Sommaire

Notations mathématiques

X, Y,	Variables aléatoires
xi, yi, ε_i	Réalisations des variables aléatoires X, Y,ε
X_n	Echantillon (aléatoire)
x_n	Echantillon (observé)
$\mathcal{L}$	Loi (générique) d'une variable aléatoire
$\mathcal{N}(0, 1)$	Loi gaussienne standard
$\mathcal{N}(\mu,\sigma^2)$	Loi gaussienne (normale) d'espérance μ et de variance σ^2
$\mathcal{T}(n)$	Loi de Student à n degrés de liberté
$\chi2(n)$ ou $\chi2n$	Loi du $\chi2$ à n degrés de liberté
$\mathcal{F}(n,m)$	Loi de Fisher à n et m degrés de liberté
σ^2	Variance d'une variable aléatoire
E(Y)	Esperance théorique de la variable aléatoire Y
Var(Y)	Variance théorique de la variable aléatoire Y
$\xrightarrow{p}$	Symbole de convergence en probabilité
$\hat{\sigma}$	Estimateur de σ
p	Proportion théorique
$\hat{p}$	Estimateur d'une proportion (ou d'une probabilité)
$IC_{1-\alpha(\theta)}$	Intervalle de confiance (aléatoire) de niveau de confiance $1-\alpha$ pour θ
$ic_{1-\alpha(\theta)}$	Intervalle de confiance (réalisé) de niveau de confiance $1-\alpha$ pour θ
$1-\alpha$	Niveau de confiance d'un intervalle de confiance
$\mathcal{H}_1$	Assertion d'intérêt dans les tests d'hypothèses
$\mathcal{H}_0$	Hypothèse dite nulle, contraire de H_1
α	Niveau de signification ou risque de première espèce dans les tests
R	Coefficient de corrélation empirique aléatoire de Pearson
r	Coefficient de corrélation empirique réalisé de Pearson
β_0, β_1	Coefficients inconnus d'un modèle de régression linéaire simple
$\hat{\beta}_0$, $\hat{\beta}_1$	Estimations des coefficients inconnus d'un modèle de régression linéaire simple

Introduction

La théorie de la statistique est composée de deux parties complémentaires, une composante théorique ainsi qu'une composante appliquée. La composante théorique s'appuie sur la théorie des probabilités et d'analyse. C'est la partie qui développe, innove et nourrie en continue cette théorie. Tandis que la composante appliquée s'intéresse plus à la résolution des problèmes réels en utilisant une base de données. Cette composante s'appuie sur la théorie statistique, la programmation, et les outils informatiques. La statistique appliquée est utilisée dans presque tous les domaines de l'activité humaine: ingénierie, management, économie, biologie, informatique et autres.

Dans le cadre de mon travail, j'ai commencé mon étude par une partie théorique traitant les méthodes d'estimation et les qualités des estimateurs. J'ai donnée deux méthodes classiques pour trouver le meilleur estimateur d'un paramètre donnée à savoir; La méthode des vraisemblances et la méthode des moments. Cette partie englobe également plusieurs tests paramétriques ainsi qu'une partie de modélisation en utilisant la régression linéaire simple. Dans la deuxième partie de ce manuscrit j'ai traité une base de données en appliquant les résultats de la première partie sous le langage de programmation **R.**

Chapitre 1 : Estimation ponctuelle

1. Introduction :

Dans ce chapitre, on suppose que les données $x_1,...,x_n$ sont n réalisations indépendantes d'une même variable aléatoire sous-jacente X. Il est équivalent de supposer que $x_1,...,x_n$ sont les réalisations de variables aléatoires $X_1,...,X_n$ indépendantes et de même loi que X. Nous adopterons ici la seconde formulation, qui est plus pratique à manipuler. Les techniques de statistique descriptive comme l'histogramme ou le graphe de probabilités permettent de faire des hypothèses sur la nature de la loi de probabilité des X_i. Des techniques statistiques plus sophistiquées comme les tests non paramétriques. Les tests d'adéquation permettent de valider ou non ces hypothèses. On supposera ici que ces techniques ont permis d'adopter une famille de lois de probabilité bien précise (par exemple, loi normale, loi de Poisson, etc.) pour la loi des X_i où les paramètres de la loi sont supposés connus ou inconnus.

On notera θ le paramètre inconnu. Le problème traité dans ce chapitre est celui de l'estimation du paramètre θ.

Il s'agit de donner, au vu des observations $x_1;\ldots;x_n$, une approximation ou une évaluation de que l'on espère la plus proche possible de la vraie valeur inconnue. On pourra proposer une unique valeur vraisemblable qu'on appelle (estimation ponctuelle dans ce chapitre).

2. Méthodes d'estimations :

Il existe de nombreuses méthodes pour estimer un paramètre θ. Dans cette section, nous ne nous intéressons qu'aux deux méthodes d'estimation les plus usuelles, la méthode des moments et la méthode du maximum de vraisemblance.

Mais il faut d'abord définir précisément la nature d'une estimation et surtout d'un estimateur.

2.1. Définition d'un estimateur :

Un **estimateur** d'une grandeur ϴ est une statistique T_n à valeurs dans l'ensemble θ. Où θ est l'ensemble des valeurs possible de ϴ.

Une **estimation** de ϴ est une réalisation t_n de l'**estimateur** T_n

Un **estimateur** est donc une variable aléatoire alors qu'une **estimation** est une valeur déterministe.

2.2. Méthode des moments :

Principe de la méthode :

La méthode des moments est relativement intuitive. Il semble naturel d'estimer les moments par leur version empirique :

- L'espérance E(X) correspond à une moyenne théorique et peut être estimée par la moyenne observée $X_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

- La variance $Var(X) = E(X- E(X))^2 = E(X^2) -(E(X))^2$ peut être estimée par la variance observée $S^2_n = \frac{1}{n}\sum_{i=1}^{n}(Xi - Xn)^2$

$$= \frac{1}{n}\sum_{i=1}^{n} Xi^2 - (Xn)^2 \ .$$

Lorsque l'on souhaite estimer un paramètre θ, on essaie de l'exprimer en fonction de E(X) et de Var(X) : θ = g(E(X),Var(X)). On sait ensuite que E(X) et de Var(X) peuvent être estimés respectivement par X_n et S^2_n.

On propose donc d'estimer θ par $\hat{\theta} = g(X_n, S^2_n)$.

Ce principe se généralise avec l'ensemble des moments d'ordre p.

Application :

Si $X_1,\ldots\ldots, X_n$ suivent la loi géométrique de paramètre $E(X_i) = \frac{1}{p}$ donc par définition $p = \frac{1}{E(X)} = g(E(X))$.

L'estimateur des moments de p est donc $\hat{p} = g(\bar{X}) = \frac{1}{\bar{X}}$.

2.3. Méthode de maximum de vraisemblance :

Principe de la méthode :

La vraisemblance d'un modèle et d'un échantillon correspond à la probabilité d'avoir obtenu cet échantillon lorsqu' on a ce modèle. Ainsi, si on suppose que le modèle est $\{p\theta, \theta \in \theta\}$, la vraisemblance des observations $x_1,\ldots,x_n$ s'écrit sous la forme :

$$£(\theta, \{x_1,\ldots,x_n\}) = \begin{cases} p\theta(x1), \ldots . , p\theta(xn) \; si\ on\ a\ une\ loi\ discrete, \\ f\theta(x1), \ldots . , f\theta(xn) \; si\ on\ a\ une\ loi\ continue\ , \end{cases}$$

Avec $p\theta$ et $f\theta$ respectivement fonction de masse et de densité associées à $F\theta$.

Le principe de l'estimation par maximum de vraisemblance est basé sur la plus grande probabilité d'obtenir les observations. Plus cette probabilité grande plus le modèle est proche de la réalité.

Ainsi, on retient le modèle pour lequel la vraisemblance de notre échantillon est la plus élevée : $\hat{\theta}_n = \arg\max_{\theta} £(\theta, \{x_1,\ldots, x_n\})$.

En pratique, le problème ci-dessus est compliqué à résoudre directement en raison de la présence du produit, pour remédier à ce problème il suffit de prendre le logarithme :

$$\hat{\theta}_n = \arg\max_{\theta} £(\theta, \{x_1,\ldots,x_n\})$$

Pour trouver le maximum, on resoud l'équation du premier ordre :

$$\frac{\partial \log £(\theta,\{x1,\ldots,xn\})}{\partial \theta}\Big|_{\theta=\hat{\theta}_n} = 0$$

La théorie nous dit que la solution de cette équation nous donne toujours un maximum.

On obtient $\hat{\theta}_n$ sous la forme $\hat{\theta}_n = g(x_1,\ldots..x_n)$.

L'estimateur du maximum de vraisemblance est alors $\hat{\theta}_n = g(x_1,\ldots..x_n)$ et l'estimation du maximum de vraisemblance est obtenue en remplaçant $x_1,\ldots..x_n$ par leurs valeurs numériques dans $\hat{\theta}_n = g(x_1,\ldots..x_n)$.

