees* of the function:

```
void addDuration(in int startHour, in int startMinute,
                in int durationHour, in int durationMinute,
                out int resultHour, out int resultMinute)
{
    resultHour = startHour + durationHour;
    resultMinute = startMinute + durationMinute;

    if (resultMinute > 59) {
        ++resultHour;
    }

    resultHour %= 24;
}
```

Observe that the function has other problems. For example, `resultMinute` may end up being greater than 59. The following function calculates the minute value correctly and makes sure that the function's output guarantees are enforced:

```
void addDuration(in int startHour, in int startMinute,
                in int durationHour, in int durationMinute,
                out int resultHour, out int resultMinute)
{
    resultHour = startHour + durationHour;
    resultMinute = startMinute + durationMinute;

    resultHour += resultMinute / 60;
    resultHour %= 24;
    resultMinute %= 60;

    assert((resultHour >= 0) && (resultHour <= 23));
    assert((resultMinute >= 0) && (resultMinute <= 59));
}
```

4. Good luck.

71.33 Unit Testing

- The first thing to do is to compile and run the program to ensure that the tests actually work and indeed fail:

```
$ dmd deneme.d -ofdeneme -w -unittest
$ ./deneme
core.exception.AssertError@deneme(11): unittest failure
```

The line number 11 indicates that the first one of the tests has failed.

For demonstration purposes let's write an obviously incorrect implementation that passes the first test by accident. The following function simply returns a copy of the input:

```
dstring toFront(dstring str, in dchar letter)
{
    dstring result;
```

```

    foreach (c; str) {
        result ~= c;
    }

    return result;
}

unittest
{
    immutable str = "hello"d;

    assert(toFront(str, 'h') == "hello");
    assert(toFront(str, 'o') == "ohell");
    assert(toFront(str, 'l') == "llheo");
}

void main()
{}

```

The first test passes but the second one fails:

```

$ ./deneme
core.exception.AssertError@deneme.d(17): unittest failure

```

Here is a correct implementation that passes all of the tests:

```

dstring toFront(dstring str, in dchar letter)
{
    dchar[] firstPart;
    dchar[] lastPart;

    foreach (c; str) {
        if (c == letter) {
            firstPart ~= c;
        } else {
            lastPart ~= c;
        }
    }

    return (firstPart ~ lastPart).idup;
}

unittest
{
    immutable str = "hello"d;

    assert(toFront(str, 'h') == "hello");
    assert(toFront(str, 'o') == "ohell");
    assert(toFront(str, 'l') == "llheo");
}

void main()
{}

```

The tests finally pass:

```

$ ./deneme
$

```

This function can now be modified in different ways under the confidence that its tests will have to pass. The following two implementations are very different from the first one but they too are correct according to the tests.

- An implementation that takes advantage of `std.algorithm.partition`:

```
import std.algorithm;

dstring toFront(dstring str, in dchar letter)
{
    dchar[] result = str.dup;
    partition!(c => c == letter, SwapStrategy.stable)(result);

    return result.idup;
}

unittest
{
    immutable str = "hello"d;

    assert(toFront(str, 'h') == "hello");
    assert(toFront(str, 'o') == "ohell");
    assert(toFront(str, 'l') == "llheo");
}

void main()
{}
```

Note: The `=>` syntax that appears in the program above creates a lambda function. We will see lambda functions in later chapters.

- The following implementation first counts how many times the special letter appears in the string. That information is then sent to a separate function named `repeated()` to produce the first part of the result. Note that `repeated()` has a set of unit tests of its own:

```
dstring repeated(size_t count, dchar letter)
{
    dstring result;

    foreach (i; 0..count) {
        result ~= letter;
    }

    return result;
}

unittest
{
    assert(repeated(3, 'z') == "zzz");
    assert(repeated(10, 'é') == "éééééééééé");
}

dstring toFront(dstring str, in dchar letter)
{
    size_t specialLetterCount;
    dstring lastPart;

    foreach (c; str) {
        if (c == letter) {
            ++specialLetterCount;
        } else {
            lastPart ~= c;
        }
    }

    return repeated(specialLetterCount, letter) ~ lastPart;
}

unittest
{}
```

```

immutable str = "hello"d;

assert(toFront(str, 'h') == "hello");
assert(toFront(str, 'o') == "ohell");
assert(toFront(str, 'l') == "llheo");
}

void main()
{}

