Université de Saida Dr Moulay Tahar
Faculté des sciences économiques
Commerciales et de gestion

Polycopié de cours

Initiation à l'utilisation du logiciel

IBM SPSS Statistics

Destiné aux étudiants de 3ème année management

Année universitaire 2021-2022

Mourad MADOUNI

Maître de conférences classe B

Table des matières

Troisième chapitre : Les exercices d'application

Table des figures

Introduction

Durant cette révolution numérique moderne, on a pu constater l'émergence d'un nombre important de logiciel de traitement des données. Désormais, ces logiciels nous permettent, d'une manière facile et très agréable, le traitement des données de masse (big data) en un temps record. Cette grande diversité se manifeste par des logiciels payants et libres, tel que IBM SPSS Statistics, PSPP, R, Stata, SAS, Minitab, Ms Excel, Google sheet, Statistica, Epi info, XLSTAT, Statgraphics, le Sphinx, etc.

En outre, les multiples livres et surtout les pages web des grandes universités anglosaxonnes (Kent State University, Sheffield University, Sherbrooke University, University of Pennsylvania, etc.) offrent des tutoriels détaillés pour l'apprentissage de ces logiciels.

Néanmoins, tous ces avantages qui nous sont offerts par ces logiciels resterons toujours conditionnés par la maîtrise de la partie théorique en l'occurrence les bases de la statistique en plus de la qualité des bases de données.

Ce polycopié de cours « initiation à l'analyse de données avec IBM SPSS Statistics » est destiné aux étudiants de troisième année licence et aux étudiants de master en phase de préparation de projet de fin de cycle. Ce modeste travail est scindé en trois parties. La première partie vise à réviser et consolider tous les fondements théoriques des statistiques descriptives et la deuxième partie présente toute les fonctionnalités correspondantes du logiciel IBM SPSS Statistics. La dernière partie sera consacrée à des exercices d'application.

Dans un souci pédagogique, on n'a pas pu traiter toutes les fonctionnalités du logiciel car ce polycopié sera suivi par deux autres travaux qui seront consacrés aux statistiques inférentielles et aux statistiques avancées (corrélation, régression linéaire et régression logistique).

Enfin, vous pouvez envoyer un mail à l'adresse suivante mourad.madouni@univ-saida.dz pour vous communiquer toutes les bases de données traitées dans les exemples de ce polycopié et je suis toujours dans l'attente de vos remarques et commentaires pour améliorer ce travail.

Mourad MADOUNI.

Premier chapitre

Rappel sur les statistiques descriptives

I. Les définitions de base

1. Définition de la statistique

La **statistique** est une branche des mathématiques qui a pour objet l'analyse et l'interprétation de données quantifiables (Dress, 2007).

La statistique est l'étude de la collecte des données (observations), leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous (Goldfarb & Pardoux, 2011). C'est à la fois une science, une méthode et un ensemble de techniques. On en distingue souvent deux sous-branches (Dodge, 2007) :

- ✓ La **statistique descriptive**, dont le but est la description des données exhaustives à l'aide de moyens appropriés, qu'ils correspondent à des valeurs calculées (moyenne, médiane, écart-type, quartile, etc.) ou à des représentations graphiques (histogramme, secteur, etc.). Par exemple, l'étude de la distribution des salaires dans une entreprise rentre dans ce cadre.

- ✓ La **statistique inférentielle**, dont le but est d'effectuer des estimations et des prévisions à partir d'un sous-ensemble de la population. Elle est complémentaire à la statistique descriptive. Par exemple : l'étude statistique de l'efficacité d'un médicament.
  Statistique inférentielle = statistique déductive

2. Les domaines de l'utilisation de la Statistique

Les statistiques sont aujourd'hui utilisées dans tous les secteurs d'activité. Partout où l'on dispose de données.

- Industrie : fiabilité, contrôle qualité, etc.
- Économie et finance : sondages, enquête d'opinion, assurance, marketing.
- Santé, environnement, biologie, agronomie, etc.
- Architecture, génie civil.

3. Définition de la population

Ensemble d'unités statistiques de même nature sur lequel on recherche des informations quantifiables (Dodge, 2007). Il s'agit de l'univers de référence lors de l'étude d'un problème statistique donné.

Exemples :

- ✓ Personnes d'un pays (unité statistique = personne).
- ✓ Ensemble de la production d'une usine (unité statistique = produit).
- ✓ Ensemble des prix d'articles de consommation (unité statistique = prix).

4. Définition de l'échantillon

Sous-ensemble d'une population sur lequel on effectue une étude statistique. Son étude vise généralement à tirer des conclusions relatives à la population dont il est issu (Dodge, 2007).

Exemples :

Fraction de la population d'un pays (1000 personnes / 45 millions)

5. Définition de l'échantillonnage

Ensemble des opérations destinées à former l'échantillon ou bien la sélection d'une partie de la population pour étudier certaines caractéristiques de l'échantillon (Dodge, 2007).

6. Définition d'un échantillon représentatif

Un échantillon est dit représentatif s'il possède les mêmes caractéristiques de la population que l'on souhaite étudier. Cette représentativité doit surtout se faire sur les caractéristiques pouvant influencer les réponses. Faute de représentativité, les résultats obtenus sur un échantillon ne peuvent être généralisés à la population étudiée.

7. Variable

7.1. Définition

Caractéristique mesurable à laquelle on peut attribuer plusieurs valeurs différentes (Veysseyre, 2014).

7.2. Les types de variables

7.2.1. Variable quantitative

Modalités = valeurs numériques

- ✓ **Continue** (peut prendre toutes les valeurs d'un intervalle) : poids, revenu, longueur, âge, temps, etc.
- ✓ **Discontinue ou discrète** : (ne peut prendre que des valeurs isolées) : nombre d'enfants dans une famille.

- Variables continues

On dit qu'une variable est continue si elle peut supposer un nombre infini de valeurs réelles. La distance, l'âge et la température sont des exemples d'une **variable continue**.

***Nota* :** Pour en faciliter la manipulation, on groupe habituellement les variables continues à l'aide d'« intervalles de classe ».

- Variables discrètes

Contrairement à une variable continue, une **variable discrète** ne peut revêtir qu'un nombre défini de valeurs réelles. Le nombre d'enfants par classe est un exemple. On peut aussi grouper des variables discrètes. Encore une fois, on groupe des variables discrètes pour en faciliter la manipulation.

7.2.2. Variable qualitative

Modalités = grandeurs non quantifiables

- ✓ **Dichotomique** : valeurs binaires (vrai/faux, oui/non, Homme/Femme)
- ✓ **Catégorielle** : les modalités sont des catégories (couleurs, genre, etc.)
  - **Nominale** : rouge, orange, vert, bleu, indigo, violet
  - **Ordinale** : faible, moyen, fort

- **Variables nominales**

Les variables nominales présentent des catégories que l'on nomme avec un nom. Par exemple : homme ou femme, le nom de la voiture, une couleur.

- **Variables ordinales**

Les variables ordinales sont des catégories qui sont naturellement ordonnées. Ça peut être le classement à une course, par exemple ou le résultat à un questionnaire sur une échelle de Likert

1 : pas du tout d'accord,

2 …

5 : Tout à fait d'accord.

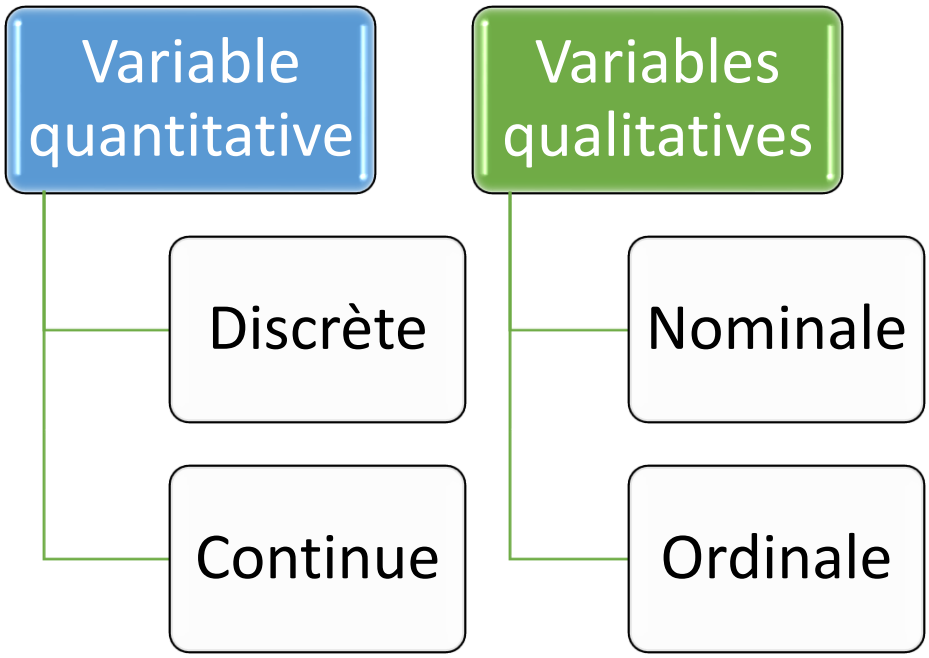


Figure 1 Les types de variable

II. Séries statistiques associées à un caractère quantitatif discret

1. Paramètres de position

1.1. Moyenne[1]

On appelle moyenne (Dodge, 2007) d'une série statistique d'effectif total N, le réel $\bar{X}$

$$\bar{X} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{N}$$

xi : Valeurs du caractère (ex : Age= 20 ans).
ni : Effectifs (nombre de personnes ayant l'âge 20 ans).
N : Effectif total.

A la place des effectifs (ni), on peut aussi utiliser les fréquences :

$$\bar{X} = f_1 x_1 + f_2 x_2 + \cdots + f_k x_k$$

Fréquences :

$$f_i = \frac{n_i}{N}$$

Fréquences en pourcentages :

$$f_i = \frac{n_i}{N} \times 100$$

Exemple : Les notes sur 20 obtenues lors d'un devoir de statistique dans une classe sont les suivantes :

10, 8, 11, 9, 12, 10, 8, 10, 7, 9, 10, 11, 12, 10, 8, 9, 10, 9, 10, 11.

La population étudiée est la classe et les individus sont les élèves. L'effectif total est égal à 20 et la note obtenue au devoir est le caractère **discret** que l'on étudie.

La série statistique définie par les effectifs est la suivante :

Valeurs du caractère (notes) xi	7	8	9	10	11	12
Effectifs (nombre d'élèves ayant la note) ni	1	3	4	7	3	2

La série statistique définie par les fréquences en pourcentage est la suivante :

Valeurs du caractère (notes) xi	7	8	9	10	11	12
Fréquences en %	5%	15%	20%	35%	15%	10%

[1] Pour calculer la moyenne de 10 valeurs sur MS Excel ou Google Sheet, la fonction s'écrit comme suit :
=moyenne(A1:A10)
=average(A1 :A10)

Moyenne :

$$\bar{X} = \frac{1 \times 7 + 3 \times 8 + 4 \times 9 + 7 \times 10 + 3 \times 11 + 2 \times 12}{20} = 9{,}7$$

1.2. Médiane

Définition

L'idée générale est que la médiane[2] est une valeur du caractère qui partage la population en deux parties de même effectif (Dress, 2007) (Dodge, 2007). De façon plus précise, on appelle médiane d'une série statistique discrète toute valeur M du caractère telle qu'au moins 50% des individus aient une valeur du caractère inférieure ou égale à M et au moins 50% des individus aient une valeur du caractère supérieure ou égale à M.

- ✓ Recherche pratique de la médiane :

On range les valeurs du caractère une par une dans l'ordre croissant (chaque valeur du caractère doit apparaître un nombre de fois égal à l'effectif correspondant).

- ✓ Si l'effectif total est impair, la médiane M est la valeur du caractère située au milieu.
- ✓ Si l'effectif total est pair, la médiane M est la demi-somme des 2 valeurs situées au milieu.

Exemple 1 :
On considère la série statistique suivante :

Valeurs du caractère xi	7	8	9	10	11	14	16
Effectifs ni	2	1	1	1	2	1	2

La liste des valeurs du caractère :

7 ; 7 ; 8 ; 9 ; 10 ; 11 ; 11 ; 14 ; 16 ; 16

L'effectif total est pair : la médiane M est la demi- somme des 2 valeurs situées au milieu. D'où:

$$Médiane = \frac{10 + 11}{2}$$

Exemple 2 :
On considère la série statistique suivante :

Valeurs du caractère xi	6	8	9	12	13	17
Effectifs ni	3	1	2	1	3	3

La liste des valeurs du caractère :

6 ; 6 ; 6 ; 8 ; 9 ; 9 ; 12 ; 13 ; 13 ; 13 ; 17 ; 17 ; 17.

L'effectif total est impair : la médiane M est la valeur située au milieu. D'où, M = 12.

[2] Concernant la fonction de la médiane sur MS Excel et Google Sheet est comme suit :
=mediane(A1 :A10)

1.3. Mode

Soit X une variable quantitative discrète, on appelle **mode**[3] la valeur du caractère qui possède le plus grand effectif (Dodge, 2007) (Dress, 2007).
Exemple : Dans le tableau suivant, représentant le nombre d'étudiants par classe, le mode est '12 étudiants'.

Valeurs du caractère xi	6	8	9	12	13	17
Effectifs ni	3	1	2	7	3	3

2. Paramètres de dispersion

Ces paramètres permettent de mesurer la façon dont les valeurs du caractère sont réparties autour de la moyenne et de la médiane (Goldfarb & Pardoux, 2011).

2.1. Étendue d'une série statistique

Pour une série statistique donnée, nous pouvons calculer l'étendue « e » de la série. L'étendue vaut :

$$E = Max - Min$$

Où Max et Min sont deux valeurs extrêmes de la série : le Max est la plus grande valeur et le Min est la plus petite valeur. Cependant, l'étendue ne nous donne pas d'indication sur comment sont réparties les valeurs entre ces deux valeurs extrêmes.

Pour avoir une idée un peu plus précise de la dispersion des valeurs, on partage la série en quatre parties de même effectif. On définit ainsi les quartiles.

2.2. Quartiles

L'idée générale est de partager la population en quatre parties de même effectif.
Étant donné une série statistique de médiane M dont la liste des valeurs est rangée dans l'ordre croissant (il s'agit de la même liste que celle qu'on utilise pour déterminer la médiane) :

- Le premier quartile[4] Q_1 est la plus petite valeur du caractère pour laquelle 25% des valeurs de la série statistique lui sont inférieures ou égales.

- Le troisième quartile Q_3 est la plus petite valeur du caractère pour laquelle 75% des valeurs de la série statistique lui sont inférieures ou égales.

] Q1;Q3 [est appelé intervalle interquartile.

[3] La fonction pour calculer le mode sur MS Excel et Google Sheet est comme suit :
=mode(A1 :A10)
[4] Il faut noter qu'il existe deux méthodes pour le calcul des quartiles. La première méthode dite *exclusive* consiste à exclure la valeur de la médiane pour déterminer la valeur du Q1 tandis que la méthode inclusive prend en considération la valeur de la médiane. On peut trouver ces deux formules sur MS Excel et Google Sheet.

Remarque :

Centiles

Les centiles C_1, C_2, ..., C_{99} divisent une série statistique en 100 parties d'effectifs égaux. Ce sont les abscisses respectives des points d'ordonnée 0.01 ; 0.02 ; ... ; 0.99 sur la courbe cumulative croissante.

Par exemple le centile C_{98} est une valeur dépassée par 2 % des observations ; les centiles n'ont de sens que si on dispose d'un grand nombre (plusieurs centaines) d'observations.

Déciles

Les déciles D_1, D_2, ..., D_9 divisent une série statistique en 10 parties d'effectifs égaux. Ce sont les abscisses respectives des points d'ordonnée 0.1 ; 0.2 ; ... ; 0.9 sur la courbe cumulative croissante.

2.3. L'écart interquartile

L'écart interquartile est la taille de l'intervalle situé au centre de la série et incluant 50% des observations :

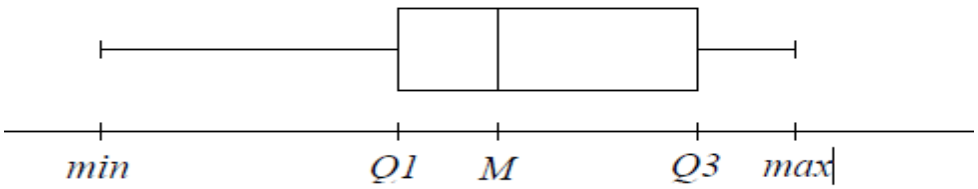
$$Ecart\ interquartile = Q_3 - Q_1$$

Plus cet écart est grand, plus la dispersion des observations est forte.

2.4. Diagramme en boite (boite à moustache ou Boxplots[5])

Définition

Le diagramme en boîtes d'une série statistique se construit alors de la façon suivante : (Les valeurs du caractère sont en abscisse - min et max représentent les valeurs minimales et maximales du caractère).



Interprétation :
25% de la population admet une valeur du caractère entre min et Q1.
25% de la population admet une valeur du caractère entre Q1 et M.
25% de la population admet une valeur du caractère entre M et Q3.
25% de la population admet une valeur du caractère entre Q3 et max.

[5] La boite à moustaches aussi appelée diagramme en boite, boite de Tukey ou box-and-whisker plot, plus simplement box plot en anglais.

Exemple 1 :

Valeurs du caractère xi	7	8	9	10	11	14	16
Effectifs ni	2	1	1	1	2	1	2

La liste des valeurs du caractère :

$$7 ; 7 ; 8 ; 9 ; 10 ; 11 ; 11 ; 14 ; 16 ; 16$$

L'effectif de chaque sous-série est impair : Q1 = 8 (valeur située au milieu de la sous-série inférieure) et Q3 = 14 (valeur située au milieu de la sous-série supérieure).

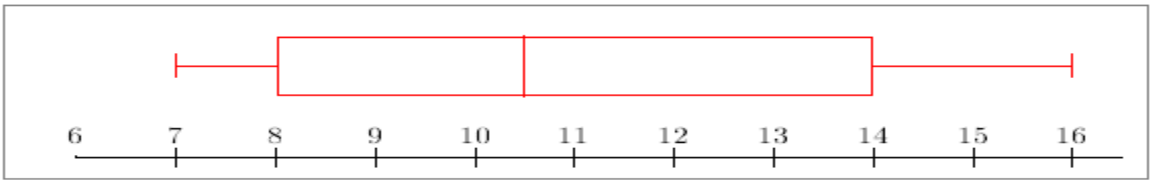


Figure 2 Boîte à moustaches du premier exemple

Exemple 2 :
La liste des valeurs du caractère :

Valeurs du caractère xi	6	8	9	12	13	17
Effectifs ni	3	1	2	1	3	3

$$6 ; 6 ; 6 ; 8 ; 9 ; 9 ; 12 ; 13 ; 13 ; 13 ; 17 ; 17 ; 17.$$

L'effectif de chaque sous-série est pair : Q1 = 7 (demi-somme des deux valeurs situées au milieu de la sous-série inférieure) et Q3 = 15 (demi-somme des deux valeurs situées au milieu de la sous-série supérieure).

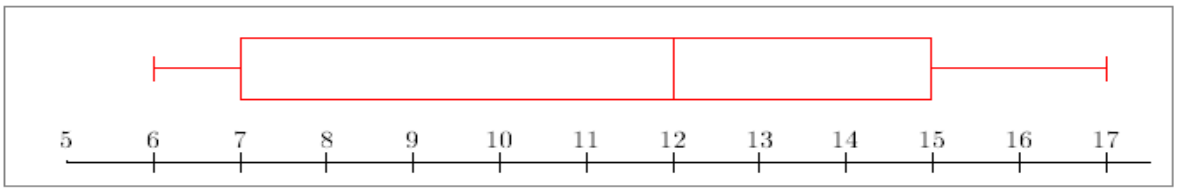


Figure 3 Boîte à moustaches du second exemple

2.5. Variance

La variance[6] mesure la distance des réalisations de la variable par rapport à la moyenne. La variance est définie comme un moment d'ordre 2.

Soit la série statistique définie dans le tableau suivant :

[6] Le concept de la variance et de l'écart type ont été développé par le statisticien britannique Sir Ronald Aylmer Fisher.

Valeurs	x_1	x_2	x_3	…	x_P
Effectifs	n_1	n_2	n_3	…	n_p
Fréquences	f_1	f_2	f_3	…	f_P

Soit $\bar{x}$ la moyenne de cette série.

Le réel $V = \frac{1}{N}[n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + n_3(x_3 - \bar{x})^2 + \cdots + n_P(x_P - \bar{x})^2]$ est appelé **variance** de cette série statistique.

La racine carrée de la variance $\sigma = \sqrt{V}$ est l'écart-type de cette série.

2.6. Écart-type

L'écart type est la racine carrée de la variance.

$$\sigma = \sqrt{Var(x)}$$

La **variance** et **l'écart type** permettent de mesurer la **« dispersion »** des valeurs de la série autour de la moyenne. L'écart-type est un indicateur de dispersion. Il nous informe sur la manière dont les individus se répartissent autour de la moyenne. Sont-ils tous à peu près identiques, concentrés autour de la moyenne ? ou au contraire, sont-ils dispersés entre des valeurs très basses et des valeurs très hautes ?

Interprétation : L'écart-type est l'écart moyen à la moyenne pour tous les individus. Si celui-ci est faible, les individus formulent des réponses similaires, si celui-ci est fort les variations sont fortes dans la population étudiée.

Nota : Le niveau qui permet de repérer un fort écart-type est 1/2 moyenne. Si l'écart-type est supérieur à 0,5 moyenne (0,5 *moyenne), on peut donc considérer que les variations sont fortes. Si les valeurs de la série possèdent une unité de mesure par exemple l'âge (années), l'écart type s'exprime dans la même unité à savoir (années). Par contre la variance n'a pas d'unité de mesure.

Exercice 1 :
Pour illustrer nos propos, prenons un exemple simple. 21 étudiants reçoivent une note sur 20 en statistique, les voici :

1 - 3 - 3 - 4 - 4 - 6 - 7 - 8 - 9 - 9 - 9 - 9 - 9 - 10 - 10 - 15 - 17 - 18 - 19 - 20 - 20

- Moyenne = 10 (210/21)
- Ecart-type = 5,93.

Dans ce cas, *la dispersion est forte* puisque l'écart-type est supérieur à 5 = (1/2 moyenne).

Exercice 2 :
1°) Soit la série statistique répertoriant la taille en mètres de 100 serpents.

Taille (en mètres)	1,5	2	2,5	3	3,5	4	4,5
Effectifs	8	10	25	32	19	4	2

La taille moyenne est :

$$\bar{X} = \frac{1{,}5 \times 8 + 2 \times 10 + 2{,}5 \times 25 + 3 \times 32 + 3{,}5 \times 19 + 4 \times 4 + 4{,}5 \times 2}{100} = 2{,}82$$

La variance

$$V = \frac{1}{100}[8(1{,}5 - 2{,}82)^2 + 10(2 - 2{,}82)^2 + 25(2{,}5 - 2{,}82)^2 + \cdots + 2(4{,}5 - 2{,}82)^2]$$

- Moyenne = 2,82
- Écart-type = 0,665 mètres.

Dans ce cas, *la dispersion est faible* puisque l'écart-type est inférieur à 1,41 = (1/2 moyenne).

3. Les paramètres de forme (Les coefficients de Fisher)

Les coefficients de Fisher regroupent deux coefficients : *le coefficient d'asymétrie* et le *coefficient d'aplatissement*. Ces deux mesures sont des caractéristiques de forme et s'appliquent sur des variables quantitatives mesurées sur une échelle d'intervalles ou de rapports. Ils permettent de rendre compte de manière chiffrée de certains aspects de la forme graphique que peuvent prendre une distribution et sont tous deux sans dimension.

Rappel :

Loi de probabilité aussi appelée *loi de Laplace-Gauss*, la loi *normale* (Thierry Ancelle, 2015b), ainsi dénommée par Pearson, au sens de *naturelle*, intervient dans l'étude de phénomènes quantitatifs aléatoires *continus* soumis à de multiples causes agissant additivement et indépendamment l'une de l'autre et dont la répartition des valeurs s'étale autour de leur moyenne.

Si X est la variable aléatoire soumise à une telle loi, on recherche la probabilité que X prenne ses valeurs dans un intervalle donné.

3.1. Coefficient d'asymétrie ou Skewness

Le coefficient d'asymétrie de Fisher est le moment centré d'ordre 3 (Thierry Ancelle, 2017b) (zedstatistics, 2019a).

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Le coefficient d'asymétrie de Fisher, noté ou *S*, se définit comme étant le rapport entre le moment centré d'ordre 3 (m_3) et le cube de l'écart-type (s^3) :

$$g_1 = \frac{m_3}{S^3} \qquad S = \frac{\mu_3}{\sigma^3}$$

Ou on peut écrire :
Le coefficient d'asymétrie de Fisher est dit également « skewness » ou « gamma un ».

Interprétation :

- ✓ Lorsque la distribution est symétrique, le coefficient de Skewness est nul.
- ✓ Lorsque la distribution possède une forte queue vers la droite (étalement à droite), le coefficient de Skewness est positif (les + l'emportent) donc **l'asymétrie est gauche**.
- ✓ Lorsque la distribution possède une forte queue vers la gauche, le coefficient de Skewness est négatif (les - l'emportent) **donc l'asymétrie est droite**.

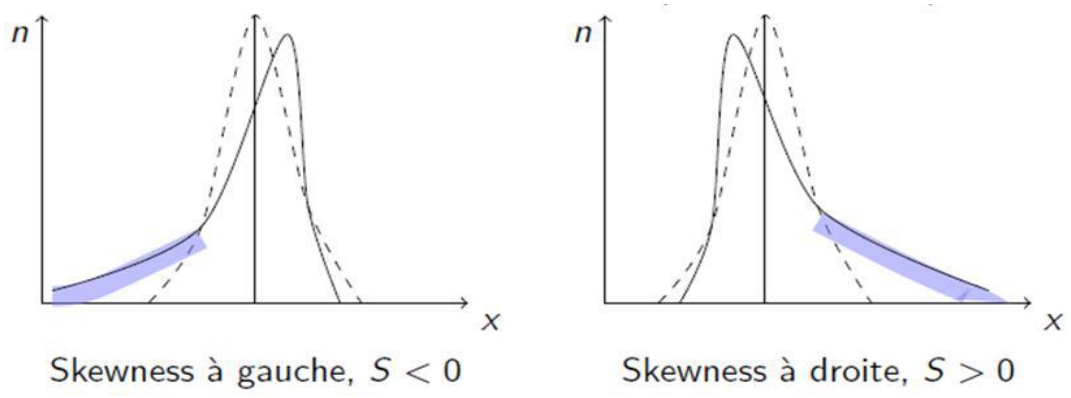


Figure 4 Le coefficient d'asymétrie

Souvent, l'analyse du diagramme en bâtons – ou de l'histogramme permet de se rendre compte du caractère symétrique ou non d'une distribution. L'examen de la boîte à moustaches permet aussi de se faire une idée sur cette question selon que la boîte et les moustaches sont symétriques ou, au contraire, de plus petite amplitude à gauche (asymétrie à gauche) ou à droite (asymétrie à droite).