Application :

Si $X_1,\ldots\ldots, X_n$ suivent une loi normale $\mathcal{N}(m, \sigma^2)$, La fonction de vraisemblance du modèle est :

$$£(m,\sigma^2, \{x_1,\ldots., x_n\}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma 2}} e^{-\frac{(xi-m)^2}{2\sigma 2}}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma 2}}\right)^n \exp\left(-\frac{1}{2\sigma 2}\sum_{i=1}^{n}(xi-m)^2\right)$$

La log-vraisemblance vaut :

$$\text{Log } £(m,\sigma^2, \{x_1,\ldots., x_n\}) = \frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma 2}\sum_{i=1}^{n}(xi-m)^2) = 0$$

Les équations du premier ordre s écrivent :

$$\frac{\partial \log £(m,\sigma 2,\{x1,\ldots,xn\})}{\partial m} = \frac{1}{\sigma 2}\sum_{i=1}^{n}(xi-m) = 0$$

$$\frac{\partial \log £(m,\sigma 2,\{x1,\ldots,xn\})}{\partial \sigma^2} = -\frac{n}{2\sigma 2}+\frac{1}{2\sigma 4}\sum_{i=1}^{n}(xi-m)^2 = 0$$

La solution de ce système est m =$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} xi$ et $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(xi - m)^2$

Les estimateurs du maximum de vraisemblance sont donc $\hat{m} = \bar{X}_n$ et $\hat{\sigma}^2 = S^2_n$

Si $X_1, \ldots\ldots, X_n$ sont de loi géométrique de paramètre p alors la fonction de vraisemblance du modèle est :

£ (p, $\{x_1,\ldots,x_n\}$) = $\prod_{i=1}^{n} p(1-p)^{xi-1} = p^n(1-p)^{\sum xi - n}$

La log-vraisemblance vaut :

£ (p, $\{x_1,\ldots,x_n\}$) = nlog(p)+($\sum xi - n$)log(1-p)

L'équation du premier ordre s'écrit :

$\frac{\partial log£\ (p,\{x1,\ldots.,xn\})}{\partial p} = \frac{n}{p} = \frac{\sum xi - n}{1-p} = 0$

Ou encore $\frac{1}{p} = \frac{(\overline{X}n-1)}{1-p}$ dont la solution est p = $\frac{1}{\bar{X}n}$.

L'estimateur du maximum de vraisemblance de p est donc $\hat{p}_n = \frac{1}{\bar{X}n}$

Remarque :

La méthode des moments et la méthode du maximum de vraisemblance donnent les mêmes résultats pour quelques lois de probabilité. Pour d'autre les deux méthodes fournissent des estimateurs différents.

C'est le cas de la loi gamma. On a deux estimateurs différents pour chaque paramètre.

On doit donc se demander quel est le meilleur d'entre eux. Cela amène à se poser la question de la qualité et de l'optimalité d'un estimateur, ce qui fait l'objet de la section suivante.

3. Qualité d'un estimateur :

Pour qu'une statistique puisse être considerée comme un 'bon' estimateur d'un paramétre ou d'une fonction,on souhaite qu'elle posséde certaines qualités.

Les plus importantes d'entre elles sont données ci-dessous.

3.1. Biais :

Le biais d'un estimateur $\hat{\theta}$ est la différence entre la valeur attendue de cet estimateur et la valeur théorique du paramètre θ qu'il estime.

Un estimateur est dit non biaisé lorsque $E(\hat{\theta})-\theta = 0$ tandis qu'il est biaisé dans le cas contraire. L'absence de biais garantit donc qu'en moyenne l'estimateur restitue la valeur correcte du paramètre. Notons qu'un estimateur sera dit asymptotiquement non biaisé lorsqu'il est biaisé mais que ce biais tend vers 0 lorsque la taille de l'échantillon augmente.

On a donc :

$\hat{\theta}$ non biaisé $\Leftrightarrow E(\hat{\theta})-\theta = 0$

$\hat{\theta}$ asymptotiquement non biaisé $\Leftrightarrow \lim_{n\to\infty} E(\hat{\theta}) - \theta = 0$.

3.2. Efficacité :

On appelle efficacité d'un estimateur la quantité :

$$Eff(\hat{\theta}_n) = \frac{\frac{\partial E[\hat{\theta}]}{\partial \theta}}{\square n(\theta) var(\hat{\theta})} \quad \text{avec} \quad \square(\theta) = var[\frac{\partial £ (\theta,\{x1,....,xn\})}{\partial \theta}]$$

On a $0 \leq Eff(\hat{\theta}_n) \leq 1$.

$\hat{\theta}$ est dit estimateur efficace si et seulement si $Eff(\hat{\theta}_n) = 1$.

$\hat{\theta}$ est dit asymptotique efficace si et seulement si $\lim_{n\to\infty} Eff(\hat{\theta}_n) = 1$.

Si $\hat{\theta}_n$ est un estimateur sans biais de $\hat{\theta}$, $Eff(\hat{\theta}_n) = \frac{1}{n(\theta)var(\hat{\theta})}$.

Si la valeur de la borne de cramer-Rao ($\frac{1}{n(\theta)}$) est très grande il est impossible d'estimer correctement $\hat{\theta}$ car tous les estimateurs sans biais possibles auront une forte variance.

On peut donc juger de la qualité d'un estimateur sans biais en calculant son efficacité.

3.3. Convergence :

L'estimateur $\hat{\theta}_n$ converge en moyenne quadratique vers θ si et seulement si son erreur quadratique moyenne ($E((\hat{\theta}_n - \theta)^2)$) tend vers 0 quand n tend vers l'infini :

$$\hat{\theta}_n \longrightarrow \theta \Leftrightarrow \lim_{n\to\infty} E((\hat{\theta}_n - \theta)^2) = 0.$$

Si $\hat{\theta}_n$ est sans biais, il sera convergent en moyenne quadratique si et seulement si sa variance tend vers 0 quand n tend vers l'infini.

L'estimateur $\hat{\theta}_n$ est convergent s'il converge en probabilité vers θ

Soit : $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0 \qquad \forall \varepsilon > 0.$

4. Propriétés des estimateurs des moments et de maximum de vraisemblance :

4.1. Propriétés des estimateurs des moments :

Propriété 1 :

La moyenne empirique $\bar{X}_n$ est un estimateur sans biais et convergent en moyenne quadratique de E(X).

Démonstration:

On peut monter facilement que $\bar{X}_n$ est un bon estimateur de

θ = E(X) :

$E(\bar{X}_n) = E(\frac{1}{n}\sum_{i=1}^{n} Xi)$

$= \frac{1}{n}\sum_{i=1}^{n} E(Xi) = \frac{1}{n} n\, \theta = \theta.$

Donc $\bar{X}_n$ est un estimateur sans biais de θ = E(X).

La variance de $\bar{X}_n$ est :

$Var(\bar{X}_n) = var(\frac{1}{n}\sum_{i=1}^{n} Xi)$

$= \frac{1}{n^2}\sum_{i=1}^{n} var(Xi)$

$= \frac{var(X)}{n}$

Car les X_i sont indépendantes, donc la variance de leur somme égale à la somme de leurs variances, qui sont toutes égales à var(X).

$Var(\bar{X}_n)$ tend vers 0 quand n tend vers l'infini.

Propriété 2 :

La variance estimée $S'^2_n = \frac{1}{n-1}\sum_{i=1}^{n}(Xi - \bar{X}n)^2$ est un estimateur sans biais et convergent en moyenne quadratique de var(X).

Démonstration :

On considère maintenant l'estimation de la variance de la loi des X_i par la variance empirique de l'échantillon

$S^2_n = \frac{1}{n}\sum_{i=1}^{n}(Xi - \bar{X}n)^2 = \frac{1}{n}\sum_{i=1}^{n}(Xi)^2 - (\bar{X}n)^2$

Déterminons le biais de cet estimateur :

$E(S^2) = E(\frac{1}{n}\sum_{i=1}^{n}(Xi)^2 - \overline{(X}n)^2)$

$= \frac{1}{n}\sum_{i=1}^{n} E((Xi)^2) - E(\overline{(X}n)^2)$

$= E(X^2) - E(\bar{X}n^2)$

$= var(X) + E(X)^2 - var(\bar{X}n) - E(\bar{X}n)^2$

$= var(X) + E(X)^2 - \frac{var(X)}{n} - E(X)^2$

$= (1 - \frac{1}{n})var(X)$

$= \frac{n-1}{n}var(X) \neq var(X)$

Donc contrairement à ce qu'on pourrait croire, la variance empirique S^2_n n'est pas un estimateur sans biais de var(X). Cet estimateur n'est qu'asymptotiquement sans biais.

En revanche, on voit que $E(\frac{n-1}{n}S^2_n) = \frac{n-1}{n}E(S^2_n) = var(X)$, on pose donc $S'^2_n = \frac{n-1}{n}S^2_n = \frac{1}{n-1}\sum_{i=1}^{n}(Xi - \bar{X}n)^2$, S'^2_n appelée variance estimée de l'échantillon. Le résultat précédent montre que c'est un estimateur sans biais de var(X).

Par ailleurs on montre que :

$var(S'^2_n) = \frac{1}{n(n-1)}[(n-1)E((X-E(X))^4) - (n-3)var(X)^2]$ qui tend vers 0 quand n tend vers l'infini.

Remarque :

Pour cela la commande var(X) en **R** donne la variance estimée, et non pas la variance empirique de X.

4.2. Propriétés des estimateurs de maximum de vraisemblance :

Propriété 1 :

Si les X_i sont indépendantes et de même loi dépendant d'un paramètre réel θ. On a :

$\hat{\theta}_n$ converge presque surement vers θ.

Propriété 2 :

$\sqrt{\mathcal{I}n(\theta)}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$,(avec $\mathcal{I}n(\theta) = E((\frac{\partial ln\pounds(\theta,\{x1,\ldots,xn\})}{\partial\theta})^2)$) ce qui signifie que quand n est grand, $\hat{\theta}_n$ est approximativement de loi $\mathcal{N}(\theta, \frac{1}{\mathcal{I}n(\theta)})$. On déduit que $\hat{\theta}_n$ est asymptotiquement Gaussien, sans biais et efficace.