```

71.34 Contract Programming

- The `unittest` block can be implemented trivially by copying the checks that are already written in `main()`. The only addition below is the test for the case when the second team wins:

```

int addPoints(in int goals1,
             in int goals2,
             ref int points1,
             ref int points2)
in
{
    assert(goals1 >= 0);
    assert(goals2 >= 0);
    assert(points1 >= 0);
    assert(points2 >= 0);
}
out (result)
{
    assert((result >= 0) && (result <= 2));
}
body
{
    int winner;

    if (goals1 > goals2) {
        points1 += 3;
        winner = 1;
    } else if (goals1 < goals2) {
        points2 += 3;
        winner = 2;
    } else {
        ++points1;
        ++points2;
        winner = 0;
    }

    return winner;
}

unittest
{
    int points1 = 10;
    int points2 = 7;
    int winner;

    // First team wins
    winner = addPoints(3, 1, points1, points2);
    assert(points1 == 13);
    assert(points2 == 7);
    assert(winner == 1);

    // Draw
    winner = addPoints(2, 2, points1, points2);
    assert(points1 == 14);
}

```

```

assert(points2 == 8);
assert(winner == 0);

// Second team wins
winner = addPoints(0, 1, points1, points2);
assert(points1 == 14);
assert(points2 == 11);
assert(winner == 2);
}

void main()
{
    // ...
}

```

71.35 Structs

1. One of the simplest designs is to use two **dchar** members:

```

struct Card
{
    dchar suit;
    dchar value;
}

```

2. It would be as simple as printing the two members side by side:

```

void printCard(in Card card)
{
    write(card.suit, card.value);
}

```

3. Assuming that there is already a function called **newSuit()**, **newDeck()** can be implemented by calling that function for each suit:

```

Card[] newDeck()
out (result)
{
    assert(result.length == 52);
}
body
{
    Card[] deck;

    deck ~= newSuit('♠');
    deck ~= newSuit('♥');
    deck ~= newSuit('♦');
    deck ~= newSuit('♣');

    return deck;
}

```

The rest of the work can be accomplished by the following **newSuit()**, which constructs the suit by combining the suit character with each value of a string:

```

Card[] newSuit(in dchar suit)
in
{
    assert((suit == '♠') ||
           (suit == '♥') ||
           (suit == '♦') ||
           (suit == '♣'));
}
out (result)

```



```

{
    assert(result.length == 13);
}
body
{
    Card[] suitCards;

    foreach (value; "234567890JQKA") {
        suitCards ~= Card(suit, value);
    }

    return suitCards;
}

```

Note that the functions above take advantage of contract programming to reduce risk of program errors.

- Swapping two elements at random would make the deck become more and more shuffled at each repetition. Although it is possible to pick the same element by chance, swapping an element with itself does not have any effect other than missing an opportunity towards a more shuffled deck.

```

void shuffle(Card[] deck, in int repetition)
{
    /* Note: A better algorithm is to walk the deck from the
     * beginning to the end and to swap each element
     * with a random one that is picked among the
     * elements from that point to the end.
     *
     * It would be even better to call randomShuffle() from
     * the std.algorithm module, which already applies the
     * same algorithm. Please read the comment in main() to
     * see how randomShuffle() can be used.
     */

    foreach (i; 0 .. repetition) {
        // Pick two elements at random
        immutable first = uniform(0, deck.length);
        immutable second = uniform(0, deck.length);

        swap(deck[first], deck[second]);
    }
}

```

The function above calls **std.algorithm.swap**, which simply swaps the values of its two **ref** parameters. It is effectively the equivalent of the following function:

```

void mySwap(ref Card left,
            ref Card right)
{
    immutable temporary = left;
    left = right;
    right = temporary;
}

```

Here is the entire program:

```

import std.stdio;
import std.random;
import std.algorithm;

struct Card
{
    dchar suit;
    dchar value;
}