Exemple :

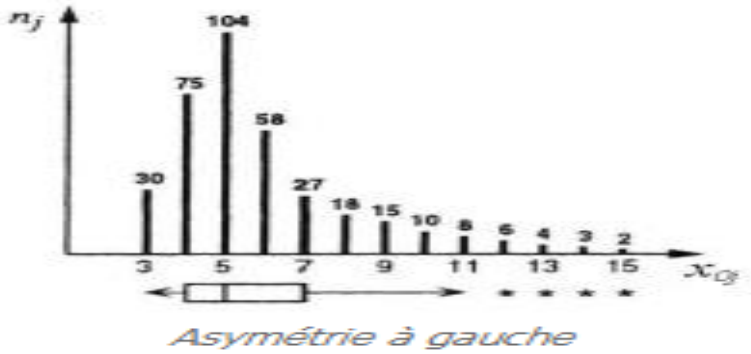


Figure 5 Diagramme en bâtons et une boite à moustaches montrant une asymétrie à gauche

Le diagramme en bâtons et la boîte à moustaches ci-dessous permettent de se rendre compte aisément que la distribution observée présente une asymétrie gauche.

3.2. Le coefficient d'aplatissement de Fisher ou Kurtosis

Le coefficient d'aplatissement ou Kurtosis est le moment centré d'ordre 4 (Thierry Ancelle, 2017a) (zedstatistics, 2019b) (Dodge, 2007).

$$K = \frac{\mu_4}{\sigma^4}$$

Fisher propose d'étudier K' ce qui permet de faire référence à une distribution particulière qui est la loi normale pour laquelle K vaut 3. Les logiciels statistiques vous donnent la valeur de K'.

$$K' = K - 3$$

Le Kurtosis mesure l'aplatissement (ou la planéité) d'une distribution. Il donne une information sur **les QUEUES de distribution.**

En effet, ce coefficient est grand quand il y a beaucoup de valeurs éloignées de la moyenne. Plus la valeur de ce coefficient est grande, plus la distribution est pointue.

Interprétation

- ✓ Si la valeur obtenue est nulle, on dit que la répartition des observations est de type gaussien ou normal (i.e. la courbe des fréquences à la forme d'une cloche ressemblant à celle d'une loi Normale).

- ✓ Dans le cas où le coefficient d'aplatissement est positif, on obtient une courbe moins aplatie que celle de la loi normale donc plus d'observations que dans une distribution gaussienne.

- ✓ A l'inverse s'il est négatif, la représentation de la courbe est plus aplatie qu'une loi normale donc moins d'observations que dans une distribution gaussienne.

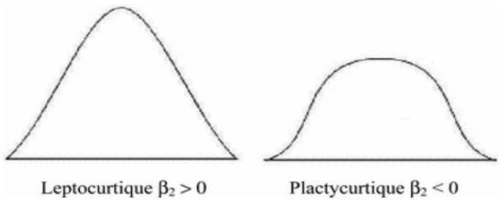


Figure 6 Le coefficient d'aplatissement

4. Représentations graphiques

Le diagramme en bâtons : cette représentation est utilisée lorsque le caractère de l'étude est discret : la longueur du segment ou de la barre est proportionnelle à l'effectif.

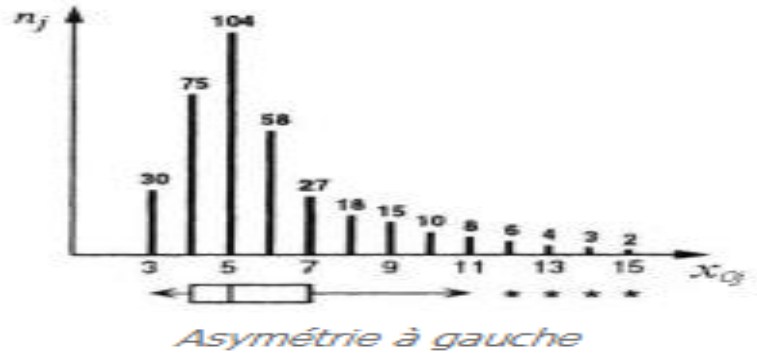


Figure 7 Diagramme en bâtons

III. Séries statistiques associées à un caractère quantitatif continu

La seule différence par rapport aux caractères discrets, c'est que les valeurs du caractère sont regroupées dans des intervalles (appelés classes du caractère).

Tableau 1 Tableau de fréquence pour une variable quantitative continue

Classe numéro i $[b_i; b_{i+1}[$	Centres c_i	Effectifs n_i	Fréquences f_i	Fréquences cumulées croissantes F_i
$[b_1; b_2[$	c_1	n_1	f_1	F_1
$[b_2; b_3[$	c_2	n_2	f_2	F_2
...	…	…	…	…
$[b_p; b_{p+1}[$	c_p	n_p	f_p	F_p

Une classe est un intervalle fermé à gauche et ouvert à droite, du type :

$$[b_i; b_{i+1}[$$

Le centre d'une classe est : $C_i = \frac{b_i + b_{i+1}}{2}$

L'amplitude d'une classe : $a_i = b_{i+1} - b_i$

1. Paramètres de position

1.1. Moyenne

Dans le cas d'une *variable statistique continue*, où les valeurs sont regroupées dans des intervalles, on calcule la moyenne en choisissant le centre des intervalles pour chaque valeur de la variable (Dodge, 2007).

$$\bar{X} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{N}$$

N : Effectif total
x : Centre de la classe
n : Effectif de la classe

Exemple :

Notes	$[0\,;5[$	$[5\,;8[$	$[8\,;12[$	$[12\,;15[$	$[15\,;20[$
Effectifs	10	8	12	11	9
Centre de la classe (X_i)	2,5	6,5	10	13,5	17,5

$$\bar{X} = \frac{2{,}5 \times 10 + 6{,}5 \times 8 + 10 \times 12 + 13{,}5 \times 11 + 17{,}5 \times 9}{50} = 10{,}06$$

1.2. La Classe modale

Soit X une variable quantitative continue, on appelle **classe modale** la classe du caractère qui possède le plus grand effectif.
Exemple : Dans le tableau suivant, représentant les notes d'une classe, la classe modale est la classe [8 ; 12[.

Notes	[0 ; 5[	[5 ; 8[	[8 ; 12[	[12 ; 15[	[15 ; 20[
Effectifs	10	8	12	11	9

1.3. La classe médiane

Les fréquences cumulées (ou les effectifs cumulés) permettent de déterminer dans quelle classe [ai ; ai+1[se situe la médiane. La classe ainsi obtenue est appelée **classe médiane**. Une fois qu'on a la classe médiane, on trouve la médiane par interpolation linéaire.

Pour calculer l'abscisse de ce point, on utilise le théorème de Thalès.

Pour trouver la médiane, on trace :

- Soit la courbe des effectifs cumulés et on doit trouver l'abscisse du point d'ordonnées N/2 de cette courbe (où N est l'effectif total).
- Soit la courbe des fréquences cumulées et on doit trouver l'abscisse du point d'ordonnés 0,5 de cette courbe.

Exemple 1 :

Notes	Effectifs	Effectifs cumulés
[0 ; 5[	10	10
[5 ; 8[	8	18
[8 ; 12[	12	30
[12 ; 15[	11	41
[15 ; 20[	9	50
	50	

Utilisons la colonne des effectifs cumulés pour déterminer la médiane : il y a 50 notes, 50 % de l'effectif total c'est 25, la médiane est ici la note correspondant à l'effectif cumulé 25. La médiane se trouve donc dans la classe médiane [8;12[.

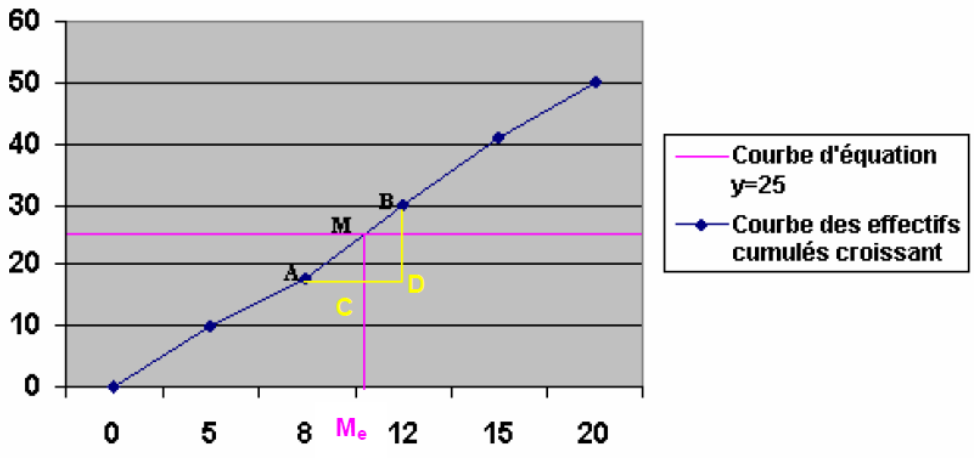


Figure 8 Polygone des effectifs cumulés

Théorème de Thalès
On applique le théorème de Thalès dans le triangle ABD, on a :

$$\frac{AM}{AB} = \frac{AC}{AD} = \frac{MC}{BD}$$

D'où en particulier :

$$\frac{AC}{AD} = \frac{MC}{BD} \Leftrightarrow \frac{M_e - 8}{12 - 8} = \frac{25 - 18}{30 - 18} \Leftrightarrow M_e = 10{,}33$$

Définition de l'amplitude

En statistiques, l'amplitude représente l'écart entre la valeur de début et la valeur de fin de la classe (ou ensemble de données). L'amplitude montre l'étendue d'une série statistique. Si elle est élevée, alors les valeurs de la série sont éloignées les unes des autres, si l'intervalle est faible, alors les valeurs de la série sont très proches les unes des autres.

2. Paramètres de dispersion

2.1. Étendue d'une série statistique

Pour une série statistique donnée, nous pouvons calculer l'étendue E de la série. L'étendue vaut :

$$E = Max - Min$$

Où Max et Min sont deux valeurs extrêmes de la série : Max est la plus grande valeur et Min est la plus petite.

2.2. Quartiles

Le paramètre Q_1 permet de dire que 25 % environ de la population étudiée a une modalité inférieure à la Q_1 et 75 % une modalité supérieure à la Q_1. Le paramètre Q_3 permet de dire que 75 % environ de la population étudiée a une modalité inférieure à la Q_3 et 25 % une modalité supérieure à la Q_3.

Les nombre Q_1, Q_2, Q_3 correspondent aux effectifs cumulés n/4, n/2 et 3n/4.

Les quartiles se calculent alors par **interpolation linéaire.**

Notes	Effectifs	Effectifs cumulés
$[\mathbf{0}\,;\mathbf{5}[$	10	10
$[\mathbf{5}\,;\mathbf{8}[$	8	18
$[\mathbf{8}\,;\mathbf{12}[$	12	30
$[\mathbf{12}\,;\mathbf{15}[$	11	41
$[\mathbf{15}\,;\mathbf{20}[$	9	50
	50	

Exemple :
Les nombre Q_1, Q_2 , Q_3 correspondent aux effectifs cumulés n/4 , n/2 et 3n/4 (soit 12,5 ; 25 et 37,5).

Construisons le polygone des effectifs cumulés croissants :

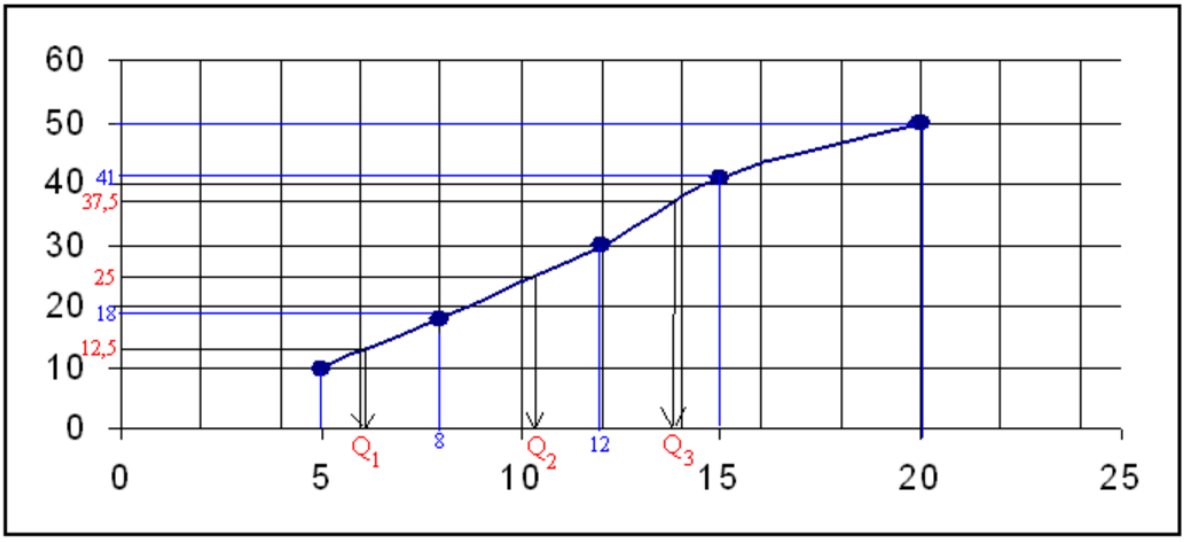


Figure 9 La courbe des effectifs cumulés croissant

Théorème de Thalès

On applique le théorème de Thalès dans le triangle ABD, on a :

$$\frac{Q_1 - 5}{8 - 5} = \frac{12{,}5 - 10}{18 - 10} \Leftrightarrow \frac{Q_1 - 5}{3} = \frac{2{,}5}{8} \Leftrightarrow Q_1 = 5 + 3 \times \frac{2{,}5}{8} \approx 5{,}94$$

$$\frac{Q_2 - 8}{12 - 8} = \frac{25 - 18}{30 - 18} \Leftrightarrow M_e = 10{,}33$$

$$\frac{Q_3 - 12}{15 - 12} = \frac{37{,}5 - 30}{41 - 30} \Leftrightarrow \frac{Q_3 - 12}{3} = \frac{7{,}5}{11} \Leftrightarrow Q_1 = 12 + 3 \times \frac{7{,}5}{11} \approx 14{,}05$$

2.3. L'écart interquartile

L'écart interquartile est la taille de l'intervalle situé au centre de la série et incluant 50% des observations :

$$Ecart = Q_3 - Q_1$$

Plus cet écart est grand, plus la dispersion des observations est forte.

2.4. La boite à moustache

Pour construire la boîte à moustaches, il faut :

- La valeur minimale de la série : 0
- Le premier quartile $Q_1 = 5{,}94$
- Le second quartile ou la médiane $Q_2 = 10{,}33$
- Le troisième quartile $Q_3 = 14{,}05$
- La valeur maximale de la série : 20

2.5. Variance et écart-type

Soit $\bar{x}$ la moyenne de cette série.

Le réel $V = \frac{1}{N}[n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + n_3(x_3 - \bar{x})^2 + \cdots + n_P(x_P - \bar{x})^2]$ est appelé **variance** de cette série statistique. Avec x : centre de la classe

La racine carrée de la variance $\sigma = \sqrt{V}$ est l'écart-type de cette série.

3. Les coefficients de Fisher

Même principe que les variables quantitatives discrètes.

4. Représentation graphique

L'histogramme : représentation utilisée lorsque l'étude porte sur un caractère continu : les rectangles ont pour bases les amplitudes des classes et leurs aires sont proportionnelles aux effectifs (Dodge, 2007) (Thierry Ancelle, 2021).

IV. Série statistique associée à un charactère qualitatif

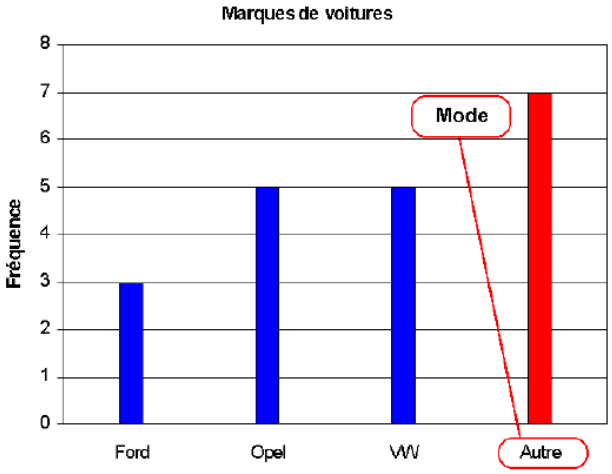
1. Paramètres de position

1.1. Mode

Le mode est le seul paramètre de position qui s'applique à tous les types de variables, qu'elles soient quantitatives ou qualitatives.

A. Variable qualitative nominale

Une étude réalisée par enquêteur posté à la sortie d'un parking d'une grande surface. Les informations concernant les marques de voitures son reprises à la figure 10. La valeur qui présente la plus grande fréquence est la marque « Autre ». Autrement dit, la marque de voiture la plus rencontrée est « Autre » ; « Autre » est le mode du tableau et de l'histogramme.

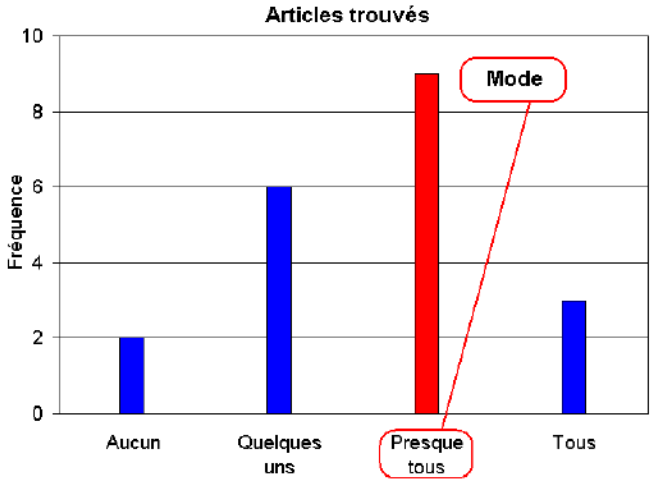


Voiture		
Valeurs de la variable x_i	Fréquence absolue n_i	Fréquence relative f_i
Ford	3	15 %
Opel	5	25 %
Volkswagen	5	25 %
Autre	**7**	**35 %**
Nbre observ.	**20**	**100 %**

Figure 10 Mode d'une variable qualitative nominale

B. Variable qualitative ordinale

Poursuivons avec l'appréciation sur la proportion d'articles que le client souhaitait trouver dans le magasin.



Articles trouvés		
Valeurs de la variable x_i	Fréquence absolue n_i	Fréquence relative f_i
Aucun	2	10 %
Quelques-uns	6	30 %
Presque tous	**9**	**45 %**
Tous	3	15 %
Nbre observ.	**20**	**100 %**

Figure 11 Mode d'une variable qualitative ordinale

La réponse « Presque tous » est celle qui recueille le maximum de suffrages. Cette réponse est le mode de cette étude.

1.2. Médiane

La médiane ne s'applique qu'aux variables qui admettent une relation d'ordre, c'est-à-dire aux variables que l'on peut ordonner ou classer. C'est le cas de toutes les variables quantitatives et qualitatives ordinales.

Articles				
A	A	Q	Q	Q
Q	Q	Q	P	P
P	P	P	P	P
P	P	T	T	T

Figure 12 Articles trouvés

Classons les vingt réponses qui concernent les articles trouvés (de gauche à droite et de haut en bas).

Puisque le nombre d'observations est pair la valeur de la médiane se trouve entre les deux centrales (la $10^{ème}$ et la $11^{ème}$). Nous avons de la chance : il se trouve que ces deux valeurs sont identiques. La réponse médiane est donc « Presque tous ». Si les deux valeurs avaient été différentes, nous aurions formulé notre réponse disant, par exemple, que la médiane se trouve entre « quelques-uns et presque tous ».

1.3. Moyenne

Il est impossible de définir une moyenne puisque les variables qualitatives n'admettent pas l'addition.

2. Paramètres de dispersion

Les paramètres de dispersion caractérisent l'étalement des observations autour d'un paramètre de position. On ne détermine pas de paramètres de dispersion pour les variables nominales.

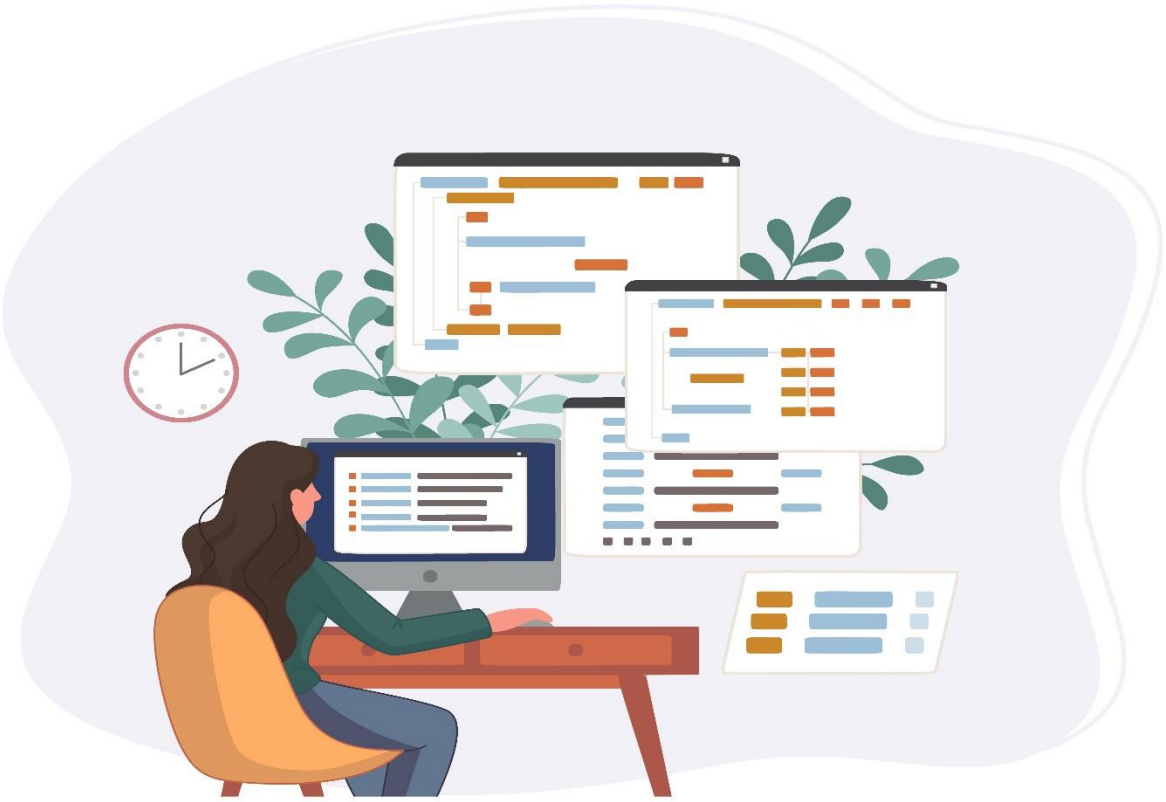
2.1. Étendue ou plage ou amplitude

Comme il a été dit plus haut, il n'est pas possible de calculer la plage d'une variable qualitative. Toutefois dans le cas d'une variable qualitative ordinale, nous pourrons exprimer la plage par une phrase de type : les valeurs se répartissent tous entre la valeur « satisfaisant » et la valeur presque « parfait ».

2.2. L'écart interquartile

La définition implique le calcul d'une différence, ce qui est impossible dans le cas des variables nominales. Toutefois, dans le cas d'une variable qualitative ordinale, nous pourrons exprimer l'écart-interquartile par une phrase de type : « *Les premiers 25% des valeurs observées se*

situent en-deçà de la valeur « satisfaisant » tandis que les derniers 25% se situent en-deçà de la valeur « presque parfait ».



Deuxième chapitre

La description du logiciel

IBM SPSS STATISTICS

I. Introduction

1. Description du logiciel IBM SPSS STATISTICS

IBM SPSS STATISTICS (IBM, 2019) (*Statistical Package for the Social Sciences*, Ensemble des programmes statistiques pour les sciences sociales) est un logiciel commercial pour le traitement et l'analyse statistique de données que ce soit dans le domaine économique et de gestion, biologie, ainsi que le secteur médical.

C'est aussi le nom de la société qui le revend (SPSS Inc), basée à Chicago. En 2009, la compagnie est rachetée par IBM pour 1,24 milliard $. La première version a été lancée en 1968. Actuellement, on est à la version 28 (IBM, 2019) (Wikipedia contributors, 2022). On peut trouver d'autres logiciels de traitement statistiques: Stata, SAS, R, minitab, XLSTAT, le sphinx, statistica, statgraphics, epi info, Nvivo. etc.

La version de IBM SPSS Statistics utilisée dans le cadre de ce polycopié est la version 26.

2. A quoi sert le logiciel IBM SPSS STATISTICS

Il a pour but de :

- Décrire et simplifier les données.
- Analyser les données.
- Organiser les résultats (sous forme de tableaux et de graphiques modernes).
- Présenter les résultats obtenus dans des tableaux et des graphiques modernes.

Par conséquent, il est convivial et surtout flexible.

3. Dans quels domaines IBM SPSS Statistics est utilisé ?

-Industrie : fiabilité, contrôle qualité, etc.
-Économie et finance, marketing, etc.
-Santé, environnement, biologie, etc.
-Génie civil et architecture.
-Partout où l'on dispose de données.

Les données peuvent être des réponses à un questionnaire d'opinion, des mesures provenant d'expériences en laboratoire, de variables socio-économiques extraites de fichiers de renseignements, etc. Les analyses se font à partir des **données saisies.** La qualité des analyses statistiques est fonction de **la qualité des données saisies.**

II. Présentation de l'interface IBM SPSS STATISTICS

Le lancement d'IBM SPSS STATISTICS s'effectue par un simple clic sur l'option **IBM SPSS Statistics 26**, qui vous amène à l'écran principal.

