Annexe 6 : Utilisation de SPSS

Cet annexe contient une introduction à l'utilisation du logiciel d'analyse statistique SPSS. Chaque section qui suit complète les sections des notes de cours en montrant les commandes à utiliser pour réaliser les tests décrits dans les notes. Par exemple, la section 5 du présent document complète la section 5 des notes de cours. Notez qu'il n'y a pas de sections 7 et 13 puisque dans ces sections, il n'est pas nécessaire d'utiliser un logiciel d'analyse statistique.

1. Les bases de SPSS

1.0. Qu'est-ce que SPSS?

SPSS signifie *Statistical package for the Social Sciences*. Il s'agit d'un logiciel dont les premières versions datent des années soixante (sans doute l'un des plus anciens). À l'origine, il s'agissant d'un programme "open source", c'est à dire que n'importe qui pouvait ajouter de nouvelles commandes, et les possesseurs du logiciel recevaient un pamphlet supplémentaire décrivant cette commande. Dans les années 80, le logiciel a cessé d'être "open source" et est maintenant la propriété exclusive de SPSS inc. Néanmoins, les propriétaires sont très agressifs, sortant de nouvelles versions régulièrement (en moins de 8 ans, nous sommes passé de SPSS v. 6 à SPSS v.12). À partir de la version 7, SPSS est devenu un produit pour Windows.

SPSS est un produit très dispendieux, et la licence dure généralement une année seulement. Il existe aussi une version étudiante beaucoup plus accessible. Malheureusement, elle est incomplète et ne peut pas être utilisée pour les besoins du cours.

L'objectif de SPSS est d'offrir un logiciel intégré pour réaliser la totalité des tests statistiques habituellement utilisés en sciences sociales et en psychologie. De fait, SPSS est un logiciel très complet. Dans le cours, nous ne verrons qu'une faible partie de ses possibilités.

Cependant, c'est un logiciel dont le noyau central est âgé. Certaines des normes de syntaxes que nous verrons plus loin était habituelles pour les programmeurs en langage Fortran des années soixante. Il ne reste plus beaucoup de ces programmeurs et aujourd'hui, ces règles de syntaxes semblent arbitraires. De plus, il existe maintenant des langages de manipulation de fichier de données (par exemple, SQL) qui sont nettement plus performants et plus intuitifs que les commandes offertes par SPSS. Malgré ces défauts, SPSS reste un logiciel utilisé dans presque toutes les universités du monde, et il est tellement complet qu'on utilise rarement un autre logiciel.

D'autres produits qui remplissent le même rôle que SPSS existent, tel SAS, SYSTAT, statistica ou encore S-PLUS et son clone gratuit, le R-project. Tous sont très différents d'utilisation mais atteignent le même but: faire des tests statistiques sans devoir connaître les formules par cœur. D'autres logiciels, tel le chiffrier Excel, peuvent aussi faire quelques tests statistiques, mais ils sont limités quant au nombre de données permises, et ne sont pas plus faciles à utiliser (malgré les apparences…).

Figure 1 : Démarrage de SPSS

1.1. Comment démarrer SPSS

SPSS installe dans le menu "Démarrer : Programme" une ligne pour démarrer le logiciel, comme c'est le cas dans la Figure 1.

La localisation exacte peut varier d'un ordinateur à l'autre (ainsi que l'image de fond!).

1.2. Les fenêtres de SPSS

Lorsque SPSS démarre, il ouvre une fenêtre principale qui ressemble un peu à une fenêtre de chiffrier (tel Excel). Il peut aussi ouvrir d'autres fenêtres, comme celles que l'on voit dans la Figure 2.

Une session typique sur SPSS aura toujours ces trois fenêtres. Elles sont:

a. fenêtre de données (SPSS Data Editor)

Cette fenêtre permet d'entrer des données, de les modifier ou de les effacer. Il est rare que l'on va taper les données manuellement dans SPSS car il y a trop d'erreurs de saisie possibles. On va plutôt ouvrir un fichier déjà existant (souvent généré par les instruments de mesures lors d'une expérience).

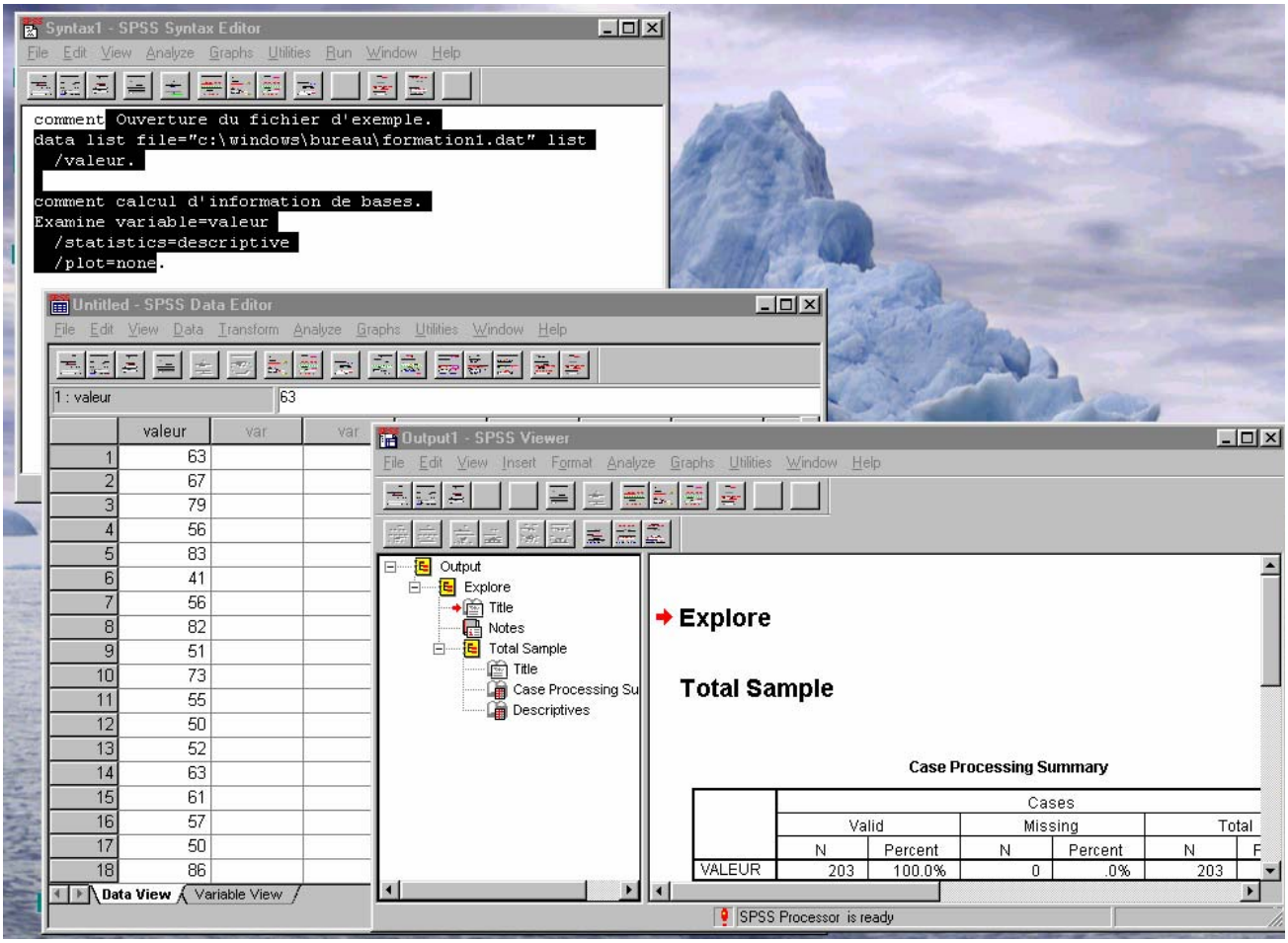


Figure 2 : Les trois fenêtres de SPSS

Quand vous fermez cette fenêtre, vous quittez SPSS. SPSS vous demande toujours si vous voulez sauvegarder les données: Répondez toujours non. Si vous dites oui, il va créer une copie supplémentaire de votre fichier de données, copie inutile qui en plus ne peut plus être lue par un traitement de texte, tel Bloc-notes.

b. fenêtre de syntaxe (SPSS Syntax Editor)

Cette fenêtre permet d'écrire les commandes d'analyses statistiques. Elle fonctionne comme un traitement de texte simple. Lorsqu'une commande est complète, on peut l'exécuter en allant dans le menu "Run : Current" (ou encore en tapant Ctrl-R). Pour obtenir une fenêtre de syntaxe vide, aller dans le menu "Fichier : Nouveau : syntaxe".

c. fenêtre des résultats (SPSS Viewer)

Cette fenêtre apparaît après qu'une commande d'analyse a été effectuée, et contient les résultats de cette analyse. Les résultats proprement dit apparaissent à droite alors qu'à gauche, on voit une table des matières des résultats générée par SPSS. Les résultats peuvent être imprimés tels quels, ou encore, on peut faire copier-coller vers un autre logiciel (tel votre traitement de texte). Parfois, le résultat est très long et SPSS n'en montre qu'une partie, suivi d'un triangle rouge. Pour voir la suite, double-cliquez sur le texte, puis étendez la taille de la zone de texte vers le bas jusqu'à ce que vous ne voyez plus de texte.

1.3. Comment organiser les données dans SPSS

Les données doivent être organisées de la façon suivante:

- Chaque ligne représente un participant; si le participant a été observé plusieurs fois, on doit retrouver plusieurs colonnes, une par observation.
- Chaque colonne représente une variable, soit:
- La ou les variables indépendantes identifiant la condition expérimentale, le numéro de sujets, et possiblement des informations sur le sujet (tel le sexe, l'âge, etc.) selon les besoins;
- La (ou les) variable dépendante, c'est à dire le résultat de la mesure. Si vous mesurez la taille, il y aura une colonne pour la taille; si vous cherchez à mesurer le Q.I, une colonne contiendra le résultat du sujet.
- Vous pouvez utiliser autant de colonnes que désiré. De plus, il peut y avoir des colonnes avec de l'information qui ne servira pas mais que vous avez récolté quand même.
- L'ordre des colonnes n'est pas important.

Dans SPSS (nous verrons comment plus loin), vous pouvez donner les noms que vous voulez à vos colonnes. Utiliser des noms significatifs (tel **sexe**, **age**, **qi**, etc.) de préférence à des noms sans signification (tel **toto**, **patate**, ou encore **v1**). Par contre, SPSS ne permet pas l'utilisation de symboles spéciaux dans le nom d'une colonne (le point, la virgule, le dièse, etc.), ne permet pas d'utiliser plus que huit lettres, et finalement, un nom de colonne doit commencer par une lettre.

1.4. Règles d'utilisation de la syntaxe

SPSS fonctionne de deux façons: soit avec les menus, ou encore en mode syntaxe. Le mode de fonctionnement avec la souris et les menus peut sembler plus intuitif mais dans les faits, est beaucoup plus difficile à apprendre (il faut des douzaines de clics à droite et à gauche pour arriver à un résultat). De plus, les menus ne donnent pas accès à toutes les commandes SPSS. Finalement, si vous devez faire la même analyses sur plusieurs fichiers de données, vous ne pouvez pas enregistrer la séquence de clics.

Par opposition, dans le mode syntaxe, il faut dactylographier suivant une syntaxe rigide la ou les commandes que l'on veut effectuer sur les données. Il est avantageux de procéder directement dans la fenêtre de syntaxe. D'une part, cette méthode est beaucoup plus flexible: Plus de commandes sont disponibles de cette façon. Aussi, une fois la syntaxe faite pour une opération, il est facile d'enregistrer les commandes et de les réutiliser pour différents fichiers de données.

Voici les règles générales pour écrire des commandes dans SPSS :

- Chaque nouvelle commande se trouve en tête de ligne, précédée d'aucun espace.
- Les options qui suivent une commande débute sur la ligne suivante et sont précédées d'au moins un espace et d'une barre oblique (/).

- Chaque commande doit ABSOLUMENT se terminer par un point.
- Lorsqu'on spécifie un nom de fichier, il doit être "entre guillemets".
- SPSS ne fait pas de différence entre les lettres majuscules et minuscules. Vous pouvez taper les commandes autant dans l'une que l'autre casse.
- De plus, entre les commandes, vous pouvez insérer des lignes vides. SPSS les ignore, mais peuvent améliorer la lisibilité de vos commandes quant il y en aura plusieurs dans une fenêtre.

Pour exécuter une commande, il faut sélectionner la commande à exécuter puis choisir dans le menu "Run : Current" ou encore Ctrl-R ou encore utiliser le bouton 'run' (►).

Nous verrons au cours de la session un grand nombre de commandes. Toutes respectent la syntaxe décrite ci-haut. On appelle parfois un ensemble de commandes SPSS un *script* ou un *script d'analyse* et parfois aussi *un programme d'analyse*. Vous pouvez enregistrer votre script pour le modifier plus tard.

Notons en passant que vous pouvez aussi enregistrer la fenêtre de résultats mais que très souvent, il en résulte un fichier énorme (que vous ne pourrez sauvegarder sur une disquette). Cependant, tant et aussi longtemps que vous avec votre script et vos données, vous pouvez toujours exécuter le script à nouveau pour revoir les résultats.

1.5. Comment ouvrir un fichier de données avec SPSS

L'ouverture d'un fichier de données déjà existant est sans doute l'étape la plus cruciale. Pour ce faire, nous allons vraiment détailler avec soin la procédure à suivre. Il faut premièrement obtenir le fichier, savoir où il se trouve sur votre ordinateur avant d'écrire la commande SPSS qui pourra l'ouvrir.

a. Obtenir un fichier de données.

Il existe plusieurs façon d'obtenir un fichier de données. La plus simple est de dactylographier soit-même le fichier. Pour se faire, ouvrez un traitement de texte simple (tel *notepad* ou *Bloc-note*) et entrer les données en respectant les règles de la section 1.3. Entre les données, vous pouvez utiliser un ou plusieurs espaces ou encore une tabulation, comme dans l'exemple qui suit à la Figure 3.

Ici, nous avons des données sur deux colonnes. Dans cette étude (fictive), nous avons demandé à des québécois d'origine française (code 1) et à des québécois d'origine asiatique (code 2) de nommer le plus de cousins, cousines, oncles, tantes et autres membres de leur famille en 3 minutes. Chaque ligne est un participant différent (15 personnes ont participés à cette étude); la colonne un nous indique l'origine du participant, et la colonne deux, le nombre de membres de sa famille qu'il a pu nommer en trois minutes.

Le fichier ne doit contenir que les résultats. De plus, vous devez le sauvegarder et quitter l'application avant d'essayer de l'ouvrir avec SPSS.

Une autre façon est d'avoir obtenu le fichier par une disquette ou sur Internet. Dans le premier cas, il est préférable de copier le fichier sur le disque dur pour accélérer l'accès au fichier.

Plusieurs des fichiers utilisés dans le cours seront aussi disponible sur le site web du cours. Pour récupérer un fichier, il suffit souvent de cliquer sur le lien, et votre explorateur va vous demander où vous souhaitez enregistrer le fichier. S'il ne pose pas la question, appuyez

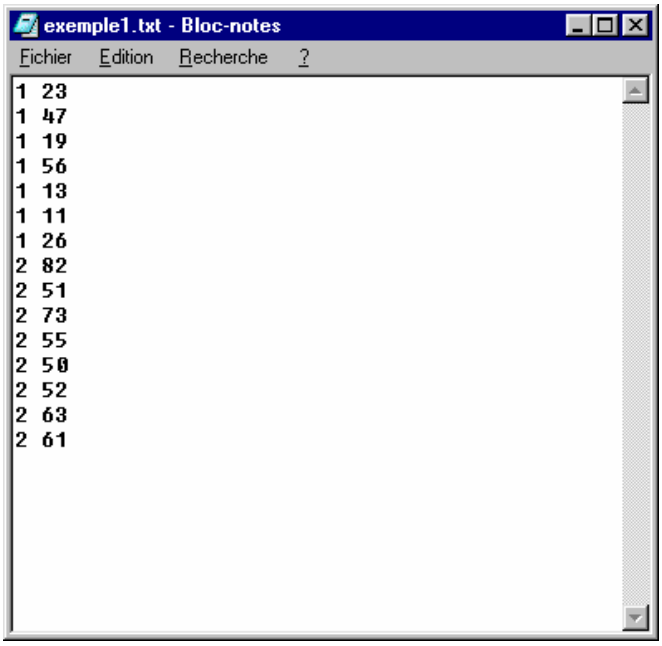


Figure 3 : Exemple d'un fichier de données

sur la touche CTRL avant de cliquer sur le fichier pour forcer votre explorateur à prendre une copie du fichier. Alternativement, vous pouvez aussi utiliser le bouton de droite de la souris pour enregistrer le lien.

b. Localiser le fichier

Dans SPSS, il faut indiquer la localisation complète du fichier en plus de son nom. Pour les gens familiarisés avec le DOS, il s'agit d'un exercice facile. Pour les autres, ça peut sembler plus difficile. Nous proposons ci-bas deux façons possibles de demander à Windows de nous localiser un fichier.

b.1. En explorant le disque dur de votre ordinateur.