```

```

}

void printCard(in Card card)
{
    write(card.suit, card.value);
}

Card[] newSuit(in dchar suit)
in
{
    assert((suit == '♠') ||
           (suit == '♥') ||
           (suit == '♦') ||
           (suit == '♣'));
}
out (result)
{
    assert(result.length == 13);
}
body
{
    Card[] suitCards;

    foreach (value; "234567890JQKA") {
        suitCards ~= Card(suit, value);
    }

    return suitCards;
}

Card[] newDeck()
out (result)
{
    assert(result.length == 52);
}
body
{
    Card[] deck;

    deck ~= newSuit('♠');
    deck ~= newSuit('♥');
    deck ~= newSuit('♦');
    deck ~= newSuit('♣');

    return deck;
}

void shuffle(Card[] deck, in int repetition)
{
    /* Note: A better algorithm is to walk the deck from the
     * beginning to the end and to swap each element
     * with a random one that is picked among the
     * elements from that point to the end.
     *
     * It would be even better to call randomShuffle() from
     * the std.algorithm module, which already applies the
     * same algorithm. Please read the comment in main() to
     * see how randomShuffle() can be used.
     */

    foreach (i; 0 .. repetition) {
        // Pick two elements at random
        immutable first = uniform(0, deck.length);
        immutable second = uniform(0, deck.length);

        swap(deck[first], deck[second]);
    }
}

void main()
{

```

```

Card[] deck = newDeck();

shuffle(deck, 100);
/* Note: Instead of the shuffle() call above, it would be
 *      better to call randomShuffle() as in the
 *      following line:
 *
 * randomShuffle(deck);
 */

foreach (card; deck) {
    printCard(card);
    write(' ');
}

writeln();
}

```

71.36 Variable Number of Parameters

- For the `calculate()` function to be able to take variable number of parameters, its parameter list must include a slice of `Calculation` followed by `...`:

```

double[] calculate(in Calculation[] calculations ...)
{
    double[] results;

    foreach (calculation; calculations) {
        final switch (calculation.op) {
            case Operation.add:
                results ~= calculation.first + calculation.second;
                break;

            case Operation.subtract:
                results ~= calculation.first - calculation.second;
                break;

            case Operation.multiply:
                results ~= calculation.first * calculation.second;
                break;

            case Operation.divide:
                results ~= calculation.first / calculation.second;
                break;
        }
    }

    return results;
}

```

Each calculation is evaluated inside a loop and their results are appended to a slice of type `double[]`.

Here is the entire program:

```

import std.stdio;

enum Operation { add, subtract, multiply, divide }

struct Calculation
{
    Operation op;
    double first;
    double second;
}

```

```

double[] calculate(in Calculation[] calculations ...)
{
    double[] results;

    foreach (calculation; calculations) {
        final switch (calculation.op) {
            case Operation.add:
                results ~= calculation.first + calculation.second;
                break;

            case Operation.subtract:
                results ~= calculation.first - calculation.second;
                break;

            case Operation.multiply:
                results ~= calculation.first * calculation.second;
                break;

            case Operation.divide:
                results ~= calculation.first / calculation.second;
                break;
        }
    }

    return results;
}

void main()
{
    writeln(calculate(Calculation(Operation.add, 1.1, 2.2),
                    Calculation(Operation.subtract, 3.3, 4.4),
                    Calculation(Operation.multiply, 5.5, 6.6),
                    Calculation(Operation.divide, 7.7, 8.8)));
}

```

The output:

```
[3.3, -1.1, 36.3, 0.875]
```

71.37 Function Overloading

1. The following two overloads take advantage of the existing `info()` overloads:

```

void info(in Meal meal)
{
    info(meal.time);
    write('- ');
    info(addDuration(meal.time, TimeOfDay(1, 30)));

    write(" Meal, Address: ", meal.address);
}

void info(in DailyPlan plan)
{
    info(plan.amMeeting);
    writeln();
    info(plan.lunch);
    writeln();
    info(plan.pmMeeting);
}