1. L'environnement IBM SPSS STATISTICS

1.1. La fenêtre éditeur de données

La fenêtre Éditeur de données (Kent State University Libraries., 2022b) présente le contenu d'un fichier de données que vous avez préalablement sélectionné. Vous pouvez créer de nouvelles feuilles de données ou modifier des données préexistantes.

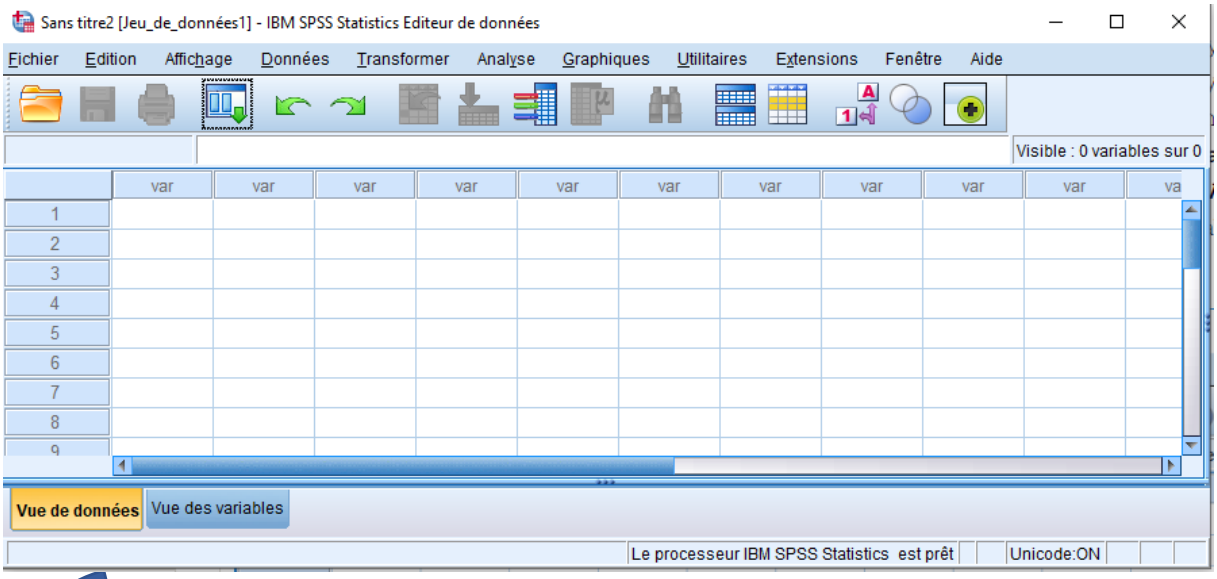


Figure 13 Editeur de données

Cette fenêtre comprend deux onglets :

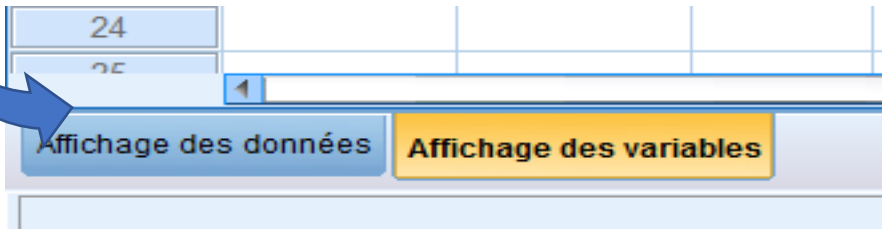


Figure 14 Les onglets de l'éditeur de données

1.1.1. L'onglet affichage des données

Cet onglet permet de voir la banque de données, où les cas sont présentés en lignes et les variables sont en colonne. Chaque cellule présente la valeur que prend une variable pour un cas donné. Le fichier, lorsque mis sous l'option affichage des données, se présente sous la forme suivante :

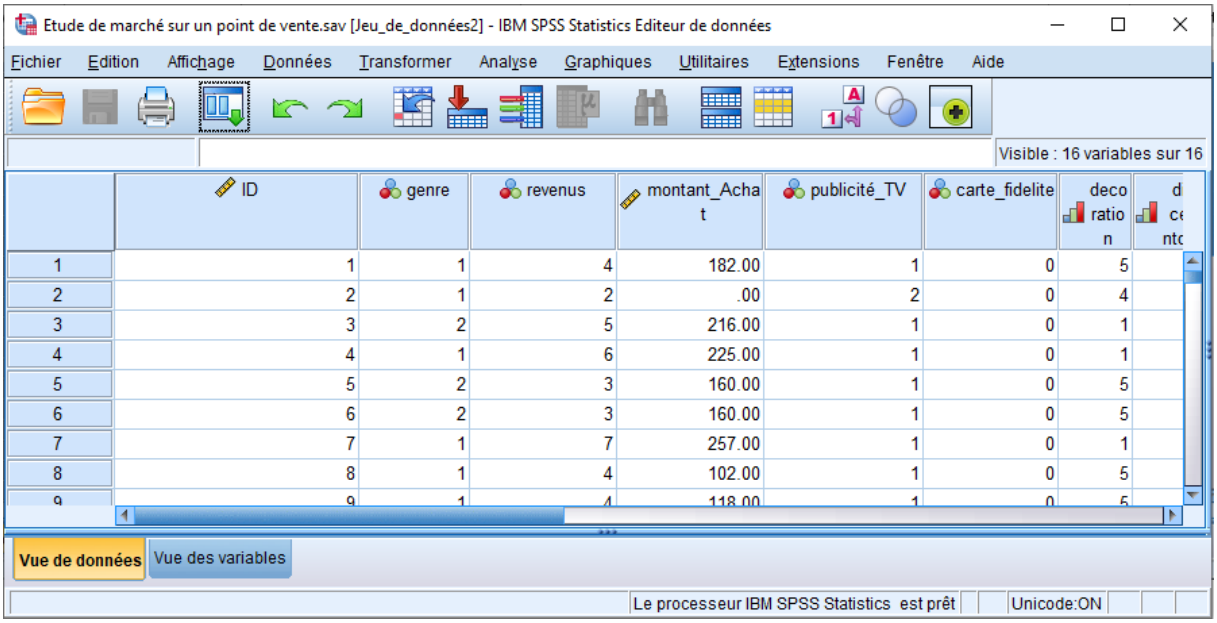


Figure 15 Editeur de données

1.1.2. L'onglet affichage des variables

Permet de voir toutes les variables présentes de la banque de données, leurs noms, ce qu'elles représentent, leurs valeurs manquantes, leurs valeurs possibles, les libellés qui les désignent. Le fichier, lorsque mis sous l'option affichage des variables, se présente sous la forme suivante :

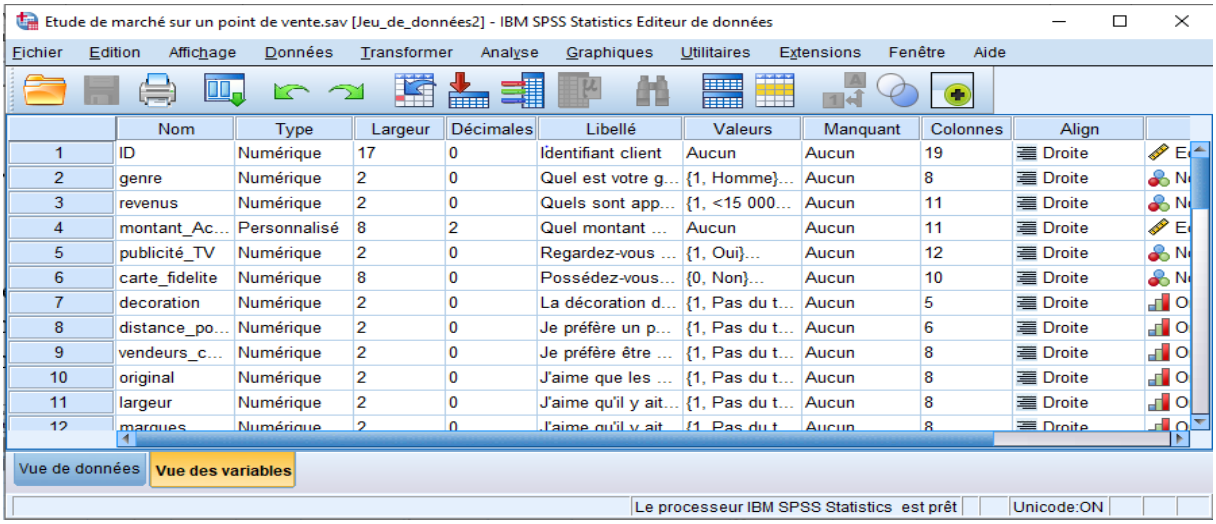


Figure 16 Editeur de variables

1.2. La fenêtre de résultats « sortie »

La fenêtre de résultats Sortie enregistre les résultats des opérations effectuées : tableaux, statistiques et diagrammes obtenus tout au long de votre session de travail. IBM SPSS STATISTICS ouvre automatiquement cette fenêtre et y inscrit l'ensemble des résultats ainsi que le détail des opérations effectuées. La fenêtre de résultats se présente sous la forme suivante :

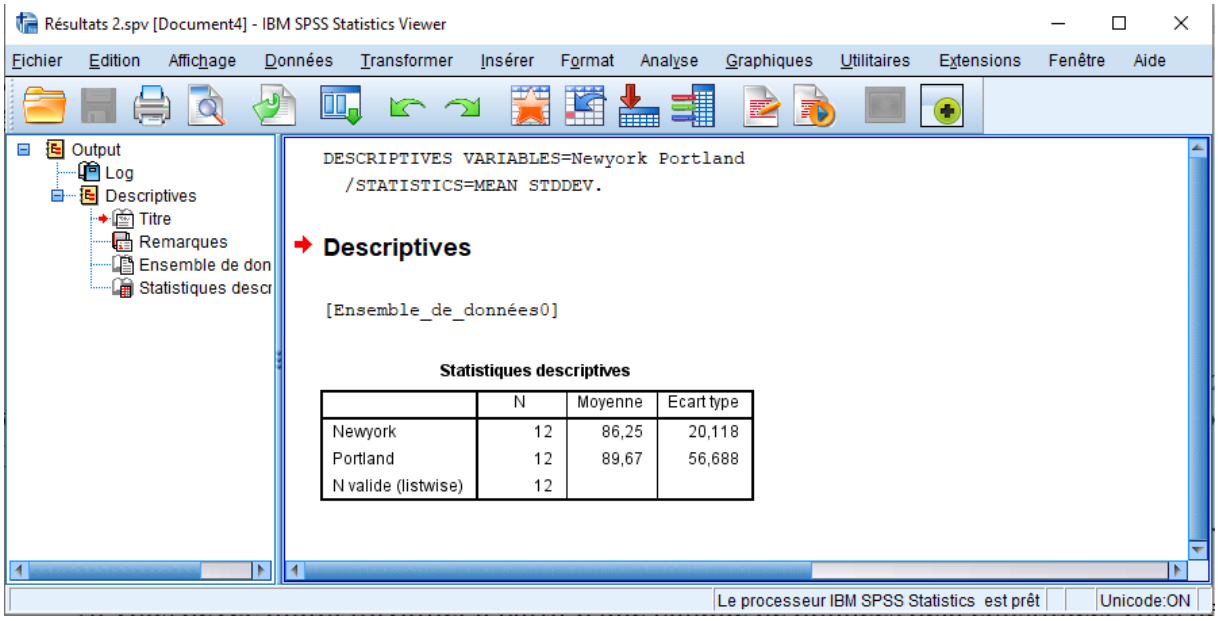


Figure 17 Fenêtre sortie

2. Les fichiers dans IBM SPSS STATISTICS

2.1. Les fichiers de données

Les fichiers de données contiennent les banques de données avec lesquelles vous aurez à travailler. Ces fichiers se présentent sous la forme d'une feuille quadrillée remplie de chiffres ou de lettres répartis en colonnes et en rangées.

-Si vous désirez créer une nouvelle banque de données, il ne vous reste plus qu'à y entrer les informations pertinentes en respectant les normes qui vous auront été communiquées.

-Si vous devez plutôt travailler à partir d'une banque de données déjà constituée, vous devez alors **ouvrir** le fichier en question en suivant la procédure suivante :

Procédure pour ouvrir un fichier de données

Cliquez sur Fichier dans le menu principal entête de IBM SPSS STATISTICS

- Ouvrir => Données (la liste des fichiers (extension **.SAV**) de données lisibles pour IBM SPSS STATISTICS apparait).

2.2. Les fichiers de résultats

Procédure pour ouvrir un fichier de résultat

Cliquez sur Fichier dans le menu principal entête de IBM SPSS STATISTICS

- Ouvrir => Sortie (cliquez sur le nom du fichier choisi (extension **.spv**) qui apparait sous cette rubrique puis cliquez sur => Ouvrir)

2.3. Le fichier syntaxe

Procédure pour ouvrir un fichier de syntaxe sur IBM SPSS Statistics.

Cliquez sur Fichier dans le menu principal entête de IBM SPSS STATISTICS

- Ouvrir => Syntaxe (cliquez sur le nom du fichier choisi (extension .sps) qui apparait sous cette rubrique puis cliquez sur => Ouvrir)

3. Description des principales icônes de la barre d'outils d'IBM SPSS STATISTICS

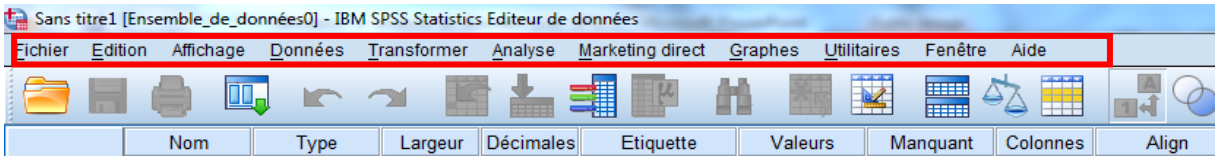


Figure 18 Barre d'outils d'IBM SPSS Statistics

Le menu **FICHIER / FILE** permet la gestion des fichiers (p. ex., ouvrir un nouveau fichier, fermer, enregistrer, etc.).

Le menu **ÉDITION / EDIT** permet d'effectuer les opérations de traitement de texte (p. ex., copier, couper, coller, sélectionner, etc.).

Le menu **AFFICHAGE / VIEW** permet de définir les options de l'écran.

Le menu **DONNÉES / DATA** traite de tout ce qui est lié à la gestion de la barre de données (p. ex., définir ou insérer une variable, trier les données, etc.).

Le menu **TRANSFORMER / TRANSFORM** présente les différentes opérations de transformation possibles sur les variables de la barre de données
(p. ex., recodification, catégorisation, création d'indices, etc.).

Le menu **ANALYSE / ANALYZE** permet d'accéder à toutes les analyses statistiques que IBM SPSS STATISTICS rend possibles (p. ex., analyses descriptives, corrélations, etc.).

Le menu **GRAPHES / GRAPHS** présente tous les types de graphiques que IBM SPSS STATISTICS permet de créer (p. ex., histogrammes, boîtes à moustaches, courbes, etc.).

Le menu **OUTILS / UTILITIES** comprend les utilitaires du programme (p. ex., informations sur les fichiers, informations sur les variables, etc.).

Le menu **FENÊTRE / WINDOWS** permet la gestion des fenêtres.

Le menu **AIDE / HELP** propose des rubriques d'aide à l'utilisation de IBM SPSS.

Voici un tableau récapitulant tous les raccourcis de la barre outil(Kent State University Libraries., 2022b).

Tableau 2 Tableau résumant tous les raccourcis sur IBM SPSS Statistics

	Ouvrez un fichier de données. Équivalent à Fichier > Ouvrir > Données.
	Imprimer le contenu de la fenêtre d'affichage des données active. Non recommandé. Équivalent à Fichier > Imprimer.
	Affiche la liste des fenêtres de dialogue les plus récemment utilisées. À utiliser lorsque vous devez réexécuter une analyse.
	Enregistrez l'ensemble de données actif. Équivalent à Fichier > Enregistrer ou Ctrl + S
	Équivalent à Edition > Annuer (dans les menus déroulants) ou Ctrl + Z
	Équivalent à Edition > Rétablir (dans les menus déroulants) ou Ctrl + Y.
	Accédez à un cas spécifique (ligne) dans l'ensemble de données actif. Équivalent à Edition > Aller à la casse (Observation ou à la variable).
	Aller à une variable spécifique (colonne) dans l'ensemble de données actif. Équivalent à Modifier > Aller à la variable
	Affichez le nom de la variable, les libellés, le type, le niveau de mesure, les codes de valeur manquante et les libellés de valeur pour toutes les variables dans la fenêtre active. Équivalent à Utilitaires > Variables
	Recherchez une valeur ou une observation dans l'ensemble de données, ou recherchez et remplacez une valeur ou une observation dans l'ensemble de données. S'active uniquement lorsqu'une cellule de la fenêtre Affichage des données est sélectionnée. Équivalent à Édition > Rechercher et éditer > Remplacer, ou Ctrl + F et Ctrl + H, respectivement.

	Exécuter des statistiques descriptives (à l'aide de la procédure Fréquences) sur *la ou (les) variable(s) sélectionnée(s)*. Les statistiques affichées sont déterminées par le réglage du niveau de mesure variable. Les variables nominales et ordinales sont résumées avec un tableau de fréquence ; les variables d'échelle sont résumées en utilisant la moyenne, la médiane, l'écart type, la plage, le minimum et le maximum. S'active uniquement lorsqu'une cellule ou une colonne de la fenêtre Affichage des données est sélectionnée. Équivalent à Analyser > Statistiques descriptives > Fréquences.
	Insérer un cas entre deux cas existants. Équivalent à Édition > Insérer des cas
	Insérez une nouvelle variable entre deux variables existantes. Par défaut, les nouvelles variables créées de cette manière sont des variables numériques d'échelle. Équivalent à Édition > Insérer une variable.
	Stratifiez vos analyses en fonction d'une variable catégorielle. Par exemple, si la variable Genre est sélectionnée dans le fichier fractionné, l'exécution de statistiques descriptives sur toute autre variable produira des descriptions pour les hommes et les femmes séparément. Équivalent à Données > Scinder un Fichier.
	Extrayez un ensemble d'observations dans un nouveau fichier de données en fonction de certains critères ou appliquez une variable de filtre. Équivalent à Données > Sélectionner des observations
	Activez ou désactivez l'affichage des données brutes ou de libellé de valeur dans la fenêtre Affichage des données. Équivalent à Affichage > Libellés de valeur
	Sélectionnez ou désélectionnez les ensembles de variables à afficher dans la fenêtre active. Plusieurs ensembles peuvent être sélectionnés à la fois. Équivalent à Utilitaires> Utiliser des ensembles variables. Notez que vous devez d'abord définir un ensemble de variables (Utilitaires> Définir des ensembles de variables) pour que cela soit utile
	Affiche toutes les variables de l'ensemble de données actif. Ne s'active que si Use Variable Sets a été utilisé. Équivalent à Utilitaires > Afficher toutes les variables.

4. Sauvegarder les fichiers dans IBM SPSS Statistics

Procédure pour enregistrer un fichier de données

Cliquez sur Fichier dans le menu principal entête de IBM SPSS STATISTICS

- Fichier => Enregistrer sous (remplacez le nom par défaut par un nom de votre choix => Enregistrer).

ATTENTION : Chaque nouvelle sauvegarde de votre travail à l'intérieur d'un fichier qui garde son nom « écrase » l'ancienne version !!! Soyez donc sûr de votre coup, particulièrement si vous avez fait plusieurs manipulations de variables, avant de sauvegarder un fichier nouveau ou modifié sous le nom d'un ancien fichier car vous risquez ainsi de perdre une partie du travail antérieurement réalisé, les nouvelles données prenant la place des anciennes.

5. La syntaxe dans IBM SPSS STATISTICS

La fenêtre de programmation ou de commandes, appelée **Syntaxe** (fichier de type .SPS), permet d'écrire des commandes directement ou indirectement (par le bouton Coller) dans les boites de dialogue).

Cette caractéristique du logiciel IBM SPSS STATISTICS est extrêmement intéressante et importante, car elle permet de garder une trace écrite des commandes exécutées par les boites de dialogue. On peut alors les sauvegarder pour utilisation ultérieure, les modifier ou les exécuter à répétition sans toujours passer par les boites de dialogue.

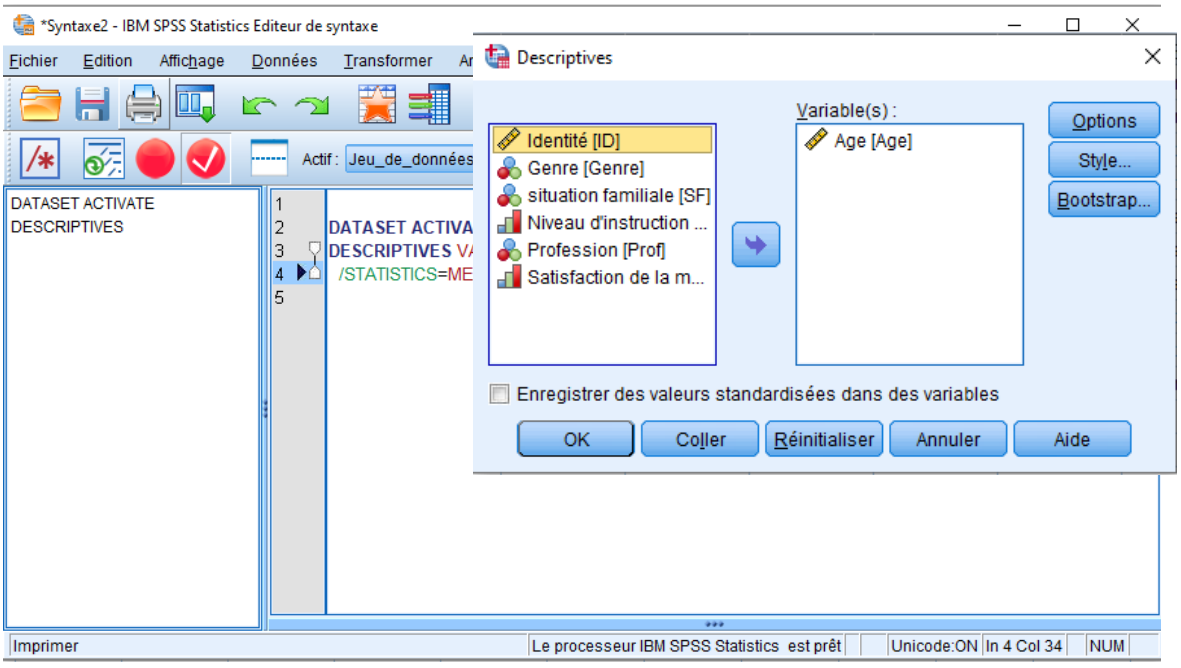


Figure 19 Fichier syntaxe

Il s'agit simplement d'utiliser les fonctions des menus et lorsque toutes les commandes et sous-commandes ont été vérifiées, cliquer sur le bouton « COLLER » situé au bas à gauche avant de cliquer sur OK. Une fenêtre Éditeur de syntaxe s'ouvrira automatiquement.

Lorsqu'une commande est complète, on peut l'exécuter en allant dans le menu « Exécuter ». Pour obtenir une fenêtre de syntaxe vide, aller dans le menu "Fichier : Nouveau : syntaxe".

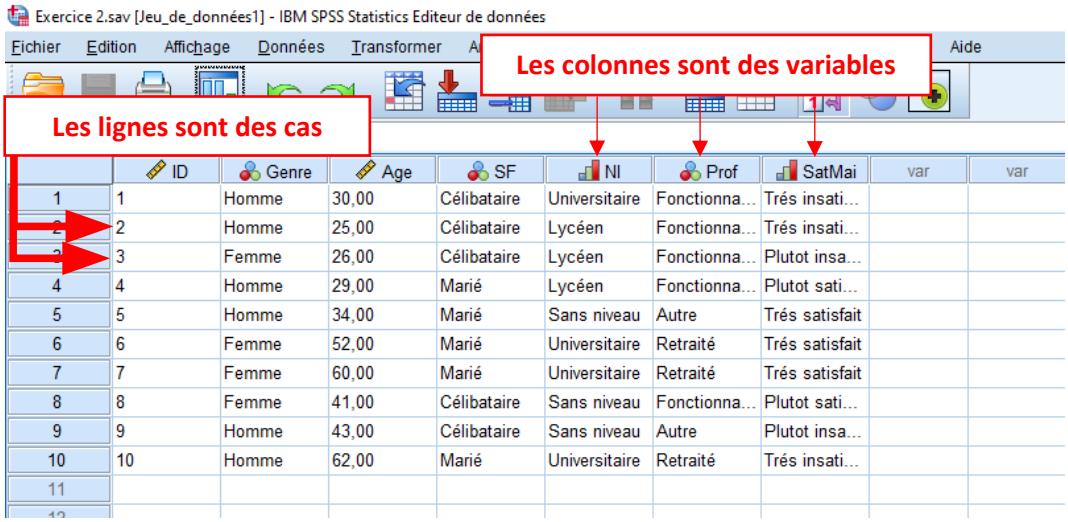
III. Créer ou transformer un fichier de données

1. La codification des données

Dans un fichier de données, chaque cas (individu, répondant, dossier...) est représenté par une ligne contenant les données pour un ensemble de variables. Les variables, pour leur part, figurent en colonnes (chaque variable occupe l'espace d'une colonne). La rencontre d'une variable et d'un cas se présentera donc sous forme d'une cellule dans laquelle l'information se trouve consignée (Kent State University Libraries., 2022f).

Le fichier de données comprendra uniquement des données brutes. Ces dernières peuvent être numériques (composées de chiffres ou de codes chiffrés) ou alphanumériques (combinaison de lettres et de chiffres). On préférera toujours, dans la mesure du possible, associer une forme numérique aux données. En effet, IBM SPSS statistics travaille beaucoup plus facilement avec des chiffres qu'avec des lettres.

Dans le cas des données de niveau de mesure nominale ou ordinale, il s'agit alors d'attribuer un code numérique aux données. Par exemple, pour indiquer le genre d'un individu, on préférera coder les réponses de la façon suivante : garçon : "1" ; fille : "2", plutôt que garçon : "G" ; filles : "F".