Remarque :

En général, l'estimateur de maximum de vraisemblance est meilleur que l'estimateur de moments au sens où $var_v(\hat{\theta}_n) \leq var_m(\hat{\theta}_n)$ c'est au moins vrai asymptotiquement.

5. Remarque :

Le fait que l'estimateur de maximum de vraisemblance soit asymptotiquement sans biais et efficace fait que si on a beaucoup de données, on est pratiquement certains que la méthode du maximum de vraisemblance est la meilleure méthode d'estimation possible. C'est pourquoi cette méthode est considérée comme globalement la meilleure. Elle est utilisée plus fréquemment ce qui n'a pas été le cas pour la méthode des moments.

Chapitre 2 : Mise en place des tests statistiques sous R

1. Introduction :

Le logiciel **R** est un logiciel de statistique crée par Ross Ihaka et Robert Gentleman. Il est à la fois un langage informatique et un environnement de travail : les commandes sont exécutées grâce à des instructions codées dans un langage relativement simple, les résultats sont affichés sous forme de texte et les graphiques sont visualisés directement dans une fenêtre qui leur est propre.

Ce logiciel sert à manipuler des données, à tracer des graphiques et à faire des analyses statistiques sur ces données telles que :

-Statistique descriptive ;

-Tests d'hypothèses ;

-Analyse de la variance ;

-Méthodes de régression linéaire (simple et multiple) ;

2. Installation de R et ses packages :

2.1. Installation de R sous Microsoft Windows :

Commencez par télécharger le logiciel **R** (fichier **R**-x -win.exe où x est le numéro de la dernière version disponible) à l'aide de votre navigateur web usuel à l'adresse suivante : http://cran.r-project.org/bin/windows/base/ Enregistrez ensuite ce fichier exécutable sur le bureau de Windows puis double-cliquez sur le fichier **R**-x -win.exe dont voici l'icône : 

Le logiciel s'installe alors et vous n'avez plus qu'à suivre les instructions qui s'affichent et à conserver les options proposées par défaut.

Lorsque l'icône  est ajoutée sur le bureau, l'installation peut être considérée comme terminée.

2.2. Installation de packages :

Le logiciel **R** contient un ou plusieurs packages les uns sont une partie de l'installation basique et les autres peuvent être installés directement à partir d'un site internet CRAN (compréhensive **R** archive network) (http://cran.r-project.org/web/packages), on peut aussi créer des nouveaux packages.

Il existe plusieurs moyens pour installer un package que nous présentons ci-dessous :

Installation à partir d'un fichier situé sur le disque :

Vous pouvez par exemple télécharger depuis le site mentionné ci-dessus le fichier : R2HTML numero.zip et l'enregistrer sur le bureau de Windows.

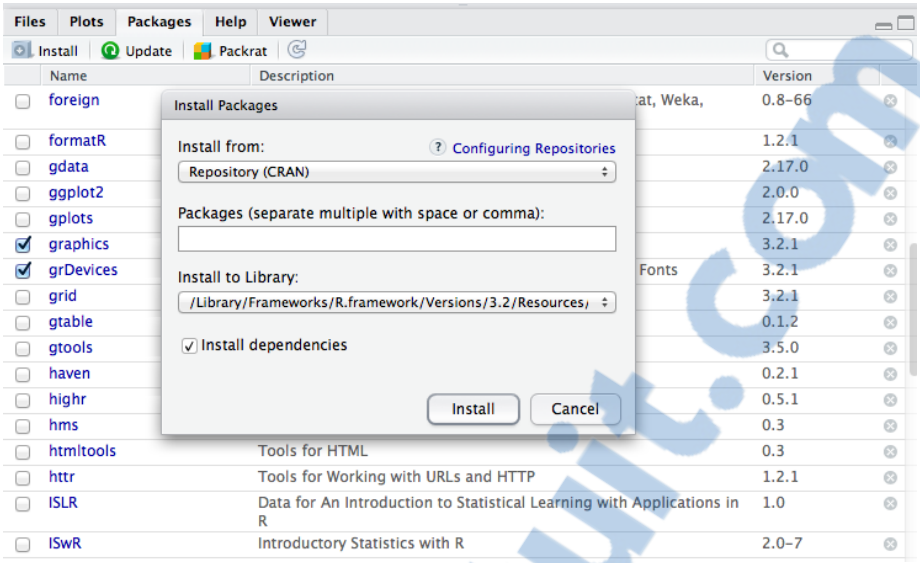
Pour installer ce package, commencez par lancer le logiciel **R** en double-cliquant sur son icône 

Ensuite, allez dans le menu Packages, puis dans le sous-menu Installer Sélectionnez alors le fichier R2HTML numero.zip situé sur le bureau de Windows, Puis cliquez sur « Ouvrir ».

Installation directement depuis l'Internet :

A partir du menu Packages, puis dans le sous-menu Installer le(s) package(s)...

Sélectionnez un miroir (CRAN mirror) et le nom du package.



Installation depuis la ligne de commande :

On peut se passer des menus de l'interface graphique de **R**. C'est par exemple utile sous Unix/Linux où le logiciel **R** ne possède pas d'interface graphique. Pour cela, tapez directement dans la console de **R** les commandes suivantes :

– pour des packages dont les fichiers *.zip sont situés sur votre disque dur :

```
install.packages(choose.files(),repos = NULL)
```

– pour un package (par exemple Rcmdr) dont le fichier est sur le site internet CRAN :

```
install.packages("Rcmdr")
```

3. Les données dans R :

Comme la plupart des langages informatiques, **R** dispose des types de données classiques. Selon la forme des données saisies, **R** sait d'ailleurs reconnaitre automatiquement le type de ces données. Une des grandes forces de **R** réside aussi dans la possibilité d'organiser les données de façon structurée.

3.1. Nature (type) des données :

Les fonction smode() et typeof(), renvoyant des valeurs identiques à de rares subtilités près non détaillées ici permettent de gérer le « type » des données.

Voilà maintenant les divers types de données (aussi appelés modes).

➢ Table 1 – Les différents types de données en **R**.

Type de données	**Type sous R**	**Présentation**
réel (entier ou non)	Numeric	2.54
Complexe	Complex	2+3i
Logique (vrai /faux)	Logical	TRUE ou FALSE
Manquant	Logical	NA
Texte (chaine)	Character	"texte"
Binaires	Raw	1B

Remarque :

1- La fonction class() est plus générale puisqu'elle permet de gérer à la fois le type et la structuration des données.

2- Ne pas confondre NA avec le mot réservé NaN signifiant (not a number) :

```
> 0/0
[1] NaN
```

3- TRUE et FALSE peuvent être saisis de manière plus succincte en tapant respectivement T et F. Mais cette approche n'est pas recommandée :> T

```
[1] TRUE
> F
```

```
[1] FALSE
```

3.2. Structures de données :

R offre la possibilité d'organiser (de structurer) les différents types de données définies précédemment. La fonction class() permettra de manipuler des structures. Nous présentons les plus utiles :

Les vecteurs (vector) :

Cette structure de données est la plus simple. Elle représente une suite de données de même type. La fonction permettant de créer ce type de structure (c'est-à-dire les vecteurs) est la fonction c() (pour collection ou concaténation).

D'autres fonctions comme seq() ou bien les deux points « : » permettent aussi de créer des vecteurs. Notez que lors de la création d'un vecteur, il est possible de mélanger des données de plusieurs types différents. **R** se charge alors d'opérer une conversion implicite vers le type de données le plus fréquent comme vous pouvez le constater dans les exemples ci-dessous :

```
>c(1,3,2,5)
[1] 1 3 2 5
>seq(from=0,to=1,by=0.1)
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
>seq(from=0,to=20,length=6)
[1]  0  4  8 12 16 20
```

Il est possible de « nommer » les éléments d'un vecteur à l'aide de la fonction names().

```
>vec<- c(4, 5, 6, 1, 9, 4, 8, 2, 0)
>names(vec) <- letters[1:9] // 9 premières lettres de l'alphabet.
>vec
a b c d e f g h i
4 5 6 1 9 4 8 2 0
```

Les matrices (matrix), les tableaux (arrays)

Ces deux notions généralisent la notion de vecteur puisqu'elles représentent des suites à double indice pour les matrices et à multiples indices pour les tableaux (array). Ici aussi les éléments doivent avoir le même type, sinon des conversions implicites seront effectuées.