En utilisant l'explorateur de fichier de Windows, vous pouvez localiser votre fichier de données (ici, exemple1.txt), comme dans la Figure 4.

Dans la section "Adresse", vous voyez la localisation du fichier, c'est à dire dans notre cas: **C:\WINDOWS\Bureau**. Autrement dit, le fichier se trouve sur le disque dur (C:), dans un dossier WINDOWS, et sous-dossier Bureau. Les barres obliques inversées sont des séparateurs entre les noms de dossiers.

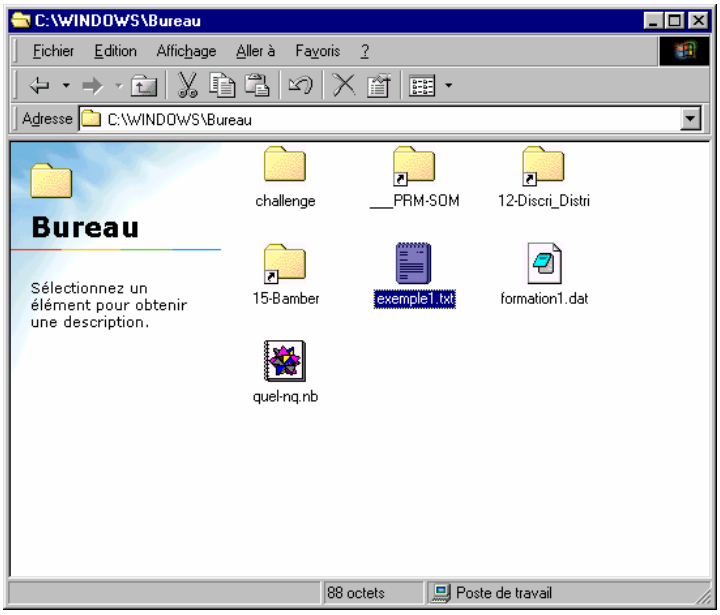


Figure 4 : Localisation du fichier de données

b.2. En faisant une recherche sur votre disque dur:

Si vous ne trouvez vraiment pas le fichier de données, faites une recherche sur votre disque dur en utilisant dans le menu Démarrer, l'option Rechercher (Figure 5), puis en entrant le nom du fichier recherché (exemple1.txt, voir Figure 6).

Dans les deux cas, nous obtenons le même résultat.

c. Ouverture du fichier avec SPSS

Commencez par démarrer SPSS puis ouvrez une fenêtre de syntaxe vide en allant dans le menu "File : New : Syntax". Dans cette fenêtre, nous allons entrer une commande dont le but est d'ouvrir des fichiers de données. Cette commande s'appelle `data list`.

Figure 5 : Rechercher un fichier

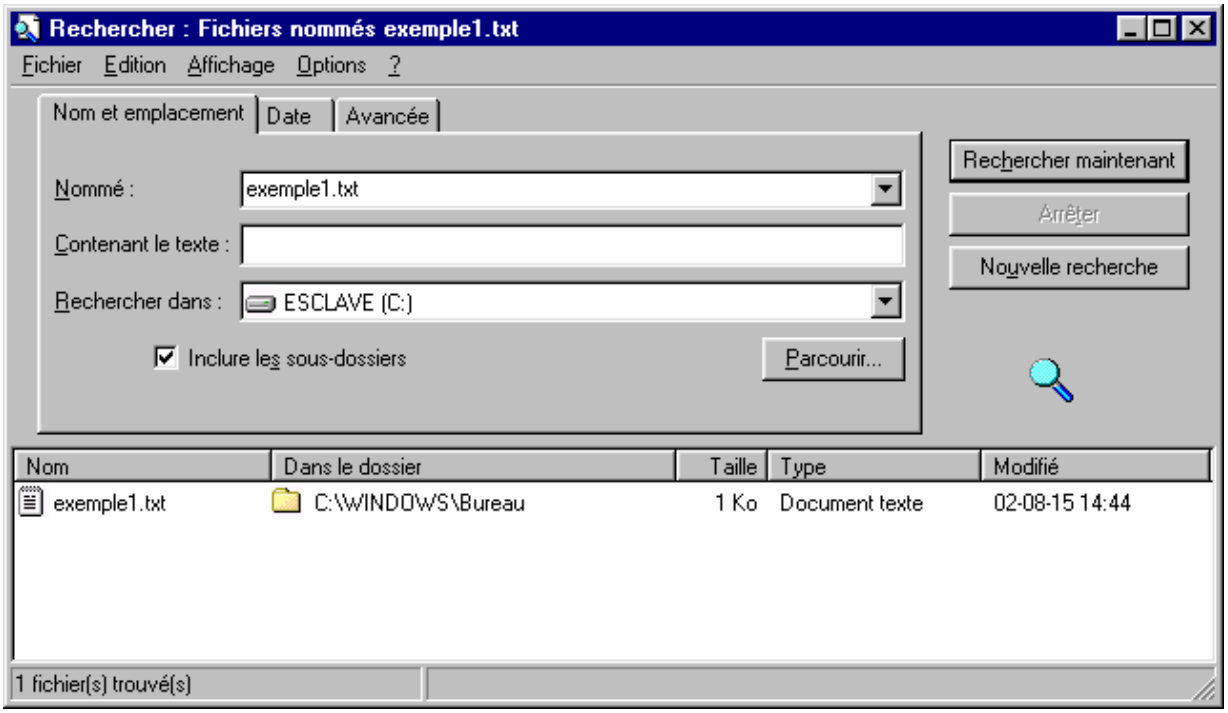


Figure 6 : Entré du nom de fichier à rechercher

Commande et options:

```
o   data list file="localisation\nom du fichier" list
o          /nomcol1 nomcol2 etc.
o   execute.
```

Notez la présence du point final et le fait que la seconde ligne est distante de quelques espaces et débute par une barre oblique.

Dans cette commande, vous devez spécifier la localisation du fichier (voir point précédent) ainsi que le nom complet du fichier (souvent, les noms de fichier se termine par .dat ou encore .txt; cette partie doit aussi se trouver dans le nom de fichier). De plus, vous devez nommer vos colonnes.

Si vous exécutez la commande `data list` (Ctrl-R) et qu'elle a bien fonctionné, vous ne verrez pas de résultat dans la fenêtre de résultats puisque cette commande ne produit rien. Cependant, après avoir exécuter la commande `execute`, vous verrez vos données dans la fenêtre de données.

Notez que sur les ordinateurs ayant Windows en français, il arrive parfois que le point décimal ne soit pas reconnu. Si tel est le cas, vous devez ouvrir le fichier avec un traitement de texte et remplacer tous les points par des virgules. Pour gagner du temps, utilisez la fonction Chercher et remplacer qui se trouve dans le menu "Édition" d'un grand nombre de traitements de texte.

1.6. Exemple complet

Vous allez trouver deux fichiers sur le site Web de ce cours, à la section pour la formation SPSS. Essayez de récupérer le fichier nommé Exemple1.dat et de l'ouvrir.

Sur le site Web, récupérez le fichier (voir Figure 7) et notez sa localisation sur le disque dur. Notez que pour récupérer un fichier, il ne suffit pas de cliquer dessus. Il faut utiliser le bouton de gauche de la souris et choisir « enregistrer sous ».

Dans une fenêtre de syntaxe, entrez la commande que l'on voit dans la Figure 8.

Notez que **origine** et **famille** sont des noms arbitraires que j'ai choisi. Vous auriez pu choisir des noms de colonnes différentes. Cependant, les noms ont huit lettres ou moins.

J'ai aussi exécuté la commande `execute`. Parfois, SPSS mets des commandes dans une file d'attente. La commande `execute` permet de vider la file d'attente et d'exécuter toutes les

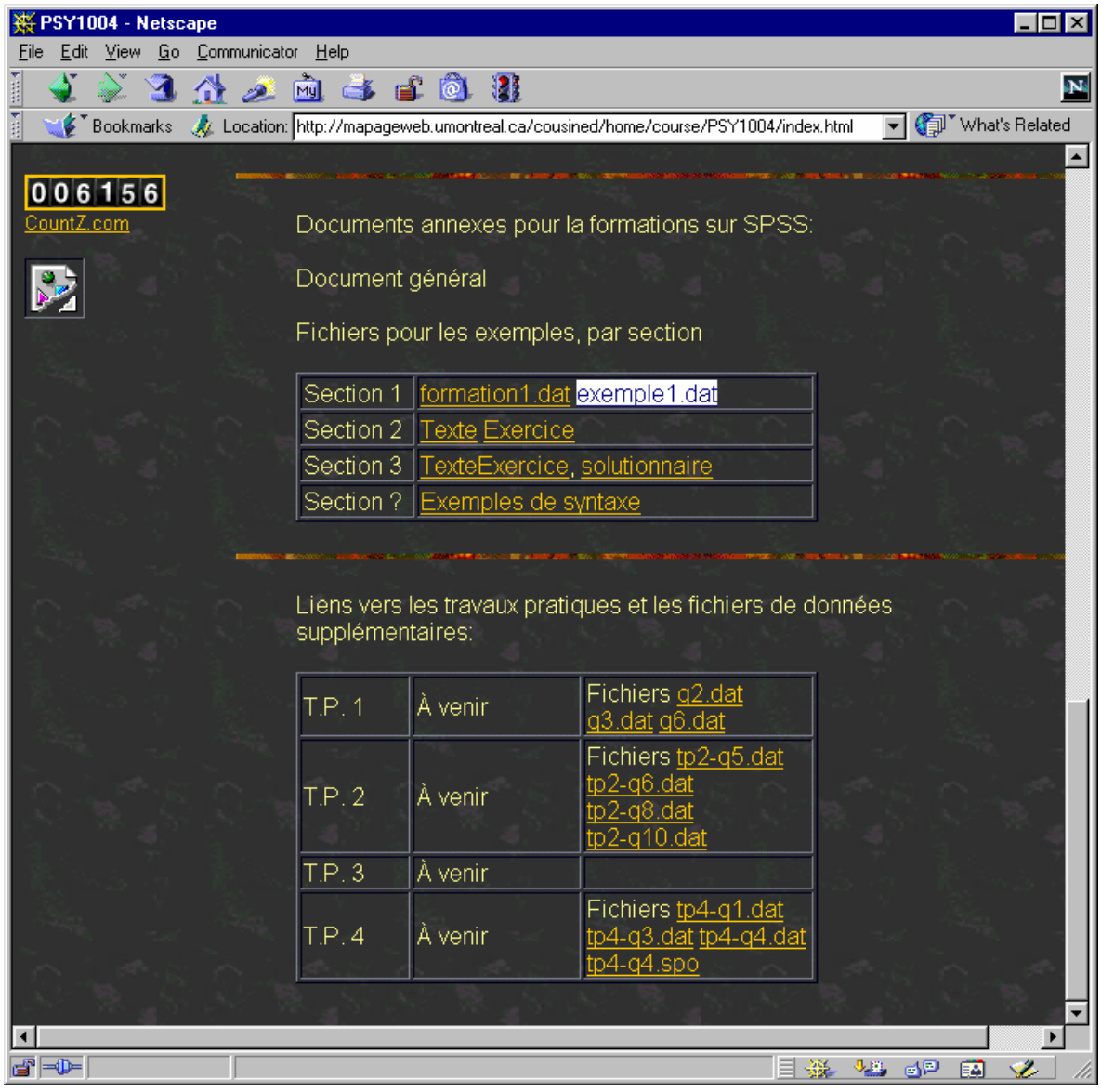


Figure 7 : Récupérer un fichier sur le site Web

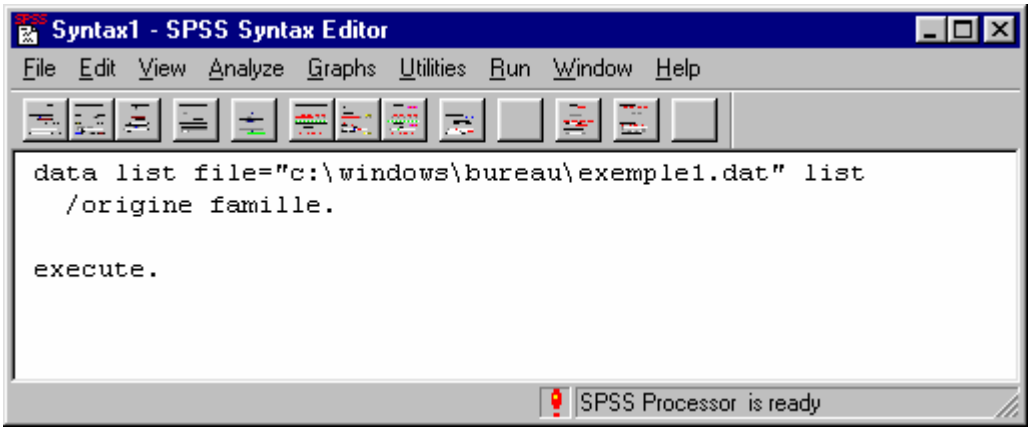


Figure 8 : Ouvrir un fichier de donnéesavec la syntaxe

commandes en attente maintenant.

Pour vous assurer que la commande a bien fonctionné, vérifiez que la fenêtre des données contient bien les bonnes données, comme c'est le cas dans la Figure 9.

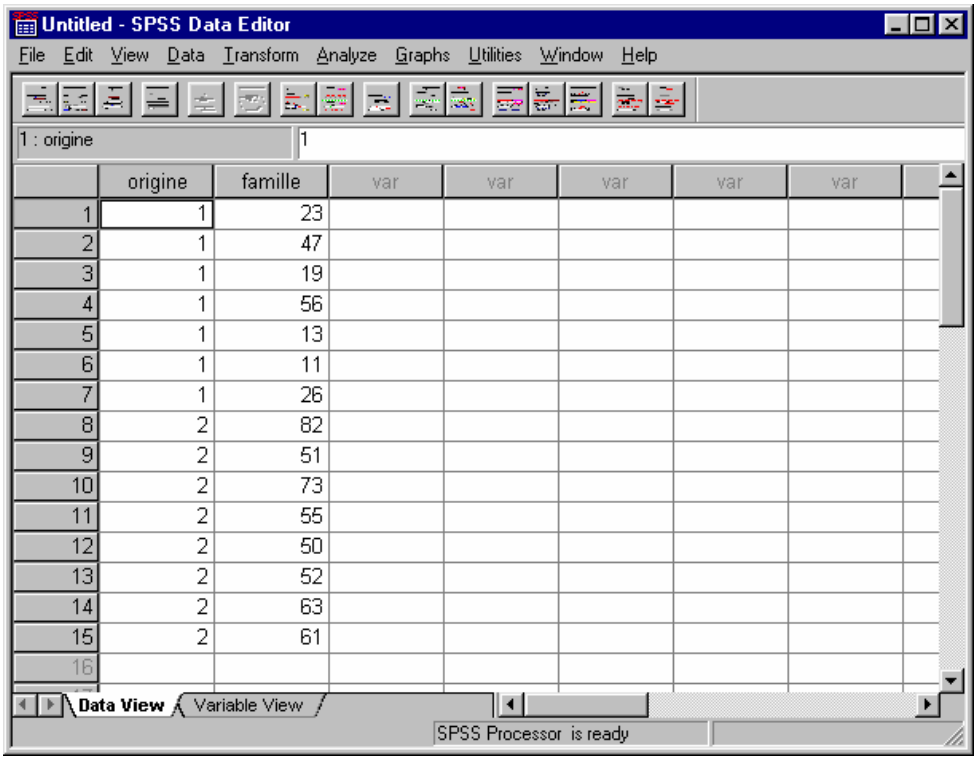


Figure 9 : Résultat de l'ouverture d'un fichier dans SPSS

2. Examiner les données

2.0. Quelques commandes SPSS

Nous continuons avec d'autres commandes SPSS qui s'écrivent dans la fenêtre de syntaxe et qui doivent être exécutées avec "Run : Current" ou Ctrl-R. N'oubliez pas le point final qui fait partie des commandes.