	ID	Genre	Age	SF	NI	Prof	SatMai	var	var
1	1	Homme	30,00	Célibataire	Universitaire	Fonctionna...	Trés insati...		
2	2	Homme	25,00	Célibataire	Lycéen	Fonctionna...	Trés insati...		
3	3	Femme	26,00	Célibataire	Lycéen	Fonctionna...	Plutot insa...		
4	4	Homme	29,00	Marié	Lycéen	Fonctionna...	Plutot sati...		
5	5	Homme	34,00	Marié	Sans niveau	Autre	Trés satisfait		
6	6	Femme	52,00	Marié	Universitaire	Retraité	Trés satisfait		
7	7	Femme	60,00	Marié	Universitaire	Retraité	Trés satisfait		
8	8	Femme	41,00	Célibataire	Sans niveau	Fonctionna...	Plutot sati...		
9	9	Homme	43,00	Célibataire	Sans niveau	Autre	Plutot insa...		
10	10	Homme	62,00	Marié	Universitaire	Retraité	Trés insati...		
11									

Figure 20 Les lignes et les colonnes dans l'éditeur de données

Avant de saisir les données, nous devons définir les variables concernées. Cette opération porte l'appellation de « **définition du masque de saisie** ». Un masque de saisie est un fichier de données qui comporte uniquement la définition des variables et qui est prêt à la saisie des données (Yergeau & Poirier, 2021b).

Le bouton Nom : c'est le nom de la variable. Le nom des variables n'apparaît pas automatiquement en entête de colonne, il faut l'inscrire soi-même. Par défaut, les variables ajoutées sont nommées : « **v00001** », « **v00002** », « **v00003** », etc., ce qui complique leur reconnaissance par le chercheur, surtout s'il lui arrive de revenir au fichier de données, quelque temps après.

Le nom que peut prendre une variable dans IBM SPSS Statistics obéit à certaines contraintes :

- **L'espace entre les caractères n'est pas permis**
- **Le premier caractère ne doit jamais être un chiffre**
- **Les caractères suivants ne sont pas autorisés : () - , ; : ! / ? * + = & # $ ‘ « ~ []**
- **Deux variables différentes ne peuvent pas avoir le même nom**
- **Les caractères suivants peuvent être utilisés sans problème : . _**

Il est très recommandé d'utiliser les petits tirets « _ » pour définir les variables à nom composé, comme dans « **niv_sco** » pour « **Niveau Scolaire** ». Quand un nom de variable est saisi, IBM SPSS Statistics dresse automatiquement la liste des caractéristiques en rapport. C'est-à-dire qu'il remplit l'ensemble des cases situées sur la même ligne que le nom de la variable.

Le bouton Type : permet de modifier le type de données de la variable, c’est-à-dire le genre de données qui sera assigné à la variable (numérique, alphabétique, date, monétaire, etc.). Toutes les possibilités se présentent comme suit :

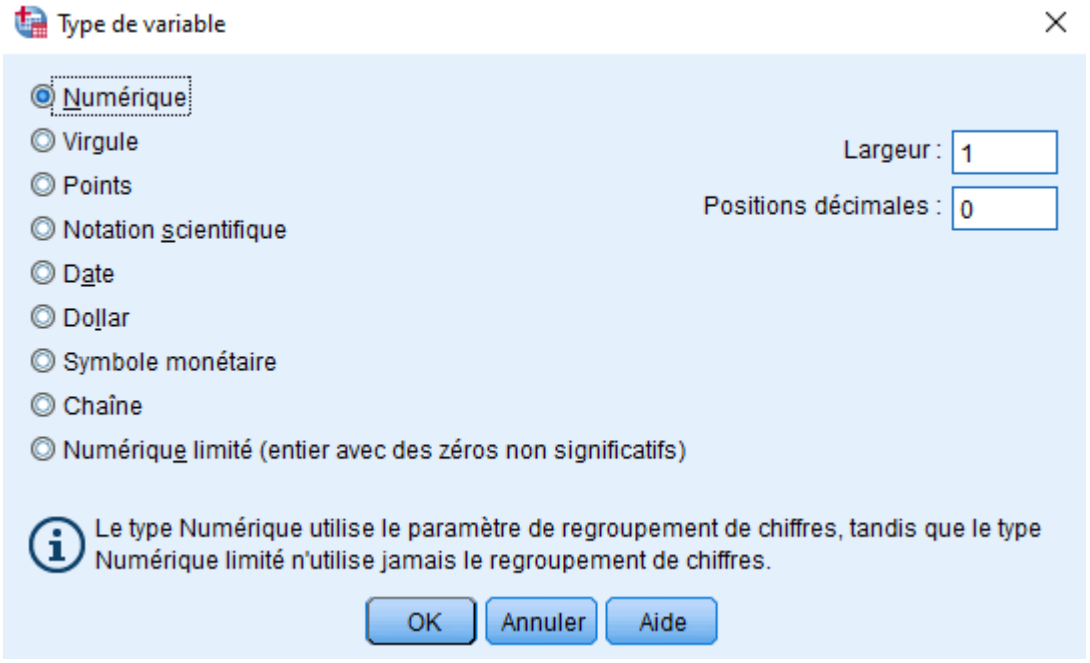


Figure 21 Le bouton type de variables

- **Numérique** : Variable dont les valeurs sont des nombres.

- **Virgule** : Variable numérique dont les valeurs sont affichées avec des virgules toutes les trois positions, le point servant de séparateur décimal. Ce système est utile pour la saisie des nombres trop élevés : exemple : **952,457,235.12** au lieu de 952457235,12.

- **Point** : Variable numérique dont les valeurs sont affichées avec des points toutes les trois positions, la virgule servant de séparateur décimal. Exemple : **952.457.235,12** au lieu de 952457235,12.

- **Notation scientifique** : Variable numérique dont les valeurs sont affichées avec un **E intégré et un exposant de puissance. Exemple** : 300 000 est saisi 3.E+05

- **Dollar :** variable numérique affichée avec le signe dollar ($), avec des virgules toutes les trois positions, le point servant de séparateur décimal. Vous pouvez entrer des valeurs de données avec ou sans le signe dollar.

- **Date** : Variable numérique dont les valeurs sont affichées dans l'un des formats de date ou d'heure possibles.

- **Symbole monétaire** : Variable numérique dont les valeurs sont affichées dans l'un des formats de symbole monétaire que l'on définit au préalable (Dollar ou devise personnalisée).

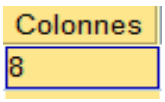
- **Chaîne (String)** : Variable dont les valeurs ne sont pas numériques et ne sont donc pas utilisées pour les calculs. Ces variables peuvent contenir n'importe quel caractère dans la limite du nombre de caractères défini. Les variables chaînes établissent une distinction entre les majuscules et les minuscules. Ces variables sont également connues sous le terme de variables alphanumériques.

- **Numérique limité (entier avec des zéros non significatifs) :** variable dont les valeurs sont limitées à des entiers non négatifs. Les valeurs sont affichées avec des signes zéro complétant la largeur maximale de la variable. Les valeurs peuvent être saisies en notation scientifique.

Le bouton Largeur : la largeur des variables permet de définir le nombre des caractères à saisir. Elle doit être supérieure ou égale à « **1** ». Pour une variable numérique, il s'agit de définir le nombre des caractères avant la virgule. Le nombre de caractères après la virgule permet de définir les décimales. Dans IBM SPSS Statistics, les variables numériques doivent avoir une largeur strictement supérieure aux décimales. Pour les variables alphanumériques, la case des décimales devient inactive.

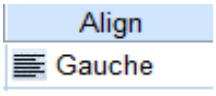
Le bouton Libellé : puisque le nom d'une variable peut contenir des contraintes, il arrive de lui joindre un libellé plus explicite qui rappellera exactement ce que la variable représente.

- **Le bouton Décimal** : la largeur de décimal si elle existe dans la variable.



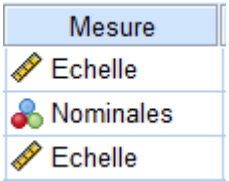
- **Le bouton Colonnes :** permet de déterminer la largeur de la colonne.

- **Le bouton Alignement** : sert à aligner les données à l'intérieur des cellules. Pour modifier l'alignement du texte, cliquez sur l'une des trois dispositions disponibles.



- **Le bouton Mesure :** permet de fixer l'échelle de mesure.

Il existe 3 types d'échelle de mesure : l'échelle nominale (**NOMINAL**), l'échelle ordinale (**ORDINAL**), et l'échelle d'intervalle (**ÉCHELLE / SCALE)**.



2. L'Echelle de mesure

Les tests statistiques mis en œuvre pour mesurer ces relations seront sélectionnés en fonction de l'objectif de l'enquête (Mesurer à l'aide d'un questionnaire) et en fonction des variables collectées. Les variables sont de deux types :

- **Qualitatives** : leurs modalités, c'est-à-dire la manière dont les observations sont regroupées, ne peuvent être calculées.
- Q**uantitatives** : leurs modalités sont mesurables et les tests envisageables sont nombreux.

2.1. Échelles de mesures nominales

Une échelle est nominale si elle définit simplement l'appartenance d'un élément à une modalité, classe ou catégorie non hiérarchique. Une échelle nominale comporte un certain nombre de catégories, dont la seule propriété est qu'elles sont toutes différentes les unes des autres (Stafford & Bodson, 2006) (Carricano & Poujol, 2009).

- Chaque observation se trouve dans une catégorie et une seule.
- Ce sont des échelles non ordonnées

Par exemple, dans la question suivante, chacune des réponses est indépendante et, même si on associait 1 = Blanc, 2 = Bleu, 3 = Rouge, etc., il serait impensable de faire une moyenne à partir de ces chiffres.

2.2. Échelles de mesures ordinales

Cette échelle possède deux propriétés : **l'identification et l'ordonnancement.** C'est-à-dire les nombres représentent des modalités ordonnées ou classées par ordre de grandeur. Dans une échelle ordinale, les catégories la composant sont munies d'une structure d'ordre, établie en fonction d'un critère donné. (Stafford & Bodson, 2006) (Carricano & Poujol, 2009) Par exemple, le rang dans la famille : le premier né, le deuxième, etc.

Notez de 1 à 5 la qualité du lait
(1 étant la note la plus faible, 5 la note la plus élevée)
Echelle non interchangeable

1	2	3	4	5

2.3. Échelles de mesures métriques

L'échelle d'intervalle

Cette échelle possède trois propriétés : **l'identification, l'ordonnancement et l'égalité des intervalles.** C'est une échelle métrique dont le zéro est fixé arbitrairement. Dans cette échelle, un zéro ne correspond pas à une « absence de la variable d'intérêt » et les nombres qui y sont associés ont une signification (p. ex. : la température : le zéro correspond au point de congélation de l'eau et non pas à l'absence de température). Cette échelle permet de déterminer l'intervalle entre les observations et de comparer ces intervalles. Il est ainsi possible de déterminer si deux intervalles sont ou ne sont pas de même étendue (Stafford & Bodson, 2006) (Carricano & Poujol, 2009).

Variables de ratios ou de rapport

Les variables de ratios sont des variables d'intervalles avec un zéro naturel. Par exemple pour la durée d'un test, à 0, il n'y pas de temps.

L'échelle métrique est la plus couramment utilisée, même si pour ces mesures d'attitudes les intervalles ne sont pas toujours équidistants. Appartiennent à cette catégorie, l'échelle d'Osgood ou l'échelle de Stapel, qui ont pour but de conduire à l'élaboration de profils de répondants, l'échelle d'intensité de Likert ou échelle d'accord, les échelles d'intention.(Carricano & Poujol, 2009)

Echelle d'Osgood

Avez-vous trouvé que le goût du produit X était

Mauvais	1	2	3	4	5	**Bon**

Echelle de Stapel

Choisissez un nombre positif si vous pensez que le mot décrit bien le produit X,
Un nombre négatif si vous pensez que le mot ne décrit pas bien le produit X,
En notant de +5 à -5

Bon	…
Utile	…
Pratique	…
etc	…

Echelle de Likert

Pas du tout d'accord	1	2	3	4	5	**Tout à fait d'accord**

Echelle d'intention

Si la marque M lançait ce type de produit

Je n'achèterai pas ce produit	1	2	3	4	5	**J'achèterai ce produit**

Exemple (Échelle d'intervalles) : Lorsque la température augmente de 10 degrés Celsius à 20 degrés Celsius, on ne peut pas dire qu'il fait 2 fois plus chaud, parce que 10 degrés Celsius correspond à 50 degrés Fahrenheit et 20 degrés Celsius correspond à 68 degrés Fahrenheit, ce qui ne correspond pas au double de la température. Alors on peut seulement dire qu'il fait 10 degrés de plus.

Exemple (Échelle de rapports) : dans le cas d'une variable mesurée à l'aide d'une échelle de rapports, on peut affirmer qu'un élève qui a 2 emplois à deux fois plus d'emplois qu'un élève qui n'a qu'un seul emploi.

Remarque :
Les échelles de mesure avec SPSS
· ***l'échelle nominale*** : **(Nominale)**
Ex : (féminin, masculin); (marié, veuf, célibataire, ..); (conforme, non-conforme)

· ***l'échelle ordinale*** **: (Ordinale)**
Ex : (faible, moyen, élevé); (-18, 18 à 24, 25 à 29; 30 à 34); (bon, moyen, élevé)

· ***l'échelle d'intervalles et l'échelle de rapport : (Échelle)***
Ex : (Écarts entre 2 modalités : Écart de 10°C (10°C et 20°C) ou (15°C et 25°C)
Ex : (le chiffre d'affaires d'une entreprise, nombre d'heures de travail, …).

- **Les données QUALITATIVES sont : nominales, ordinales.**
- **Les données QUANTITATIVES sont : nominales, ordinales, échelles.**

3. Comment coder les réponses ?

- **Comment coder les réponses à réponses courtes ?**

Exemple :
Quelle est votre nationalité ? _____

Solution : Codez les réponses ouvertes avec des valeurs numérique (1 = Algérienne, 2 = française, etc.) en faisant une liste.

- **Comment coder les réponses multiples ?**

Exemple :
Quelles occupations/loisirs avez-vous régulièrement ? (*Plusieurs réponses sont possibles*)
Activités culturelles (expositions, concerts, lecture, université 3e âge, etc.)
Activités physiques (natation, marche, club de sport, etc.)
Bricolage (tricot, jardinage, etc.)
Animal domestique
Instrument de musique
Internet, e-mail
Autre : _____

Solution :

- Créez une variable pour chaque catégorie (par exemple loisir1 - loisir6), codé par 0 = n'a pas ce loisir, 1 = a ce loisir.
- Pour les réponses ouvertes (« autre : »), créez soit une variable alphanumérique (Chaîne de caractère), par exemple. add_lois, soit une variable numérique en faisant une liste des loisirs qui apparaissent.

Comment coder les réponses ouvertes ?
Exemple : Qu'est-ce que ces loisirs vous apportent ?
_____

Solution :

- Regrouper l'information en catégorie grâce à l'analyse de contenu.

4. Ajouter un cas ou une variable à une banque de données déjà constituée

Il est toujours possible d'ajouter une ligne (un cas) ou une colonne (une variable) à un fichier de données déjà constitué. Pour ce faire :

5. Nommer une variable

Le nom de la variable n'apparaît pas automatiquement en entête de colonne, il faut que vous l'inscriviez vous-même (par défaut, les variables sont appelées v00001, v00002, v00003...v0000z).

Procédure pour nommer plus précisément une variable

Mettez-vous en mode **Vue des variables**

- ⇨ La liste de variables et de leurs caractéristiques, comme son nom, son type, sa largeur, les décimales, son libellé, ses valeurs, les valeurs manquantes, etc. apparaît.
- ⇨ Compléter ou corriger chacune des cases afin de bien nommer et coder vos données.

6. Libeller une variable

Libellé : Il arrive que vous vouliez joindre des libellés à vos variables afin qu'elles soient plus explicites.

Nom	Type	Largeur	Décimales	Libellé	Valeurs	M
ID	Numérique	2	0	Identité	Aucun	Aucu
Genre	Numérique	1	0	Genre	{1, Homme}...	Aucu
Age	Numérique	4	2	Age	Aucun	Aucu
SF	Numérique	1	0	situation familiale	{1, Célibatai...	Aucu
Niv_ist	Numérique	1	0	Niveau d'instruction	{1, Universit...	Aucu
Prof	Numérique	1	0	Profession	{1, Fonction...	Aucu
Sat_Mai	Numérique	1	0	Satisfaction de la maison	{1, Trés ins...	Aucu

Figure 22 Libellés de variables

7. Libeller les valeurs de la variable

Il est possible que vous ayez attribué une valeur numérique à des informations de type nominal ou ordinal (ainsi, garçon=1 et fille=2). Si c'est le cas, vous voudrez sans doute associer un libellé à ces valeurs numériques de remplacement, histoire de vous souvenir à quelle valeur correspond en réalité chaque code numérique.

Procédure pour libeller les valeurs numériques d'une variable nominale ou ordinale

Cliquez sur le petit carré bleu apparaissant dans la case appropriée.

Nom	Type	Largeur	Décimales	Libellé	Valeurs	Manquant	Colc
ID	Numérique	2	0	Identité	Aucun ...	Aucun	8
Genre	Numérique	1	0	Genre	{1, Homme}...	Aucun	8

- ⇨ Apparait une boîte de dialogue nommée « Libellés de valeurs ».
- ⇨ Inscrivez le code numérique dans le rectangle de **valeurs** et la valeur nominale correspondante dans le rectangle **Libellé**.
- ⇨ Cliquez sur **ajouter**.
- ⇨ Recommencez l'opération jusqu'à ce que toutes les valeurs numériques de la variable aient leurs libellés.
- ⇨ Cliquez sur OK.

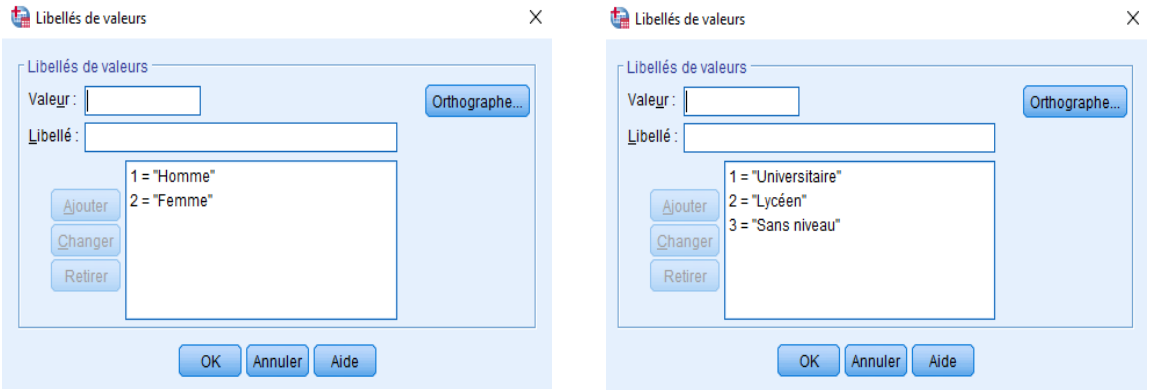


Figure 23 Boîte de dialogue : Libellés de valeurs

Remarque 1 :

Dans le cas de variables proportionnelles, comme par exemple l'âge du répondant ou le salaire familial brut calculé sur une base annuelle, il n'est pas utile de chercher à libeller les valeurs de la variable puisque les données portent de facto l'étiquette à laquelle elles correspondent (20 ans, 30 ans, 40 ans... 25,999 DZD, 59,999 DZD, 100,000 DZD. L'apposition d'un libellé précisant le nom de la variable demeure toutefois de mise.

Remarque 2 :

Il est préférable de codifier avec des chiffres et non pas avec des lettres. Pour les variables de chaîne ordinales, l'ordre alphabétique des valeurs chaîne est supposé refléter l'ordre des catégories. Par exemple, pour une variable de chaîne comportant des valeurs *Faible*, *Moyen*, *Elevé*, l'ordre des catégories est interprété comme *Elevé*, *Faible* ou *Moyen*, ce qui ne correspond pas à l'ordre correct. En règle générale, il est recommandé d'utiliser les codes numériques pour représenter les données ordinales.

8. Indiquer la présence de valeurs manquantes

Lorsqu'il est question de traitements statistiques des données, une chose dont il faut s'assurer c'est que les valeurs manquantes (les répondants pour lesquels nous ne possédons pas

l'information concernant une ou plusieurs variables) ne soient pas prises en compte. Ces données seront tout simplement considérées manquantes et retirées des analyses. Les valeurs manquantes sont définies et comptabilisées pour chacune des variables séparément (en effet, une donnée peut ne pas être disponible pour un répondant et pas pour un autre tout comme le fait de ne pas posséder l'information pour un répondant sur une variable ne signifie pas que l'information est manquante pour les autres variables concernant le même répondant) (van den Berg, 2022).

8.1. Comment coder les valeurs manquantes ?

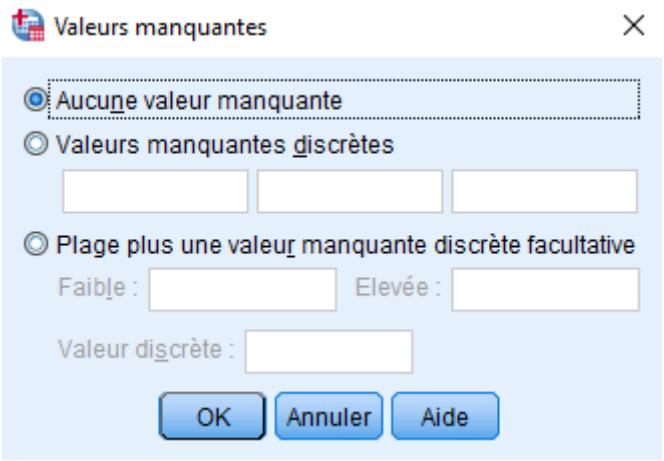


Figure 24 Boîte de dialogue valeurs manquantes

- Dès qu'on a entré une donnée, toutes les cellules des autres variables numériques de ce cas sont désignées par un point (= « ***missing ou manquant*** ») qui est remplacé quand on entre une valeur. Garder le point dans la cellule signifie que la valeur pour cette cellule est manquante et ce type de donnée manquante n'a pas à être défini comme telle car elle est reconnue automatiquement par IBM SPSS Statistics comme valeur manquante (Kent State University Libraries., 2022a).
- Garder la cellule vierge pour les variables alphanumériques n'est pas considéré par IBM SPSS Statistics comme valeur manquante. Il faut la définir comme telle dans la vue des variables. Pour cela, entrez un espace dans « **discret missing values ou valeurs manquantes discrètes** ».
- Entrer un chiffre en dehors de l'étendue de valeurs valables (p. ex. 8 ou 88). Il faut définir ces valeurs dans la vue des variables. Pour cela, entrez un espace dans « **discret missing values** ».

8.1.1. Aucune valeur manquante

Signale qu'il n'y a pas de valeurs manquantes pour cette variable. Il s'agit de la valeur par défaut de cette sous-commande.

8.1.2. Valeurs manquantes discrètes

Trois rectangles sont ici prévus afin de permettre de définir jusqu'à trois valeurs manquantes pour une même variable. Par exemple, les valeurs **99** et **200** sont manquantes pour la variable âge.

On peut aussi prévoir différents types de valeurs manquantes et leur attribuer différents codes permettant de les distinguer. Par exemple, on attribuera la valeur « 7 » ou « NSP » à ceux qui, en réponse à une question, indiquent « ne sait pas » ; la valeur « 8 » à ceux qui refusent de répondre, et la valeur « 9 » pour toute information vraiment manquante.

8.1.3. Plage plus une valeur manquante facultative

Elle vous permet de déclarer toutes les valeurs comprises entre une borne inférieure (faible) et une borne supérieure (élevée) comme étant autant de valeurs manquantes. Par exemple, on pourrait vouloir précises que toutes les valeurs comprises entre 75 et 99 pour la variable âge du répondant correspondent à autant de valeurs manquantes.

8.1.4. Valeur discrète

Reprend les propriétés de l'option précédente tout en permettant d'ajouter la déclaration d'une valeur discrète comme étant **aussi manquante.**

Nota : Il faut appliquer l'exercice sur les valeurs manquantes pour voir la différence dans le calcul d'une variable (calcul d'une moyenne) en présence des valeurs manquantes.

9. Recoder une variable

La boîte de dialogue Recodage des variables vous permet de réaffecter les valeurs de variables existantes ou de fusionner des plages de valeurs existantes dans des nouvelles valeurs. Par exemple, vous pourriez fusionner des salaires dans des catégories de plages de salaires.

Vous pouvez recoder des variables numériques et de chaîne. Si vous sélectionnez plusieurs variables, elles doivent toutes être du même type. Vous ne pouvez pas recoder ensemble des variables numériques et de chaîne.