L'instruction suivante :

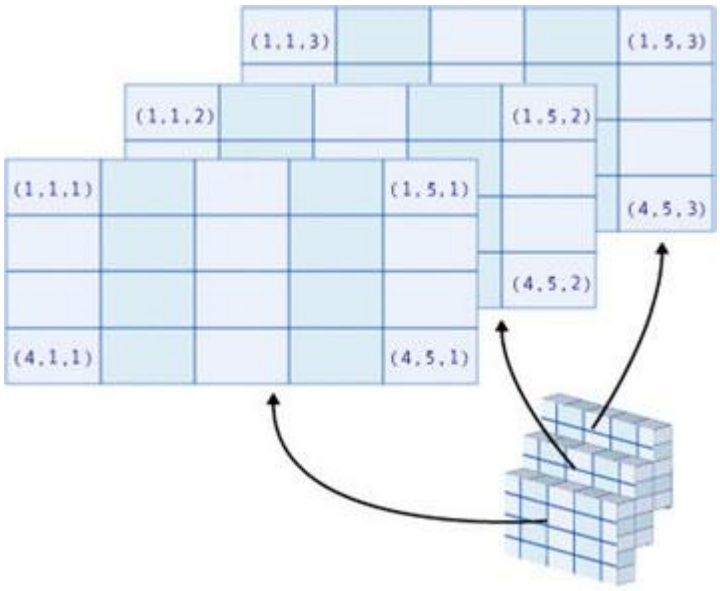
```
> X <- matrix(1:12,nrow=4,ncol=3,byrow=TRUE)
> X
[,1] [,2] [,3]
[1,]   1    2 3
[2,] 4 5 6
[3,] 7 8 9
[4,] 10 11 12
```

Permet de créer (et stocker dans la variable X) une matrice comportant quatre lignes (row signifie ligne) et trois colonnes remplies par lignes successives (byrow=TRUE) avec les éléments du vecteur 1:12 (c'est-à-dire les douze premiers entiers).

De la même manière, il est possible de créer une matrice remplie par colonnes successives (byrow=FALSE).

```
> Y <- matrix(1:12,nrow=4,ncol=3,byrow=FALSE)
> Y
[,1] [,2] [,3]
[1,] 1 5 9
[2,] 2 6 10
[3,] 3 7 11
[4,] 4 8 12
```

La fonction array() permet de créer des matrices multidimensionnelles à plus de deux dimensions comme cela est illustré sur la figure suivante (pour un array ayant trois dimensions).



➢ Figure 1 : Illustration d'une array

```
> X <- array(1:12,dim=c(2,2,3))

> X
, , 1
[,1] [,2]
[1,] 1 3
[2,] 2 4
, , 2
[,1] [,2]
[1,] 5 7
[2,] 6 8
, , 3
[,1] [,2]
[1,]  9   11
[2,]  10   12
```


Les listes (list) :

La structure du langage **R** la plus souple et à la fois la plus riche est celle de la liste. Contrairement aux structures précédentes, les listes permettent de regrouper dans une même structure des données de types différents sans pour

autant les altérer. De façon générale, chaque élément d'une liste peut ainsi être un vecteur, une matrice, un *array* ou même une liste.

Voici un exemple :

```
> A <- list(TRUE,-1:3,matrix(1:4,nrow=2),c(1+2i,3), "Une chaîne de c
aractères")
> A
[[1]]
[1] TRUE

[[2]]
[1] -1  0  1  2  3

[[3]]
     [,1] [,2]
[1,]    1    3
[2,]    2    4

[[4]]
[1] 1+2i 3+0i

[[5]]
[1] "Une chaîne de caractères"
```

Le tableau individus × variables (data.frame) :

Le tableau individus × variables est la structure par excellence en statistique. Cette notion est exprimée dans **R** par le data.frame. Conceptuellement, c'est une matrice dont les lignes correspondent aux individus et les colonnes aux variables (ou caractères) mesurées sur ces derniers. Chaque colonne représente une variable particulière dont tous les éléments sont du même type. Les colonnes de la matrice-données peuvent être nommées.

Voici un exemple :

```
> IMC <- data.frame(Sexe=c("H","F","H","F","H","F"),
Taille=c(1.83,1.76,1.82,1.60,1.90,1.66), Poids=c(67,58,66,48,75,55),
row.names=c("Rémy","Lol","Pierre","Domi","Ben","Cécile"))
> IMC
Sexe Taille Poids
Rémy H 1.83 67
LolF 1.76 58
Pierre H 1.82 66
Domi F 1.60 48
Ben H 1.90 75
Cécile F 1.66 55
```

Remarque :

La fonction str() permet d'afficher la structure de chacune des colonnes d'un *data.frame*.

Les dates :

R permet de structurer les données constituées par des dates, au moyen de la fonction as.Date().

Par exemple :

```
> dates <- c("27/02/92","27/02/92","14/01/92","28/02/92","01/02/92")
> dates <- as.Date(dates, "%d/%m/%y")
>dates
[1] "1992-02-27" "1992-02-27" "1992-01-14" "1992-02-28"
[5] "1992-02-01"
```

➢ Table 2 – Les différentes structures de données en **R**

Structure des données	**Instruction R**	**Description**
Vecteur	C( )	Suite d'éléments de même nature.
Matrice	Matrix( )	Tableau à deux dimensions dont les éléments sont de même nature.
tableau multidimensionnel	Array( )	Plus général que la matrice ; tableau à plusieurs dimensions.
Liste	List( )	Suite de structures R de nature différente et quelconque.
tableau individus × variables	Data.frame( )	Tableau à deux dimensions dont les lignes sont des individus et les colonnes des variables (numériques ou facteurs).Les colonnes peuvent être de nature différente, mais doivent avoir la même longueur. Les éléments à l'intérieur d'une même colonne sont tous de la même nature.
Dates	As.date( )	Vecteur de dates.

4. Intervalles de confiance et tests d'hypothèses :

Cette partie se veut être un catalogue des fonctions **R** le plus couramment utilisées afin d'obtenir les intervalles de confiance observés pour les paramètres classiques : moyenne, proportion, variance. Nous présentons également un catalogue des fonctions **R** permettant d'effectuer les tests d'hypothèses les plus classiques.

4.1. Notations :

Le tableau 3 présente les notations nécessaires à la définition des intervalles de confiance et des tests d'hypothèses introduits dans ce chapitre. Nous présentons dans le tableau 4 les notations des différents quantiles qui seront utilisés.

- Table 3 : Notations sur les estimations de paramètres classiques.

Paramètre	**Notation**	**Estimateur**	**Estimation**	**Fonction R**
Moyenne	μ	$\bar{X}$	$\bar{x}$	Mean( )
Variance	σ^2	$\hat{\sigma}^2$	$\hat{\sigma}^2$	Var( )
Proportion	P	$\hat{p}$	$\hat{p}$	Mean( )
Corrélation	m_e	$\hat{m}_e$	$\hat{m}_e$	Cor( )

- Table 4 : Notations des différents quantiles d'ordre p

Loi	**Notation**	**Fonction R**
Normale : $\mathcal{N}(0, 1)$	u_p	qnorm(p)
Student à n d.d.l. : $\mathcal{T}(n)$	t^n_p	qt(p,df=n)
Khi-deux à n d.d.l. : $\mathcal{X}^2(n)$	q^n_p	qchisq(p,df=n)
Fisher à n et m d.d.l.: $\mathcal{F}(n,m)$	$f^{n,m}_p$	qf(p,df1=n,df2=m)

d.d.l. : degrés de liberté

4.2. Intervalle de confiance :

On dispose d'un échantillon $\mathbf{X}_n = (X_1, \ldots, X_n)$ de variables aléatoires suivant une loi dépendant d'un certain paramètre θ inconnu que l'on cherche à estimer.

Un intervalle de confiance aléatoire de niveau (de confiance) $1-\alpha$ pour θ est la donnée de deux variables aléatoires $A := a(\mathbf{X}_n; \alpha)$ et $B := b(\mathbf{X}_n; \alpha)$ telles que :

$$P[A \leq \theta \leq B] = 1 - \alpha.$$

Les variables aléatoires A et B constituent les bornes de cet intervalle de confiance aléatoire, que l'on note en général

$$IC_{1-\alpha(\theta)} = [A, B].$$

Lorsque l'échantillon est effectivement observé, et que l'on dispose des données $(x_1, \ldots, x_n)$, on notera

$$ic_{1-\alpha(\theta)} = [a, b].$$

Le tableau 5 présente les intervalles de confiance sous **R** :

- table 5 : Résumé sur les intervalles de confiance.

Type	**Fonction de validité**	**Fonction R**
Moyenne	n>30 ou normalité	t.test(x)$conf
Variance	Normalité	sigma2.test(x)$conf
Proportion	$np \geq 5$ et $n(1-p) \geq 5$ aucune	prop.test(x)$conf binom.test(x)$conf
Corrélation	binormale	cor.test(x)$conf

4.3. Tests d'hypothèses :

Objectif :

Il s'agit de faire un choix entre plusieurs hypothèses possibles sans disposer d'informations suffisantes pour que le choix soit sûr.

On met en avant une hypothèse, dite hypothèse nulle et notée (H_0). On souhaite vérifier si (H_0) est vraie, alors que deux hypothèses seulement sont

possibles : (H_0) et une hypothèse alternative (H_1) qui peut prendre trois formes selon la nature de la question posée ou l expérience menée.

$H_1 : \theta \neq \theta_0$ (test bilatéral)

$\theta < \theta_0$ (test unilatéral à gauche)

$\theta > \theta_0$ (test unilatéral à droite)

Risques :

L'information étant incomplète, toute décision est associée à une prise de risque.

Si on décide que (H_0) est fausse, le risque de se tromper est note α et s'appelle risque de première espèce.

Si on décide que (H_0) est vraie, le risque de se tromper est note β et s'appelle risque de deuxième espèce.

Le seuil de signification des tests présentés ci-après sera fixé à $\alpha = 5\%$.