Dans ce qui suit, la présence d'accolades { } indique une section optionnelle d'une commande. De plus, *nomcol* représente le nom d'une de vos colonnes, telles que vous les avez nommé dans la commande `data list`.

a. COMMENT

Cette commande permet de mettre des commentaires dans le fichier syntaxe. Leur exécution ne produit absolument aucun résultat. Syntaxe:

```
o   comment votre-commentaire-ici.
```

b. DATA LIST

Nous avons déjà vu cette commande qui permet d'ouvrir un fichier de données. Cette commande est très importante et est toujours faite en premier.

```
o   data list file="localisation\nom du fichier" list
o          /nomcol1 nomcol2 etc.
```

c. EXAMINE

Permet d'obtenir des statistiques descriptives sur les données. Entre autres, elle calcule la moyenne et la variance de votre échantillon, le score le plus faible, le plus élevé, etc. Cette commande produit une longue sortie dans la fenêtre de résultats. Nous en verrons un exemple plus loin. L'option `Plot=none` est pour indiquer à la commande de ne pas produire de graphes. Ceux-ci sont assez grossiers; nous verrons à la section 2.2 comment faire de meilleurs graphes.

```
o   examine variable=nomcol  {by nomcol}
o          /statistics=descriptive
o          /plot=none.
```

La section optionnelle `by` permet d'indiquer si vous voulez des statistiques pour différents sous-groupes. Nous verrons un exemple à la section 2.3.

d. EXECUTE

Nous avons déjà vu la commande `execute` à la section précédente. Elle permet d'exécuter les commandes précédentes maintenant (lorsque nous l'exécutons avec Ctrl-R) si elles ont été mises dans une file d'attente. Il n'y a pas d'option à la commande:

```
o   execute.
```

e. QUANTILE

Il est parfois utile de diviser les données en quantiles, par exemple dans le cas de quartile (4), on veut les 25% des plus petits scores dans un groupe, les 25% suivant dans un deuxième groupe, les 25% suivant dans un troisième groupe, et les données restantes dans un dernier groupe. Dès lors, on peut vouloir la valeur qui sépare les groupes, c'est à dire à partir de quel score on passe du premier quartile au second, etc. La commande est

```
o   frequencies variable=nomcol
o     /ntile=nbre-de-séparations
o     /format=notable.
```

f. SELECT IF

Sert à sélectionner un sous-ensemble de lignes du fichier; SPSS effectue les analyses subséquentes sur ces données uniquement. Cette commande est utilisée pour éliminer les données aberrantes sans avoir à les effacer physiquement du fichier. Il s'agit en fait d'une façon de filtrer les données.

```
o   select if nomcol condition valeur.
```

Dans cette commande, *condition* peut être soit >, <, >=, <= ou encore <>, qui signifient respectivement: plus grand, plus petit, plus grand ou égal, plus petit ou égal, et différent de.

Par exemple, si vous savez qu'une donnée aberrante s'est glissée dans la colonne QI de votre fichier, un résultat trop grand pour être crédible (supérieur à 280!), vous pouvez l'omettre des analyses qui vont suivre en faisant

```
o   select if qi < 280.
```

qui ne va conserver que les lignes dans le fichier où la colonne qi est inférieure à 280. Notez que `select if` n'efface rien dans le fichier d'origine. Si plus tard, vous ouvrez à nouveau ce fichier, n'oubliez pas d'omettre les données aberrantes à nouveau en ré-exécutant la commande `select if`. Vous pouvez avoir plusieurs commandes `select if` à la suite si vous désirez filtrer les données suivant plusieurs critères.

g. IGRAPH

La commande `igraph` permet de faire des graphiques. Cependant, cette commande est très complexe, et possède une bonne vingtaine d'options. Plutôt que d'apprendre ces options par cœur, il est préférable d'utiliser les menus. Dans la section suivante, nous montrons comment faire des graphes avec les menus.

2.1. Comment faire des graphiques avec SPSS

Pour réaliser un graphe avec SPSS, il faut d'abord ouvrir un fichier de données (avec la commande `data list`) et si nécessaire, avoir filtré les données aberrantes (voir l'exemple complet à la section 2.3). De façon, générale, il faut aller dans le menu "Graphes : Interactif"

(voir Figure 10). Parmi les différents graphes possibles, il y a le graphe des histogrammes qui est très important, puis les graphes de moyennes, soit Bar, Line, Dot. que nous verrons après.

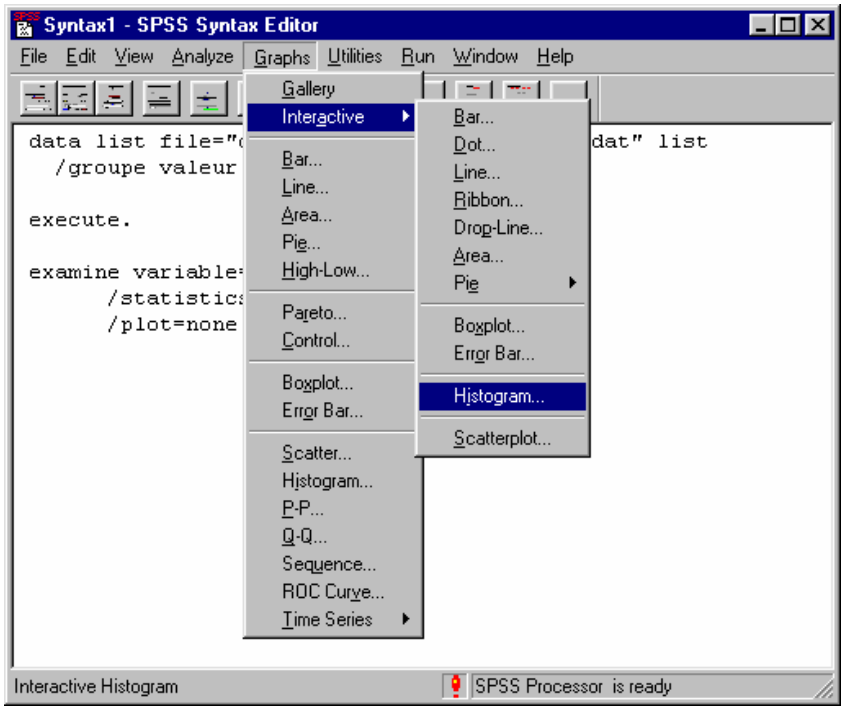


Figure 10 : Le menu Graphs : Interactive

a. Graphe des histogrammes.

Ce graphe est très utile pour détecter les données aberrantes dans le fichier qui devraient être filtrées. Il est généralement fait en tout premier par pure précaution. Si aucune

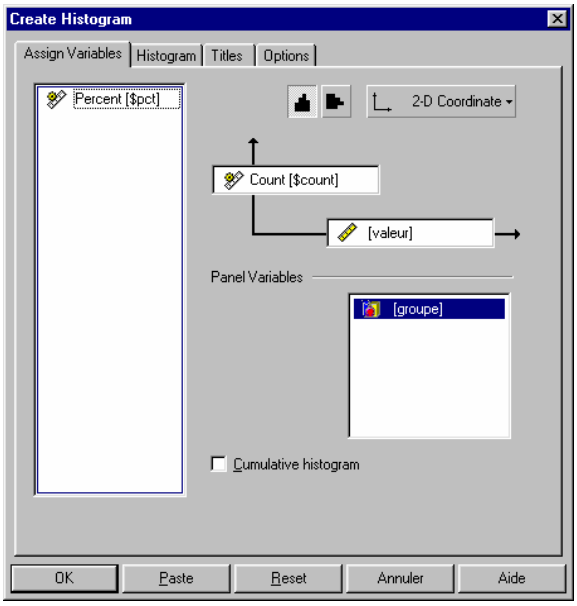


Figure 11 : Graphe des histogrammes

anomalie n'est visible, le graphe ne servira probablement plus.

Pour faire un graphe des histogrammes, ouvrir un fichier de données avec la syntaxe, puis aller dans le menu "Graphes : Interactif : Histogrammes". La boîte de dialogue illustrée à la Figure 11 apparaît (version anglaise illustrée).

Sur le côté gauche, on voit le nom des colonnes qui existent dans votre fichier de données (telles que définies par votre commande `data list`). À droite, vous pouvez choisir ce que représente l'axe horizontal en glissant un *nomcol* de la gauche vers cet espace avec votre souris. Sur l'axe vertical se trouve déjà la mention: *Count*: c'est la façon pour SPSS de faire un décompte dans chaque catégorie de valeur (n'y touchez pas).

Si jamais vous avez plusieurs groupes de sujets dans votre fichier, vous pouvez demander à obtenir un graphe par groupe, dans différents panneaux (*panels*). Pour ce faire, il vous faut un *nomcol* qui indique le groupe d'appartenance de chaque observation (que j'ai appelé `groupe` dans mon exemple; voir section 2.3). À l'occasion lorsque vous utilisez les panneaux, vous verrez le message de la Figure 12: Appuyez sur "convert".

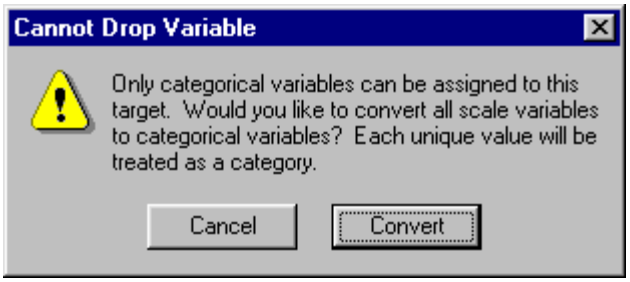


Figure 12 : Convertir en variable categorielle

La Figure 13 montre le résultat que vous devriez obtenir dans la fenêtre de résultats. Le graphique n'est pas très beau puisque nous n'avons que 15 observations au total (7 à gauche et 8 à droite). De façon générale, un graphique des histogrammes est plus fiable si nous avons une centaine d'observations par groupe. Dans ce cas-ci, il est impossible de dire s'il existe des données aberrantes.

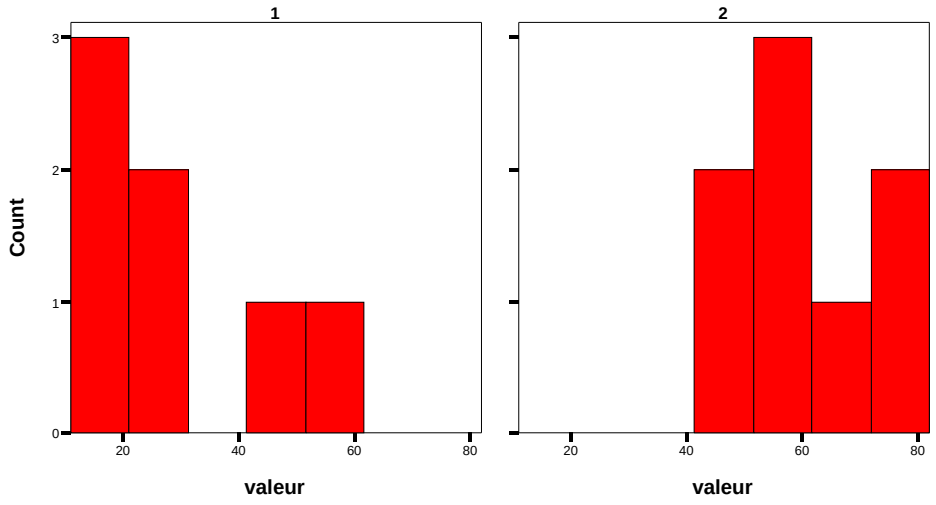


Figure 13 : Résultat de la commande Graphes : Histogrammes

b. Graphes des moyennes

Ayant vérifié les données brutes pour la présence de données aberrantes, l'étape suivante est de faire un résumé des données en regardant les moyennes des groupes. On voudrait donc un seul graphique montrant la moyenne des différents groupes. Pour ce faire, il faut un graphique des moyennes. SPSS nous offre différentes présentations (bar, dot et line), mais toutes fonctionnent sur le même principe. Allez dans "Graphs : Interactive : Bar", et vous verrez le dialogue visible dans la Figure 14.

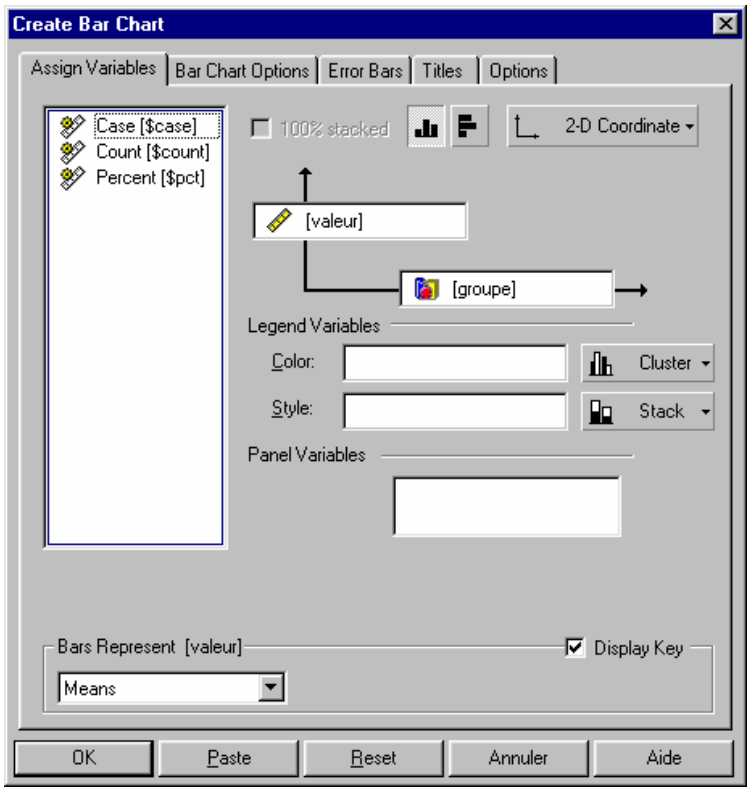


Figure 14 : Dialogue pour obtenir un graphique des moyennes

Dans l'axe vertical, glissez avec la souris un *nomcol* qui représente le score de vos participants que vous voulez comparer au travers des groupes. Vérifiez qu'en bas, la boîte de choix "Bars represent" montre bien "Means". Sur l'axe horizontal, indiquez le *nomcol* qui permet de savoir dans quel groupe chaque sujet se trouve. [En utilisant le bouton de droite de la souris sur l'axe horizontal, j'ai changé [groupe] de "scale" à "categorical" pour que le résultat soit plus joli].

Avant de faire "Ok", allez sur l'onglet "Error Bars" pour cocher "Display Error Bars" et choisissez "Standard Error of mean" dans la liste (nous verrons plus tard ce que représente l'erreur standard ou l'erreur type), comme on le voit dans la Figure 15.