```

Here is the entire program that uses all of these types:

```
import std.stdio;
```

```

struct TimeOfDay
{
    int hour;
    int minute;
}

void info(in TimeOfDay time)
{
    printf("%02s:%02s", time.hour, time.minute);
}

TimeOfDay addDuration(in TimeOfDay start,
                    in TimeOfDay duration)
{
    TimeOfDay result;

    result.minute = start.minute + duration.minute;
    result.hour = start.hour + duration.hour;
    result.hour += result.minute / 60;

    result.minute %= 60;
    result.hour %= 24;

    return result;
}

struct Meeting
{
    string    topic;
    size_t    attendanceCount;
    TimeOfDay start;
    TimeOfDay end;
}

void info(in Meeting meeting)
{
    info(meeting.start);
    write('-');
    info(meeting.end);

    printf(" \"%s\" meeting with %s attendees",
           meeting.topic,
           meeting.attendanceCount);
}

struct Meal
{
    TimeOfDay time;
    string    address;
}

void info(in Meal meal)
{
    info(meal.time);
    write('-');
    info(addDuration(meal.time, TimeOfDay(1, 30)));

    write(" Meal, Address: ", meal.address);
}

struct DailyPlan
{
    Meeting amMeeting;
    Meal    lunch;
    Meeting pmMeeting;
}

void info(in DailyPlan plan)
{
    info(plan.amMeeting);
    writeln();
}

```

```

    info(plan.lunch);
    writeln();
    info(plan.pmMeeting);
}

void main()
{
    immutable bikeRideMeeting = Meeting("Bike Ride", 4,
                                        TimeOfDay(10, 30),
                                        TimeOfDay(11, 45));

    immutable lunch = Meal(TimeOfDay(12, 30), "İstanbul");

    immutable budgetMeeting = Meeting("Budget", 8,
                                       TimeOfDay(15, 30),
                                       TimeOfDay(17, 30));

    immutable todaysPlan = DailyPlan(bikeRideMeeting,
                                     lunch,
                                     budgetMeeting);

    info(todaysPlan);
    writeln();
}

```

That `main()` function can also be written with only object literals:

```

void main()
{
    info(DailyPlan(Meeting("Bike Ride", 4,
                          TimeOfDay(10, 30),
                          TimeOfDay(11, 45)),

                  Meal(TimeOfDay(12, 30), "İstanbul"),

                  Meeting("Budget", 8,
                        TimeOfDay(15, 30),
                        TimeOfDay(17, 30))));

    writeln();
}

```

71.38 Member Functions

1. Potentially negative intermediate values make `decrement()` slightly more complicated than `increment()`:

```

struct TimeOfDay
{
    // ...

    void decrement(in Duration duration)
    {
        auto minutesToSubtract = duration.minute % 60;
        auto hoursToSubtract = duration.minute / 60;

        minute -= minutesToSubtract;

        if (minute < 0) {
            minute += 60;
            ++hoursToSubtract;
        }

        hour -= hoursToSubtract;

        if (hour < 0) {

```

```

        hour = 24 - (-hour % 24);
    }
}
// ...
}

```

2. To see how much easier it gets with `toString()` member functions, let's look at the `Meeting` overload of `info()` one more time:

```

void info(in Meeting meeting)
{
    info(meeting.start);
    write('-');
    info(meeting.end);

    writef(" \"%s\" meeting with %s attendees",
           meeting.topic,
           meeting.attendanceCount);
}

```

Taking advantage of the already-defined `TimeOfDay.toString`, the implementation of `Meeting.toString` becomes trivial:

```

string toString()
{
    return format("%s-%s \"%s\" meeting with %s attendees",
                 start, end, topic, attendanceCount);
}

```

Here is the entire program:

```

import std.stdio;
import std.string;

struct Duration
{
    int minute;
}

struct TimeOfDay
{
    int hour;
    int minute;

    string toString()
    {
        return format("%02s:%02s", hour, minute);
    }

    void increment(in Duration duration)
    {
        minute += duration.minute;

        hour += minute / 60;
        minute %= 60;
        hour %= 24;
    }
}

struct Meeting
{
    string    topic;
    int       attendanceCount;
    TimeOfDay start;
    TimeOfDay end;
}