Exemple 1 :

Les valeurs de la variable Age sont codées en valeurs continues et on veut les recoder en 2 catégories (moins de 29 ans et Plus de 30 ans). Pour recoder les valeurs d'une variable
1. A partir des menus, sélectionnez :
Transformer > **Recoder des variables...**
2. Sélectionnez les variables que vous désirez recoder.
3. Cliquez sur **Anciennes et nouvelles valeurs** et spécifiez comment recoder les valeurs.
4. Spécifiez une ancienne valeur et une nouvelle valeur.
4.a. Entrez la valeur **29** dans « Plage, du MINIMUM à la valeur » afin de signifier la nouvelle première catégorie d'âge ensuite inscrivez 1 c'est-à-dire « Lowest thru 29 -> 1 ».
5. Cliquez sur **Ajouter** pour placer la spécification

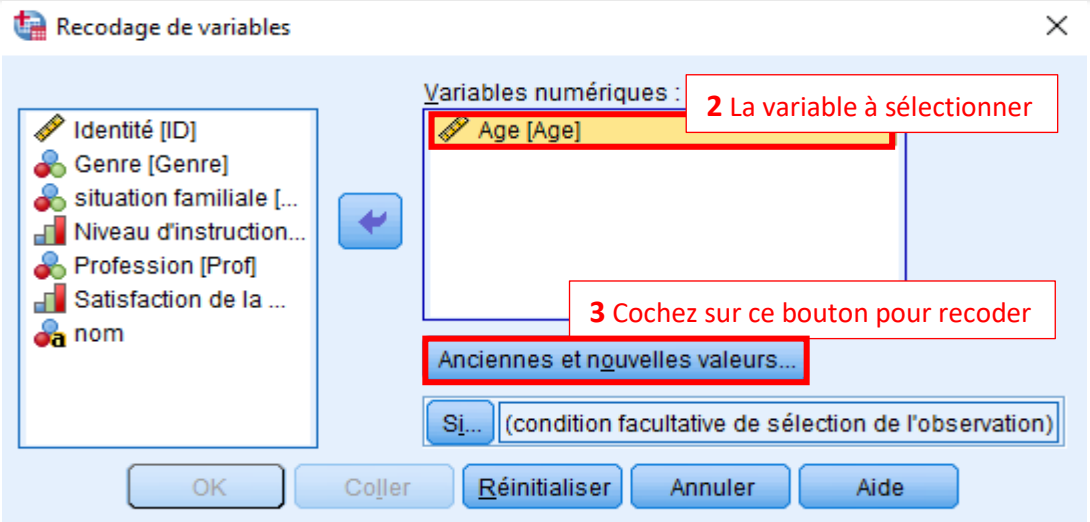


Figure 25 Boite de dialogue "Recodage de variables"

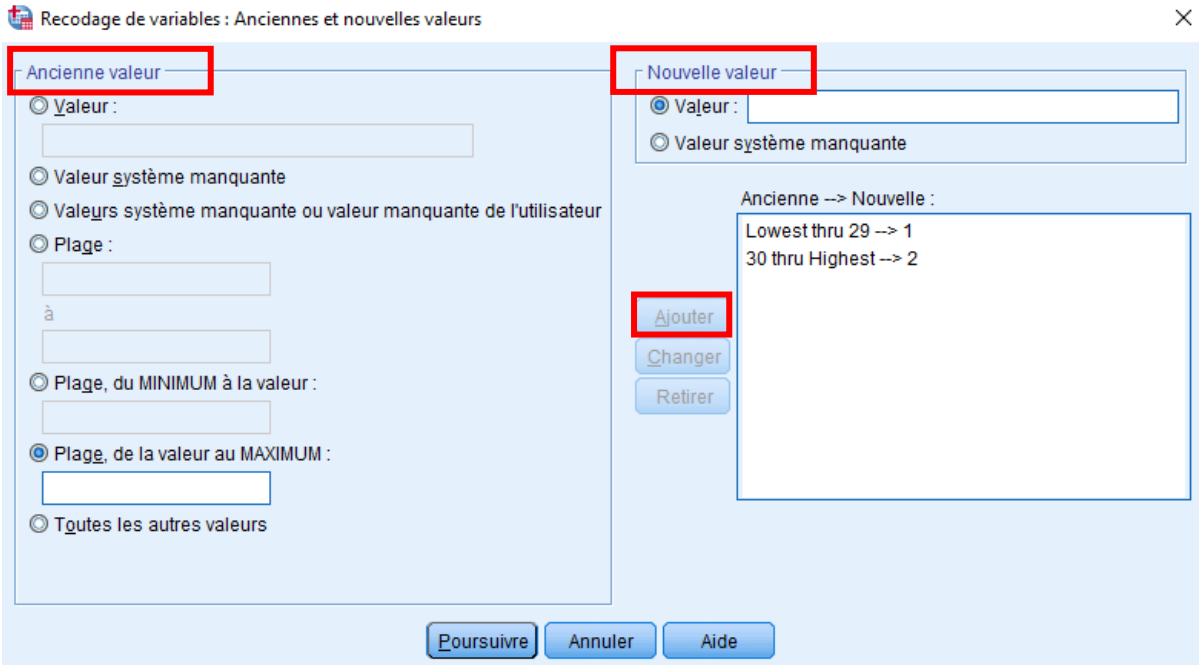
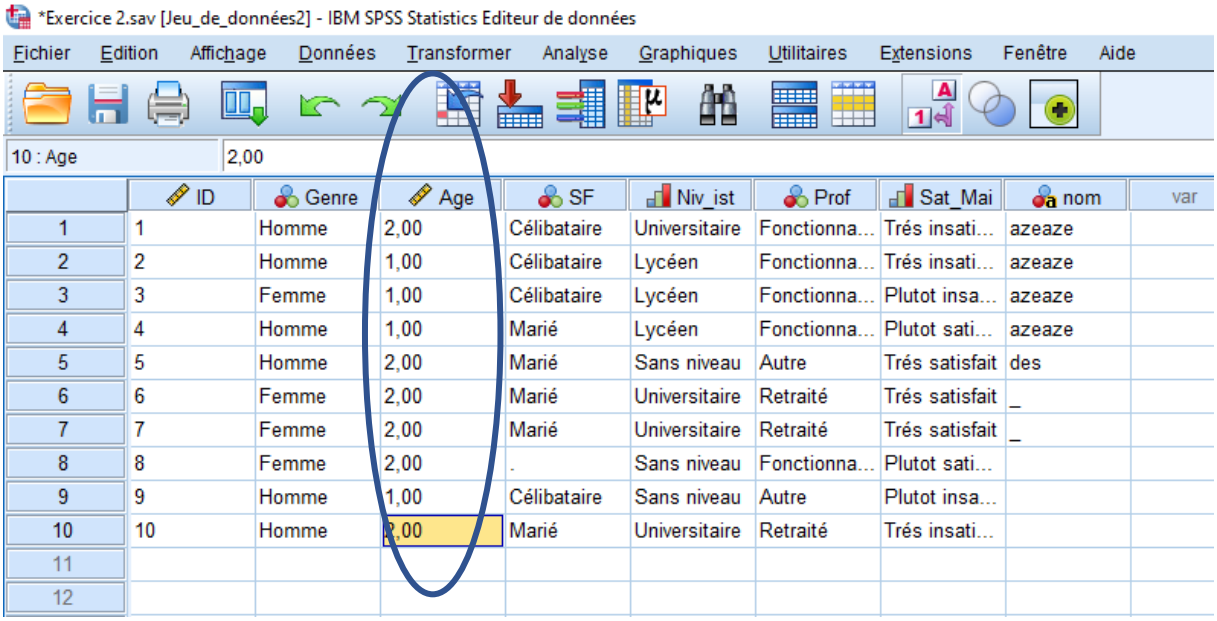


Figure 26 Boite de dialogue ancienne et nouvelles valeurs

6. On met notre fichier sous vue des données.



	ID	Genre	Age	SF	Niv_ist	Prof	Sat_Mai	nom	var
1	1	Homme	2,00	Célibataire	Universitaire	Fonctionna...	Trés insati...	azeaze	
2	2	Homme	1,00	Célibataire	Lycéen	Fonctionna...	Trés insati...	azeaze	
3	3	Femme	1,00	Célibataire	Lycéen	Fonctionna...	Plutot insa...	azeaze	
4	4	Homme	1,00	Marié	Lycéen	Fonctionna...	Plutot sati...	azeaze	
5	5	Homme	2,00	Marié	Sans niveau	Autre	Trés satisfait	des	
6	6	Femme	2,00	Marié	Universitaire	Retraité	Trés satisfait	_	
7	7	Femme	2,00	Marié	Universitaire	Retraité	Trés satisfait	_	
8	8	Femme	2,00	.	Sans niveau	Fonctionna...	Plutot sati...		
9	9	Homme	1,00	Célibataire	Sans niveau	Autre	Plutot insa...		
10	10	Homme	2,00	Marié	Universitaire	Retraité	Trés insati...		
11									
12									

Figure 28 Nouvel affichage de la variable âge

7. Enfin, il faut libeller les valeurs
(1: Moinds de 29 ans ; 2 : Plus de 30 ans).

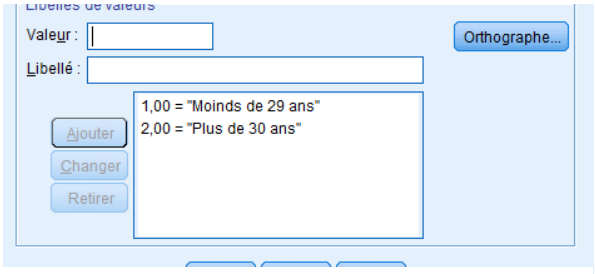


Figure 27 Libellés des nouvelles valeurs

Exemple 2 :
Les valeurs de la variable Revenu sont codées en 7 catégories et on veut les recoder en 3 catégories.

1 : < 15 000
2 : 15 000 et 24 999
3 : 25 000 et 49 999
4 : 50 000 et 74 999
5 : 75 000 et 99 999
6 : 100 000 et 149 999
7 : > 150 000

1 : < 49 999
2 : 50 000 et 99 999
3 : > 100 000

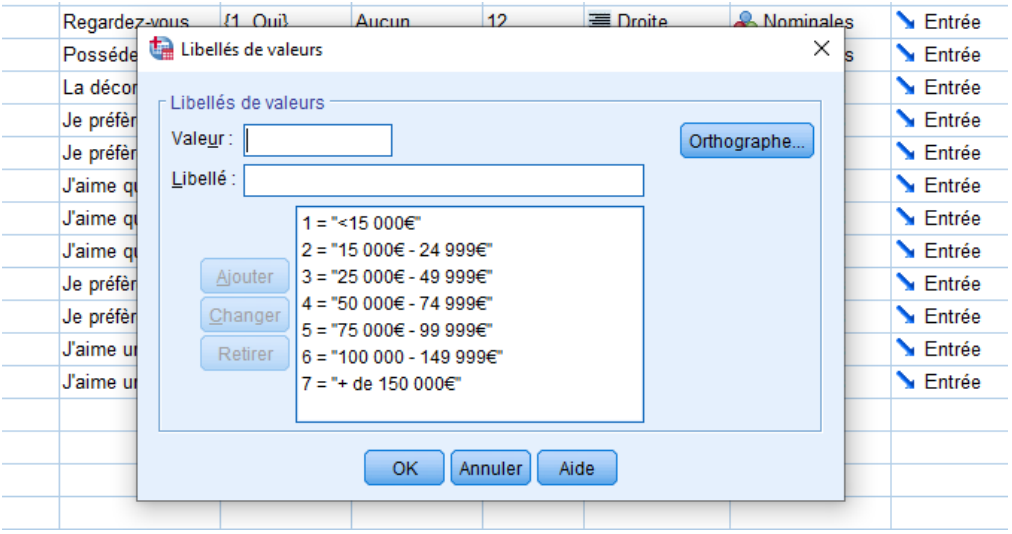


Figure 29 Anciennes codification de la variable Revenu

Pour recoder les
valeurs d'une variable
1. A partir des menus, sélectionnez :
Transformer > **Recoder des variables...**
2. Sélectionnez les variables que vous désirez recoder.
3. Cliquez sur **Anciennes et nouvelles valeurs** et spécifiez comment recoder les valeurs.

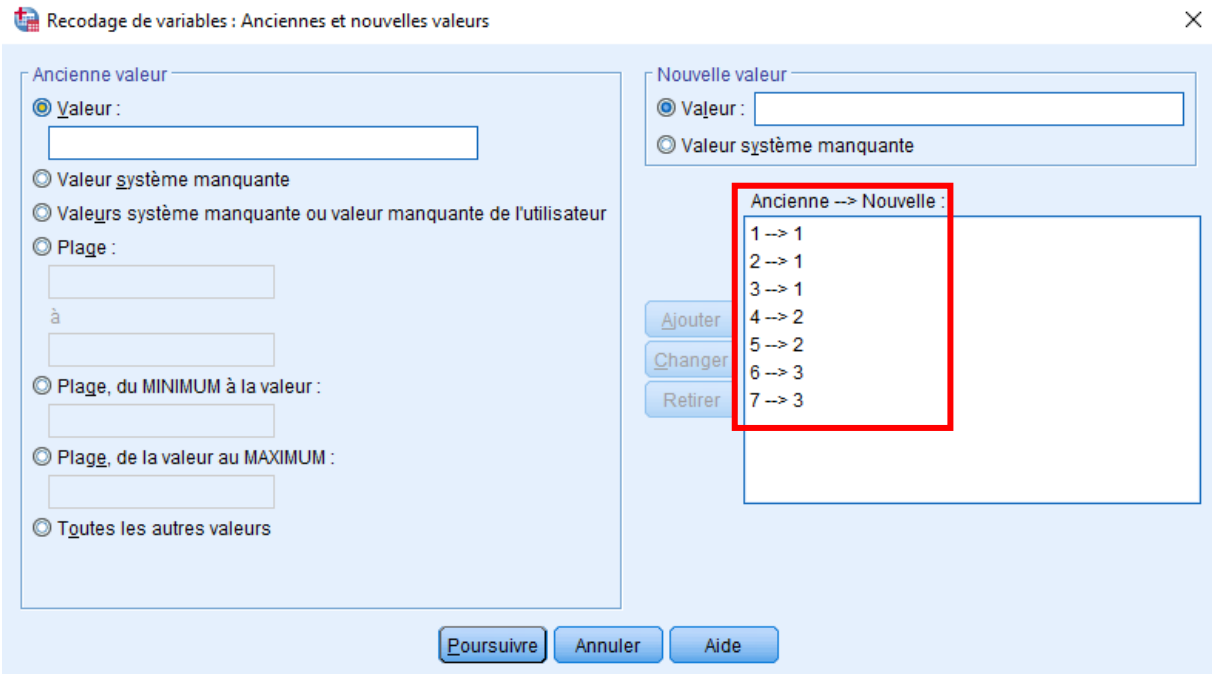


Figure 30 Recodage de la variable Revenu

4. Si on met notre base de données en mode Vue de données, on trouvera les nouvelles valeurs de la variable revenu (1, 2 et 3). Alors, on procède en dernière étape par nommer les nouvelles valeurs.

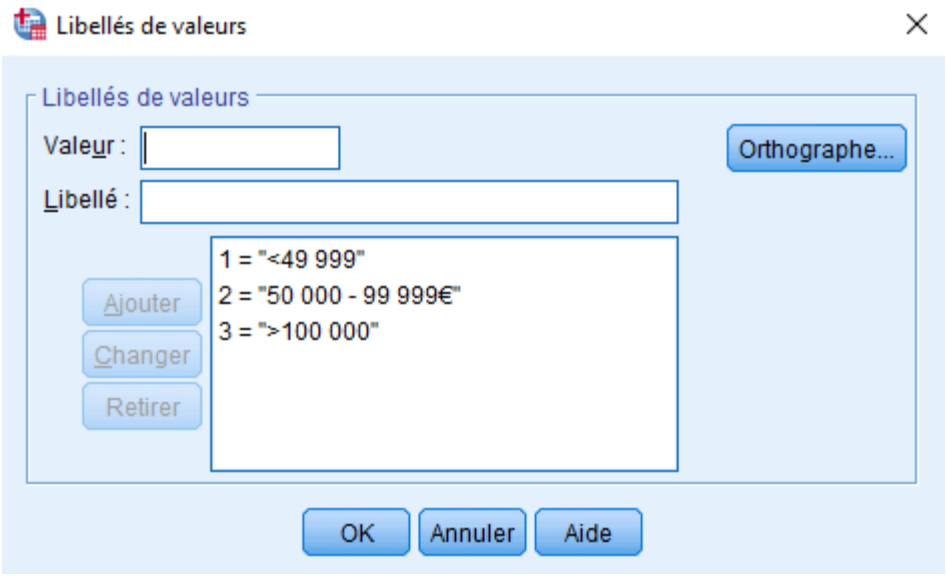


Figure 31 Nouveau recodage des valeurs de la variable Revenu

10. Créer une nouvelle variable

Procédure pour créer une variable

Cliquez sur Transformer sur la barre de menu.

- ⇨ Cliquez sur Création de variables ;
- ⇨ Sélectionnez la variable à transformer dans le rectangle de gauche ;
- ⇨ Cliquez sur la flèche pour insérer le rectangle de droite. Le nom de la variable apparaitra dans le rectangle **Variable numérique ->Variable de destination.**
- ⇨ Inscrivez dans le rectangle de droite le nom de la nouvelle variable crée, ainsi que son libellée
- ⇨ Cliquez sur **changer** ensuite sur **Ancienne et nouvelles valeurs.**
- ⇨ Donner les nouvelles valeurs (même procédure que recoder les variables).

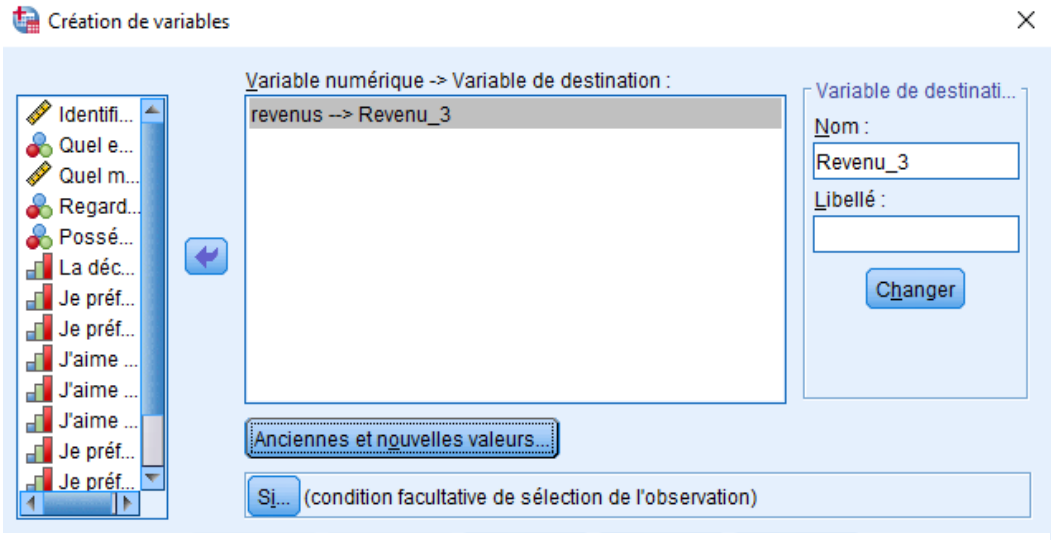


Figure 32 Boîte de dialogue création de variables

Remarque

L'option RECODAGE DE VARIABLE est aussi disponible. Toutefois, il est fortement suggéré de ne jamais recoder la variable initiale en tant que telle, car une fois les transformations effectuées, il n'est plus possible de revenir en arrière et de retrouver les valeurs telles qu'initialement colligées. Il vaut donc mieux garder une version intacte de la variable initiale et ne travailler qu'avec des « duplicata » pour tout ce qui touche les transformations des données.

IV. Calculer les statistiques descriptives d'une variable

1. Les mesures de tendances centrales et de dispersion

Les mesures de tendances centrales et de dispersion sont tout particulièrement pertinentes lorsqu'il s'agit de décrire des variables de niveau de mesure proportionnelle ou intervalle.

- Première méthode : Analyses →Statistiques descriptives→ Descriptives (*analyse descriptive simple*).
- Deuxième méthode : Analyses →Statistiques descriptives→ Fréquences (analyse exhaustive plus l'option des diagrammes).

Procédure pour obtenir les mesures de tendances centrales et de dispersion

Cliquez sur **Analyse** à partir du menu principal.

- ⇨ Cliquez sur **Statistiques descriptives,** ensuite sur Fréquences ;
- ⇨ Sélectionnez la variable à étudier.
- ⇨ Cochez sur « *afficher les tables de fréquences* » si vous voulez voir les tableaux.
- ⇨ Cliquez sur Statistiques pour générer (les fractiles, les paramètres de position, de dispersion et de forme).
- ⇨ Cliquez sur Graphiques si vous désirez afficher les graphiques circulaires, à barres, l'histogramme.
- ⇨ Enfin, cliquez sur OK.

La deuxième méthode simple pour générer les statistiques descriptives se réalise comme suit :

Analyse →Statistiques descriptives→ Descriptives.

Nota : Sur la boîte de dialogue suivante, on peut sélectionner le nom d'une variable et en cliquant sur le bouton droit de la souris, on peut afficher les noms ou bien les libellés des variables et même faire un tri des noms.

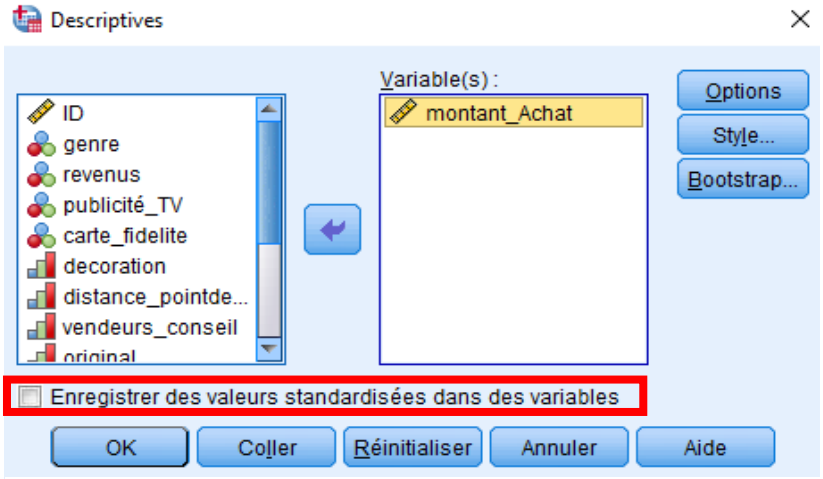


Figure 33 Boîte de dialogue pour les statistiques descriptives

Le score Z

En cochant l'option « **Enregistrer des valeurs standardisées dans des variables »** dans la boite de dialogue, IBM SPSS Statistics crée une nouvelle variable qui contient la conversion en score Z (Thierry Ancelle, 2015b). Cette nouvelle variable porte le même nom que la variable originale, mais commence par « **Z** ». Dans l'exemple ci-dessous, la nouvelle variable contenant les valeurs Z transformées se nommera **ZAGE**.

Objectif : le score Z est utilisé pour comparer une valeur à une population, si le score Z est inférieur ou égal aux valeurs de référence selon un intervalle de confiance, on peut conclure la justesse d'une méthode ou de la valeur dans le niveau de gamme considéré.

$$\boldsymbol{Score\ Z = \frac{Valeur - Moyenne}{écart - type}}$$

Le score standardisé permet de savoir à combien d'écart-type une observation se situe de la moyenne, c'est simplement l'équivalent de sa distance à la moyenne exprimée en écart-type. Lorsqu'une échelle de mesure d'un score est transformée en score Z, la **moyenne** est toujours de **0** et **l'écart type** est toujours égal à **1**. De plus, lorsque le **score brut** est au-**dessus** de la moyenne, le score Z est **positif.** Lorsque le **score brut** est au-**dessous** de la moyenne, le score Z est **négatif.**

Score Z (écart-type)	Valeur de p (Probabilité)	Niveau de confiance
<-1,65 ou >+1,65	<0, 10	90%
<-1,96 ou >+1,96	<0,05	95%
<-2,58 ou >+2, 58	<0,01	99%

Figure 34 Les intervalles de confiances du score Z

Exemple 1 :

Calcule du score Z d'un laboratoire par rapport aux 10 résultats obtenus lors d'un essai inter laboratoires : X_i =10,5. Le risque alpha est de 5%.

10,3	9,4	10,1	10,4	9,9	10,5	10,2	9,5	10,2	9,6

Solution :

$$Z_i = \frac{X_i - \bar{X}}{S}$$

$X_i = 10{,}5$; $\bar{X} = 10{,}01$; $S = 0{,}39$ et donc $Z = 1{,}256$

Conclusion :

En résumé, le score Z donne deux informations capitales sur la position d'une observation à l'intérieur d'une distribution :

1) l'observation est-elle au-dessus (+) ou en-dessous de la moyenne (-),

2) à quelle distance en écart- type se situe l'observation de la moyenne.

Le score Z est inferieur à 1,96 pour un intervalle de confiance de 95%. Donc la valeur est satisfaisante.

Exemple 2 :

Le tableau montre le nombre d'observations valides pour chaque variable choisie. La ligne « N valide » représente le nombre d'observations pour lesquelles il y a une valeur valide pour toutes les variables étudiées dans la procédure[7].

Statistiques descriptives

	N	Minimum	Maximum	Moyenne	Ecart type
Quel montant moyen dépensez-vous par mois dans ce type de point de vente ?	400	.00	444.00	153.5100	91.14782
N valide (liste)	400				

ID	montant_Achat	Zmontant_Achat
1	182.00	,31257
2	.00	-1,68419
3	216.00	,68559
4	225.00	,78433

Figure 35 Le score Z de la variable "Montant-achat"

Le premier participant a dépensé pour ses achats un montant de 182,00 dollars américains. En examinant le score Z correspondant, on peut dire que :

1) cette observation est supérieure à la moyenne, en raison de la valence positive (0,31257) de la valeur Z,
2) cette observation se situe à 0, 31257 écart-type au-dessus de la moyenne.

Le troisième participant n'a rien dépensé. Cette valeur

1) est inférieure à la moyenne,
2) se situe plus précisément à -1,68419 écart-type au-dessous de la moyenne.

Exemple 3 (comprendre l'intérêt du score Z) :

Nous avons vu précédemment que la transformation en **score Z** implique essentiellement d'exprimer la distance par rapport à la moyenne d'un score brut donné en termes d'unité d'écart-type. Comme les paramètres de distribution (moyenne et écart-type) des scores bruts varient d'un instrument à l'autre et d'un échantillon à l'autre pour un même instrument, il est très pratique d'effectuer cette transformation qui donne une information qui s'interprète toujours de la même manière, peu importe les paramètres de distribution du score brut.

L'échantillon normatif de référence évalué avec l'indice de Masse corporelle (IMC) présente une distribution ayant pour moyenne (X) 28,6516 et un écart-type (s) de 4,59175. Le seuil d'une obésité sévère est défini par un indice supérieur à 40,5 (xi).

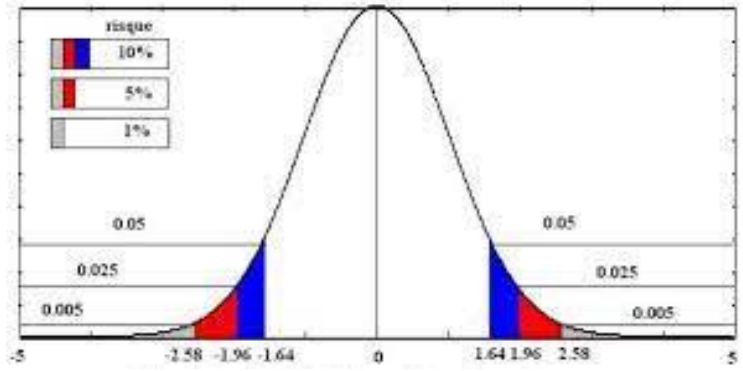
[7] Pour travailler sur la même base de données, veuillez écrire un mail à l'adresse suivante (mourad.madouni@univ-saida.dz) pour recevoir la base de données.

Comment exprimer le seuil d'obésité sévère sous forme de score Z ? Appliquons la formule présentée dans la section précédente.

$$Z_i = \frac{X_i - \bar{X}}{S} = \frac{40{,}5 - 28{,}6516}{4{,}59175} = +2{,}58$$

Un IMC de 40,5 équivaut donc à un score Z de + 2,58, ce qui revient à dire que le seuil d'obésité sévère se situe à 2,58 écart-types après la moyenne de l'échantillon normatif. C'est habituellement le seuil employé dans l'évaluation du critère d'obésité.