- Table 6 : Les différents types de risque

	H_0 vrai	**H_1 vrai**
Ne pas rejeté H_0	1- α	β
Rejeté H_0	α	1- β

Tests de moyenne :

• Comparaison de la moyenne théorique à une valeur de référence (cas d'un échantillon) :

✓ ***Descriptif du test*** : Soit une variable quantitative X de moyenne théorique μ et de variance σ^2. A partir d'un échantillon de taille n, on veut comparer la moyenne théorique μ à une valeur de référence μ_0. Les hypothèses du test sont$H_0 : \mu = \mu_0$ et$H_1 : \begin{cases} \mu < \mu 0 \text{ ou} \\ \mu > \mu 0 \text{ ou} \\ \mu \neq \mu 0 \end{cases}$. Sous H_0 la statistique du test est :

$$T=\sqrt{n}\left(\frac{\bar{X}-\mu 0}{\hat{\sigma}}\right) \sim \mathcal{T}(n-1).$$

✓ ***Conditions de validité***: Normalité des données ou taille d'échantillon grande (n > 30).

✓ ***Instruction R*** : Il est possible d'utiliser la fonction t.test().

• Comparaison de deux moyennes théoriques (cas de deux échantillons) :

✓ ***Descriptif du test*** : On considère deux variables quantitatives X_1 et X_2 (qui mesurent la même caractéristique, mais dans deux populations différentes). On suppose que X_1 a pour moyenne théorique μ_1 et pour variance σ^2_1 et X_2 a pour moyenne théorique μ_2 et pour variance σ^2_2. à partir des estimations calculées sur deux échantillons de tailles respectives n_1 et n_2 issus des deux populations, on veut comparer μ_1 et μ_2. Les hypothèses du test sont $H_0 : \mu_1 = \mu_2$ et $H_1 : \begin{cases} \mu 1 < \mu 0 \text{ ou} \\ \mu 1 > \mu 0 \text{ ou} \\ \mu 1 \neq \mu 0 \end{cases}$. Sous H_0 la statistique du test est :

$$T= \frac{\overline{X1}-\overline{X2}}{Sc\sqrt{\frac{1}{n1}+\frac{1}{n2}}} \sim \mathcal{T}(n_1+n_2-2)$$

Avec ici $Sc^2 = \frac{(n1-1)\hat{\sigma}^2 1+(n2-1)\hat{\sigma}^2 2}{n1+n2-2}$, $\hat{\sigma}^2_1$ et $\hat{\sigma}^2_2$ les estimateurs des variances des variables X_1 et X_2.

✓ ***Conditions de validité***: Normalité des variables X_1 et X_2 et variances égales.

✓ ***Instruction R :*** Il est possible d'utiliser la fonction t.test().

• Cas des échantillons appariés :

✓ ***Descriptif du test*** : On veut comparer les moyennes théoriques de deux variables aléatoires X_1 et X_2 sur la base de deux échantillons appariés. Pour cela, on travaille avec la variable aléatoire déférence $D = X_1 - X_2$, et l'on compare la moyenne théorique $\delta = \mu_1 - \mu_2$ de D à la valeur de référence 0. On se retrouve donc dans le cas du test de moyenne à un échantillon. Les hypothèses du test sont $H_0 : \mu_1 - \mu_2 = 0$ et $H_1 : \begin{cases} \mu1 - \mu2 > 0 \ ou \\ \mu1 - \mu2 < 0 \ ou \\ \mu1 - \mu2 \neq 0 \end{cases}$. Sous H_0 la statistique du test est :

$$T = \sqrt{n}\frac{D}{\hat{\sigma}} \sim \mathcal{T}(n-1).$$

✓ ***Conditions de validité***: Normalité des données ou taille d'échantillon grande (n > 30).

✓ ***Instruction R :*** Il est possible d'utiliser la fonction t.test() avec le paramètre paired=TRUE**.**

Tests de variance :

• Comparaison de la variance théorique à une valeur de référence (cas d'un échantillon) :

✓ ***Descriptif du test*** **:** Soit σ^2 la variance d'un caractère quantitatif X. Les hypothèses du test sont $H_0 : \sigma^2 = \sigma^2_0$ versus $H_1 : \begin{cases} \sigma^2 < \sigma^2 0 \text{ ou} \\ \sigma^2 > \sigma^2 0 \text{ ou} \\ \sigma^2 \neq \sigma^2 0 \end{cases}$ .Sous H_0 la statistique du test est :

$$X^2 = \frac{(n-1)\hat{\sigma}^2}{\hat{\sigma}^2 0} \sim \mathcal{X}^2(n-1)$$

✓ ***Conditions de validité***: Le caractère X est distribué suivant une loi normale.

✓ ***Instruction R :*** Il est possible d'utiliser la fonction sigma2.test()**.**

• Comparaison de deux variances théoriques (cas de deux échantillons) :

✓ ***Descriptif du test*** **:** Ce test est souvent utile comme préalable à d'autres tests comme celui de la comparaison de deux moyennes dans le cas de faibles effectifs. En effet, dans ce cas, la statistique n'est pas la même suivant que les variances de X_1 (variable concernant le premier échantillon) et de X_2(variable concernant le second échantillon) peuvent être considérées comme égales ou non. Les hypothèses du test sont $H_0 : \sigma^2_1 = \sigma^2_2$ versus $H_1 : \sigma^2_1 \begin{cases} \sigma^2 1 < \sigma^2 2 \text{ ou} \\ \sigma^2 1 = \sigma^2 2 \text{ ou} \\ \sigma^2 1 > \sigma^2 2 \end{cases}$

. La statistique de test sous H_0 est :

$$F = \frac{\hat{\sigma}^2 1}{\hat{\sigma}^2 2} \sim \mathcal{F}(n_1\text{-}1, n_2\text{-}1)$$

✓ ***Conditions de validité***: Normalité deX_1 et X_2.

✓ ***Instruction R :*** Il est possible d'utiliser la fonction var.test().

Tests de proportion :

• Comparaison d'une proportion théorique à une valeur de référence (cas d'un échantillon) :

✓ ***Descriptif du test*****:** Soit p la fréquence inconnue d'un caractère dans une population donnée. On observe des données de présence/absence de ce caractère sur les individus d'un échantillon de taille n de cette population. Les hypothèses du test que nous considérons sont $H_0 : p = p_0$ et $H_1 : \begin{cases} p < p0 \text{ ou} \\ p > p0 \text{ ou} \\ p \neq p0 \end{cases}$. Sous H_0 la statistique du test est :

$$U = \frac{\hat{p} - p0}{\sqrt{\frac{p0(1-p0)}{n}}} \sim \mathcal{N}(0,1).$$

✓ ***Conditions de validité***: L'échantillon doit être suffisamment grand, il faut vérifier que $np_0 \geq 5$ et $n(1-p_0) \geq 5$.

✓ ***Instruction R:*** Il est possible d'utiliser la fonction prop.test()**.**

• Comparaison de deux proportions théoriques (cas de deux échantillons) :

✓ ***Descriptif du test*** **:** Soit p_1 (respectivement p_2) la proportion inconnue d'individus présentant un certain caractère dans une population P_1 (respectivement P_2). On désire comparer p_1 et p_2. Pour cela, on utilise les fréquences (notées $\hat{p}_1$ et $\hat{p}_2$) d'apparition de ce caractère dans deux échantillons représentatifs respectivement des deux populations de taille n_1 et n_2. Les hypothèses du test sont : $H_0 : p_1 = p_2$ et $H_1 : \begin{cases} p1 < p2 \text{ ou} \\ p1 > p2 \text{ ou} \\ p1 \neq p2 \end{cases}$. Sous H_0 la statistique du test est :

$$U=\frac{\hat{p}1-\hat{p}2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n1}+\frac{\hat{p}(1-\hat{p})}{n2}}} \sim \mathcal{N}(0,1)$$

Avec $\hat{p}=\frac{n1\hat{p}1+n2\hat{p}2}{n1+n2}$.

✓ ***Conditions de validité***: Grands échantillons et il faut vérifier si $n_1\hat{p} \geq 5$, $n_1(1-\hat{p}) \geq 5$, $n_2\hat{p} \geq 5$ et $n_2 (1- \hat{p}) \geq 5$.

✓ ***Instruction R :*** La fonction à utiliser est prop.test().

Le tableau suivant résume tous les tests qui ont été présentés.

➢ table 7 : Les tests usuels.

Nature	**Données**	**Condition de validité**	**Fonction R**
Moyenne	1 échantillon	-n>30 ou normalité	t.test(x,…)
	2 échantillons	-normalité et variances égaux	t.test(x,y,…)
	2 échantillons	-normalité n>30 ou	t.test(x,y,var.egal=F)
	2 ech .apparies	-normalité	t.test(x,y,paired=T)

Variance	1 échantillon 2 échantillons 2 échantillons	-normalité -normalité -grand échantillon	Sigma2.test(x,…) Var.test(x,y,…) Asymp.test(x,y,…)
Proportion	1 échantillon 1 échantillon 2 échantillons	np≥5 et n(1-p)≥5 grand échantillon	Prop.test(x,…) Binom.test(x,…) Prop.test(x,y,…)

5. Régression linéaire :

✓ ***Objectif :*** On cherche à « expliquer » les variations d'une variable quantitative Y par une variable explicative X également quantitative.

✓ ***Le modèle:*** Il s'écrit sous la forme

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Où ε représente le bruit du modèle suppose gaussien d'espérance nulle et de variance Var $(\varepsilon_i) = \sigma^2$. Les paramètres (inconnus) du modèle de régression sont β_0, β_1 et σ^2.