```

```

    string toString()
    {
        return format("%s-%s \"%s\" meeting with %s attendees",
                      start, end, topic, attendanceCount);
    }
}

struct Meal
{
    TimeOfDay time;
    string    address;

    string toString()
    {
        TimeOfDay end = time;
        end.increment(Duration(90));

        return format("%s-%s Meal, Address: %s",
                      time, end, address);
    }
}

struct DailyPlan
{
    Meeting amMeeting;
    Meal    lunch;
    Meeting pmMeeting;

    string toString()
    {
        return format("%s\n%s\n%s",
                      amMeeting,
                      lunch,
                      pmMeeting);
    }
}

void main()
{
    auto bikeRideMeeting = Meeting("Bike Ride", 4,
                                   TimeOfDay(10, 30),
                                   TimeOfDay(11, 45));

    auto lunch = Meal(TimeOfDay(12, 30), "İstanbul");

    auto budgetMeeting = Meeting("Budget", 8,
                                   TimeOfDay(15, 30),
                                   TimeOfDay(17, 30));

    auto todaysPlan = DailyPlan(bikeRideMeeting,
                                  lunch,
                                  budgetMeeting);

    writeln(todaysPlan);
    writeln();
}

```

The output of the program is the same as the earlier one that has been using `info()` function overloads:

```

10:30-11:45 "Bike Ride" meeting with 4 attendees
12:30-14:00 Meal, Address: İstanbul
15:30-17:30 "Budget" meeting with 8 attendees

```


71.39 Operator Overloading

- The following implementation passes all of the unit tests. The design decisions have been included as code comments.

Some of the functions of this struct can be implemented to run more efficiently. Additionally, it would be beneficial to also *normalize* the numerator and denominator. For example, instead of keeping the values 20 and 60, the values could be divided by their *greatest common divisor* and the numerator and the denominator can be stored as 1 and 3 instead. Otherwise, most of the operations on the object would cause the values of the numerator and the denominator to increase.

```
import std.exception;
import std.conv;

struct Fraction
{
    long num; // numerator
    long den; // denominator

    /* As a convenience, the constructor uses the default
     * value of 1 for the denominator. */
    this(long num, long den = 1)
    {
        enforce(den != 0, "The denominator cannot be zero");

        this.num = num;
        this.den = den;

        /* Ensuring that the denominator is always positive
         * will simplify the definitions of some of the
         * operator functions. */
        if (this.den < 0) {
            this.num = -this.num;
            this.den = -this.den;
        }
    }

    /* Unary -: Returns the negative of this fraction. */
    Fraction opUnary(string op)() const
    if (op == "-")
    {
        /* Simply constructs and returns an anonymous
         * object. */
        return Fraction(-num, den);
    }

    /* ++: Increments the value of the fraction by one. */
    ref Fraction opUnary(string op)()
    if (op == "++")
    {
        /* We could have used 'this += Fraction(1)' here. */
        num += den;
        return this;
    }

    /* --: Decrements the value of the fraction by one. */
    ref Fraction opUnary(string op)()
    if (op == "--")
    {
        /* We could have used 'this -= Fraction(1)' here. */
        num -= den;
        return this;
    }

    /* +=: Adds to the right-hand fraction to this one. */
    ref Fraction opOpAssign(string op)(in Fraction rhs)
    if (op == "+")
```

```

{
    /* Addition formula: a/b + c/d = (a*d + c*b)/(b*d) */
    num = (num * rhs.den) + (rhs.num * den);
    den *= rhs.den;
    return this;
}

/* -=: Subtracts the right-hand fraction from this one. */
ref Fraction opOpAssign(string op)(in Fraction rhs)
if (op == "-")
{
    /* We make use of the already-defined operators += and
     * unary - here. Alternatively, the subtraction
     * formula could explicitly be applied similar to the
     * += operator above.
     *
     * Subtraction formula: a/b - c/d = (a*d - c*b)/(b*d)
     */
    this += -rhs;
    return this;
}

/* *=: Multiplies the fraction by the right-hand side. */
ref Fraction opOpAssign(string op)(in Fraction rhs)
if (op == "*")
{
    /* Multiplication formula: a/b * c/d = (a*c)/(b*d) */
    num *= rhs.num;
    den *= rhs.den;
    return this;
}

/* /=: Divides the fraction by the right-hand side. */
ref Fraction opOpAssign(string op)(in Fraction rhs)
if (op == "/")
{
    enforce(rhs.num != 0, "Cannot divide by zero");

    /* Division formula: (a/b) / (c/d) = (a*d)/(b*c) */
    num *= rhs.den;
    den *= rhs.num;
    return this;
}

/* Binary +=: Produces the result of adding this and the
 * right-hand side fractions. */
Fraction opBinary(string op)(in Fraction rhs) const
if (op == "+")
{
    /* Takes a copy of this fraction and adds the
     * right-hand side fraction to that copy. */
    Fraction result = this;
    result += rhs;
    return result;
}

/* Binary -=: Produces the result of subtracting the
 * right-hand side fraction from this one. */
Fraction opBinary(string op)(in Fraction rhs) const
if (op == "-")
{
    /* Uses the already-defined -= operator. */
    Fraction result = this;
    result -= rhs;
    return result;
}

/* Binary *: Produces the result of multiplying this
 * fraction with the right-hand side fraction. */
Fraction opBinary(string op)(in Fraction rhs) const
if (op == "*")