Grâce aux propriétés de la distribution normale, il est possible de connaître le pourcentage d'observations qui se situe au-delà du seuil de 40,5 ou de Z=+ 2,58. Ceci nous permet de savoir quel est le pourcentage de l'échantillon normatif qui présente un indice supérieur à 40,5 (ou à Z=+2,58). Le graphique suivant montre une distribution normale centrée ou standardisée. On emploie ce terme car l'échelle de mesure est transformée en score Z qui remplace l'échelle de mesure de l'indice de masse corporelle brut. On remarque que le pourcentage d'observations ayant un indice supérieur au seuil de Z=+ 2,58 est d'environ 0,5%.



Ceci revient à dire qu'il y a 0,5% de l'échantillon normatif qui présente un indice supérieur à 40,5.

1.1. La boîte de dialogue « descriptives »

Après avoir choisi sa variable pour le calcul des statistiques descriptives, en cochant sur le bouton ***options***, on peut trouver toutes les statistiques descriptives calculables de (la) ou les variables en question.

1.1.1. Erreur standard de la moyenne (ou erreur type de la moyenne) :

La plupart du temps, on ne veut pas connaître la moyenne et l'écart-type de notre échantillon, mais **estimer (statistique inférentielle)** la moyenne et l'écart-type de la population entière à partir de notre échantillon "représentatif".

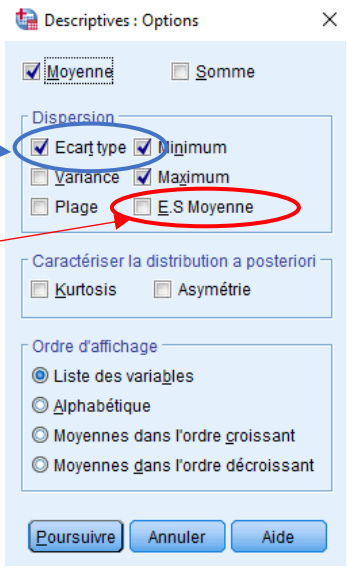
On estime la moyenne du caractère dans la population en calculant la moyenne du caractère dans l'échantillon. Donc la moyenne estimée et la moyenne sont synonymes.

Pour la variance, c'est plus compliqué : on estime la variance du caractère dans la population en mesurant la variance du caractère dans l'échantillon et en multipliant le résultat obtenu par (n-1) où n est le nombre des individus constituant l'échantillon. Il en résulte que l'erreur-type (ou standard error) se calcul comme l'écart-type mais en prenant **n-1** comme nombre de l'échantillon. Donc **l'erreur-type** est **l'estimation de l'écart-type du paramètre dans la population** (Thierry Ancelle, 2015a)**.**

ES de la moyenne= écart-type / racine carré du nombre d'individu constituant l'échantillon

$$Standard\ deviation = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

$$Standard\ Error = \frac{Standard\ Deviation}{\sqrt{n}}$$



Explication :

La première différence entre l'écart-type (standard deviation) et l'erreur-type (standard error), c'est que l'écart-type s'applique à des données, alors que l'erreur-type s'applique à la statistique de la moyenne.

A chaque fois, qu'un échantillon est pris, sa moyenne va servir à estimer la moyenne de la population. Bien-sûr, toutes les moyennes des échantillons ne sont pas identiques. Il existe une variabilité. Certains sont plus proches de la réalité que d'autre. Cette variabilité des résultats entre les échantillons est donnée par l'erreur-type. Ainsi un intervalle à l'intérieur duquel la moyenne de la population se tient pourra être estimé[8].

1.2. La boîte de dialogue « Fréquences »

A travers le bouton fréquences, on peut calculer toutes les statistiques descriptives avec plus d'options possibles comme : les quartiles, les graphiques.

Attention, les statistiques concernant la distribution des valeurs d'une variable : mesures de tendances centrales, de dispersion ou de position se calculent sur les données brutes nettoyées et à l'exclusion des valeurs manquantes.

Figure 36 La boîte de dialogue "Fréquences"

Remarque :

Points de césure pour : vous pouvez diviser votre échantillon en plus de quatre groupes. Vous n'avez qu'à identifier le nombre de groupes et vous obtiendrez les percentiles pour chacun de ceux-ci.

Centiles : vous pouvez également identifier les valeurs des centiles choisis (par exemple, 8ème, 15ème, 36ème et 72ème). Vous écrivez ces centiles dans la boîte prévue

[8] On peut calculer cet intervalle de confiance par le chemin suivant :
Analyse-> Statistiques descriptives -> Explorer.

à cet effet, puis vous cliquez sur **Ajouter**. Vous obtiendrez les valeurs des observations qui correspondent à ces centiles.

V. Les représentations graphiques

Procédure pour créer un graphique

Cliquez sur **Analyse** à partir du menu principal.

- ⇨ Cliquez sur **Statistiques descriptives,** ensuite sur Fréquences ;
- ⇨ Sélectionnez la variable pour laquelle vous désirer une représentation graphique.
- ⇨ Cliquez sur le bouton **Graphiques**.

L'écran de dialogue suivant apparaîtra :

Figure 37 L'options Graphiques à partir du bouton Fréquences

- Sélectionnez le graphique de votre choix.

  On peut choisir le graphique à barres pour les variables quantitatives discrètes et l'histogramme pour les variables quantitatives continues. Quant au graphique circulaire (parfois appelé aussi camembert, secteur), il sera approprié pour les variables qualitatives.

Remarque :

- Pour réaliser d'autres types de graphiques, choisissez plutôt l'option Graphiques du menu principal.
- Une fois le graphique choisi apparus dans la fenêtre Sortie de Résultats : *Sélectionnez le graphique en cliquant deux fois dessus, la fenêtre Editeur de graphiques devrait maintenant présenter le graphique, lequel peut être modifié à votre manière*.

1. Box-plot (boîte à moustache)

Le Box-plot est un moyen de représenter graphiquement l'aspect de la distribution d'une variable, comme la médiane et la dispersion.

1.1. Box plot simple :

- **Box plots pour une variable dans différents sous-groupes :**

Procédure pour créer un Box plot

- ⇨ Cliquez sur **Graphes**.
- ⇨ Cliquez sur **Boite de dialogue ancienne version**.
- ⇨ Cliquez sur Boites à moustaches.
- ⇨ Choisissez **« Simple »** et **« Récapitulatifs pour groupes d'observations ».**
- ⇨ Dans **Variable**, entrez la variable **quantitative** considérée pour laquelle vous voulez obtenir la médiane et la dispersion. Ex : **montant_achat**.
- ⇨ Dans l'axe des catégories, entrez la variable de groupement, par exemple le **genre**.

- ⇨ Dans **étiqueter les observations par**, si on a une variable d'identification des sujets, on peut l'y entrer pour identifier les valeurs extrêmes (quand on laisse cette option vide, IBM SPSS Statistics utilise le numéro de ligne). Ex : numéro de la fiche du questionnaire.
- ⇨ Dans lignes et colonnes on glisse d'autre variable qu'on veut étudier autre que la variable qui est dans l'axe des modalités. Il est toutefois important de noter que vous obtiendrez beaucoup d'informations et que ces dernières rendent l'interprétation du graphique plus *complexe*.
- ⇨ Si vous cochez variables emboitées ou non, vous aurez le même graphe dans les deux cas.

Exemple

1 **2**

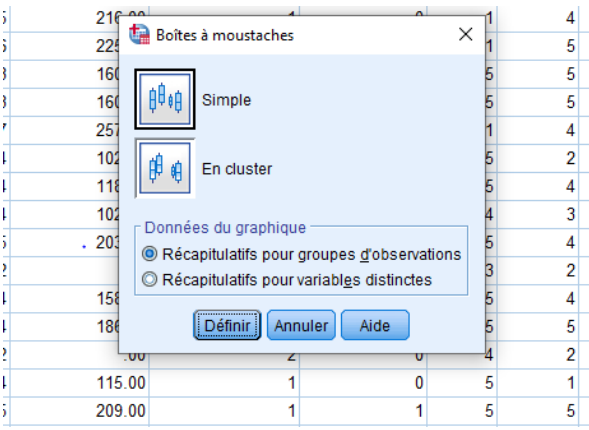


Figure 38 La boîte de dialogue "Boîte à moustaches"

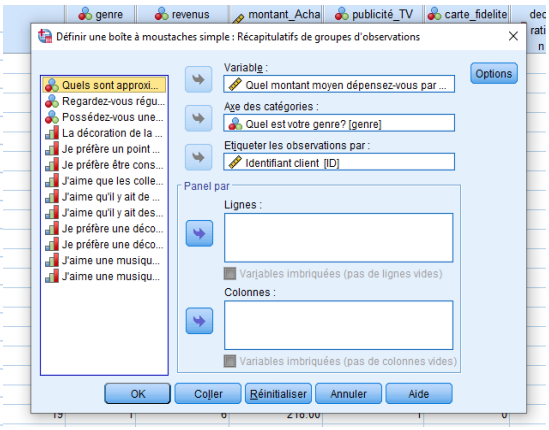
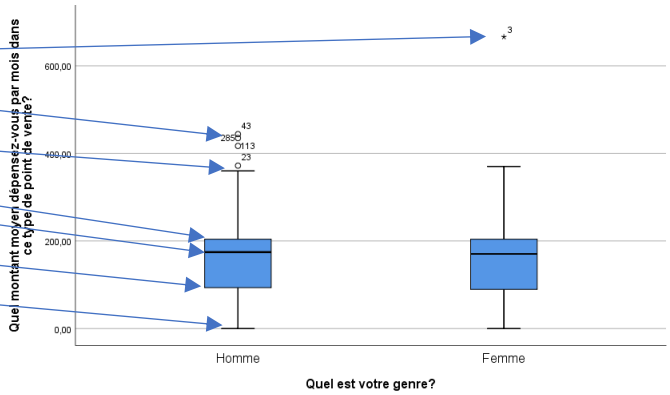


Figure 39 Récapitulatif pour groupes d'observations

1- Les valeurs extrêmes[9].
2- Les valeurs aberrantes.
3- Le maximum.
4- Le 1er quartile.
5- Le 2ème quartile (la médiane)
6- Le 3ème quartile.
7- Le minimum.



[9] Les valeurs extrêmes sont désignées par une étoile dans le box plot et les valeurs aberrantes sont représentées par des cercles. Toutes les valeurs qui se trouvent au-delà de 1,5 de l'intervalle interquartile sont des valeurs aberrantes, alors que les valeurs extrêmes se situent au-delà de 3 fois l'intervalle interquartile (Simple Learning Pro, 2015) (Tukey, 1972).

- **Box plots pour plusieurs variables dans tout l'échantillon :**

Procédure pour créer un Box plot

- ⇨ Cliquez sur **Graphes**.
- ⇨ Cliquez sur **Boite de dialogue ancienne version**.
- ⇨ Cliquez sur Boites à moustaches.
- ⇨ Choisissez **« Simple »** et **« Récapitulatifs pour variables distinctes ».**
- ⇨ Dans « **les cases représentent »**, entrez la/les variable(s) **quantitative(s)** considérée(s)
- ⇨ pour laquelle vous voulez obtenir la médiane et la dispersion. Ex : **montant_achat**.

1

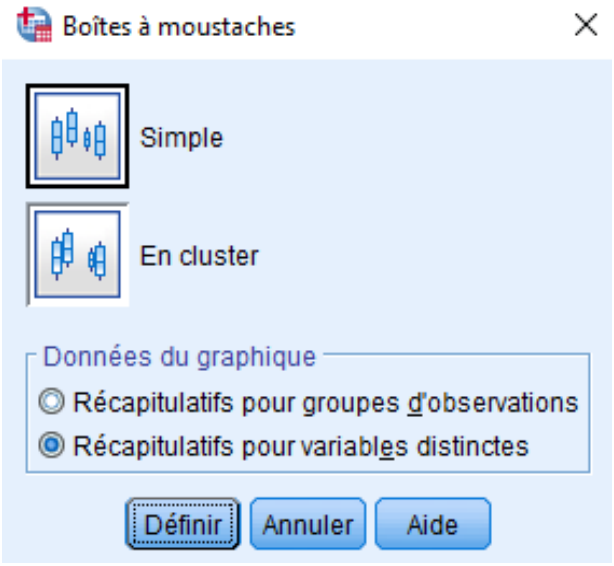


Figure 40 La boîte de dialogue pour la création du box plot

2

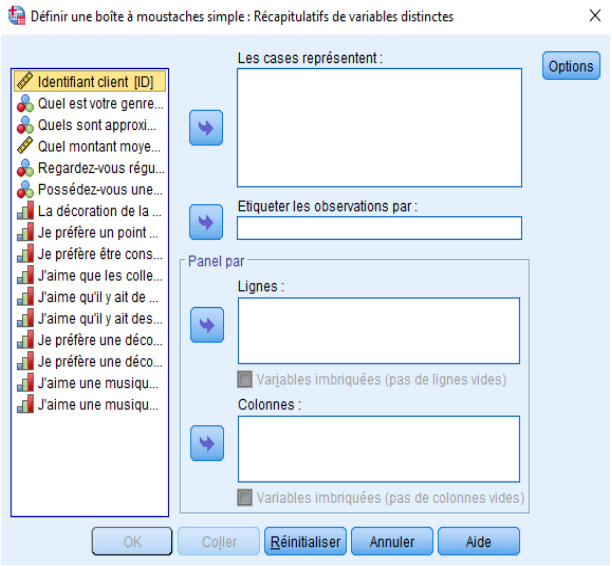


Figure 41 Récapitulatifs pour variables distinctes

1.2. Box plot en cluster

- **Box plots pour une variable dans différents sous-groupes :**

Procédure pour créer un Box plot pour une variable dans différents sous-groupes

- ⇨ Cliquez sur **Graphes**.
- ⇨ Cliquez sur **Boite de dialogue ancienne version**.
- ⇨ Cliquez sur Boites à moustaches.
- ⇨ Choisissez **« En cluster[10] »** et **« Récapitulatifs pour groupes d'observations ».**
- ⇨ La deuxième boite de dialogue s'ouvre.
- ⇨ Insérez **la variable continue** à décrire dans la boite **Variable**.
- ⇨ Insérez la variable **catégorielle (des catégories)** dans **l'Axe des catégories**.
- ⇨ Insérez la variable qui sépare les groupes en sous-groupes dans la boite **Définir les clusters par.**
- ⇨ Vous pouvez choisir une variable qui remplacera le numéro de l'observation sur les étiquettes (**Étiqueter les observations par**).

1

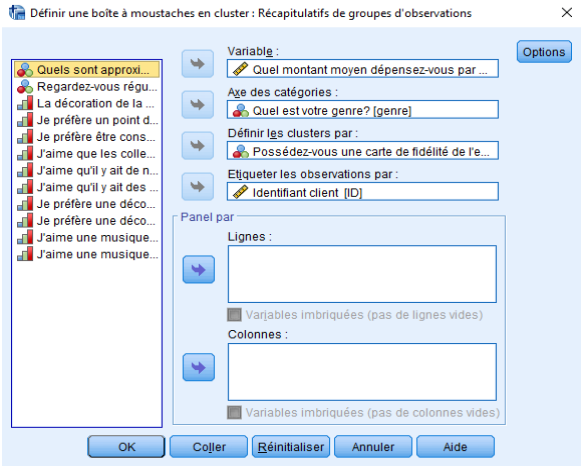


Figure 43 La boîte de dialogue pour la création de box-plot en cluster

2.

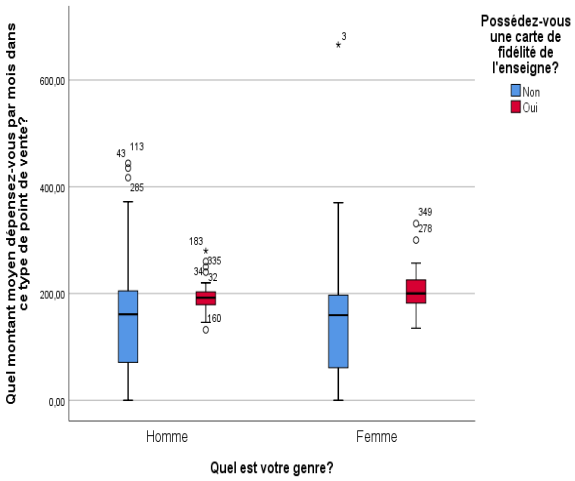


Figure 42 Le box-plot en cluster

Interprétation :

Pour les clients qui détiennent une carte de fidélité, la médiane est inférieure à 200 dollars, ce qui suppose une distribution asymétrique vers les valeurs basses des montants dépensés sur le point de vente. De plus, la médiane semble généralement plus basse pour les hommes que pour les femmes, ce qui indique que les hommes dépensent moins d'argent.

L'écart interquartile varie en fonction de la possession de la carte de fidélité. On note une grande variabilité des dépenses pour ceux qui possèdent une carte par rapport à autres.

[10] Dans les anciennes versions de IBM SPSS Statistics, les boîtes à moustaches en cluster sont dénommés juxtaposés.

Enfin, le graphique montre la présence de plusieurs valeurs extrêmes et aberrantes. Il est possible que certaines soient dues à des erreurs d'entrées de données.

1.3. Quelles informations retire-t-on du graphique (Box plot) ?

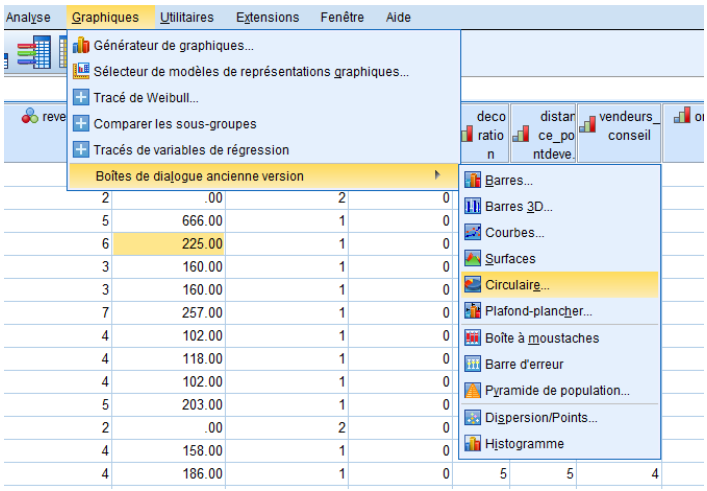
Nous pouvons avoir une idée de la tendance centrale des valeurs de chaque boite en observant la position de la médiane. Si la médiane n'est pas au centre, on peut juger la symétrie de la distribution (aplatissement et asymétrie).

Par la longueur de la boite, il est possible d'estimer la variabilité des valeurs pour chaque sous-groupe. Enfin, la longueur des « moustaches » donne une idée de la taille de la queue de la distribution.

2. Diagramme circulaire

Pour créer un diagramme circulaire, cliquez sur **Graphes** dans la barre d'outils, puis sur **Boites de dialogue ancienne version** dans le premier menu déroulant. Ensuite, sélectionnez **Circulaire** dans le deuxième.

1



2

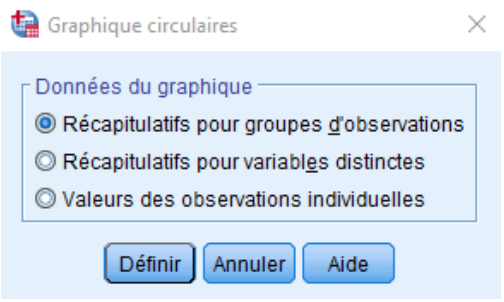


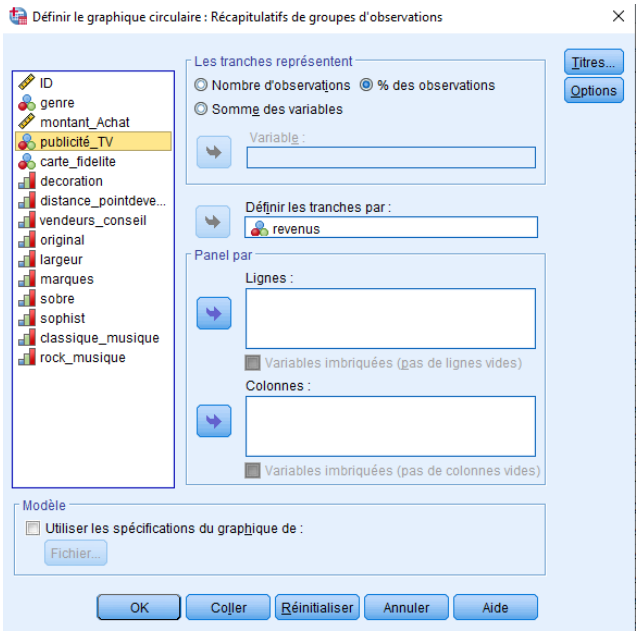
Figure 44 La boîte de dialogue pour la création des graphiques circulaires

Figure 45 La procédure Graphiques

1-Récapitulatifs pour groupes d'observations : les pointes illustrent le nombre d'occurrences en proportion pour chaque valeur de la variable catégorielle. C'est l'option par défaut.

Vous cliquez ensuite sur le bouton et vous allez voir une deuxième boite de dialogue.

Vous insérez la variable à décrire dans **Définir les tranches par.**



Vous pouvez choisir plus précisément ce que vous voulez que les pointes (les parts) représentent :

- Le Nombre d'observations ou le % d'observations : vous obtenez le même graphique dans les deux cas.
- Si vous voulez personnaliser votre graphe (couleurs, titre, etc.), cliquez deux fois le graphique circulaire.

Remarque : Généralement, on choisit la présentation en pourcentage : % des observations.

Interprétation d'un exemple :

L'interprétation du diagramme circulaire est relativement simple. Dans le graphique suivant, vous pouvez observer que 15% des répondants ont un salaire inférieur à 25.000 euros. Ensuite, vous voyez que les répondants qui disposent d'un salaire moyen (entre 25.000 et 100.000) représentent près de 69%, alors que les plus aisés représentent 16,5%.

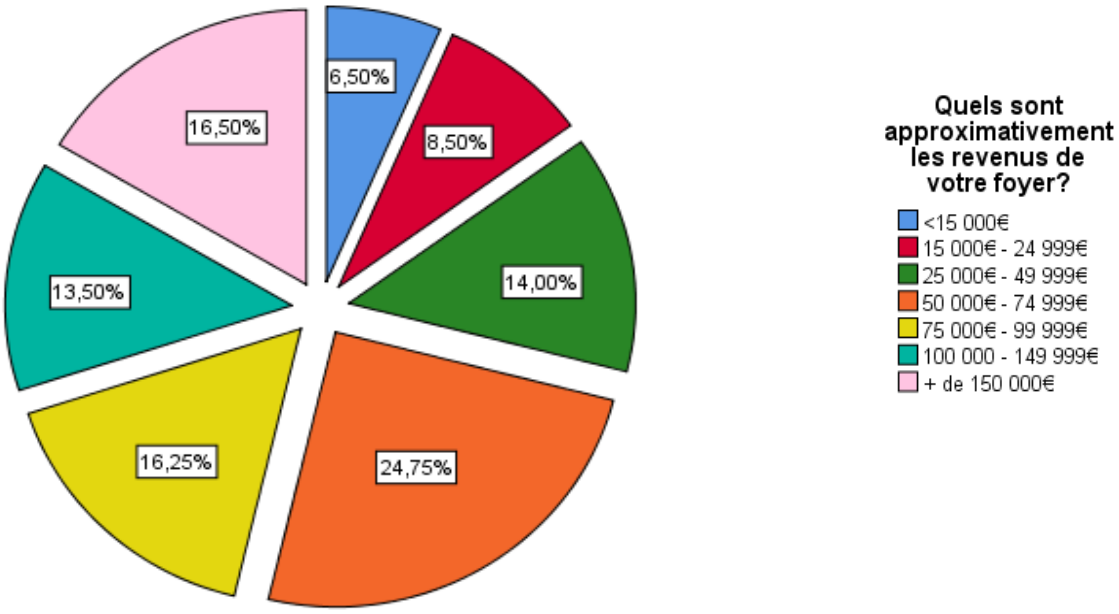


Figure 46 Graphique circulaire représentant la variable "Revenu"

2-Récapitulatifs pour variables distinctes : les pointes (les parts) montrent la somme de toutes les observations pour les variables choisies.

3-Valeurs des observations individuelles : le graphique présente chaque observation.

VI. Scinder un fichier

Scinder un fichier[11] vous permet de scinder un fichier de données en plusieurs classes (catégories), en fonction des valeurs d'une ou plusieurs variables de regroupement. A partir du moment où vous avez scinder (séparé) votre fichier, tous les résultats demandés sur IBM SPSS Statistics seront séparés selon les valeurs de la variable de regroupement (Kent State University Libraries., 2022e). Par exemple, si vous scinder votre fichier par la variable ***genre***, tous les commandes pour les statistiques et les graphiques seront présentés en fonction du genre.

Tableau 3 Les statistiques descriptives de la variable Montant dépensé en fonction du genre

Quel est votre genre ?		N	Minimum	Maximum	Moyenne	Écart type
Homme	Quel montant moyen dépensez-vous par mois dans ce type de point de vente ?	204	.00	444.00	155.8922	95.30929
	N valide (liste)	204				
Femme	Quel montant moyen dépensez-vous par mois dans ce type de point de vente ?	196	.00	370.00	151.0306	86.77929
	N valide (liste)	196				

Si vous sélectionnez plusieurs variables de regroupement, les observations sont regroupées pour chaque variable au sein des modalités de la variable précédente dans la liste des variables de regroupement.

Par exemple, si vous sélectionnez *genre* comme première variable de regroupement, les observations seront classées en fonction de chaque modalité de la variable genre.

• Vous pouvez spécifier jusqu'à 8 variables de regroupement.

• Les observations doivent être triées en fonction des valeurs des variables de regroupement, suivant l'ordre dans lequel les variables sont présentées dans la liste des variables de regroupement. Si le fichier de données n'est pas trié, sélectionnez **Tri suivant les variables de regroupement**.

[11] Pour exécuter la commande scinder un fichier par la syntaxe, on écrit : Split file by « nom de la variable de séparation » et pour enlever la séparation, on écrit : SPLIT FILE off.
En outre, pour pouvoir vérifier s'il y a une séparation (si le fichier est scindé) on écrit SHOW SPLIT.