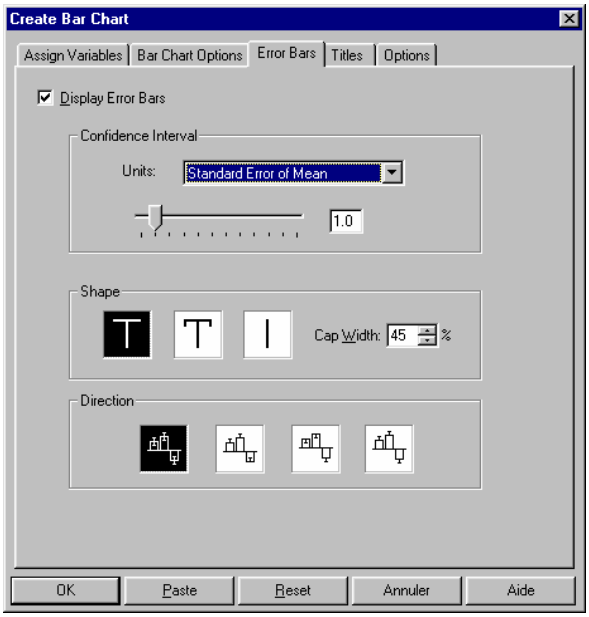


Figure 15 : Spécifier les barres d'erreurs dans un graphe des moyennes

Appuyez sur "Ok" pour obtenir le graphe. La Figure 16 présente des exemples de graphes auquel j'ai ajouté des titres et sous-titres en utilisant l'onglet "Titles":

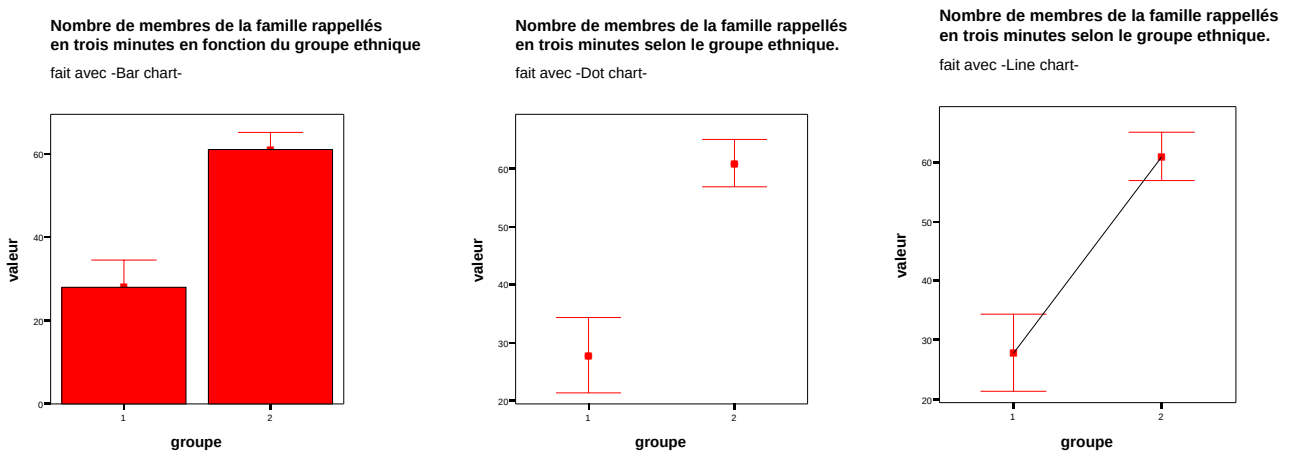


Figure 16 : Trois exemples de graphiques des moyennes.

2.2. Exemples complets

Nous allons réaliser deux exemples.

a. Le fichier Formation1.dat

Allez sur le site web du cours et trouvez le fichier **formation1.dat** se trouvant vers le bas de la page. Le début du fichier est illustré dans la Figure 17.

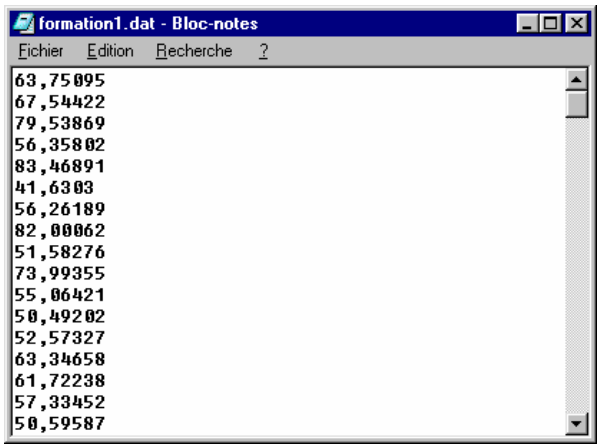


Figure 17 : Une partie des données du fichier Formation1.dat

Le fichier ne contient qu'une colonne de valeurs, le poids en kilogrammes de 203 poches de cafés faites en Amérique centrale par des ouvriers. On a constaté que le poids varie considérablement d'une poche à l'autre et les acheteurs aimeraient savoir quel est le poids moyen d'une poche. Votre but est donc de vérifiez le fichier pour voir s'il n'existe pas d'anomalie, puis de sortir des statistiques décrivant les poches de cafés.

Comment allez-vous procéder?

Premièrement, il faut vérifier les données brutes pour s'assurer qu'il n'existe pas d'anomalie. Pour ce faire, ouvrez le fichier de données puis faites un graphique des histogrammes.

```
o   comment ouverture du fichier.
o   data list file="c:\windows\bureau\formation1.dat" list
o     /valeur.
o   execute.
```

Notez que `"c:\windows\bureau\formation1.dat"` est la localisation du fichier sur mon ordinateur, mais que cela peut varier. De plus, j'ai choisi d'appeler la colonne de poids "`valeur`" mais "`poids`" aurait été aussi bon, sinon meilleur.

Puis faites le graphe des histogrammes avec les menus pour vérifier la présence de données aberrantes (voir le dialogue de la Figure 18).

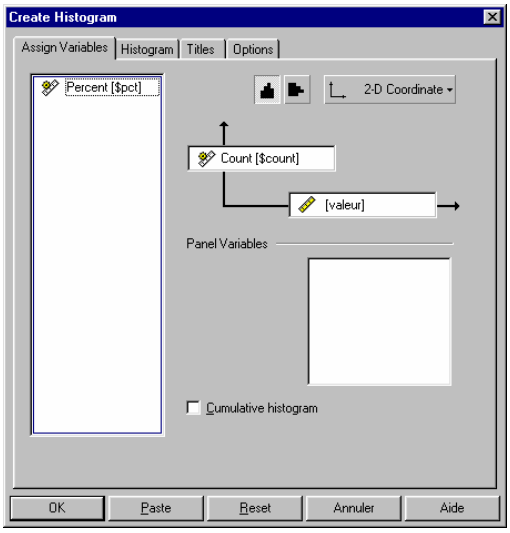


Figure 18 : Création d'un graphique en histogramme

Vous devriez obtenir la Figure 19. Que peut-on en conclure?

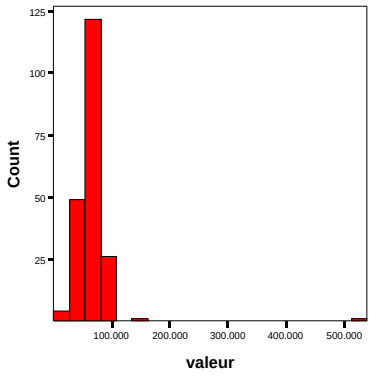


Figure 19 : Graphique des histogrammes des données du fichier Formation1.dat

Bien qu'on ne la voie presque pas, il existe une donnée aberrante passée 500 kilogrammes. Vérification faite auprès des manutentionnaires de café dans le vieux port, il est impossible qu'une poche puisse être aussi lourde, et il s'agit sans aucun doute d'une erreur dans la saisie de données. Il faut donc omettre cette valeur dans les analyses suivantes. Pour ce faire, filtrer le fichier de données avec la commande `select if`:

```
o   comment élimine la donnée aberrante à 500 kg.
o   select if valeur < 500.
```

Puis on refait l'histogramme, que l'on voit dans la Figure 20.

Cette fois-ci, on voit mieux la dispersion des données, mais il existe encore deux points isolés, un supérieur à 125 kg, et l'autre autour de 0 kg. Les deux sont visiblement hors-norme, et nous allons les omettre des analyses suivantes:

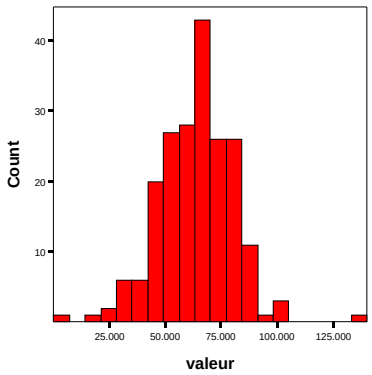


Figure 20 : Histogramme après l'enlèvement de données aberrantes

```
o   comment élimine les deux données aberrantes à 130 et 0 kg.
o   select if valeur < 125.
o   select if valeur > 2.
```

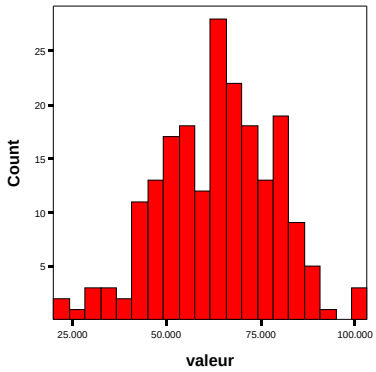


Figure 21 : Les données du fichier Formation1.dat après l'enlèvement de toutes les données aberrantes

Voici en Figure 21 le graphique des histogrammes final:

Nous passons aux statistiques descriptives avec la commande:

```
o   comment génère les statistiques descriptives.
o   examine variable=valeur
o     /statistics=descriptive
o     /plot=none.
```

Voici dans la Figure 22 le résultat qui apparaît dans la fenêtre de résultats.

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
VALEUR	200	100.0%	0	.0%	200	100.0%

Descriptives

			Statistic	Std. Error
VALEUR	Mean		63.51307	1.073235
	95% Confidence Interval for Mean	Lower Bound	61.39669	
		Upper Bound	65.62944	
	5% Trimmed Mean		63.77569	
	Median		64.24151	
	Variance		230.367	
	Std. Deviation		15.177838	
	Minimum		20.073	
	Maximum		103.169	
	Range		83.096	
	Interquartile Range		21.97413	
	Skewness		-.193	.172
	Kurtosis		.026	.342

Figure 22 : Les statistiques descriptives des données du fichier Formation1.dat

Nous apprenons qu'après avoir traité 200 cas, le poids moyen est de 63.5 ± 1.1 kg. L'écart type dans le poids est de 15.2 kg, ce qui est beaucoup. (Rappelons que) Nous avons éliminé trois données aberrantes, une inférieure à 2 kg et les deux autres supérieures à 125 kg.

b. Exemple 2: le QI pour les gens de la ville et de la campagne.

Le fichier de données s'appelle QI.dat et se trouve sur le site web. Le fichier contient deux colonnes, la première indiquant par un 1 ou un 2 la provenance des personnes (1=ville, 2=campagne), et la seconde, la mesure du QI obtenue sur un test standardisé.

Examinez le fichier pour la présence de données aberrantes, puis calculez des statistiques pour chacun des habitants de la ville et de la campagne.

La seule différence ici est que nous avons deux groupes de sujets. Pour faire le graphe des histogrammes, nous allons utiliser deux panneaux. Finalement, pour la commande `examine`, nous allons séparer les statistiques selon le groupe d'appartenance.

La syntaxe est:

```
o   comment ouverture du fichier de données.
o   data list file="c:\windows\bureau\QI.dat" list
o     /groupe qi.
o   execute.
o
o   comment calcul des stats de base.
o   examine variable=qi by groupe
o     /statistics=descriptive
o     /plot=none.
```

Les noms de colonne `groupe` et `qi` sont totalement à mon choix, puis j'utilise ces noms dans les commandes subséquentes. Avant de faire la commande `examine`, nous avons

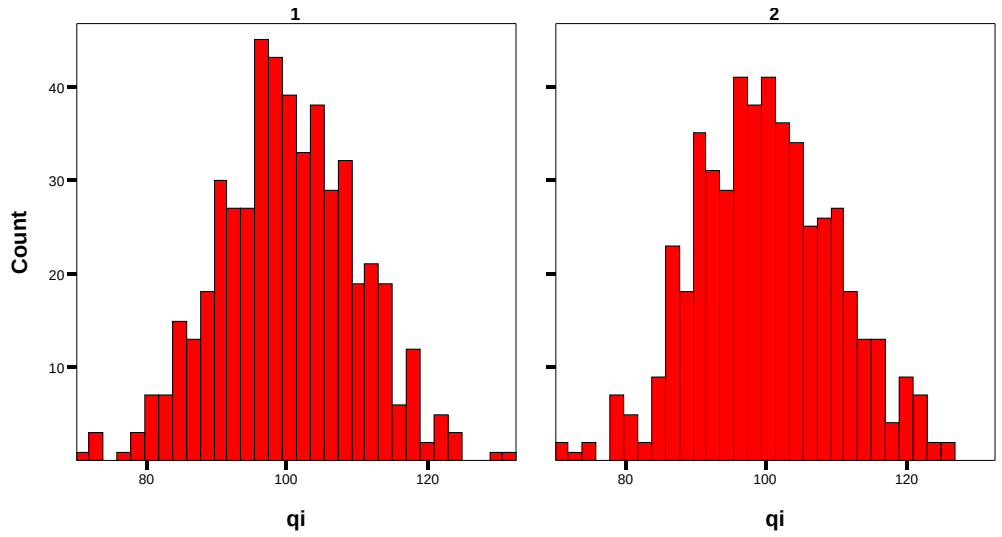


Figure 23 : Le graphe des histogrammes pour le QI

examiné les données pour la présence possible de données aberrantes, et il ne semble pas y en avoir d'extrêmes (voir Figure 23).

La fenêtre des résultats est imprimée dans la Figure 24. Elle vient en deux sections, une pour l'ensemble des données, peu importe le groupe d'appartenance, l'autre section divisée selon le groupe. On voit dans la seconde section que la moyenne du QI des deux groupes est très semblable, 100.1 ± 0.4 vs. 99.9 ± 0.4. Bien qu'on ne puisse rien conclure sans un test statistique, il semble en tout cas que si différence il y a, elle est très faible.

Explore

Total Sample

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Perc
QI	1000	100.0%	0	.0%	1000	100.

Descriptives

			Statistic	Std. Error
QI	Mean		100.03	.314
	95% Confidence Interval for Mean	Lower Bound	99.42	
		Upper Bound	100.65	
	5% Trimmed Mean		100.01	
	Median		99.70	
	Variance		98.618	
	Std. Deviation		9.931	
	Minimum		70	
	Maximum		133	
	Range		63	
	Interquartile Range		13.89	
	Skewness		.038	.077
	Kurtosis		-.037	.155

GROUPE

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	GROUPE	N	Percent	N	Percent	N	Percent
QI	1	500	100.0%	0	.0%	500	100.0%
	2	500	100.0%	0	.0%	500	100.0%

Descriptives

	GROUPE			Statistic	Std. Error
QI	1	Mean		100.13	.441
		95% Confidence Interval for Mean	Lower Bound	99.26	
			Upper Bound	100.99	
		5% Trimmed Mean		100.11	
		Median		99.70	
		Variance		97.378	
		Std. Deviation		9.868	
		Minimum		70	
		Maximum		133	
		Range		63	
		Interquartile Range		13.48	
		Skewness		.017	.109
		Kurtosis		.056	.218
	2	Mean		99.94	.447
		95% Confidence Interval for Mean	Lower Bound	99.06	
			Upper Bound	100.82	
		5% Trimmed Mean		99.91	
		Median		99.73	
		Variance		100.038	
		Std. Deviation		10.002	
		Minimum		71	
		Maximum		126	
		Range		55	
		Interquartile Range		13.87	
		Skewness		.059	.109
		Kurtosis		-.114	.218

Figure 24 : Statistique descriptives

3. Transformation des données

3.0. Quelques commandes SPSS

Il peut être nécessaire parfois de transformer les données en cotes *z*. Il s'agit alors de réaliser une transformation linéaire, dans laquelle on soustrait par la moyenne, puis on divise par l'écart type du groupe. Pour réaliser une transformation linéaire (ou non linéaire, voir section 11), la commande SPSS est:

```
o   compute nomcol = (nomcol - mu ) / sigma.
```

où mu est une valeur spécifique (qui peut être trouvée avec la commande explore) et sigma est aussi une valeur spécifique. Par exemple:

```
o   compute valeur2 = (valeur - 100 ) / 10.
```

va créer une nouvelle colonne, nommée valeur2, qui contient les données de valeur normalisée suivant une moyenne de 100 et un écart type de 10.