```

```

{
    /* Uses the already-defined *= operator. */
    Fraction result = this;
    result *= rhs;
    return result;
}

/* Binary /: Produces the result of dividing this fraction
 * by the right-hand side fraction. */
Fraction opBinary(string op)(in Fraction rhs) const
if (op == "/" )
{
    /* Uses the already-defined /= operator. */
    Fraction result = this;
    result /= rhs;
    return result;
}

/* Returns the value of the fraction as double. */
double opCast(T : double)() const
{
    /* A simple division. However, as dividing values of
     * type long would lose the part of the value after
     * the decimal point, we could not have written
     * 'num/den' here. */
    return to!double(num) / den;
}

/* Sort order operator: Returns a negative value if this
 * fraction is before, a positive value if this fraction
 * is after, and zero if both fractions have the same sort
 * order. */
int opCmp(const ref Fraction rhs) const
{
    immutable result = this - rhs;
    /* Being a long, num cannot be converted to int
     * automatically; it must be converted explicitly by
     * 'to' (or cast). */
    return to!int(result.num);
}

/* Equality comparison: Returns true if the fractions are
 * equal.
 *
 * The equality comparison had to be defined for this type
 * because the compiler-generated one would be comparing
 * the members one-by-one, without regard to the actual
 * values that the objects represent.
 *
 * For example, although the values of both Fraction(1,2)
 * and Fraction(2,4) are 0.5, the compiler-generated
 * opEquals would decide that they were not equal on
 * account of having members of different values.
 */
bool opEquals(const ref Fraction rhs) const
{
    /* Checking whether the return value of opCmp is zero
     * is sufficient here. */
    return opCmp(rhs) == 0;
}
}

unittest
{
    /* Must throw when denominator is zero. */
    assertThrown(Fraction(42, 0));

    /* Let's start with 1/3. */
    auto a = Fraction(1, 3);

    /* -1/3 */