Procédure pour scinder un fichier

- ⇨ Cliquez sur **Données**.
- ⇨ **Scinder un fichier.**
- ⇨ Cliquer sur **comparer les groupes ou organiser la sortie par groupes**[12] (les deux options donnent les mêmes résultats).
- ⇨ Sélectionner la variable considéré dans critère de regroupement (une variable **qualitative**).
- ⇨ Choisir trier le fichier par variables de regroupement ou le fichier est déjà trié (les deux options donnent les mêmes résultats).
- ⇨ Cliquer sur OK.
- ⇨ Les résultats (statistiques, graphiques, etc.) seront affichés séparés dans la fenêtre Sortie.

VII. Pondérer les observations

La pondération des données vous permet d'associer un poids à chaque observation, c'est-à-dire les individus, plutôt que d'être affecté d'un même poids dans les analyses, se voient attribués des poids différents selon le genre et la catégorie d'âge auxquels ils appartiennent par exemple. (Kent State University Libraries., 2022g)

Exemple :

Si on dispose de ce tableau croisé entre la variable « genre » et une autre variable nominale « possédez-vous une carte de fidélité ?».

	Oui, je possède une carte	Non, je ne possède pas une carte	Total
Femmes	15	30	45
Hommes	14	21	35
Total	29	51	80

Les variables sur IBM SPSS Statistics seront codées comme suit :

- La variable genre est codée sur IBM SPSS Statistics: 1 : Femmes, 2 : Hommes.
- La variable « carte de fidélité » est codée par 1 : Oui, 2 : Non.
- On ajoute une variable « Fréquence » pour inscrire les effectifs du tableau croisé.

[12] La seule différence entre les deux options est comme suit : La première option nous sépare les résultats dans un même tableau alors que la deuxième option « Organiser la sortie par groupe » nous donne les résultats dans des tableaux séparés.

1. Vue des variables

Nom	Type	Largeur	Valeurs	Manquant	Colonnes	Align	Mesure
Genre	Numérique	8	{1,00, Fem...	Aucun	8	Droite	Nominales
Carte_fidélité	Numérique	8	{1,00, Oui}...	Aucun	8	Droite	Nominales
Fréquence	Numérique	8	Aucun	Aucun	8	Droite	Echelle

2. Vue de données

Genre	Carte_fidélité	Fréquence
Femmes	Oui	15,00
Femmes	Non	30,00
Hommes	Oui	14,00
Hommes	Non	21,00

Pour terminer, il faut pondérer les observations par fréquences pour permettre à IBM SPSS Statistics de relier les données avec les fréquences.

Procédure pour pondérer les observations

- ⇨ Cliquez sur **Données** =>
- ⇨ Pondérer les observations =>
- ⇨ Il faut cocher sur « Pondérer les observations par valeur de fréquence : »
- ⇨ Insérer la variable « fréquence » dans notre exemple.
- ⇨ OK =>
- ⇨ Désormais, on peut réaliser tous les tests statistiques[13].

VIII. Sélectionner les observations

La sélection des données *vous* permet de faire apparaitre (ou filtrer) uniquement les données que vous désirez ainsi que toute votre analyse statistique et graphique sur le set de données que vous avez choisi. (Yergeau & Poirier, 2021c)

Procédure pour sélectionner les observations

- ⇨ Cliquez sur **Données** =>
- ⇨ Sélectionner des observations =>
- ⇨ Une nouvelle boîte de dialogue s'affiche avec plusieurs possibilités de sélection.
- ⇨ Selon une condition logique (ex : si on veut sélectionner uniquement les femmes et ces dernières sont codées par 1, on met en condition Genre=1).
- ⇨ OK =>

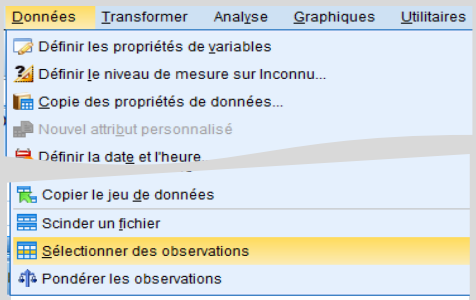


Figure 47 Le menu déroulant pour sélectionner les observations

[13] L'option « pondérer les observations » nous permet de lier les fréquences avec les autres variables sinon la variable fréquence sera considérée comme une variable indépendante comme l'âge ou le poids par exemple.

Il existe plusieurs options pour sélectionner des observations :

- La première option est la plus couramment utilisée : **Selon une condition logique.** Tout d'abord, il faut cliquer sur le bouton **Si** ensuite une nouvelle boîte de dialogue va s'afficher afin d'énoncer les conditions.

- Vous pouvez taper les conditions arithmétiques (<, >, =, <=, >=, <>) et les nombres.

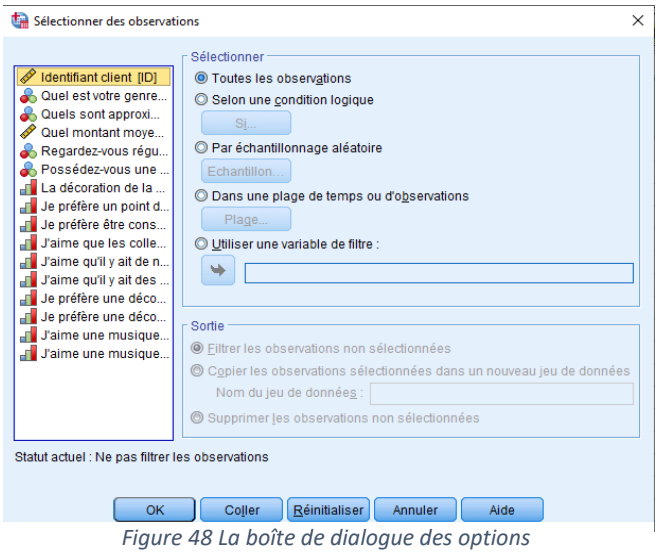


Figure 48 La boîte de dialogue des options pour Sélectionner des observations

Exemple : Pour choisir les femmes de plus de 20 ans, on entrerait dans la boîte :

Genre = 1 AND[14] age > 20

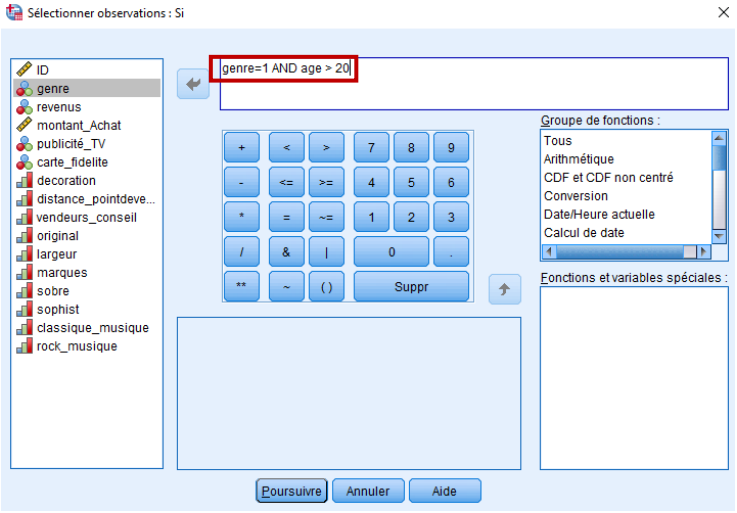


Figure 49 La boîte de dialogue de la condition logique

- Les autres options de sélection

- **Toutes les observations** : comme son nom l'indique vous utiliser toutes les observations.
- **Par échantillonnage aléatoire** : cette option procède par une sélection aléatoire des observations, soit en pourcentage (*Environ __ % de toutes les observations*), soit en précisant un nombre d'observations parmi les premières observations (*Exactement _ observations à partir des premières _ Observations*).

[14] L'opérateur AND implique que toutes les conditions doivent être remplies pour l'expression globale soit vraie. Par contre, l'opérateur OR implique au moins une des conditions doit être remplie pour que l'expression soit vraie.

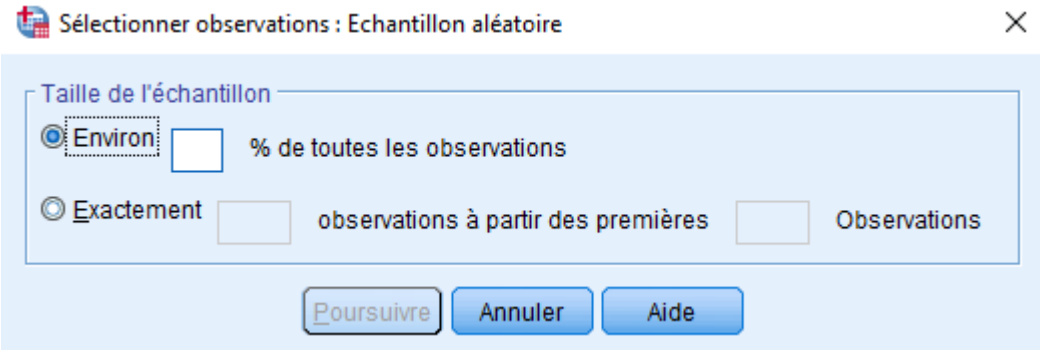


Figure 50 L'option de la sélection : échantillonnage aléatoire

- **Dans une plage de temps ou d'observations :** cette option nous permet de sélectionner des observations dans un intervalle de temps ou d'observations. Par exemple, on peut sélectionner uniquement les observations qui se situent entre la 5ème observation et la 17ème observation.
- **Utiliser une variable de filtre** : A travers cette technique, on peut procéder à une sélection en ne gardant que les observations qui ont des valeurs valides pour une variable filtre, par exemple : *poids*.

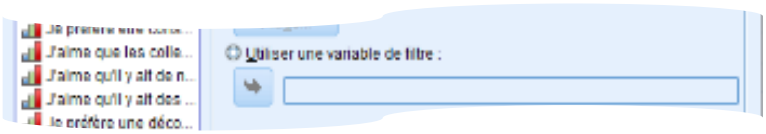


Figure 51 L'option : utiliser une variable de filtre

IX. Calculer une variable

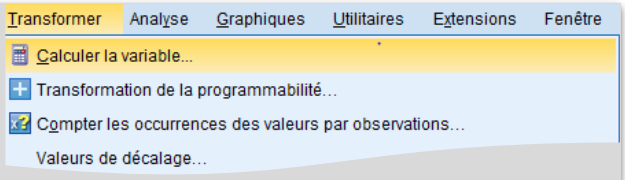
IBM SPSS Statistics nous offre une autre possibilité très intéressante pour calculer une variable. Par exemple :

- Calculer le score total d'une échelle de mesure (eg :FINDRISC[15] score pour les sciences médicales) ;
- Calculer la moyenne ou la somme d'une série de variables existantes (eg : calculer la moyenne des notes de chaque étudiant, calculer l'indice de masse corporelle pour chaque patient).
- Créer une variable qui contient la puissance ou la racine carrée d'une variable existante.
- Calculer des dates. Etc.

La commande **Calculer la variable** ou **COMPUTE** sert à *créer de nouvelles variables* sur la base de fonctions arithmétiques, statistiques ou logiques (Yergeau & Poirier, 2021a) (Kent State University Libraries., 2022c).

[15] Finish Diabetes Risk Score (https://canadiantaskforce.ca/tools-resources/diabete-de-type-2/diabete-de-type-2-findrisc-pour-cliniciens/?lang=fr). Site consulté le 04/03/2022.

Procédure pour calculer une variable



- ⇨ Cliquez sur **Transformer** =>
- ⇨ Calculer la variable =>
- ⇨ Une nouvelle boîte de dialogue s'affiche pour insérer une fonction avec la possibilité d'ajouter une condition au calcul.
- ⇨ Poursuivre =>
- ⇨ OK =>

Exemple :

Admettant qu'on dispose d'une classe de 13 élèves qui ont passé trois examens et on veut calculer la moyenne de chaque personne.

1 : On doit d'abord donner un nom à cette nouvelle variable. On peut également spécifier le type de la variable (numérique ou chaine) ainsi que son libellé dans le bouton en bas du nom **Type et libellé…**

2 et 3 : Dans la case **Expression numérique,** on introduit notre formule de calcul en utilisant la calculatrice. Dans le cas du calcul de la moyenne des élèves, on écrit **(note1+note 2+note3)/3**

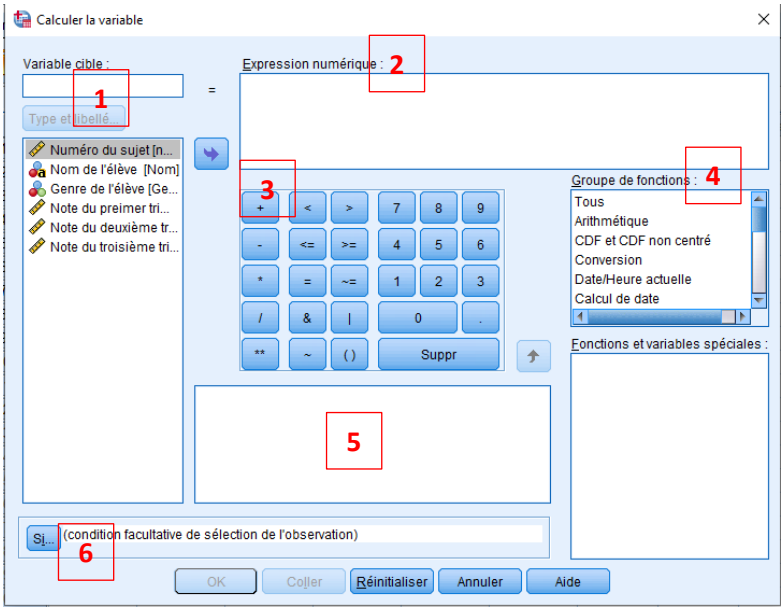


Figure 52 Boîte de dialogue : Calculer la variable

4 : IBM SPSS Statistics nous offre à travers la case **Groupe de fonctions** une autre possibilité pour calculer une variable. Dans le cas du calcul de la moyenne des élèves, on sélectionne **Statistiques** et on choisit la fonction Mean dans **Fonction et variables spéciales**. On fait un double clique sur Mean ensuite on insère les arguments (note1, Note2 et Note 3) à la place des points d'interrogation.

5 : Ce champ nous donne une description détaillée de chaque fonction choisie.

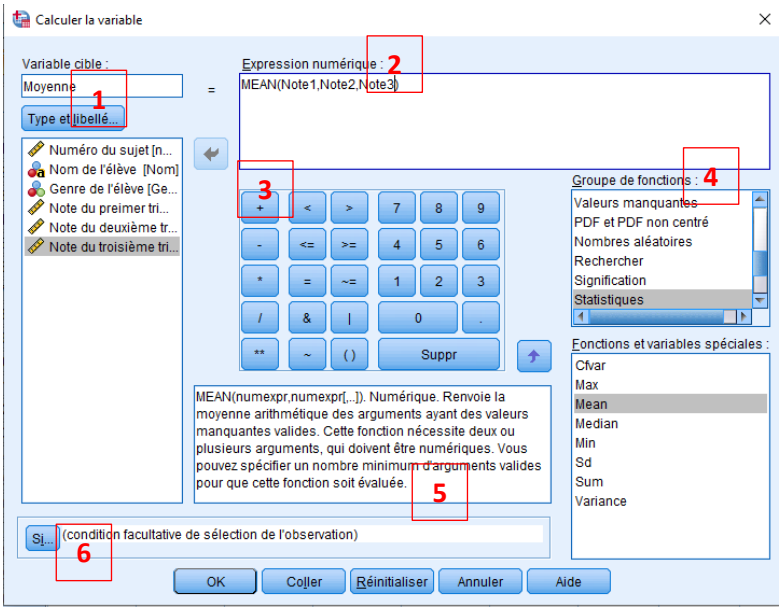


Figure 53 Boîte de dialogue : calculer la variable

6 : Le bouton **SI** nous permet d'inclure une condition dans le calcul de la nouvelle variable. Par, exemple, si on veut calculer la moyenne uniquement pour les filles. Il faut cocher **Inclure lorsque l'observation remplit une condition** ensuite on introduit notre condition (genre=2).

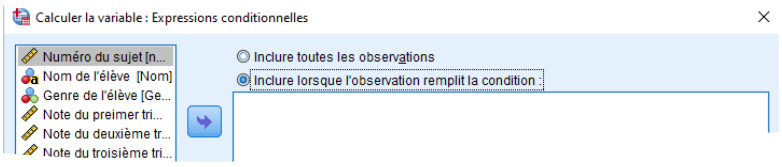


Figure 54 La boîte de dialogue : Expression conditionnelle SI

Nota : calcul d'une moyenne en présence de valeurs manquantes

Il faut noter que les trois méthodes suivantes sont identiques dans le cas où on ne dispose pas de valeurs manquantes. Cependant, en présence de valeurs manquantes, nous aurons des résultats différents (van den Berg, 2022).

- Première méthode **(Note1+Note2+Note3)/3**
  Avec cette méthode, IBM SPSS Statistics ne donne aucun résultat.
- **SUM(Note1,Note2,Note3)/3**
  Dans ce cas, la moyenne calculée prend les valeurs manquantes comme étant des zéros.
- **Mean(Note1,Note2,Note3)**
  Cette méthode ne prend pas en considération les cases des valeurs manquantes.

Le bouton double étoile qui se trouve dans la calculatrice signifie « la fonction puissance ». Par exemple, dans le calcul de l'indice de masse corporelle (IMC). On peut mettre dans l'expression numérique ou bien dans la syntaxe :

```
DATASET ACTIVATE Jeu_de_données2.
COMPUTE IMC=Poids/Taille ** 2.
EXECUTE.
```

X. Les tableaux croisés

Le tableau croisé appelé aussi parfois « tableau de contingence » examine la relation entre deux variables catégorielles. Il décrit donc la ventilation de chaque catégorie d'une variable en fonction d'une autre variable catégorielle (Kent State University Libraries., 2022d) (Yergeau & Poirier, 2021d). L'expression tableau de contingence a été introduite par le statisticien britannique Karl Pearson.

Procédure pour générer un tableau croisé

- ⇨ Cliquez sur **Analyse** =>
- ⇨ Statistiques descriptives
- ⇨ Tableaux croisés=>
- ⇨ Insérer la variable indépendante dans **ligne(s)**
- ⇨ Insérer la variable dépendante dans **Colonne(s)**
- ⇨ Dans le bouton **Statistiques** à droite, on trouve tous les tests statistiques (Khi deux, coefficients de contingence, etc.)
- ⇨ Le bouton **Cellules** est utilisé pour afficher les cellules en pourcentage ou en effectifs.

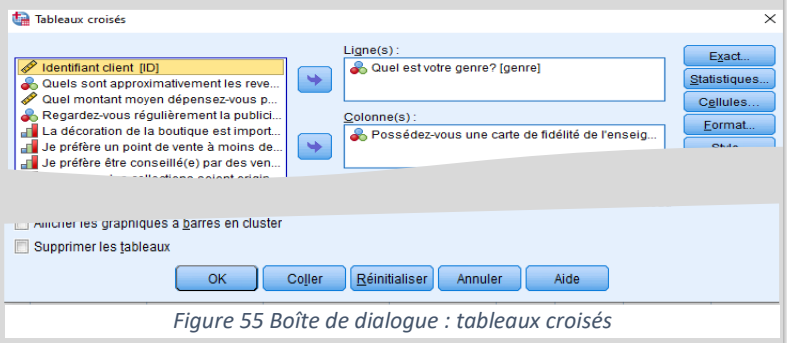


Figure 55 Boîte de dialogue : tableaux croisés

⇨ On peut également afficher les graphiques à barres.
⇨ OK =>

En prenant l'exemple du fichier « Etude de marché sur un point de vente ». En choisit de croisé la variable « genre » avec « possédez-vous une carte de fidélité ? ». Deux variables qualitatives. Voici le tableau croisé 2x2.

*Tableau 4 Tableau croisé: Genre * Carte de fidélité*

		Possédez-vous une carte de fidélité de l'enseigne?		
		Non	Oui	Total
Quel est votre genre?	Homme	169	35	204
	Femme	152	44	196
Total		321	79	400

Remarque :

Le bouton **EXACT…** n'est pas vraiment utilisée pour les statistiques descriptives.

L'option **Statistiques** est utilisée pour les statistiques inférentielles. Elle permet de choisir le type de test que vous désirez utiliser afin d'évaluer s'il y a des différences significatives entre les groupes.

Troisième chapitre
Les exercices d’application

Exercice N°01 – Codification des données, statistiques descriptives, comparer les moyennes et trier les observations et les variables

On veut faire une étude statistique sur une classe de 12 élèves. La série statistique est la suivante :

N°	Noms	Genre	Note
1	F	F	10
2	G	G	8
3	H	G	12
4	R	F	13
5	Z	G	15
6	A	F	18
7	S	G	9
8	F	G	15
9	H	G	7
10	K	F	13
11	L	F	9
12	M	F	16

1- Quelles sont les variables de cette série statistique ?
2- Définissez les types de variables de cette série statistique et préciser leurs modalités ?
3- Entrez les variables de cette série statistique sur le logiciel IBM SPSS avec codification.
4- Entrez les données de cette série statistique sur le logiciel IBM SPSS.
5- Calculer par le logiciel la moyenne, la médiane, l'écart type, le minimum, le maximum et les tableaux d'effectifs de la variable considérée.
6- Créer le tableau de fréquences pour la variable **Genre**. Détaillez les étapes.
7- Le test de différence de moyenne peut se faire ? Expliquer.

Si oui, Comparez les moyennes (moyennes et écart type). Interprétez.

8- Dans un nouveau fichier, trier les observations (note par ordre croissant).
9- Dans un nouveau fichier, trier les variables par Nom.

Solution de l'exercice N°01

1- On à deux variables : le genre et la note.
2- Définition et modalités :

Variables	Type	Modalités
Genre	Qualitative dichotomique nominale	2 modalités : masculin ou féminin
Note	Quantitative discrète	La note varie entre 0 et 20.

3- Codage dans IBM SPSS Statistics

Nom	Type	Largeur	Décimales	Libellé	Valeurs	Manquant	Colonnes	Align	Mesure
Prénom	Chaîne	1	0		Aucun	Aucun	8	Gauche	Nominales
Genre	Numérique	1	0		{1, Féminin}...	Aucun	8	Gauche	Nominales
Note	Numérique	4	2		Aucun	Aucun	8	Gauche	Echelle

4- Voir IBM SPSS.

5- La variable **genre** est Qualitative nominale donc il est impossible de calculer les paramètres de position et de dispersion.

La variable **note** est quantitative donc il est possible de calculer les paramètres de position et de dispersion.

Statistiques descriptives : Note						
N Valide	Manquante	Moyenne	Médiane	Ecart-type	Minimum	Maximum
12	**0**	**12,0833**	**12,5000**	**3,50216**	**7,00**	**18,00**

A partir de la fenêtre éditeur des données, on va sélectionner

Analyse => Statistiques descriptives =>Effectifs =>Glisser la variable **note** =>Statistiques Cocher moyenne, médiane, minimum, maximum, écart-type => Poursuivre =>OK

On trouve les résultats dans la fenêtre Sortie.

6-A partir de la fenêtre éditeur des données, on va sélectionner :

- Analyse =>Statistiques descriptives =>Effectifs

Cocher la case **afficher les tableaux d'effectifs**.

Décocher la moyenne, médiane, minimum, maximum, écart-type qui était sélectionnée précédemment =>Poursuivre =>Ok.

7-On trouve les résultats dans la fenêtre Sortie. **Genre**

	Effectif s	Pourcentage	Pourcentage valide	Pourcentage cumulé
Féminin	6	50	50	50
Masculin	6	50	50	100
Total	12	100	100	

Oui, le test de comparaison de moyennes peut se faire puisqu'on a :

- Une variable DÉPENDANTE de type PROPORTIONNELLE (QUANTITATIVE) qui est la **note**.
- Une variable INDÉPENDANTE de type NOMINALE DICHOTOMIQUE (c'est-à-dire présentant seulement deux modalités ou catégories) qui est le **genre**.

- Comparaison des moyennes :

A partir de la fenêtre éditeur des données on va sélectionner :

- Analyse => Comparer les moyennes =>Moyennes. =>Sélectionner la variable dépendante. =>Sélectionner la variable indépendante =>Option =>Vérifier que la moyenne, écart type et le nombre d'observations sont intégrées. Poursuivre => Ok.

- Les résultats sont dans la fenêtre Sortie.

Genre	Moyenne	N	Ecart-type
Féminin	13,1667	6	3,43026
Masculin	11,0000	6	3,52136
Total	12,0833	12	3,50216

Commentaire :

- La moyenne du genre féminin est de 13,16 est légèrement supérieur à celle du genre masculin de valeur 11,00.
-Les écart-types sont presque égaux pour les 2 genres.
-La moyenne totale 12,08 et l'écart-type pour les 2 genres 3,50.

8- Dans la fenêtre Editeur de données cliquer sur :

Données =>Trier les observations =>Glisser la variable dans Trier par (Note) =>Choisir ordre croissant =>Ok. Voir la **fenêtre Editeur de données**

9- Dans la fenêtre Editeur de données cliquer sur :

Données =>Trier les variables =>Glisser la variable dans Tier par =>Choisir le paramètre de tri dans colonnes d'affichage des variables (Nom) =>Choisir ordre croissant =>Ok. Voir la **fenêtre Editeur de données**

Exercice N°02 – Codification, tableau d'effectif et graphique

Voici un questionnaire :

1 Genre

- Homme
- Femme

2 Quel est votre âge ?

……..années

4 Niveau d'instruction

- Universitaire
- Lycéen
- Sans niveau

3 Situation familiale

Célibataire

Marié

5 Profession

- Fonctionnaire
- Retraité
- Autres

6 Êtes-vous satisfaits de l'état de votre magasin ? (Classez par ordre de 1 à 4).

- Très insatisfait
- Plutôt insatisfait

-Plutôt satisfait

- Très satisfait

Voilà les données :

N	Genre	Age	Situation familiale	Niveau d'instruction	Profession	Satisfaction maison
1	Homme	30	Célibataire	Universitaire	Fonctionnaire	Très insatisfait
2	Homme	25	Célibataire	Lycéen	Fonctionnaire	Très insatisfait
3	Femme	26	Célibataire	Lycéen	Fonctionnaire	Plutôt insatisfait
4	Homme	29	Marié	Lycéen	Fonctionnaire	Plutôt satisfait
5	Homme	34	Marié	Sans niveau	Autre	Très satisfait
6	Femme	52	Marié	Universitaire	Retraité	Très satisfait
7	Femme	60	Marié	Universitaire	Retraité	Très satisfait
8	Femme	41	Célibataire	Sans niveau	Fonctionnaire	Plutôt satisfait
9	Homme	43	Célibataire	Sans niveau	Aucun	Plutôt insatisfait
10	Homme	62	Marié	Universitaire	Retraité	Très insatisfait

1-Codifier le questionnaire en précisant le type de chaque variable.
2-Entrez les variables dans SPSS.
3- Entrez les données dans SPSS.
4-Analyser les effectifs de toutes les variables (création des tableaux de fréquences).
5-Tracer les graphs de toutes les variables avec explication du choix du graphe.