✓ ***Estimation des paramètres β_0 et β_1 par la méthode des moindres carrées :*** Notre objectifs est d'obtenir une droite qui s'approche le mieux des points expérimentaux. Pour atteindre ce but nous utiliserons la méthode dite des moindres carrées qui consiste à minimiser la somme des carrés des écarts des valeurs observées Y_i à la droite. Cette droite s'appelle droite de régression empirique notons

$$\hat{Y}_i = b_0 + b_1 X_i$$

Posons $e_i = Y_i - \hat{Y}_i$, $i = 1,\ldots n$

La somme des carrés des écarts pour l'ensemble des points est :

$\sum_{i=1}^{n} e^2_i = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$

On cherche b_0 et b_1 tels que la somme soit minimale. Les dérivées partielles par rapports à b_0 et b_1 sont :

$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0$

$\sum_{i=1}^{n} Xi\ e_i = \sum_{i=1}^{n} Xi(Y_i - b_0 - b_1 X_i) = 0$

Donc les estimateurs des 3 paramètres (β_0, β_1 et σ^2) :

$$b_1 = \frac{\sum XiYi - \frac{(\sum Xi)(\sum Yi)}{n}}{\sum X^2 i - \frac{(\sum Xi)^2}{n}} \quad \text{et} \quad b_0 = \bar{Y} - b_1\bar{X} \quad \text{et } S^2 = \frac{\sum(Yi - \hat{Y}i)^2}{n-2}$$

$$= \frac{\sum(Xi - \bar{X})(Yi - \bar{Y})}{\sum(Xi - \bar{X})^2}$$

✓ ***Tableau d'analyse de la variance :*** En régression linéaire simple, le test de Fisher (dont la statistique se lit dans F-statistic) est équivalent au test de Student associé à la pente de la régression. Nous avons la relation suivante F-statistic=t^2 et les valeurs-p des tests sont égales. Le test de Fisher est souvent associé à une table d'analyse de la variance que vous obtenez en utilisant la fonction anova().

- Table 8 : Liste des principales fonctions **R** permettant l'analyse d'une régression linéaire simple.

Instruction R	**Description**
plot(Y~X)	graphe du nuage de points
lm(Y~X)	estimation du modèle linéaire
summary(lm(Y~X))	description des résultats du modèle
abline(lm(Y~X))	trace la droite estimée
confint(lm(Y~X))	intervalle de confiance des paramètres de régression

predict()	fonction permettant d'obtenir des prédictions
plot(lm(Y~X))	analyse graphique des résidus

Remarque :

La solution dans **R** est donnée par ce que l'on appelle la valeur-p (aussi appelée p-valeur). Cette valeur-p est fournie par tout logiciel de statistique, et l'utilisateur saura alors comparer ce risque à un seuil de signification α. Plus la valeur-p est faible, plus la décision d'accepter H_1 est grande.

6. Application :

Cette partie présente un jeu de données provenant du département américain de la santé et des services humains, centre national des statistiques de sante, troisième enquête nationale pour l'examen de la santé et de l'alimentation. Ce jeu de donnée s'accompagne d'une problématique qui permettra de mieux comprendre le contexte de l'étude. Nous montrerons comment il est possible d'utiliser les déférentes fonctionnalités du logiciel **R** afin d'importer, de manipuler et d'effectuer les analyses statistiques adéquates sur ce jeu de données.

Calcul de la moyenne et de la variance de chaque grandeur pour les deux sexes :

Les données sont enregistrer dans des fichiers sous les noms « santehomme» et «santefemme» (voir Annexes).

```
>setwd("C:/statR") //définir l'espace de travail

>getwd()
[1] "C:/statR"

> pb1<-"jeu1homme.CSv"
```

```
>santehomme<-read.csv(pb1,sep=";" )//pour import et export des
                                    données et fichiers

>summary(santehomme)//(moyenne ,min ,max ,médian ,1ere et 3ere
                      quartile) de tous les grandeurs
```

age	taille	poids	ttaille
Min. :17.00	Min. :155.7	Min. : 54.20	Min. : 75.20
1st Qu.:25.75	1st Qu.:168.4	1st Qu.: 69.12	1st Qu.: 84.38
Median :32.50	Median :173.5	Median : 77.10	Median : 91.20
Mean :35.48	Mean :173.6	Mean : 78.27	Mean : 91.28
3rd Qu.:44.50	3rd Qu.:178.0	3rd Qu.: 86.03	3rd Qu.: 99.90
Max. :73.00	Max. :193.5	Max. :107.50	Max. :108.70

sys	dia	chol	imc
Min. : 95.0	Min. :44.00	Min. : 31.0	Min. :19.60
1st Qu.:111.5	1st Qu.:67.50	1st Qu.: 163.8	1st Qu.:23.73
Median :117.0	Median :75.00	Median : 282.5	Median :26.20
Mean :118.9	Mean :73.22	Mean : 395.2	Mean :26.00
3rd Qu.:125.0	3rd Qu.:81.00	3rd Qu.: 619.2	3rd Qu.:27.50
Max. :153.0	Max. :87.00	Max. :1252.0	Max. :33.20

```
> summary(santehomme$age)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.00   25.75   32.50   35.48   44.50   73.00

> summary(santehomme$sys)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   95.0   111.5   117.0   118.9   125.0   153.0


>var(santehomme$age)//calcule du variance de chaque grandeur
[1] 193.9481
>
```

```
>var(santehomme$taille)
[1] 58.78328
>
>
>diag (var(santehomme))//calcule de la variance de tous les
                        grandeurs
```

age	taille	poids	ttaille	pouls
1.939481e+02	5.878328e+01	1.426839e+02	9.725618e+01	1.276308e+02
chol	imc	jmbg	coud	poign
8.550464e+04	1.176999e+01	8.454865e+00	2.030769e-01	1.256410e-01
bras				
9.087788e+00				

Les mêmes fonctions pour les données du tableau des femmes en remplaçant « santehomme » par « santefemme ».

Comparaison de la moyenne et de la variance des deux sexes pour quelques grandeurs :

On note par SYS (la pression sanguine systolique (mmHg)), IMC (l'indice de masse corporelle (kg /m²)), DIA (pression sanguine diastolique (mmHg) et CHOL (taux de cholestérol (mg)).

Notons : m_1 : la moyenne du pression sanguine systolique pour les hommes

m_2 : la moyenne du pression sanguine systolique pour les femmes

les hypothèses du test : H_0 : $m_1 = m_2$

H_1 : $m_1 \neq m_2$

```
>t.test(santehomme$sys,santefemme$sys,var.egal=F) //comparaison de
                                                  la moyenne entre(sys)
                                                     des deux sexes
```

```
        WelchTwoSample t-test

data:  santehomme$sys and santefemme$sys
t = 2.5539, df = 64.582, p-value = 0.01302
alternativehypothesis: truedifference in meansis not equal to 0
95 percent confidence interval:
  1.765122 14.434878
sampleestimates:
mean of x mean of y
    118.9     110.8
```

On peut conclure qu'il y a une différence significative de la moyenne de la mesure de pression sanguine systolique entre les femmes et les hommes.

Notons : m_1 : la moyenne de l'indice de masse corporelle pour les hommes

m_2 : la moyenne de l'indice de masse corporelle pour les femmes

Les hypothèses du test : H_0 : $m_1 = m_2$

H_1 : $m_1 \neq m_2$

```
>t.test(santehomme$imc,santefemme$imc,var.egal=F)//comparaison de
                                                  la moyenne entre(imc)
                                                     des deux sexes
```

```
        WelchTwoSample t-test

data:  santehomme$imc and santefemme$imc
t = 0.0067354, df = 61.128, p-value = 0.9946
alternativehypothesis: truedifference in meansis not equal to 0
95 percent confidence interval:
 -2.219027  2.234027
sampleestimates:
mean of x mean of y
  25.9975   25.9900
```

On peut conclure qu'il n y a pas une différence significative de la moyenne de la mesure de l'indice de masse corporelle entre les femmes et les hommes.

Notons : m_1 : la moyenne du taux de cholestérol pour les hommes

m_2 : la moyenne du taux de cholestérol pour les femmes

Les hypothèses du test : H_0 : $m_1 = m_2$

H_1 : $m_1 \neq m_2$

```
>t.test(santehomme$chol,santefemme$chol,var.egal=F)//comparaison de
                                                    la moyenne entre(chol)
                                                        des deux sexes
```

```
        WelchTwoSample t-test

data:  santehomme$chol and santefemme$chol
t = 2.8169, df = 66.116, p-value = 0.006386
alternativehypothesis: truedifference in meansis not equal to 0
95 percent confidence interval:
  44.95439 263.74561
sampleestimates:
mean of x mean of y
  395.225   240.875
```

On peut conclure qu'il y a une différence significative de la moyenne de la mesure du taux de cholestérol entre les femmes et les hommes.

Notons : m_1 : la moyenne du pression sanguine diastolique pour les hommes

m_2 : la moyenne du pression sanguine diastolique pour les femmes

Les hypothèses du test : H_0 : $m_1 = m_2$

H_1 : $m_1 \neq m_2$

```
>t.test(santehomme$dia,santefemme$dia,var.egal=F)//comparaison de
                                                  la moyenne entre(dia)
                                                      des deux sexes
```

```
        WelchTwoSample t-test

data:  santehomme$dia and santefemme$dia
t = 2.5574, df = 73.244, p-value = 0.01262
alternativehypothesis: truedifference in meansis not equal to 0
95 percent confidence interval:
  1.33539 10.76461
sampleestimates:
mean of x mean of y
   73.225    67.175
```

On peut conclure qu'il y a une différence significative de la moyenne de la mesure de pression sanguine diastolique entre les femmes et les hommes.