```

```

assert(-a == Fraction(-1, 3));

/* 1/3 + 1 == 4/3 */
++a;
assert(a == Fraction(4, 3));

/* 4/3 - 1 == 1/3 */
--a;
assert(a == Fraction(1, 3));

/* 1/3 + 2/3 == 3/3 */
a += Fraction(2, 3);
assert(a == Fraction(1));

/* 3/3 - 2/3 == 1/3 */
a -= Fraction(2, 3);
assert(a == Fraction(1, 3));

/* 1/3 * 8 == 8/3 */
a *= Fraction(8);
assert(a == Fraction(8, 3));

/* 8/3 / 16/9 == 3/2 */
a /= Fraction(16, 9);
assert(a == Fraction(3, 2));

/* Must produce the equivalent value in type 'double'.
 *
 * Note that although double cannot represent every value
 * precisely, 1.5 is an exception. That is why this test
 * is being applied at this point. */
assert(to!double(a) == 1.5);

/* 1.5 + 2.5 == 4 */
assert(a + Fraction(5, 2) == Fraction(4, 1));

/* 1.5 - 0.75 == 0.75 */
assert(a - Fraction(3, 4) == Fraction(3, 4));

/* 1.5 * 10 == 15 */
assert(a * Fraction(10) == Fraction(15, 1));

/* 1.5 / 4 == 3/8 */
assert(a / Fraction(4) == Fraction(3, 8));

/* Must throw when dividing by zero. */
assertThrown(Fraction(42, 1) / Fraction(0));

/* The one with lower numerator is before. */
assert(Fraction(3, 5) < Fraction(4, 5));

/* The one with larger denominator is before. */
assert(Fraction(3, 9) < Fraction(3, 8));
assert(Fraction(1, 1_000) > Fraction(1, 10_000));

/* The one with lower value is before. */
assert(Fraction(10, 100) < Fraction(1, 2));

/* The one with negative value is before. */
assert(Fraction(-1, 2) < Fraction(0));
assert(Fraction(1, -2) < Fraction(0));

/* The ones with equal values must be both <= and >=. */
assert(Fraction(-1, -2) <= Fraction(1, 2));
assert(Fraction(1, 2) <= Fraction(-1, -2));
assert(Fraction(3, 7) <= Fraction(9, 21));
assert(Fraction(3, 7) >= Fraction(9, 21));

/* The ones with equal values must be equal. */
assert(Fraction(1, 3) == Fraction(20, 60));

```

```

    /* The ones with equal values with sign must be equal. */
    assert(Fraction(-1, 2) == Fraction(1, -2));
    assert(Fraction(1, 2) == Fraction(-1, -2));
}

void main()
{}

```

As has been mentioned in the chapter, string mixins can be used to combine the definitions of some of the operators. For example, the following definition covers the four arithmetic operators:

```

/* Binary arithmetic operators. */
Fraction opBinary(string op)(in Fraction rhs) const
    if ((op == "+" || (op == "-" ||
        (op == "*" || (op == "/")))
{
    /* Takes a copy of this fraction and applies the
    * right-hand side fraction to that copy. */
    Fraction result = this;
    mixin ("result " ~ op ~ "= rhs;");
    return result;
}

```

71.40 Inheritance

1. The member functions that are declared as **abstract** by superclasses must be defined by the **override** keyword by subclasses.

Ignoring the **Train** class for this exercise, **Locomotive.makeSound** and **RailwayCar.makeSound** can be implemented as in the following program:

```

import std.stdio;
import std.exception;

class RailwayVehicle
{
    void advance(in size_t kilometers)
    {
        writeln("The vehicle is advancing %s kilometers",
            kilometers);

        foreach (i; 0 .. kilometers / 100) {
            writeln(" %s", makeSound());
        }
    }

    abstract string makeSound();
}

class Locomotive : RailwayVehicle
{
    override string makeSound()
    {
        return "choo choo";
    }
}

class RailwayCar : RailwayVehicle
{
    // ...

    override string makeSound()
    {
        return "clack clack";
    }
}

```

```

}

class PassengerCar : RailwayCar
{
    // ...
}

class FreightCar : RailwayCar
{
    // ...
}

void main()
{
    auto railwayCar1 = new PassengerCar;
    railwayCar1.advance(100);

    auto railwayCar2 = new FreightCar;
    railwayCar2.advance(200);

    auto locomotive = new Locomotive;
    locomotive.advance(300);
}

```

2. The following program uses the sounds of the components of **Train** to make the sound of **Train** itself:

```

import std.stdio;
import std.exception;

class RailwayVehicle
{
    void advance(in size_t kilometers)
    {
        writeln("The vehicle is advancing %s kilometers",
            kilometers);

        foreach (i; 0 .. kilometers / 100) {
            writeln(" %s", makeSound());
        }
    }

    abstract string makeSound();
}

class Locomotive : RailwayVehicle
{
    override string makeSound()
    {
        return "choo choo";
    }
}

class RailwayCar : RailwayVehicle
{
    abstract void load();
    abstract void unload();

    override string makeSound()
    {
        return "clack clack";
    }
}

class PassengerCar : RailwayCar
{
    override void load()
    {
        writeln("The passengers are getting on");
    }
}