Avec la commande `compute`, n'importe quelle transformation peut être réalisée qui utilise des +, -, / et * (multiplication) et des noms de colonnes existants. Par exemple, cette commande totalement arbitraire:

```
o   compute valeur2 = (valeur * groupe) - 50.
```

où on suppose que les colonnes `valeur` et `groupe` existent, va créer une nouvelle colonne `valeur2`.

4. Les tests de base

4.1. Quelques commandes SPSS

a. Commande pour effectuer un test binomial sur une proportion

Ce test permet de tester l'hypothèse suivante:

$$\mathbf{H_0: p = } \textbf{\textit{probabilité}}$$

$$\mathbf{H_1: p \neq } \textbf{\textit{probabilité}}$$

où *probabilité* est une valeur fixée à priori par votre hypothèse de recherche. Pour réaliser ce test, il faut indiquer à SPSS dans quelle colonne se trouve codé le succès ou l'échec de l'observation, et quelles valeurs codent un succès, un échec (souvent 1 et 0, mais peut être une autre valeur). La syntaxe est (le signe ¬ est la virgule):

```
O   npar tests
O     /binomial(probabilité)=nomcol(val-succès¬ val-échec).
```

b. Commande pour effectuer un test de médiane sur plusieurs groupes.

Dans le test de médiane implémenté dans SPSS, on veut savoir si deux ou plusieurs groupes ont la même médiane M_0. L'hypothèse testée pour tous les groupes simultanément est:

$$\mathbf{H_0: } M_1 = M_0, M_2 = M_0, \ldots$$

$$\mathbf{H_1: } M_1 \neq M_0, M_2 \neq M_0, \ldots$$

où M_i est la médiane du groupe numéro *i*. SPSS peut calculer la médiane globale automatiquement si nous ne mettons pas une médiane M_0 entre crochet ou nous pouvons lui en indiquer une entre crochets après le mot `median`.

```
O   npar tests
O     /median { [M0] }=nomcol1 by nomcol2(val-groupe1¬ val-groupe2).
```

c. Commande pour effectuer un test des signes sur observations couplées.

Ce test permet de tester l'hypothèse suivante:

$$\mathbf{H_0: } M_1 = M_2$$

$$\mathbf{H_1: } M_1 \neq M_2$$

où M_1 est l'incidence d'une caractéristique avant et M_2, l'incidence de cette même caractéristique après. Pour réaliser ce test, le fichier doit contenir deux colonnes par sujet, une indiquant son score avant, l'autre colonne, son score après. La syntaxe est:

```
o     npar tests
o     /sign=nomcol1 with nomcol2.
```

d. Commandes manquantes

Il n'existe pas dans SPSS de commande pour tester deux proportions (du genre H_0: $p_1=p_2$); nous verrons à la section 6 une méthode alternative pour réaliser ce genre de test.

De plus, il n'existe pas de commande pour tester la médiane d'un seul groupe (du genre H_0: $M = M_0$). Une façon indirecte d'arriver à faire un test sur la médiane d'un seul groupe est de rajouter une colonne qui ne contiendrait que la valeur de la médiane à tester (M_0), puis de faire un test des signes entre la colonne contenant le score et la colonne contenant la médiane à tester.

4.2. Exemple complet

a. Exemple sur une proportion

Allez chercher le fichier **pile-ou-face.dat** se trouvant sur le site web du cours. Il contient une série de nombres, où 1 indique que nous avons obtenu un pile sur le lancer d'un 25 cents. Est-ce que cette pièce est normale ou est-elle truquée? (au seuil $\alpha = 5\%$)

L'hypothèse est que la proportion de 1 est de 50%. On fait un test de proportion:

```
o   data list file="c:\windows\bureau\pile-ou-face.dat" list
o     /succes.
o   execute.
o
o   npar tests
o     /binomial(0.5)=succes(0,1).
```

Le résultat est montré à la Figure 25. On voit qu'il y a 48 piles sur 100, une proportion de .48 comparée à la proportion de notre hypothèse de .50. Le test ne rapporte que la *signification* (*Sig.*), c'est à dire le seuil α qui aurait fallu choisir pour que le test soit significatif. Puisque nous avons choisi un seuil de 5% et que le seuil reporté (76%) est nettement plus grand, le test

NPar Tests

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (2-tailed)
SUCCES	Group 1	0	48	.48	.50	.764
	Group 2	1	52	.52		
	Total		100	1.00		

a Based on Z Approximation.

Figure 25 : Résultat du test de proportion

n'est pas significatif. La pièce est normale selon toute vraisemblance.

b. Exemple de test sur des médianes

Nous avons mesuré le revenu de trois cents personnes dans trois régions de Montréal différentes. Les régions sont (1) le Plateau Mont-Royal, (2) Westmount, et (3) la Petite Patrie. Est-ce que les revenus sont comparables? Les données se trouvent dans le fichier **revenus.dat**.

Premièrement, il faut voir que les résultats sont distribués de façon très asymétrique (comme nous le voyons avec un graphe des histogrammes, voir la Figure 26). Comme l'échantillon montre une très importante violation du postulat de normalité, il n'est pas possible de faire un `test-t` (voir section 5).

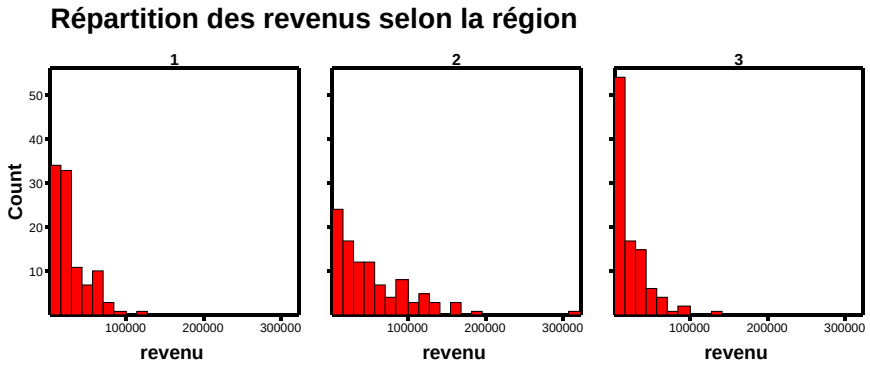


Figure 26 : Graphes des histogrammes montrant le revenu en fonction des trois régions

Nous réalisons donc un test de la médiane dans lequel SPSS va calculer la médiane générale lui-même. SPSS utilise une approche différente de celle vue en classe qui lui permet de tester simultanément plus de deux groupes à la fois (nous verrons le rationnel de cette approche à la section 6 du cours):

```
o   data list file="c:\windows\bureau\revenus.dat" list
o     /region revenu.
o   execute.
o
o   npar tests
o     /median=revenu by region(1,3).
```

Les résultats se trouvent à la Figure 27. Ce que nous voyons dans la première partie, c'est que pour la région du plateau Mont-royal (région 1), 47 personnes ont un revenu supérieur à la médiane, alors que dans la région de Westmount (région 2), 68 personnes ont un revenu qui dépasse la médiane, etc. Dans la seconde partie, on voit la médiane utilisée par SPSS, qui est la médiane de l'ensemble des données (peut importe la région): 20 704.44$. La *signification* du test est de .000. Ceci signifie qu'avec un seuil α de 5%, le test est significatif. Le revenu médian diffère significativement selon la région.

Median Test

Frequencies

		REGION		
		1	2	3
REVENU	> Median	47	68	35
	<= Median	53	32	65

Test Statistics

	REVENU
N	300
Median	20704.44
Chi-Square	22.320
df	2
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 50.0.

b Grouping Variable: REGION

Figure 27 : Résultat du test de la médiane

On peut voir que SPSS utilise un test du χ^2. Nous verrons à la section 6 ce qu'est ce test; cependant, il donne le même résultat qu'un test de la médiane basée sur la distribution binomiale.

c. Exemple avant-après

Nous avons constaté que beaucoup d'étudiants utilisent des mots creux comme, "t'sais", "euh", etc. Pour tenter de réduire l'incidence de ces mots, une méthode de correction béhavioriste a été testée sur 400 étudiants de l'UdeM. Nous avons mesuré le nombre de mots creux en 10 minutes avant la méthode, puis après. Nous voulons savoir si la méthode a réduit l'incidence de ces mots creux. Les données se trouvent dans le fichiers **mots.dat**.

```
o   data list file="c:\windows\bureau\mots.dat" list
o     /avant apres.
o   execute.
o
o   examine variables=avant
o     /statistics=descriptives
o     /plot=none.
o
o   examine variables=apres
o     /statistics=descriptives
o     /plot=none.
o
o   npar tests
o     /sign=avant with apres.
```

Les résultats se trouvent à la Figure 28. Ils montrent qu'avant le traitement, les sujets disent en moyenne 100 ± 1 mots creux par 10 minutes alors qu'après, ils en disent 111 ± 1!

Pour savoir si la méthode a significativement aggravé la situation des étudiants, un test des signes a été réalisé. Il indique que les résultats diffèrent significativement après le traitement comparé au score avant le traitement. La méthode n'est pas du tout conseillée pour réduire le nombre de mots creux.

Explore

Descriptives

			Statistic	Std. Error
AVANT	Mean		100.23	1.027
	95% Confidence Interval for Mean	Lower Bound	98.21	
		Upper Bound	102.25	
	5% Trimmed Mean		99.96	
	Median		100.00	
	Variance		422.057	
	Std. Deviation		20.544	
	Minimum		45	
	Maximum		173	
	Range		128	
	Interquartile Range		27.00	
	Skewness		.245	.122
	Kurtosis		.263	.243

Descriptives

			Statistic	Std. Error
APRES	Mean		111.41	1.036
	95% Confidence Interval for Mean	Lower Bound	109.37	
		Upper Bound	113.45	
	5% Trimmed Mean		111.54	
	Median		112.50	
	Variance		429.145	
	Std. Deviation		20.716	
	Minimum		45	
	Maximum		190	
	Range		145	
	Interquartile Range		27.00	
	Skewness		-.068	.122
	Kurtosis		.322	.243

Sign Test

Frequencies

		N
APRES - AVANT	Negative Differences	136
	Positive Differences	262
	Ties	2
	Total	400

a APRES < AVANT
b APRES > AVANT
c AVANT = APRES

Test Statistics

	APRES - AVANT
Z	-6.266
Asymp. Sig. (2-tailed)	.000

a Sign Test

Figure 28 : Résultat du test des signes

5. Les tests t

5.1. Quelques commandes SPSS

a. Test-t sur un groupe unique

Le test-t à un groupe unique peut être utilisé si le graphe des histogrammes ne montre pas d'anomalies importantes. Il permet de tester l'hypothèse:

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

```
o   t-test testval μ0
o     /variable=nomcol.
```

où μ_0 est une valeur précise choisie à priori selon l'hypothèse.

b. Test t avec des données appariées.

Le test-t avec données appariées est utilisé quand on possède une paire d'observations sur un individu, par exemple, des données avant traitement et après traitement. L'hypothèse est du genre:

$$H_0: \mu_{avant} = \mu_{après}$$

$$H_1: \mu_{avant} \neq \mu_{après}$$

où autant μ_{avant} que $\mu_{après}$ sont estimés par nos observations.

```
o   t-test pairs=nomcol1 with nomcol2
o     /missing listwise.
```

L'option `listwise` indique au logiciel quoi faire lorsqu'il manque une donnée dans une cellule. Dans ce cas-ci, il va exclure la ligne au complet, ce qui est la façon la plus prudente.

c. Test t avec deux groupes indépendants.

Permet de tester deux groupes de mesures indépendantes. Le fichier doit contenir des lignes de données pour le premier groupe, des lignes de données pour le second groupe, et une colonne supplémentaire doit contenir un numéro identifiant à quel groupe la donnée appartient. L'hypothèse est du genre:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

où autant μ_1 que μ_2 sont estimés par nos observations. Ils représentent la moyenne du groupe 1 et du groupe 2. La syntaxe est:

```
o   t-test groups=nomcol (val-grp1, val-grp2)
o     /variable=nomcol.
```

où la spécification `groups` permet d'indiquer dans quelle colonne se trouve le numéro de groupe, val-grp1 est le numéro que portent les sujets du premier groupe et val-grp2 est le numéro du second groupe à comparer.

5.2. Exemple complet

Nous revenons à l'exemple des mots creux de la section précédente. En effet, nous sommes peu satisfaits de la conclusion, et il existe peut-être une raison pour ce résultat incroyable. Entre autre, nous avons oublié de regarder le graphe des histogrammes pour la présence de données aberrantes. Les graphiques (ici, on ne peut pas utiliser *panels* et nous exécutons donc la commande deux fois) sont montrés à la Figure 29.

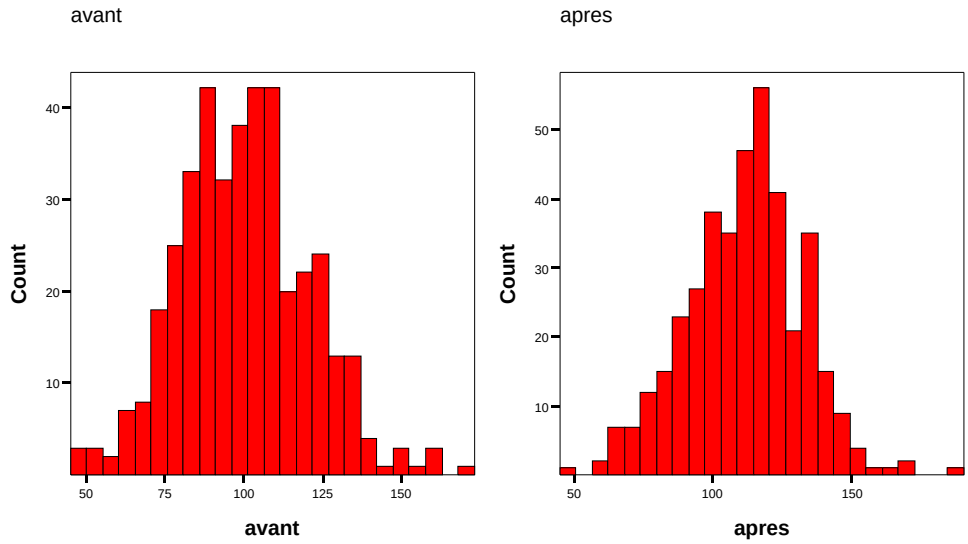


Figure 29 : Le graphe des histogrammes

Comme on peut le voir, il n'y a pas de données vraiment aberrantes. Peut-être le point isolé à droit sur le graphe "après" (situé à 190), mais la séparation n'est pas très grande. Cependant, comme nous croyons vraiment que la conclusion précédente est dans l'erreur, nous allons néanmoins omettre ce point de notre analyse.

La syntaxe complète suit, ainsi que les résultats. Vous noterez que dans la section t-test, on voit à nouveau la moyenne des deux colonnes, avant et après, la donnée supposée extrême étant enlevée dans le second cas. La moyenne de l'échantillon ne change virtuellement pas, passant de 111.41 à 111.21 après avoir éliminé une donnée.