Notons : σ^2_1 : la variance du pression sanguine systolique pour les hommes

σ^2_2 : la variance du pression sanguine systolique pour les femmes

Les hypothèses du test : H_0 : $\sigma^2_1 = \sigma^2_2$

H_1 : $\sigma^2_1 \neq \sigma^2_2$

```
>var.test(santehomme$sys,santefemme$sys)//comparaison de
                                        la variance entre(sys)
                                            des deux sexes
```

```
        F test to compare two variances

data:  santehomme$sys and santefemme$sys
F = 0.37379, numdf = 39, denomdf = 39, p-value = 0.002744
alternativehypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1976983 0.7067353
sampleestimates:
ratio of variances
         0.3737918
```

On peut conclure qu'il y a une différence significative de la variance de la mesure de pression sanguine systolique entre les femmes et les hommes.

Notons : σ^2_1 : la variance de l'indice de masse corporelle pour les hommes

σ^2_2 : la variance de l'indice de masse corporelle pour les femmes

Les hypothèses du test : H_0 : $\sigma^2_1 = \sigma^2_2$

H_1 : $\sigma^2_1 \neq \sigma^2_2$

```
>var.test(santehomme$imc,santefemme$imc)//comparaison de
                                        la variance entre(imc)
                                             des deux sexes
```

```
        F test to compare two variances

data:  santehomme$imc and santefemme$imc
F = 0.31115, numdf = 39, denomdf = 39, p-value = 0.0004201
alternativehypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1645684 0.5883021
sampleestimates:
ratio of variances
         0.3111526
```

On peut conclure qu'il y a une différence significative de la variance de la mesure de l'indice de masse corporelle entre les femmes et les hommes.

Notons : σ^2_1 : la variance du taux de cholestérol pour les hommes

σ^2_2 : la variance du taux de cholestérol pour les femmes

Les hypothèses du test : H_0 : $\sigma^2_1 = \sigma^2_2$

H_1 : $\sigma^2_1 \neq \sigma^2_2$

```
>var.test(santehomme$chol,santefemme$chol)//comparaison de
                                          la variance entre(chol)
                                               des deux sexes
```

```
        F test to compare two variances

data:  santehomme$chol and santefemme$chol
F = 2.472, numdf = 39, denomdf = 39, p-value = 0.005739
alternativehypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.307432 4.673832
sampleestimates:
ratio of variances
          2.471986
```

On peut conclure qu'il y a une différence significative de la variance de la mesure du taux de cholestérol entre les femmes et les hommes.

Notons : σ^2_1 : la variance du pression sanguine diastolique pour les hommes

σ^2_2 : la variance du pression sanguine diastolique pour les femmes

Les hypothèses du test : H_0 : $\sigma^2_1 = \sigma^2_2$

H_1 : $\sigma^2_1 \neq \sigma^2_2$

```
>var.test(santehomme$dia,santefemme$dia)//comparaison de
                                        la variance entre (dia)
                                             des deux sexes
```

```
        F test to compare two variances

data:  santehomme$dia and santefemme$dia
F = 0.59385, numdf = 39, denomdf = 39, p-value = 0.1079
alternativehypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3140867 1.1228029
sampleestimates:
ratio of variances
         0.5938497
```

On peut conclure qu'il n y a pas une différence significative de la variance de la mesure de pression sanguine diastolique entre les femmes et les hommes.

Régression linéaire simple :

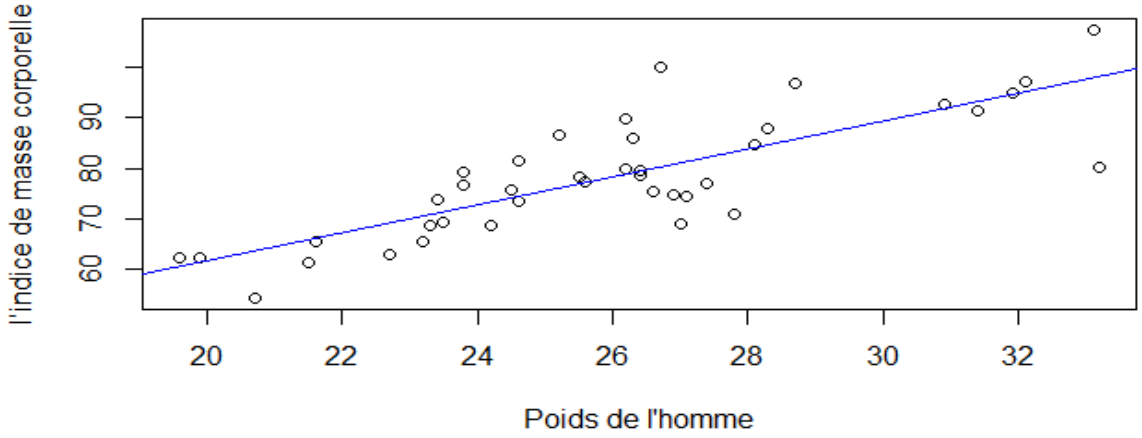
```
> coefficients(lm(santehomme$poids~santehomme$imc)) //calcul des
                              coefficients du droite de régression
 (Intercept) santehomme$imc
      5.914710       2.782971
```

La sortie **R** ci-dessus fournit les estimations par moindres carrés de β_0 et de β_1. On trouve dans l'exemple ci-dessus b_0=5.914710 et b_1=2.782971.

Nous pouvons maintenant représenter la droite de régression sur le nuage de points au moyen de la fonction abline() :

```
> plot(santehomme$poids~santehomme$imc,xlab="Poids de l'homme",ylab=
"l'indice de masse corporelle")
>
> abline(modele1,col="blue")
```

➢ Figure 2 : Représentation de la droite de régression des moindres carrés sur le nuage de points du poids de l'homme (kg) versus l'indice de masse corporelle (kg /m²).



Il est bon de noter que la fonction lm() permet une analyse complète du modèle linéaire et que on peut récupérer un résumé des calculs liés au jeu de données en utilisant la fonction summary().

```
res <- summary(lm(santehomme$poids~santehomme$imc))
> res
```

```
Call:
lm(formula = santehomme$poids ~ santehomme$imc)

Residuals:
     Min       1Q   Median       3Q      Max
-18.2093  -4.6720   0.2117   2.8046  19.8800

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.9147     8.8986   0.665     0.51
santehomme$imc   2.7830     0.3394   8.199 6.25e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 7.272 on 38 degrees of freedom
Multiple R-squared:  0.6389,      Adjusted R-squared:  0.6294
F-statistic: 67.23 on 1 and 38 DF,  p-value: 6.251e-10
```

Voici la description des différentes informations contenues dans la sortie ci-dessus.

• **Call:** Un rappel de la formule utilisée dans le modèle.

• **Residuals:** Une analyse descriptive des résidus $\hat{\varepsilon}_i = \hat{Y}_i - Y_i$.

• **Coefficients:** ce tableau comprend quatre colonnes :

- **Estimate** correspond aux estimations des paramètres de la droite de régression;

- **Std. Error** correspond à l'estimation de l'écart type des estimateurs de la droite de régression;

- **t value** correspond à la réalisation de la statistique du test de Student associé aux hypothèses $H_0 : \beta_i = 0$ et $H_1 : \beta_i \neq 0$;

- **Pr(>|t|)** correspond à la valeur-p du test de Student.

• **Signif. codes:** symboles de niveau de significativité.

• **Residual standard error:** une estimation de l'écart type du bruit σ est fournie ainsi que le degré de liberté associé $n-2$.

• **Multiple R-squared:** valeur du coefficient de détermination r^2 (pourcentage de variance expliqué par la régression).

• **Adjusted R-squared**: $r^2\sigma$ a ajusté (qui n'a pas grand intérêt en régression linéaire simple).

• **F-statistic:** correspond à la réalisation du test de Fisher associé aux hypothèses $H_0 : \beta_1 = 0$ et $H_1 : \beta_1 \neq 0$. Nous y trouvons les degrés de liberté associés (1 et n−2) ainsi que la valeur-p.