```

```

    }

    override void unload()
    {
        writeln("The passengers are getting off");
    }
}

class FreightCar : RailwayCar
{
    override void load()
    {
        writeln("The crates are being loaded");
    }

    override void unload()
    {
        writeln("The crates are being unloaded");
    }
}

class Train : RailwayVehicle
{
    Locomotive locomotive;
    RailwayCar[] cars;

    this(Locomotive locomotive)
    {
        enforce(locomotive !is null,
            "Locomotive cannot be null");
        this.locomotive = locomotive;
    }

    void addCar(RailwayCar[] cars...)
    {
        this.cars ~= cars;
    }

    override string makeSound()
    {
        string result = locomotive.makeSound();

        foreach (car; cars) {
            result ~= ", " ~ car.makeSound();
        }

        return result;
    }

    void departStation(string station)
    {
        foreach (car; cars) {
            car.load();
        }

        writeln("Departing from the %s station", station);
    }

    void arriveStation(string station)
    {
        writeln("Arriving at the %s station", station);

        foreach (car; cars) {
            car.unload();
        }
    }
}

void main()
{
    auto locomotive = new Locomotive;

```

```

auto train = new Train(locomotive);

train.addCar(new PassengerCar, new FreightCar);

train.departStation("Ankara");
train.advance(500);
train.arriveStation("Haydarpaşa");
}

```

The output:

```

The passengers are getting on
The crates are being loaded
Departing from the Ankara station
The vehicle is advancing 500 kilometers
choo choo, clack clack, clack clack
choo choo, clack clack, clack clack
choo choo, clack clack, clack clack
choo choo, clack clack, clack clack
choo choo, clack clack, clack clack
Arriving at the Haydarpaşa station
The passengers are getting off
The crates are being unloaded

```

71.41 Object

1. For the equality comparison, **rhs** being non-**null** and the members being equal would be sufficient:

```

enum Color { blue, green, red }

class Point
{
    int x;
    int y;
    Color color;

    // ...

    override bool opEquals(Object o) const
    {
        const rhs = cast(const Point)o;

        return rhs && (x == rhs.x) && (y == rhs.y);
    }
}

```

2. When the type of the right-hand side object is also **Point**, they are compared according to the values of the **x** members first and then according to the values of the **y** members:

```

class Point
{
    int x;
    int y;
    Color color;

    // ...

    override int opCmp(Object o) const
    {
        const rhs = cast(const Point)o;

```



```

        enforce(rhs);

        return (x != rhs.x
                ? x - rhs.x
                : y - rhs.y);
    }
}

```

3. Note that it is not possible to cast to type `const TriangularArea` inside `opCmp` below. When `rhs` is `const TriangularArea`, then its member `rhs.points` would be `const` as well. Since the parameter of `opCmp` is non-`const`, it would not be possible to pass `rhs.points[i]` to `point.opCmp`.

```

class TriangularArea
{
    Point[3] points;

    this(Point one, Point two, Point three)
    {
        points = [ one, two, three ];
    }

    override bool opEquals(Object o) const
    {
        const rhs = cast(const TriangularArea)o;
        return rhs && (points == rhs.points);
    }

    override int opCmp(Object o) const
    {
        auto rhs = cast(TriangularArea)o;
        enforce(rhs);

        foreach (i, point; points) {
            immutable comparison = point.opCmp(rhs.points[i]);

            if (comparison != 0) {
                /* The sort order has already been
                 * determined. Simply return the result. */
                return comparison;
            }
        }

        /* The objects are considered equal because all of
         * their points have been equal. */
        return 0;
    }

    override size_t toHash() const
    {
        /* Since the 'points' member is an array, we can take
         * advantage of the existing toHash algorithm for
         * array types. */
        return typeid(points).getHash(&points);
    }
}

```