Le t obtenu dans la dernière partie du tableau est très éloigné de zéro (-7.58) et la signification inférieure à notre seuil α de 5%. Le test plus puissant confirme ce que notre test des signes avait trouvé. On conclura donc à regret: "La méthode béhavioriste telle que testée sur notre échantillon d'étudiants de l'UdeM accroît significativement l'usage de mots creux ($t(398) = -7.58$, $p < .05$). Dans cette analyse, nous avons omis une donnée extrême (après = 190)."

```
o   data list file="c:\windows\bureau\mots.dat" list
o     /avant apres.
o   execute.
o
o   comment executer le graphe des histogrammes ici.
o
o   comment examine pour voir la valeur de la données maximale.
o   examine variable=apres
o     /statistics=descriptives
o     /plot none.
o
o   select if apres < 190.
o
o   t-test pairs=avant with apres
o     /missing=listwise.
```

Les résultats se trouvent dans la Figure 30.

Explore

Total Sample

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
APRES	400	100.0%	0	.0%	400	100.0%

Descriptives

			Statistic	Std. Error
APRES	Mean		111.41	1.036
	95% Confidence Interval for Mean	Lower Bound	109.37	
		Upper Bound	113.45	
	5% Trimmed Mean		111.54	
	Median		112.50	
	Variance		429.145	
	Std. Deviation		20.716	
	Minimum		43	
	Maximum		190	
	Range		145	
	Interquartile Range		27.00	
	Skewness		-.068	.122
	Kurtosis		.322	.243

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	AVANT	100.22	399	20.568	1.030
	APRES	111.21	399	20.363	1.019

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	AVANT & APRES	399	-.004	.939

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	AVANT - APRES	-11.00	28.998	1.452	-13.85	-8.14	-7.575	398	.000

Figure 30 : Résultat des commandes SPSS

6. Les tests de χ^2

6.1. Quelques commandes SPSS

a. Test de répartition des effectifs dans une liste de catégories.

Le test permet d'examiner comment se répartissent des observations. Le fichier contient une série d'observations qui sont classées dans une ou l'autre des catégories possibles. Par exemple, si on veut savoir si autant d'hommes que de femmes entrent dans une certaine boutique, on pourrait noter 1 pour femme, 2 pour homme, et le fichier de données ne contiendrait qu'une seule colonne avec des 1 et des 2. L'hypothèse est du genre:

$$\mathbf{H_0: } O_1 = a_1, O_2 = a_2, \textbf{ etc}$$

$$\mathbf{H_0: } O_1 \neq a_1, O_2 \neq a_2, \textbf{ etc}$$

où a_i représente l'effectif attendu pour la $i^{ième}$ catégorie. La syntaxe est de la forme:

```
o   npar tests
o     /chisquare=nomcol(valcatégorie-min, valcategorie-max)
o     /expected=val-espérée-cat-1, val-espérée-cat-2, etc.
```

ou -si l'hypothèse précise des effectifs égaux-:

```
o   npar tests
o     /chisquare=nomcol(valcatégorie-min, valcategorie-max)
o     /expected=equal.
```

Dans ces commandes, *valcatégorie-min* et *valcatégorie-max* sont les numéros de la première et de la dernière catégorie respectivement, et *val-espérée-cat-i* est l'effectif attendue selon l'hypothèse, soit a_i.

Parfois, plutôt que d'avoir un long fichier ne contenant que le numéro dans lequel on classifie l'observation, il est possible que nous ayons sous la main un fichier déjà compilé, par exemple,

sexe	nombre
1	43
2	76

où la seconde colonne indique le nombre d'observations allant dans la catégorie 1 (43) et la catégorie 2 (76). Dans ce cas, on peut faire précéder la commande `npar tests` par la commande:

```
o   WEIGHT BY nomcol.
```

où *nomcol* indique la colonne contenant le nombre d'occurrences par catégories.

b. Tester la répartition dans un tableau de contingence à deux dimensions.

La commande suivante permet d'examiner la répartition d'observations se classant suivant deux critères. On appelle le tableau qui en résulte un tableau de contingence. Dans le cas où l'on voudrait tester l'hypothèse que les effectifs marginaux (dans la marge) reflètent bien les effectifs dans chacune des catégories, on réalise un test sur un tableau de contingence, avec l'hypothèse:

H_0: absence d'interaction

H_1: présence d'interaction.

qui est un raccourci pour dire que la fréquence dans les marges doit être congruente avec la fréquence dans chacune des cellules. La syntaxe est:

```
o   crosstabs variables=nomcol1(catégo-min, catégo-max) nomcol2(catégo-
    min, catégo-max)
o     /tables=nomcol1 by nomcol2
o     /cells=count expected
o     /statistics=chisq.
```

Encore une fois, si les données sont déjà compilées et donc, qu'il existe une colonne contenant les effectifs observés dans chaque combinaison de catégories, il est possible de faire précéder la commande `crosstabs` par:

```
o   WEIGHT BY nomcol.
```

c. Commandes qui n'existent pas dans SPSS

Le test des variances n'existe pas dans SPSS.

6.2. Exemple complet

a. Exemple 1: répartition des élèves dans une petite école secondaire

Soit une étude pour examiner si le nombre d'étudiants dans les classes d'une petite école secondaire est le même. Nous avons cinq catégories (le niveau au secondaire), et les effectifs sont faciles à trouver: il s'agit du nombre d'inscrits par niveau (les données sont déjà compilées). Les données sont:

exemple :

secondaire	nombre
1	24
2	34
3	25
4	31
5	15

Premièrement, entrez ces données dans un fichier que vous sauvegardez. Puis vérifiez l'hypothèse d'une répartition égale entre les niveaux. Dans ce cas, il faut ajouter la commande `weight by nombre` avant la commande `npar tests`. La syntaxe est donc:

```
o   data list file="c:\windows\bureau\ecole.dat" list
o     /secondai nombre.
o   execute.
o
o   weight by nombre.
o   npar tests
o     /chisquare=secondai(1,5)
o     /expected=equal.
```

Les résultats que l'on voit à la Figure 31 indiquent que la répartition n'est pas significativement différente entre les niveaux ($\chi^2(4) = 8.33$, $p > .05$).

Chi-Square Test

Frequencies

	SECONDAI			
	Category	Observed N	Expected N	Residual
1	1	24	25.8	-1.8
2	2	34	25.8	8.2
3	3	25	25.8	-.8
4	4	31	25.8	5.2
5	5	15	25.8	-10.8
Total		129		

Test Statistics

	SECONDAI
Chi-Square	8.326
df	4
Asymp. Sig.	.080

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 25.8.

Figure 31 : Résultat du test de répartition sur une liste

b. Exemple 2 (tiré des notes de cours): répartition de l'attitude (positive, négative) face aux cours de statistiques selon le sexe (hommes, femmes).

Nous avons ces données:

Sexe	Attitude	Nombre
1	1	30
1	2	93

2 1 33

2 2 2

qui sont des données déjà compilées sur le nombre de femmes (sexe 1) et d'hommes (sexe 2) qui ont une attitude favorable (attitude 1) ou défavorable (attitude 2) envers la statistique. On veut savoir si le sexe influence la façon de voir les statistiques (interaction). La syntaxe est donnée par:

```
o   data list file="c:\windows\bureau\sexeXatt.dat" list
o     /sexe att nombre.
o   execute.
o
o   weight by nbre.
o   crosstabs variables=sexe(1,2) att(1,2)
o     /tables=sexe by att
o     /cells=count expected
o     /statistics=chisq.
```

Les résultats montrent que le sexe change significativement l'attitude face aux statistiques (χ^2 (1) = 55.52, $p < .05$). Le résultat se trouve à la ligne « Pearson Chi-Square ».

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
SEXE * ATT	158	100.0%	0	.0%	158	100.0%

SEXE * ATT Crosstabulation

			ATT		Total
			1	2	
SEXE	1	Count	30	93	123
		Expected Count	49.0	74.0	123.0
	2	Count	33	2	35
		Expected Count	14.0	21.0	35.0
Total		Count	63	95	158
		Expected Count	63.0	95.0	158.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	55.522	1	.000		
Continuity Correction	52.645	1	.000		
Likelihood Ratio	60.514	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	55.170	1	.000		
N of Valid Cases	158				

a Computed only for a 2x2 table

b 0 cells (.0%) have expected count less than 5. The minimum expected count is 13.96.

Figure 32 : Résultat du test de tableau de contingence

8. L'ANOVA à un facteur

8.1. Quelques commandes SPSS

a. Commande pour effectuer une ANOVA *p* à *p* groupes indépendants

Dans le cas d'un schème avec groupes indépendants, il est important d'avoir une colonne pour indiquer le groupe d'appartenance de chacun des sujets. L'ANOVA permet des tester que tous les groupes ont la même moyenne:

$$H_0: \mu_1 = \mu_2 = \mu_3 \ldots$$

$$H_0: \mu_1 \neq \mu_2 \text{ ou } \mu_2 \neq \mu_3 \text{ ou } \mu_1 \neq \mu_3 \ldots$$

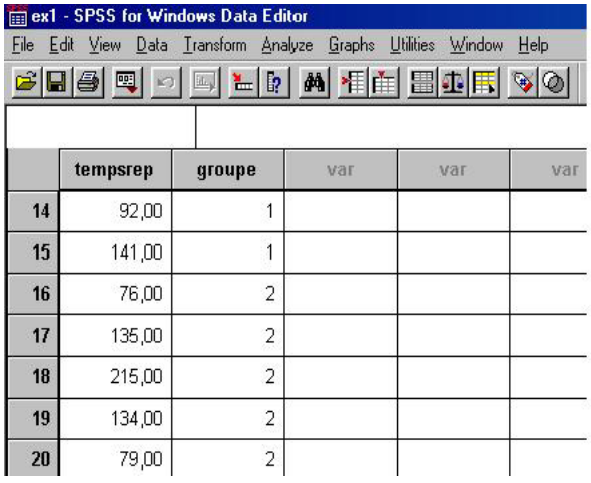
Il existe plusieurs commande dans SPSS pour faire une ANOVA à *p* groupes indépendants (`oneway`, `anova`). La commande que nous allons utiliser est `manova` car elle sera aussi utile dans les sections suivantes:

```
o   manova nomcol by nomcol(val-grp1, val-grpp)
o     /error=within
o     /design.
```

8.2. Exemple complet

Le professeur Anatomas veut savoir si l'endroit d'une lésion cérébrale affecte les temps de réponse à une tâche de calcul mental. Il teste donc un groupe de patients qui ont une lésion occipitale gauche, un groupe avec une lésion temporale droite et un groupe de personnes anormales mais sans lésion (étudiants en psychologie). Voici les temps obtenus par le premier groupe, en secondes {41, 61, 45, 121, 42, 56, 67, 98, 79, 76, 51, 68, 78, 92, 141}. Les temps de réponse du second groupe sont {76, 135, 215, 134, 79, 110, 76, 87, 98, 74, 102, 95, 99, 94, 105}. Le groupe anormal a obtenu les temps {110, 98, 75, 51, 60, 57, 78, 91, 56, 90, 41, 55, 76, 78, 99}. Y'a-t-il une différence significative entre les 3 groupes?

Votre fichier de données doit être organisé comme dans la Figure 33, avec une colonne pour l'identification du groupe:



	tempsrep	groupe	var	var	var
14	92,00	1			
15	141,00	1			
16	76,00	2			
17	135,00	2			
18	215,00	2			
19	134,00	2			
20	79,00	2			

Figure 33 : Exemple de fichier de données pour une ANOVA p

La syntaxe complète est:

```
o   data list file="c:\windows\bureau\anatomas.dat" list
o     /tempsrep groupe.
o   execute.
o
o   manova tempsrep by groupe(1, 3)
o     /error=within
o     /design.
```

Le résultat se trouve dans la Figure 34. Le groupe ayant une lésion temporale droite (moyenne = 105,267) est significativement plus lent que le groupe anormal (moyenne = 74,333) ($\underline{F}(2,42)=5,67$, $\underline{p} < .05$).

```
The default error term in MANOVA has been changed from WITHIN CELLS to
WITHIN+RESIDUAL.  Note that these are the same for all full factorial
designs.


* * * * * * A n a l y s i s   o f   V a r i a n c e * * * * * *


        45 cases accepted.
         0 cases rejected because of out-of-range factor values.
         0 cases rejected because of missing data.
         3 non-empty cells.

         1 design will be processed.

 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -


* * * * * * A n a l y s i s   o f   V a r i a n c e -- design   1 * * * * * *

 Tests of Significance for TEMPSREP using UNIQUE sums of squares

 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN CELLS           35389.87      42    842.62
 GROUPE                  9548.13       2   4774.07      5.67      .007

 (Model)                 9548.13       2   4774.07      5.67      .007
 (Total)                44938.00      44   1021.32

 R-Squared =           .212
 Adjusted R-Squared =  .175

 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

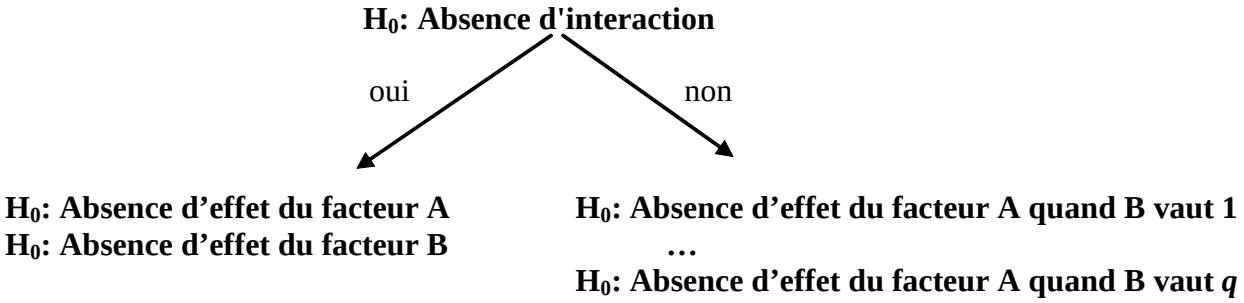
Figure 34 : Résultat d'une ANOVA p

9. L'ANOVA à groupes indépendants

9.1. Quelques commandes SPSS

a. Commandes pour effectuer une ANOVA factorielle p × q.

L'ANOVA permet de tester un ensemble d'hypothèses reliée à une situation de mesures où il y a deux variables indépendantes (ou plus). Appelons ces variables indépendantes, les facteurs A et B (des noms plus significatifs sont encouragés). De plus, supposons qu'il y a *p* niveaux pour le facteur A et *q* pour le facteur B. Les hypothèses ressemblent à:



La commande qui suit permet d'obtenir le test de l'interaction. De plus, s'il n'y a pas d'interaction, elle indique aussi le résultat pour les tests des effets principaux:

```
o   manova nomcol by fact-a(1, p) fact-b(1, q)
o     /error=within
o     /design.
```

où *nomcol* contient la mesure, *fact-A* est une colonne dans laquelle on voit le niveau du facteur A (allant de 1 à *p*) et *fact-B* est une colonne dans laquelle on voit le niveau du facteur B (1..*q*).

b. Décomposition en effets simples.

Si l'interaction est significative, il faut plutôt passer aux effets simples. Il faut alors faire une nouvelle commande dans laquelle on spécifie les effets simples désirés:

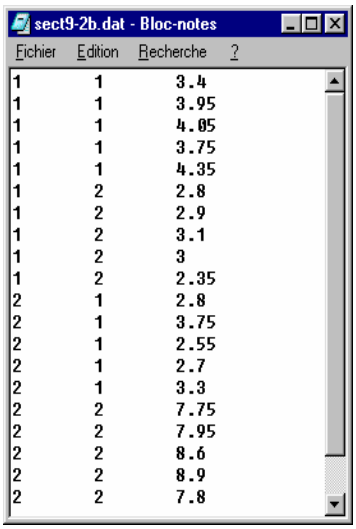
```
o   Manova nomcol by fact-a(1, p) fact-b(1, q)
o     /error=within
o     /design= fact-a within fact-b(1) … fact-a within fact-b (q).
```

Il faut demander autant d'effets simples qu'il existe de niveau de B. Alternativement (mais pas les deux), on peut vouloir les effets simples inverses, du facteur B quand le facteur A est fixé à ses différents niveaux. On écrira alors, pour les *p* niveaux de A:

```
o   manova nomcol by fact-a(1, p) fact-b(1, q)
o     /error=within
o     /design= fact-b within fact-a(1) … fact-b within fact-a (p).
```

9.2. Exemple complet (tiré des notes de cours)

Nous reproduisons l'exemple de la section 9.4(b) dont le fichier est disponible sur le site web. La Figure 35 donne le fichier de données et la syntaxe.