ANOVA :

```
> anova(lm(santehomme$poids~santehomme$imc))
Analysis of Variance Table

Response: santehomme$poids
               Df Sum Sq Mean Sq F value    Pr(>F)
santehomme$imc  1 3555.2  3555.2  67.228 6.251e-10 ***
Residuals      38 2009.5    52.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interprétation des résultats sur l’étude « santehomme » :

• Le test associé à l’intercept β_0 du modèle n’est pas significatif (valeur-p > 0.05), il est donc conseillé de ne pas garder l’intercept (β_0) dans le modèle.

• La relation linéaire entre POIDS et IMC est démontrée par le résultat du test de Student sur le coefficient β_1. La valeur-p < 0.05 nous indique une relation linéaire significative entre le poids de l’homme et l’indice de masse corporelle.

• Le pourcentage de variabilité du poids expliqué par le modèle vaut 0.6389. Ce qui veut dire que 63.89% de la variabilité du poids de l’homme est expliquée par l’indice de masse corporelle.

Conclusion

Ce PFE m'a permis d'approfondir ma formation en général et particulièrement en statistique.

Parmi ses bénéfices c'est qu'il m'a permis d'aborder et d'appréhender la problématique des estimateurs

- sur le plan théorique par la mise en évidence de deux méthodes :

  -la méthode des moments

  -la méthode de maximum de vraisemblance.

- Sur le plan pratique : de mettre en pratique mes connaissances théoriques que j'ai appris au niveau de ma formation au sien « faculté des sciences et techniques »Fès en licence mathématiques et applications.

L'un des objectifs essentiels de ce projet été l'utilisation du logiciel **R** comme moyen de traitement des données statistiques. Cet outil offre la possibilité à tout utilisateur de contribuer à son amélioration en y intégrant de nouvelles fonctionnalités ou méthodes d'analyse non encore implémentées ce qui le rend rapide et évolutif au cour du temps.

Liste des tableaux et des figures :

Annexes

age	Taille	poids	Ttaille	pouls	sys	dia	chol	imc	jmbg	Coud	poign	bras
17	163.3	52.1	67.2	76	104	61	264	19.6	41.6	6	4.6	23.6
32	168.7	67.7	82.5	72	99	64	181	23.8	42.8	6.7	5.5	26.3
25	158.2	48.9	66.7	88	102	65	267	19.6	39	5.7	4.6	26.3
55	158.2	72.6	93	60	114	76	384	29.1	40.2	6.2	5.0	32.6
27	151.4	57.7	82.6	72	94	58	98	25.2	36.2	5.5	4.8	29.2
29	161.5	55.8	75.4	68	101	66	62	21.4	43.2	6.0	4.9	26.4
25	151.9	50.7	73.6	80	108	61	126	22.0	38.7	5.7	5.1	27.9
12	160.8	70.9	81.4	64	104	41	89	27.5	41	6.8	5.5	33
41	172.5	99.2	99.4	68	123	72	531	33.5	43.8	7.8	5.8	38.6
32	156	50	67.7	68	93	61	130	20.6	37.3	6.3	5	26.5
31	169.4	85.4	100.7	80	89	56	175	29.9	42.3	6.6	5.2	34.4
19	164.6	47.8	72.9	76	112	52	44	17.7	39.1	5.7	4.8	23.7
19	160.3	61.7	85	68	107	48	8	24.0	40.3	6.6	5.1	28.4
23	169.4	82.7	85.7	72	116	62	112	28.9	48.6	7.2	5.6	34.0
40	169.7	108.1	126	96	181	102	462	37.7	33.2	7.0	5.4	35.2
23	164.3	49.4	74.5	72	98	61	62	18.3	43.4	6.2	5.2	24.7
27	165.4	54	74.5	68	100	53	98	19.8	41.5	6.3	5.3	27
45	157.2	73.4	94	72	127	74	447	29.8	40.0	6.8	5	35
41	163.3	79	92.8	64	107	67	125	29.7	38.2	6.8	4.7	33.1
56	161	82.2	105.5	80	116	71	318	31.7	38.2	6.9	5.4	39.6
22	154.2	56.4	75.5	64	97	64	325	23.8	38.2	5.9	5.0	27.0
57	161	116.1	126.5	80	155	85	600	44.9	41	8.0	5.6	43.8
24	159	48.4	70	76	106	59	237	19.2	38.1	6.1	5.0	23.6
37	153.9	68	98	76	110	70	173	28.7	38	7.0	5.1	34.3
59	161.3	74	104.7	76	105	69	309	28.5	36	6.7	5.1	34.4
40	148.8	42.8	67.8	80	118	82	94	19.3	32.1	5.4	4.2	23.3
45	152.9	72.4	99.3	104	133	83	280	31.0	31.1	6.4	5.2	35.6
52	171.7	73.8	91.1	88	113	75	254	25.1	39.4	7.1	5.3	31.8
31	161	59	74.5	60	113	66	123	22.8	40.2	5.9	5.1	27.0
32	162.8	81.6	95.5	76	107	67	596	30.9	39.2	6.2	5	32.8
23	159.3	67	79.5	72	95	59	301	26.5	39.0	6.3	4.9	31.0
23	155.7	51.2	69.1	72	108	72	223	21.2	36.6	5.9	4.7	27.0
47	147.8	88.7	105.5	88	114	79	293	40.6	37	7.5	5.5	41.2
36	160.5	56.3	78.8	80	104	73	146	21.9	38.5	5.6	4.7	25.5
34	153.7	61.2	85.7	60	125	73	149	26.0	39.9	6.4	5.2	30.9
37	165.1	64.1	92.8	72	124	85	149	23.5	37.5	6.1	4.8	27.9
18	157	56.2	72.7	88	92	46	920	22.8	39.7	5.8	5.0	26.5
29	172.7	61.5	75.9	88	119	81	271	20.7	39	6.3	4.9	27.8
48	170.2	59.1	68.6	124	93	64	207	30.5	41.6	6.0	5.3	23.0
16	144.8	45.7	68.7	64	106	64	2	21.9	33.8	5.6	4.6	26.4

age	Taille	poids	ttaille	pouls	sys	dia	chol	imc	jmbg	coud	poign	bras
58	179.8	76.7	90.6	68	125	78	522	23.8	42.5	7.7	6.4	31.9
22	168.1	65.4	78.1	64	107	54	127	23.2	40.2	7.6	6.2	31
32	182.1	81.3	96.5	88	126	81	740	24.6	44.4	7.3	5.8	32.7
31	174.5	79.7	87.7	72	110	68	49	26.2	42.8	7.5	5.9	33.4
28	171.7	69.2	87.1	64	110	66	230	23.5	40	7.3	6	30.1
46	175.8	75.7	92.4	72	107	83	316	24.5	47.3	7.1	5.8	30.5
41	168.9	61.2	78.8	60	113	71	590	21.5	43.4	6.5	5.2	27.6
56	170.7	91.4	103.3	88	126	72	466	31.4	40.1	7.5	5.6	38
20	173.5	79.5	89.1	76	137	85	121	26.4	42.1	7.5	5.5	32
54	166.6	63	82.5	60	110	71	578	22.7	36	6.9	5.5	29.3
17	160	70.9	86.7	96	109	65	78	27.8	44.2	7.1	5.3	31.7
73	173.5	84.6	103.3	72	153	87	265	28.1	36.7	8.1	6.7	30.7
52	185.7	86.7	91.8	56	112	77	250	25.2	48.4	8	5.2	34.7
25	171.7	68.6	75.6	64	119	81	265	23.3	41	7	5.7	30.6
29	172.7	95	105.5	60	113	82	273	31.9	39.8	6.9	6	34.2
17	180.3	107.5	108.7	64	125	76	272	33.1	45.2	8.3	6.6	41.1
41	155.7	80.1	104	84	131	80	972	33.2	40.2	6.7	5.7	33.1
52	193.5	100.1	103	76	121	75	75	26.7	46.2	7.9	6	32.2
32	168.4	75.3	91.3	84	132	81	138	26.6	39	7.5	5.7	31.2
20	177	62.3	75.2	88	112	44	139	19.9	44.8	6.9	5.6	25.9
20	166.1	74.5	87.7	72	121	65	638	27.1	40.9	7	5.6	33.7
29	177.8	73.7	77	56	116	64	613	23.4	43.1	7.5	5.2	30.3
18	159.8	68.9	85	68	95	58	762	27	38	7.4	5.8	32.8
26	174	65.4	79.6	64	110	70	303	21.6	41	6.8	5.7	31
33	173.5	92.8	103.8	60	110	66	690	30.9	46	7.4	6.1	36.2
55	176.3	87.9	103	68	125	82	31	28.3	41.4	7.2	6	33.6
53	175.8	78.4	97.1	60	124	79	189	25.5	42.7	6.6	5.9	31.9
28	172.7	73.4	86.9	60	131	69	957	24.6	40.5	7.3	5.7	32.9
28	182.6	79.3	88	56	109	64	339	23.8	44.2	7.8	6	30.9
37	167.9	77	91.5	84	112	79	416	27.4	41.8	7	6.1	34
40	183.9	96.8	102.9	72	127	72	120	28.7	47.2	7.5	5.9	34.8
33	185.4	89.8	93.1	84	132	74	702	26.2	48.2	7.8	6	33.6
26	172.7	78.6	98.9	88	116	81	1252	26.4	42.9	6.7	5.8	31.3
53	174.5	97.3	107.5	56	125	84	288	32.1	42.8	8.2	5.9	37.6
36	178.6	62.2	81.6	64	112	77	176	19.6	40.8	7.1	5.3	27.9
34	161.8	54.2	75.7	56	125	77	277	20.7	42.6	6.6	5.3	26.9
42	180.6	85.8	95	56	120	83	649	26.3	44.9	7.4	6	36.9
18	166.6	74.7	91.1	60	118	68	113	26.9	41.1	7	6.1	34.5
44	173.5	77.2	94.9	64	115	75	656	25.6	44.5	7.3	5.8	32.1
20	168.4	68.5	79.9	72	115	65	172	24.2	44	7.1	5.4	30.7

Le premier tableau représente les données du fichier « santefemme » et le deuxième les données du fichier « santehomme ».

Bibliographie

[1] Dalgaard P., Introductory Statistics with **R**, 2 éme édition, Springer, 2008.

[2] Pierre Lafaye de Micheaux Rémy Drouilhet Benoît Liquet, Le logiciel **R**, Maîtriser le langage Effectuer des analyses (bio) statistiques, 2 éme édition,springer,2011.

[3] Dagnelie P., Statistique théorique et appliquée, 2 éme édition, De Boeck Université, 2007

[4] R. Ihaka, R. Gentleman : R : A Language for Data Analysis and Graphics.Journal of Computational and Graphical Statistics, 5(3):299-314,1996.

[5] Pr.F. EZZAKI : Statistique inférentielle polycopié de la licence en mathématiques et applications de faculté des sciences et techniques Fès.

[6] Marc M. Triola et Mario F. Triola : Biostatistique pour les sciences de la vie et de la santé, Edition revue et corrigée.