```
data list file="C:\windows\bureau\sect9-2b.dat" list
     /a b score.
execute.

manova score by a(1,2) b(1,2)
     /design.

comment puisque nous trouvons une interaction.
manova score by a(1,2) b(1,2)
     /design= a within b(1) a within b(2).
```

Figure 35 : Fichier de données et syntaxe pour une ANOVA p × q

Les résultats de la première commande, de la seconde se trouve dans la Figure 36.

The default error term in MANOVA has been changed from WITHIN CELLS to WITHIN+RESIDUAL. Note that these are the same for all full factorial designs.

* * * * * * Analysis of Variance * * * * * *

20 cases accepted.
0 cases rejected because of out-of-range factor values.
0 cases rejected because of missing data.
4 non-empty cells.

1 design will be processed.

* * * * * Analysis of Variance -- design 1 * * * * *

Tests of Significance for SCORE using UNIQUE sums of squares

Source of Variation	SS	DF	MS	F	Sig of F
WITHIN CELLS	2.90	16	.18		
A	25.20	1	25.20	139.23	.000
B	21.12	1	21.12	116.66	.000
A BY B	48.83	1	48.83	269.77	.000
(Model)	95.14	3	31.71	175.22	.000
(Total)	98.04	19	5.16		

R-Squared = .970
Adjusted R-Squared = .965

The default error term in MANOVA has been changed from WITHIN CELLS to WITHIN+RESIDUAL. Note that these are the same for all full factorial designs.

* * * * * * Analysis of Variance * * * * * *

20 cases accepted.
0 cases rejected because of out-of-range factor values.
0 cases rejected because of missing data.
4 non-empty cells.

1 design will be processed.

* * * * Analysis of Variance -- design 1 * * * * *

Tests of Significance for SCORE using UNIQUE sums of squares

Source of Variation	SS	DF	MS	F	Sig of F
WITHIN+RESIDUAL	24.01	17	1.41		
A WITHIN B(1)	1.94	1	1.94	1.37	.258
A WITHIN B(2)	72.09	1	72.09	51.04	.000
(Model)	74.03	2	37.01	26.21	.000
(Total)	98.04	19	5.16		

R-Squared = .755
Adjusted R-Squared = .726

Figure 36 : Résultat de l'ANOVA

10. L'ANOVA à mesures répétées et mixte

10.1. Quelques commandes SPSS

Puisqu'il s'agit toujours de la même commande `manova`, nous illustrons ici seulement les différences par rapport à la commande telle que vue dans les section 8 et 9.

Deux choses changent lorsque nous utilisons un plan à mesures répétées:

(a) l'usage de l'option:

```
o     /print=signif(GG)
```

qui va afficher le Greenhouse-Geiser pour avoir une meilleure idée de la *signification* du test.

(b) l'utilisation de l'option (ici, `ws` veut dire *within subject*)

```
o     /wsfactor= nomfacteur(p)
```

En effet, dans les mesures répétées, il faut toujours une ligne par sujet. Or, ce sujet est mesuré plus d'une fois, résultant en plusieurs colonnes de mesures. Chaque colonne représente un niveau particulier du facteur répété; avec `wsfactor`, il est possible de donner un nom global à ce facteur (qui n'est pas dans aucun des noms de colonnes).

Dans les plans à mesures répétés factoriels, la décomposition des effets simples se fait sur une nouvelle option,

```
o     wsdesign= nomfacteur within nomfacteur(niveau)
```

qui indique quelle composante intra sujet de l'ANOVA nous voulons décomposer.

Dans les plans mixtes, une chose change, l'utilisation de `mwithin` plutôt que `within` dans la décomposition des effets simples (ici, `m` vient de mixte). Puisque le facteur que l'on veut tenir constant n'est pas du même genre que le facteur que l'on veut décomposer (un peut être à groupes indépendants et l'autre à mesures répétées, ou vice-versa), on indique à SPSS que la suite de la décomposition se trouve sur la ligne `design` ou `wsdesign`.

Dans la suite, nous mettons une série d'exemples de syntaxe couvrant tous les cas possibles de plans simples (à groupes indépendants ou à mesures répétées) et de plans factoriels (à groupes indépendants, à mesures répétées et mixtes: groupes indépendants et mesures répétées dans la même expérience).

```
o   COMMENT Plan 4 à 4 groupes indépendants.
o   COMMENT Les données sont mises dans la colonne nommée score.
o   COMMENT une colonne nommée A contient le no du groupe.
o
o   manova score by A(1,4)
o     /error=within
o     /design.
```

```
o   COMMENT Plan (4) à mesures répétées.
o   COMMENT Les données sont mises dans quatre colonnes.
o   COMMENT nommées ici A1 A2 A3 A4.
o
o   manova A1 A2 A3 A4
o     /wsfactor=A(4)
o     /print=signif(GG)
o     /error=within
o     /design.
```

```
o   COMMENT Plan 2 x 3 à 6 groupes indépendants.
o   COMMENT Les données sont mises dans une colonne score.
o   COMMENT deux colonnes A et B contiennent les no de groupes.
o
o   COMMENT 1: Interaction et effets principaux.
o   manova score by A(1,2) B(1,3)
o     /error=within
o     /design.
o
o   COMMENT 2a: décomposition des effets simples suivant A.
o   manova score by A(1,2) B(1,3)
o     /error=within
o     /design=A within B(1) A within B(2) A within B(3).
o
o   COMMENT -ou- 2b: décomposition des effets simples suivant B.
o   manova score by A(1,2) B(1,3)
o     /error=within
o     /design=B within A(1) B within A(2).
```

```
o   COMMENT Plan 2 x (3) à 2 groupes indépendants.
o   COMMENT il s'agit d'un plan mixte: groupes indépendants
o   COMMENT et mesures répétées dans le même plan.
o   COMMENT Les données sont mises dans les colonnes B1 B2 B3.
o   COMMENT et une colonne nommée A contient le no de groupe.
o
o   COMMENT 1: Interaction et effets principaux.
o   manova B1 B2 B3 by A(1,2)
o     /wsfactor=B(3)
o     /print=signif(GG)
o     /error=within
o     /design.
o
o   COMMENT 2a: décomposition des effets simples suivant A.
o   manova B1 B2 B3 by A(1,2)
o     /wsfactor=B(3)
o     /print=signif(GG)
o     /error=within
o     /wsdesign=mwithin B(1) mwithin B(2) mwithin B(3)
o     /design A.
o
o   COMMENT -ou- 2b: décomposition des effets simples suivant B.
o   manova B1 B2 B3 by A(1,2)
o     /wsfactor=B(3)
o     /print=signif(GG)
o     /error=within
o     /wsdesign=B
o     /design=mwithin A(1) mwithin A(2).
```

```
o   COMMENT Plan (2 x 3) à mesures répétées.
o   COMMENT Les données sont mises dans six colonnes nommées.
o   COMMENT a1b1 a2b1 a1b2 a2b2 a1b3 a2b3, le facteur A étant
    COMMENT celui qui change le plus rapidement de niveaux.
o
o   manova A1B1 A2B1 A1B2 A2B2 A1B3 A2B3
o     /wsfactor=B(3) A(2)
o     /print=signif(GG)
o     /error=within
o     /design.
o
o   COMMENT 2a: décomposition des effets simples suivant A.
o   manova A1B1 A2B1 A1B2 A2B2 A1B3 A2B3
o     /wsfactor=B(3) A(2)
o     /print=signif(GG)
o     /error=within
o     /wsdesign=A within B(1) A within B(2) A within B(3)
o     /design.
o
o   COMMENT -ou- 2b: décomposition des effets simples suivant B.
o   manova A1B1 A2B1 A1B2 A2B2 A1B3 A2B3
o     /wsfactor=B(3) A(2)
o     /print=signif(GG)
o     /error=within
o     /wsdesign=B within A(1) B within A(2)
o     /design.
```

10.2. Exemple complet (tiré des notes de cours, section 10.1(b)

Dans cette expérience, nous comparons les scores de 5 sujets ayant été mesuré 8 fois. Il y avait deux facteurs manipulés, A et B, le premier ayant deux niveaux de difficulté et le second, quatre. A et B pourrait être le moment de la journée (matin, soir) et le moment de la semaine (début, milieu, fin, week-end). Il s'agit d'un plan factoriel à mesures répétées (2 × 4). Comme on le verra, SPSS tend à produire des résultats excessivement long.

Le fichier de données doit contenir une ligne par sujet, d'où les huit colonnes montrées sur la figure suivante:

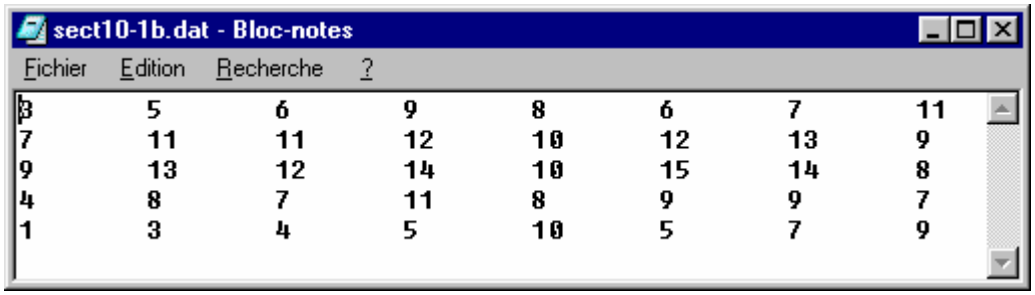


Figure 37 : Fichier de données

La syntaxe pour trouver l'interaction est la suivante:

```
o   data list file="c:\windows\bureau\sect10-1b.dat" list
o     /a1b1 a1b2 a1b3 a1b4 a2b1 a2b2 a2b3 a2b4.
o   execute.
o
o   manova a1b1 a2b1 a1b2 a2b2 a1b3 a2b3 a1b4 a2b4
o     /wsfactor=B(4) A(2)
o     /print=signif(GG)
o     /error=within
o     /design.
```

Comme on le voit dans le listing qui suit, il existe une interaction. On fait donc suivre l'analyse avec l'analyse des effets simples de B selon les niveaux de A (contrairement aux notes de cours, qui font la décomposition inverse):

```
o   COMMENT décomposition des effets simples suivant B.
o   manova a1b1 a2b1 a1b2 a2b2 a1b3 a2b3 a1b4 a2b4
o     /wsfactor=B(4) A(2)
o     /print=signif(GG)
o     /error=within
o     /wsdesign=B within A(1) B within A(2)
o     /design.
```

```
Note: there are 2 levels for the A effect.  Average tests are identical
to the univariate tests of significance.

******Analysis of Variance******
        5 cases accepted.
        0 cases rejected because of out-of-range factor values.
        0 cases rejected because of missing data.
        1 non-empty cell.

        1 design will be processed.
 ------------------------------------------

****Analysis of Variance -- design 1*****

Tests of Between-Subjects Effects.

 Tests of Significance for T1 using UNIQUE sums of squares
 Source of Variation          SS     DF       MS        F  Sig of F

 WITHIN CELLS             223.40      4    55.85
 CONSTANT                2924.10      1  2924.10    52.36     .002
 ------------------------------------------

*****Analysis of Variance -- design 1****

Tests involving 'B' Within-Subject Effect.

 Mauchly sphericity test, W =       .04652
 Chi-square approx. =          8.35162 with 5 D. F.
 Significance =                    .138

 Greenhouse-Geisser Epsilon =      .42465
 Huynh-Feldt Epsilon =             .53432
 Lower-bound Epsilon =             .33333

AVERAGED Tests of Significance that follow multivariate tests are equiv-
alent to univariate or split-plot or mixed-model approach to repeated
measures. Epsilons may be used to adjust d.f. for the AVERAGED results.
 ------------------------------------------

****Analysis of Variance -- design 1*****

 EFFECT .. B
 Multivariate Tests of Significance (S = 1, M = 1/2, N = 0)

 Test Name        Value   Exact F Hypoth. DF  Error DF  Sig. of F

 Pillais        .93482   9.56121       3.00      2.00       .096
 Hotellings   14.34182   9.56121       3.00      2.00       .096
 Wilks          .06518   9.56121       3.00      2.00       .096
 Roys           .93482
 Note.. F statistics are exact.
 ------------------------------------------

****Analysis of Variance -- design 1*****

Tests involving 'B' Within-Subject Effect.

 AVERAGED Tests of Significance for MEAS.1 using UNIQUE sums of squares
 Source of Variation          SS     DF       MS        F  Sig of F

 WITHIN CELLS             50.20     12     4.18
    (Greenhouse-Geisser)           5.10
    (Lower bound)                  4.00
 B                        35.30      3    11.77     2.81     .084
    (Greenhouse-Geisser)           1.27               2.81     .153
    (Lower bound)                  1.00               2.81     .169
 ------------------------------------------

****Analysis of Variance -- design 1*****

Tests involving 'A' Within-Subject Effect.

 Tests of Significance for T5 using UNIQUE sums of squares
 Source of Variation          SS     DF       MS        F  Sig of F

 WITHIN CELLS             27.40      4     6.85
 A                        25.60      1    25.60     3.74     .125
 ------------------------------------------
-

****Analysis of Variance -- design 1*****

Tests involving 'B BY A' Within-Subject Effect.

 Mauchly sphericity test, W =       .00545
 Chi-square approx. =         14.19008 with 5 D. F.
 Significance =                    .014

 Greenhouse-Geisser Epsilon =      .38238
 Huynh-Feldt Epsilon =             .43649
 Lower-bound Epsilon =             .33333

AVERAGED Tests of Significance that follow multivariate tests are
equivalent to
univariate or split-plot or mixed-model approach to repeated measures.
Epsilons may be used to adjust d.f. for the AVERAGED results.
 ------------------------------------------

****Analysis of Variance -- design 1*****

 EFFECT .. B BY A
 Multivariate Tests of Significance (S = 1, M = 1/2, N = 0)

 Test Name        Value   Exact F Hypoth. DF  Error DF  Sig. of F

 Pillais        .97721   28.59036       3.00      2.00       .034
 Hotellings   42.88554   28.59036       3.00      2.00       .034
 Wilks          .02279   28.59036       3.00      2.00       .034
 Roys           .97721
 Note.. F statistics are exact.
 ------------------------------------------

****Analysis of Variance -- design 1*****

Tests involving 'B BY A' Within-Subject Effect.

 AVERAGED Tests of Significance for MEAS.1 using UNIQUE sums of squares
 Source of Variation          SS     DF       MS        F  Sig of F

 WITHIN CELLS             27.40     12     2.28
    (Greenhouse-Geisser)           4.59
    (Lower bound)                  4.00
 B BY A                   42.60      3    14.20     6.22     .009
    (Greenhouse-Geisser)           1.15               6.22     .057
    (Lower bound)                  1.00               6.22     .067
 ------------------------------------------
```

Figure 38a : Première analyse pour dépister si l'interaction est significative

```
Note: there are 2 levels for the A effect.  Average tests are identical
to the univariate tests of significance.

******Analysis of Variance******
       5 cases accepted.
       0 cases rejected because of out-of-range factor values.
       0 cases rejected because of missing data.
       1 non-empty cell.

       1 design will be processed.
 ------------------------------------------

****Analysis of Variance -- design 1*****

Tests of Between-Subjects Effects.

 Tests of Significance for T1 using UNIQUE sums of squares
 Source of Variation          SS     DF       MS        F  Sig of F

 WITHIN CELLS             223.40      4    55.85
 CONSTANT                2924.10      1  2924.10    52.36      .002
 ------------------------------------------

****Analysis of Variance -- design 1*****

Tests involving 'B WITHIN A(1)' Within-Subject Effect.

 Mauchly sphericity test, W =       .24140
 Chi-square approx. =              3.86899 with 5 D. F.
 Significance =                     .568

 Greenhouse-Geisser Epsilon =       .69985
 Huynh-Feldt Epsilon =             1.00000
 Lower-bound Epsilon =              .33333

AVERAGED Tests of Significance that follow multivariate tests are equi-
valent to univariate or split-plot or mixed-model approach to repeated
measures. Epsilons may be used to adjust d.f. for the AVERAGED results.
 ------------------------------------------

****Analysis of Variance -- design 1*****

 EFFECT .. B WITHIN A(1)
 Multivariate Tests of Significance (S = 1, M = 1/2, N = 0)

 Test Name        Value    Exact F Hypoth. DF  Error DF  Sig. of F

 Pillais         .99208   83.50000       3.00      2.00       .012
 Hotellings   125.25000   83.50000       3.00      2.00       .012
 Wilks           .00792   83.50000       3.00      2.00       .012
 Roys            .99208
 Note.. F statistics are exact.
 ------------------------------------------

****Analysis of Variance -- design 1*****

Tests involving 'B WITHIN A(1)' Within-Subject Effect.

 AVERAGED Tests of Significance for MEAS.1 using UNIQUE sums of squares
 Source of Variation          SS     DF       MS        F  Sig of F

 WITHIN CELLS              7.10     12      .59
    (Greenhouse-Geisser)             8.40
    (Lower bound)                    4.00
 B WITHIN A(1)            74.15      3    24.72    41.77      .000
    (Greenhouse-Geisser)             2.10             41.77      .000
    (Lower bound)                    1.00             41.77      .003
 ------------------------------------------

****Analysis of Variance -- design 1*****

Tests involving 'B WITHIN A(2)' Within-Subject Effect.

 Mauchly sphericity test, W =       .01430
 Chi-square approx. =             11.56239 with 5 D. F.
 Significance =                     .041

 Greenhouse-Geisser Epsilon =       .38602
 Huynh-Feldt Epsilon =              .44234
 Lower-bound Epsilon =              .33333

AVERAGED Tests of Significance that follow multivariate tests are
equivalent to
univariate or split-plot or mixed-model approach to repeated measures.
Epsilons may be used to adjust d.f. for the AVERAGED results.
 ------------------------------------------

****Analysis of Variance -- design 1*****

 EFFECT .. B WITHIN A(2)
 Multivariate Tests of Significance (S = 1, M = 1/2, N = 0)

 Test Name        Value    Exact F Hypoth. DF  Error DF  Sig. of F

 Pillais         .66857   1.34483       3.00      2.00       .453
 Hotellings     2.01724   1.34483       3.00      2.00       .453
 Wilks           .33143   1.34483       3.00      2.00       .453
 Roys            .66857
 Note.. F statistics are exact.
 ------------------------------------------

****Analysis of Variance -- design 1*****

Tests involving 'B WITHIN A(2)' Within-Subject Effect.

 AVERAGED Tests of Significance for MEAS.1 using UNIQUE sums of squares
 Source of Variation          SS     DF       MS        F  Sig of F

 WITHIN CELLS             70.50     12     5.88
    (Greenhouse-Geisser)             4.62
    (Lower bound)                    4.00
 B WITHIN A(2)             3.75      3     1.25      .21      .886
    (Greenhouse-Geisser)             1.16              .21      .699
    (Lower bound)                    1.00              .21      .669
 ------------------------------------------
```

Figure 38b : Seconde analyse pour les effets simples

11. Homogénéité de la variance et transformations non linéaires

11.1. Quelques commandes SPSS

a. Pour vérifier l'homogénéité des variances dans une ANOVA.

La méthode est très simple: il s'agit de rajouter l'option:

```
o     /print=homogeneity
```

dans la commande `manova` avant l'option `design` ou `wsdesign`.

b. Pour effectuer une transformation

La commande `compute` permet de modifier la valeur contenue dans une colonne ou de rajouter une nouvelle colonne dont la valeur résulte d'un calcul. Dans les deux cas, le calcul est basé sur les colonnes déjà existantes.

```
o   compute nomcol = expression-mathematique.
```

Dans *expression-mathematique*, on peut utiliser les noms des autres colonnes, l'addition (+), la multiplication (*), etc. Par ailleurs, il existe des opérations plus avancées, tel sqrt( ), ln( ), etc, qui calcule la racine carrée (square root), le logarithme de ce qui se trouve entre les parenthèses.

Par exemple, supposons qu'il existe une colonne nommée `score`, on peut faire:

```
o   comment transformation par la racine carrée (square root).
o   compute score2=sqrt(score+0.5).
o   comment transformation logarithmique.
o   compute score2=ln(score+1).
o   comment transformation angulaire.
o   compute score2=2 * arsin(sqrt(score)).
```

qui va dans tout les cas créer une nouvelle colonne `score2`.

11.2. Exemple complet

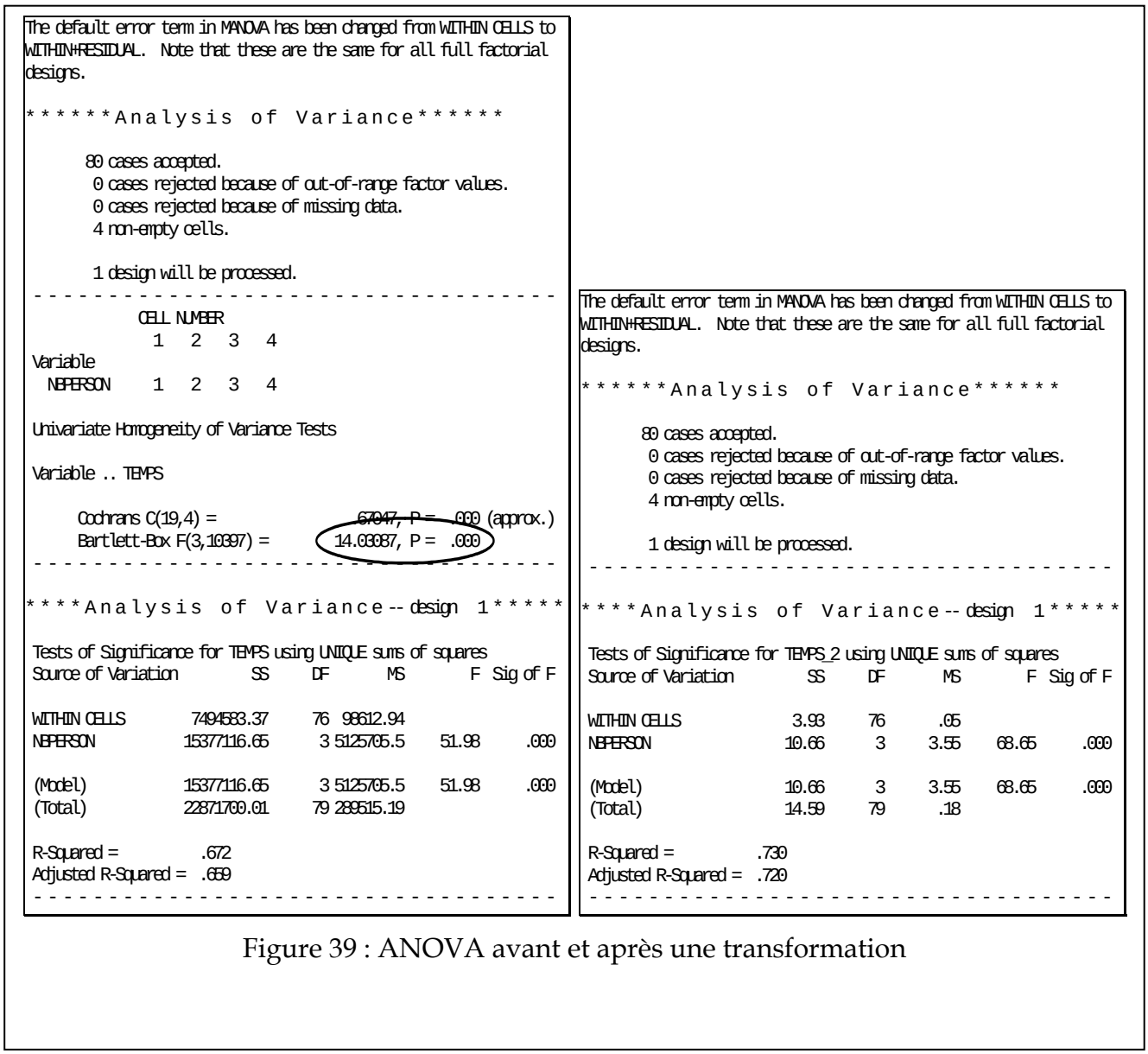
Nous démontrons l'exemple des acétates sur "Où est Charlie".

La syntaxe est:

```
o   data list file="c:\windows\bureau\charlie.dat" list
o     /nbperson temps.
o
o   comment pour avoir 4 niveaux allant de 1 à 4, je divise.
o   comment le nombre de personnage par 10.
o   compute nbperson=nbperson/10.
o   execute.
o
o   manova temps by nbperson(1,4)
o     /print=homogeneity
o     /error=within
o     /design.
o
o   compute temps2 = LN(temps+1).
o
o   manova temps2 by nbperson(1,4)
o     /error=within
o     /design.
```

Dans la première commande, nous vérifions l'homogénéité des variances avec l'option `print`. Comme les variances diffèrent significativement, nous procédons à une transformation logarithmique, puis refaisons l'analyse.

La Figure 39 présentent les résultats avant transformation et après transformation.



```
The default error term in MANOVA has been changed from WITHIN CELLS to
WITHIN+RESIDUAL.  Note that these are the same for all full factorial
designs.

******Analysis of Variance******

      80 cases accepted.
       0 cases rejected because of out-of-range factor values.
       0 cases rejected because of missing data.
       4 non-empty cells.

       1 design will be processed.
--------------------------------------------------------------------------

          CELL NUMBER
            1   2   3   4
Variable
  NBPERSON    1   2   3   4

Univariate Homogeneity of Variance Tests

Variable .. TEMPS

     Cochrans C(19,4) =              .67047, P =  .000 (approx.)
     Bartlett-Box F(3,10397) =     14.03087, P =  .000
--------------------------------------------------------------------------

****Analysis of Variance -- design 1*****

Tests of Significance for TEMPS using UNIQUE sums of squares
Source of Variation          SS     DF        MS        F  Sig of F

WITHIN CELLS          7494583.37     76  98612.94
NBPERSON             15377116.65      3 5125705.5    51.98      .000

(Model)              15377116.65      3 5125705.5    51.98      .000
(Total)              22871700.01     79 289515.19

R-Squared =             .672
Adjusted R-Squared =    .659
--------------------------------------------------------------------------
```

```
The default error term in MANOVA has been changed from WITHIN CELLS to
WITHIN+RESIDUAL.  Note that these are the same for all full factorial
designs.

******Analysis of Variance******

      80 cases accepted.
       0 cases rejected because of out-of-range factor values.
       0 cases rejected because of missing data.
       4 non-empty cells.

       1 design will be processed.
--------------------------------------------------------------------------

****Analysis of Variance -- design 1*****

Tests of Significance for TEMPS_2 using UNIQUE sums of squares
Source of Variation          SS     DF        MS        F  Sig of F

WITHIN CELLS              3.93     76       .05
NBPERSON                 10.66      3      3.55    68.65      .000

(Model)                  10.66      3      3.55    68.65      .000
(Total)                  14.59     79       .18

R-Squared =             .730
Adjusted R-Squared =    .720
--------------------------------------------------------------------------
```

Figure 39 : ANOVA avant et après une transformation

12. Corrélation et régression

12.1. Quelques commandes SPSS

Une analyse de la pente de régression permet de tester s'il existe une relation entre deux mesures. Une des deux mesures peut-être une variable indépendante, telle le sexe de l'individu ou un facteur manipulé par le chercheur ou toutes deux peuvent être des variables dépendantes, tel le temps moyen pour faire une tâche et le pourcentage d'erreurs dans cette même tâche. L'hypothèse testée est de la forme:

$$\mathbf{H_0: r_{xy} = 0}$$

$$\mathbf{H_1: r_{xy} \neq 0}$$

Si la corrélation est non nulle (rejet de H_0), le chercheur a alors le loisir de poursuivre l'analyse avec l'hypothèse subalterne:

$$\mathbf{H_0: b_{xy} = b_0}$$

$$\mathbf{H_1: b_{xy} \neq b_0}$$

où b_0 est une valeur choisie à priori par ses hypothèses de recherche. Dans SPSS, la syntaxe pour vérifier le premier test est:

```
o   regression
o     /statistics=coeff ci r
o     /dependent nomcol1
o     /method=enter nomcol2.
```

où *nomcol2* est la variable qui ira sur l'axe vertical, et *nomcol2* est la variable sur l'axe des horizontal. Cette commande permet de tester des régressions simples si vous nommez deux variables seulement et des régressions multiples si vous mettez plusieurs *nomcol* à l'option `dépendent`. Ici, `ci` signifie *Confidence interval* (affiche les deux dernières colonnes des coefficients dans le listing).

12.2. Exemple complet

Dans l'exemple de la section précédente, nous avons vu que pour détecter la présence de Charlie, le temps de détection est le plus lent quand le nombre de personnages est plus grand. Une hypothèse simple consiste à dire que le participant doit déplacer son attention d'un personnage à l'autre jusqu'à ce qu'il puisse identifier Charlie. Selon cette hypothèse, le temps pour trouver Charlie devrait croître avec le nombre de personnages et donc, on devrait obtenir une relation linéaire entre ces deux variables. Le graphe de la Figure 40 tend à confirmer notre hypothèse, et les analyses qui suivent rejettent l'hypothèse nulle qu'il n'y a pas de relation entre le nombre de personnages et le temps pour trouver Charlie.

Les résultats montrent qu'il existe une relation significative entre le nombre de personnages et le temps pour trouver Charlie ($t(78) = 12.58$, $p < .05$). Le coefficient de corrélation est de .818, soit assez important. De plus, on voit que le temps s'accroît de 39.13

ms par personnage présent sur l'affichage, une valeur qui ne diffère pas de 40 ms par personnage, $t(78) = 0.28$, $p < .05$ [le second test doit être fait manuellement]). Nous concluons que l'attention peut se porter d'un objet à un autre en un temps incroyablement court de 40 ms.

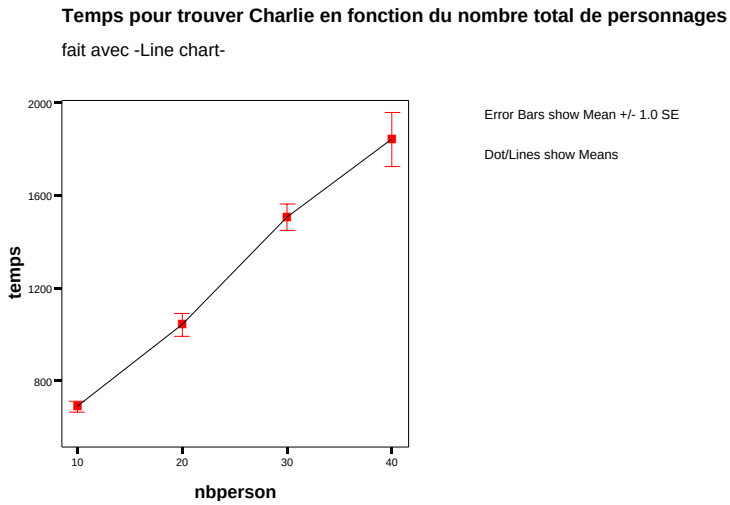


Figure 40 : Moyenne des temps de réponses

Pour vérifier la première hypothèse, nous faisons une analyse de régression. La syntaxe est:

```
o   data list file="c:\windows\bureau\charlie.dat" list
o     /nbperson temps.
o   execute.
o
o   regression
o     /statistics=coeff ci r
o     /dependent temps
o     /method=enter nbperson.
```

Les résultats sont présentés à la Figure 41.

Regression

Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	NBPERSON	.	Enter

a All requested variables entered.

b Dependent Variable: TEMPS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.818	.670	.665	311.207

a Predictors: (Constant), NBPERSON

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
Model		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	292.129	85.228		3.428	001	122.454	461.804
	NBPERSON	39.137	3.112	.818	12.576	.000	32.942	45.333

a Dependent Variable: TEMPS

Figure 41 : Résultat de la régression