Solution de l'exercice N°02

1- Codification du questionnaire :

Genre (variable qualitative nominale)

- Homme 1
- Femme 2

Quel est votre âge ? (variable quantitative continue)

……..années

Niveau d'instruction (variable qualitative ordinale)

- Universitaire 1
- Lycéen 2
- Sans niveau 3

Situation familiale (variable qualitative nominale)

- Célibataire 1
- Marié 2

Profession (variable qualitative nominale)

- Fonctionnaire 1
- Retraité 2
- Autre 3

Êtes-vous satisfaits de l'état de votre maison ? (Variable qualitative ordinale)

- Très insatisfait 1
- Plutôt insatisfait 2
- Plutôt satisfait 3
- Très satisfait 4

2- Voir SPSS.

3- Voir SPSS.

4- Analyse des effectifs des variables :

- A partir de la fenêtre éditeur des données on va sélectionner
- Analyse
- Statistiques descriptives
- Effectifs
- Sélectionner les variables
- Cocher les tableaux d'effectifs
- OK.

Tous les tableaux d'effectif sont récapitulés dans la fenêtre Sortie.

Tableau 1 : Genre

		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Homme	6	60,0	60,0	60,0
	Femme	4	40,0	40,0	100,0
	Total	10	100,0	100,0	

Tableau 2 : situation familiale					
		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Célibataire	5	50,0	50,0	50,0
	Marié	5	50,0	50,0	100,0
	Total	10	100,0	100,0	

Tableau 3 : situation familiale					
		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Célibataire	5	50,0	50,0	50,0
	Marié	5	50,0	50,0	100,0
	Total	10	100,0	100,0	

Tableau 4 : Profession					
		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Fonctionnaire	5	50,0	50,0	50,0
	Retraité	3	30,0	30,0	80,0
	Autre	2	20,0	20,0	100,0
	Total	10	100,0	100,0	

Tableau 5 : Satisfaction de la maison					
		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Très insatisfait	3	30,0	30,0	30,0
	Plutôt insatisfait	2	20,0	20,0	50,0
	Plutôt satisfait	2	20,0	20,0	70,0
	Très satisfait	3	30,0	30,0	100,0
	Total	10	100,0	100,0	

5-Les graphiques :

Chaque type de variable a sa propre représentation graphique.

Données qualitatives : graphique circulaire, graphique à barres.

Données quantitatives : l'histogramme pour une variable quantitative continue ; graphique à barres pour une variable quantitative discrète et le graphique circulaire.

A partir de la fenêtre éditeur des données on va sélectionner :

- Analyse =>Statistiques descriptives =>Effectifs =>Sélectionner les variables =>Diagrammes =>Choisir le diagramme qui convient avec le type de la variable =>Poursuivre =>OK

Les résultats sont dans la fenêtre sortie. Uniquement le graphique circulaire de la variable genre est tracé.

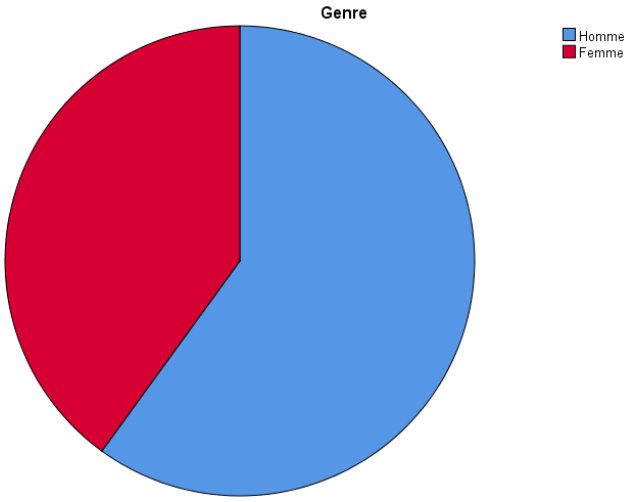


Figure 56 Graphique circulaire de la variable genre

Exercice N°03 – Pondération des observations

Voilà la série statistique suivante :

Genre	Achats du produit X	Fréquence des achats ou non achats
Féminin	Non achat	20
Féminin	Oui achat	80
Masculin	Non achat	60
Masculin	Oui achat	40

1- Définissez les variables dans cette série et leurs types.
2- Entrez les variables dans IBM SPSS Statistics.
3- Entrez les données dans IBM SPSS Statistics.
4- Afficher les tableaux de fréquences.

Solution de l'exercice N° 03

1- On a 3 variables :

- Genre : variable qualitative nominale dichotomique (féminin et masculin).
- Achat : variable qualitative nominale dichotomique (oui achat, non achat).
- Fréquence : variable quantitative continue.

2-3-Voir SPSS

Nom	Type	
Genre	Numérique	1
Achats	Numérique	1
Fréquence	Numérique	4

Align	Mesure
Droite	Nominales
Droite	Nominales
Droite	Echelle

Genre	Achats	Fréquence
1	1	20,00
1	2	80,00
2	1	60,00
2	2	40,00

4- Pour pouvoir réaliser le tableau de fréquence, il faut, tout d'abord, pondérer les observations par fréquences.

Dans la fenêtre éditeur des données cliquer sur :

=> Données

=> Pondérer les observations

=> Cocher Pondérer les observations

=> Transférez « Fréquence » dans le champ variable d'effectif.

=> OK

Ensuite, on peut afficher les tableaux de fréquence des variables genre et achats.

Exercice N°04 – Calculer une variable, sélectionner des observations

Questionnaire d'évaluation des services offerts dans une superette.

Dans le but de vérifier l'adéquation des services offerts dans notre établissement nous vous demanderons en tant que client recevant de tel service, d'évaluer la qualité des interventions dont vous avez bénéficié.

1-Numéro de sujet :…….

2-Genre :…….

3-Age : …….

4-Quel est votre niveau de satisfaction au niveau de la communication avec votre conseiller ?

1. Très insatisfait.
2. Insatisfait.
3. Ni satisfait/ni insatisfait.
4. Satisfait
5. Très satisfait

5-Quel est votre niveau de satisfaction au niveau du plan d'intervention construit par votre conseiller ?

1. Très insatisfait.
2. Insatisfait.
3. Ni satisfait/ni insatisfait.
4. Satisfait
5. Très satisfait

6-Quel est votre niveau de satisfaction au niveau des progrès réalisés en termes de votre situation professionnelle suite à l'intervention du SAV ?

1. Très insatisfait.
2. Insatisfait.
3. Ni satisfait/ni insatisfait.
4. Satisfait
5. Très satisfait

Les données du questionnaire sont comme suit :

N° sujet	Genre	Age	Q4	Q5	Q6
1	1 H	19	2	1	1
2	1	22	2	2	2
3	2 F	18	1	2	4
4	2	30	3	3	3
5	2	28	3	4	2
6	2	32	4	5	2
7	1	35	4	1	1
8	2	19	2	4	5
9	2	25	2	3	4
10	1	26	2	3	5

Questions :

1- Définissez les variables et les types de variables de cette série statistique et préciser leurs modalités ?
2- Entrez les variables de cette série statistique dans le logiciel SPSS avec codification.
3- Entrez les données de cette série statistique dans le logiciel SPSS
4- Calculer par le logiciel la moyenne, l'écart type, le minimum, le maximum de la variable considérée. Détailler les étapes. Quesque vous remarquer ?
5- Calculer la moyenne des réponses données par chaque sujet aux questions Q4, Q5, Q6. Détaillez les étapes. Quesque vous remarquer ?
6- Calculer la moyenne des réponses données par chaque sujet (uniquement les hommes) aux questions Q4, Q5, Q6. Détaillez les étapes.
7- Calculer la moyenne des réponses données par chaque sujet (uniquement les femmes âgées entre 25 ans et 30 ans) aux questions Q4, Q5, Q6. Détaillez les étapes.
8- Calculer par le logiciel la moyenne, l'écart type, le minimum, le maximum de la variable âge (uniquement les hommes âgés entre 18 ans et 30 ans). Détailler les étapes.
9- Calculer par le logiciel la moyenne, l'écart type, le minimum, le maximum de la variable âge dans un échantillon de 3 cas à partir des 7 premières cas. Détaillez les étapes.
10- Calculer par le logiciel la moyenne, l'écart type, le minimum, le maximum de la variable âge dans un échantillon entre le 4eme cas et le 8eme cas. Détaillez les étapes.
11- Refaire la question 4 par la syntaxe. Détailler les étapes.

Solution de l'exercice N°04

1- Variables et modalités :

Variables	Type	Modalités
Numéro du sujet	Qualitative nominale	
Genre	Qualitative dichotomique nominale	2 modalités : homme et femme.
Age	Quantitative continue	
Q4	Qualitative ordinale	**5 modalités :** Très insatisfait. Insatisfait. Ni satisfait/ni insatisfait. Satisfait Très satisfait
Q5	Pareille que Q4	Pareille que Q4
Q6	Pareille que Q4	Pareille que Q4

2-3- Voir SPSS.

4-A partir de la fenêtre éditeur des données on va sélectionner :

Analyse => Statistiques descriptives => Descriptives => Glisser la variable **Age** => Option. =>Cocher moyenne, minimum, maximum, écart-type =>Poursuivre =>OK

On trouve les résultats dans la fenêtre Sortie.

Statistiques descriptives de la variable *âge*					
	N	Minimum	Maximum	Moyenne	Ecart type
Age	10	18	35	25,40	5,892
N valide (liste)	10				

On remarque que le min=18, le max=35 et la moyenne 25,40 donc c'est les jeunes qui sont intéressés par les services offerts au niveau de la superette. L'écart-type est de 5,89 qui est inférieur à 0,5 moyenne (12,7). Donc les variations de l'âge sont faibles dans la population étudiée.

5-A partir de la fenêtre éditeur des données, on va sélectionner :

Transformer => Calculer la variable => Ecrivez dans la fenêtre variable cible le nom de la nouvelle variable (ex : **Résultat**).

Vous pouvez cliquer sur type et libellé (option facultative) et écrire soit un nouveau libellé ou utiliser la formule comme libellé.

=> Cliquez sur statistiques dans la fenêtre groupe de fonctions.

=> Choisissez Mean dans la fenêtre fonctions et variables spéciales.

=> Cliquer deux fois sur Mean, elle apparait automatiquement dans la fenêtre expression numérique.

=> Glissez la variable Q4, Q5, Q6 à la place des points d'interrogations.

=> Cliquer sur OK.

On trouve les résultats dans la fenêtre Sortie.

S'il y avait une très bonne satisfaction normalement la moyenne était de 5 (Q4=5, Q5=5, Q6=5 et la moyenne de cette somme Q4+Q5+Q6 est 5) et s'il y avait une bonne satisfaction normalement la moyenne était de 4 (Q4=4, Q5=4, Q6=4 et la moyenne de cette somme Q4+Q5+Q6 est 4).

Ici on remarque que toutes les valeurs de la variable résultat sont au-dessous de la valeur 5 et 4 donc en moyenne il n'y pas de satisfaction.

6-A partir de la fenêtre éditeur des données on va sélectionner :

Transformer => Calculer la variable => Ecrivez dans la fenêtre variable cible le nom de la nouvelle variable (ex : **Résultat 2**).

=> Cliquez sur statistiques dans la fenêtre groupe de fonctions.

=> Choisissez Mean dans la fenêtre fonctions et variables spéciales.

=> Cliquer deux fois sur Mean, elle apparait automatiquement dans la fenêtre expression numérique.

=> Glissez la variable Q4, Q5, Q6 à la place des points d'interrogations.

=> Cliquer sur SI

=> Cliquer sur Inclure lorsque l'observation remplit la condition Glisser la variable **genre**.

=> Ecrire =1

=> Cliquer sur Poursuivre => Cliquer sur OK.

On trouve les résultats dans la fenêtre Sortie.

7-A partir de la fenêtre éditeur des données on va sélectionner :

Transformer => Calculer la variable => Ecrivez dans la fenêtre variable cible le nom de la nouvelle variable (ex : **Résultat 3**).

=>Cliquez sur statistiques dans la fenêtre groupe de fonctions.

=> Choisissez Mean dans la fenêtre fonctions et variables spéciales.

=> Cliquer deux fois sur Mean, elle apparait automatiquement dans la fenêtre expression numérique.

=> Glissez la variable Q4, Q5, Q6 à la place des points d'interrogations.

=> Cliquer sur SI

=>Cliquer sur Inclure lorsque l'observation remplit la condition

=> Glisser la variable genre.

=> Ecrire =2.

=> Ecrire AND.

=> Glisser la variable Age.

=> Genre=2 AND Age >= 25 AND Age <= 30.

=> Cliquer sur Poursuivre.

=> Cliquer sur OK.

On trouve les résultats dans la fenêtre Sortie.

8-A partir de la fenêtre éditeur des données on va sélectionner :

Données => Sélectionner des observations. => Cliquer sur selon une condition logique.

=> Ecrire: Genre=1 AND Age >= 18 AND Age <= 30

=> Poursuivre => Ok.

=> Analyse => Statistiques descriptives => Descriptives => Glisser la variable Age => Option. => Cocher moyenne, minimum, maximum, écart-type => Poursuivre => OK

On trouve les résultats dans la fenêtre Sortie.

Statistiques descriptives de la variable *âge*

	N	Minimum	Maximum	Moyenne	Ecart type
Age	3	19	26	22,33	3,512
N valide (liste)	3				

9-A partir de la fenêtre éditeur des données on va sélectionner :

Données => Sélectionner des observations => Cliquer sur Par échantillonnage aléatoire => Cliquer sur échantillon => Cliquer sur Exactement, écrire 3 dans la première case et 7 dans la deuxième case. (ça veut dire 3 observations à partir des premières 7 observations) =>Poursuivre. => Ok

Analyse => Statistiques descriptives => Descriptives => Glisser la variable Age => Option => Cocher moyenne, minimum, maximum, écart-type => Poursuivre => OK

On trouve les résultats dans la fenêtre Sortie.

Statistiques descriptives de la variable *âge*					
	N	Minimum	Maximum	Moyenne	Ecart type
Age	3	19	32	26,33	6,658
N valide (liste)	3				

10-A partir de la fenêtre éditeur des données on va sélectionner :

Données => Sélectionner des observations => Cliquer sur dans un intervalle de temps ou d'observations => Cliquer sur intervalle => Ecrire 4 dans la case Première observation et 8 dans la case Dernière observation =>Poursuivre. => Ok

Analyse => Statistiques descriptives => Descriptives => Glisser la variable Age => Option => Cocher moyenne, minimum, maximum, écart-type => Poursuivre => OK

On trouve les résultats dans la fenêtre Sortie.

Statistiques descriptives de la variable *âge*					
	N	Minimum	Maximum	Moyenne	Ecart type
Age	5	19	35	28,80	6,058
N valide (liste)	5				

11-A partir de la fenêtre éditeur des données on va sélectionner :

Analyse => Statistiques descriptives => Descriptives => Glisser la variable Age => Option => Cocher moyenne, minimum, maximum, écart-type => Poursuivre => **Coller**

Exercice N°05 – Scinder un fichier, Box plot

Un médecin s'intéresse à trois traitements proposés dans la littérature pour aider des étudiants qui ont des problèmes d'obésité.

La première méthode (T1) consiste à donner chaque matin à l'étudiant un médicament proposé par une firme pharmaceutique (risque d'effets secondaires), la deuxième (T2) consiste à faire des exercices physiques à chaque matin et la troisième (T3) est basée sur un apprentissage nutritionnel. Ce médecin est persuadé que les méthodes 2 et 3 donnent chacune de meilleurs résultats que la première.

Pour confirmer cette hypothèse, il met en place une étude faisant intervenir 15 étudiants (5 par méthode). Il mesure l'impact sur une base du tour de taille. Plus la mesure est faible, meilleure est la condition de l'étudiant.

Voici les résultats obtenus :

Enfant	Traitement 1	Traitement 2	Traitement 3
1	95	87	95
2	100	92	94
3	102	88	93
4	99	92	93
5	122	92	92

1-Quelles sont les variables de cette série statistique ?

2-Définissez les types de variables de cette série statistique et préciser leurs modalités ?

3-Entrez les variables de cette série statistique dans le logiciel SPSS avec codification.

4-Entrez les données de cette série statistique dans le logiciel SPSS.

5-Calculer par le logiciel la moyenne, l'écart type, le minimum, le maximum de la variable considérée. Détailler les étapes.

6-Présenter graphiquement ces données par le Box-plot (les 3 box-plot doivent être dans le même graphe). Détailler les étapes.

7- Comparer la moyenne des 3 traitements. Cette comparaison confirme ou –infirme l'hypothèse des médecins ?

Solution de l'exercice N°05

1- On à 3 variables : enfant, traitement, tour de taille.
2- Définition et modalités :

Variables	Type	Modalités
Enfant	Qualitative nominale	5 modalités : E1, E2, E3, E4, E5. Avec E=enfant
Traitement	Qualitative nominale	T1, T2, T3
Tour de taille	Quantitative continu	Il n'y a pas de modalité

3- Codification des variables

Nom	Type	Largeur	Décimales	Libellé	Valeurs	Manquant	Colonnes	Align	Mesure
Enfant	Numérique	1	0	Noms de l'enfants	Aucun	Aucun	8	Droite	Nominales
Traitement	Numérique	1	0	Type de traitement	{1, Traitement 1}...	Aucun	8	Droite	Nominales
Tour_de_taille	Numérique	2	0	Niveau de concentration	Aucun	Aucun	8	Droite	Echelle

4- Voici l'aperçu de l'onglet « Vue des données ».

Enfant	Traitement	Tour_de_taille
1	1	95
2	1	100
3	1	102
4	1	99
5	1	122
1	2	87
2	2	92
3	2	88
4	2	92
5	2	92
1	3	95
2	3	94
3	3	93
4	3	93
5	3	92

5- **Important**. A partir de la fenêtre éditeur des données on va sélectionner :

Données => Scinder un fichier => Comparer les groupes => Glissez la variable traitement dans critère de regroupement => OK

Ensuite,

Analyse => Statistiques descriptives => Descriptives => Glisser la variable concentration => Option => Cocher moyenne, minimum, maximum, écart-type => Poursuivre => OK.

On trouve les résultats dans la fenêtre Sotie.

Statistiques descriptives

Type de traitement		N	Minimum	Maximum	Moyenne	Ecart type
Traitement 1	Niveau de concentration	5	95	122	103,60	10,597
	N valide (liste)	5				
Traitement 2	Niveau de concentration	5	87	92	90,20	2,490
	N valide (liste)	5				
Traitement 3	Niveau de concentration	5	92	95	93,40	1,140
	N valide (liste)	5				

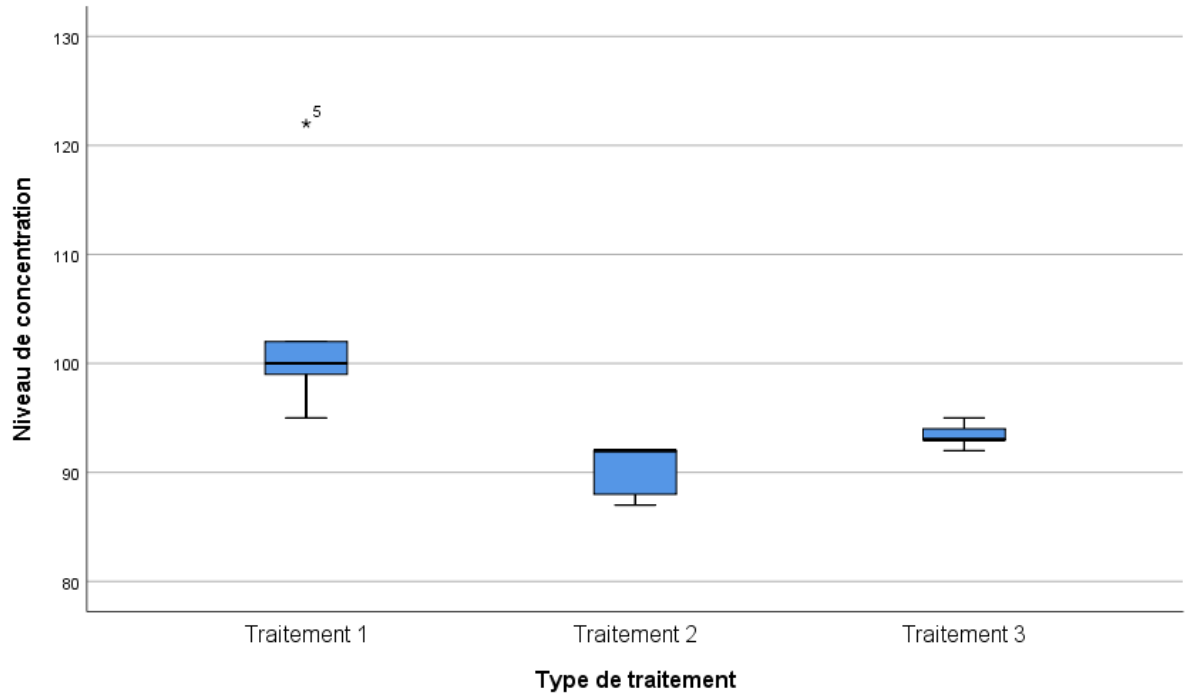
6- Avant de créer le box-plot on doit désactiver l'option scinder un fichier (pour que les 2 box plot apparaissent dans le même graphe et non pas chaque box plot dans un graphe individuel), c'est-à-dire cliquer sur :

A partir de la fenêtre éditeur de données cliquer sur :

Données => Scinder un fichier => Analyser toutes les observations, ne pas créer de groupes => OK.

Maintenant on va créer le box-plot : à partir de la fenêtre éditeur de données cliquer sur :
Graphes => Boite de dialogue ancienne version => Boites à moustaches => Simple => Récapitulatifs pour groupes d'observations => Définir => Glisser la variable concentration dans variable => Glisser la variable traitement dans axe des modalités => OK.

On trouve les résultats dans la fenêtre Sortie.



7- A partir de la fenêtre éditeur des données on va sélectionner :

Analyse => Comparer les moyennes => Moyennes => Sélectionner la variable dépendante (Concentration) => Sélectionner la variable indépendante (Traitement) => Option.
Vérifier que la moyenne, écart type et le nombre d'observations sont intégrées.
=>Poursuivre => OK.

Les résultats sont dans la fenêtre Viewer **Tableau de bord**

Rapport

Niveau de concentration

Type de traitement	Moyenne	N	Ecart type
Traitement 1	103,60	5	10,597
Traitement 2	90,20	5	2,490
Traitement 3	93,40	5	1,140
Total	95,73	15	8,319

La moyenne de T1 est de 103,60 qui est supérieure à la moyenne de T2 et T3. T2 (moyenne=90,20) et T3 (moyenne=93,40) ont des moyennes très proches l'une de l'autre.
L'analyse descriptive tend à montrer de meilleures performances en tendance centrale chez les enfants ayant bénéficié des deux méthodes non pharmaceutiques (T2 et T3) par rapport à T1 (Médicaments avec risques d'effets secondaires).

T2 (activité physique) et T3 (apprentissage nutritionnel) semblent quant à elles assez proches l'une de l'autre à leur valeur centrale. La moyenne des tours de taille est inférieure, ce qui montre leurs efficacités et confirme l'hypothèse rédigée par le médecin.

Bibliographie

Carricano, M., & Poujol, F. (2009). *Analyse de donn??es avec SPSS*. Pearson Education France.

Dodge, Y. (2007). *Statistique : Dictionnaire encyclopédique*. Springer-Verlag. https://doi.org/10.1007/978-2-287-72094-9

Dress, F. (2007). *Les probabilités et la statistique de A à Z*. Dunod.

Goldfarb, B., & Pardoux, C. (2011). *Introduction à la méthode statistique : Manuel et exercices corrigés*. Dunod.

IBM. (2019). *Your ultimate guide to SPSS Statistics vs SPSS Modeler*. https://community.ibm.com/datascience/blogs/nitin-mathur1/2019/11/14/spss-statistics-vs-modeler

Kent State University Libraries. (2022a, janvier 21). *Defining Variables—SPSS Tutorials—LibGuides at Kent State University*. https://libguides.library.kent.edu/SPSS/DefineVariables

Kent State University Libraries. (2022b, janvier 21). *LibGuides : SPSS Tutorials: The SPSS Environment*. https://libguides.library.kent.edu/SPSS/Environment

Kent State University Libraries. (2022c, février 9). *LibGuides : SPSS Tutorials: Computing Variables*. https://libguides.library.kent.edu/SPSS/ComputeVariables

Kent State University Libraries. (2022d, février 9). *LibGuides : SPSS Tutorials: Crosstabs*. https://libguides.library.kent.edu/SPSS/Crosstabs

Kent State University Libraries. (2022e, février 9). *LibGuides : SPSS Tutorials: Grouping Data*. https://libguides.library.kent.edu/SPSS/SplitData

Kent State University Libraries. (2022f, février 9). *LibGuides : SPSS Tutorials: The Data View Window*. https://libguides.library.kent.edu/SPSS/DataViewWindow

Kent State University Libraries. (2022g, février 9). *LibGuides : SPSS Tutorials: Weighting Cases*. https://libguides.library.kent.edu/SPSS/WeightCases

Simple Learning Pro. (2015, novembre 15). *The Five Number Summary, Boxplots, and Outliers (1.6)*. https://www.youtube.com/watch?v=tpToLyZibKM

Stafford, J., & Bodson, P. (2006). *L'analyse multivariée avec SPSS*. Presses de l'Universit?? du Qu??bec (PUQ).

Thierry Ancelle. (2015a, janvier 27). *Estimation d'un paramètre*. https://www.youtube.com/watch?v=m3t_iG79Pp8

Thierry Ancelle. (2015b, février 12). *Loi normale*. https://www.youtube.com/watch?v=2k-1Yi40ZSw

Thierry Ancelle. (2017a, mars 7). *Coefficient d'aplatissement ou kurtosis*. https://www.youtube.com/watch?v=Ht3KeTJigGo

Thierry Ancelle. (2017b, mars 7). *Coefficient d'asymétrie ou skewness*. https://www.youtube.com/watch?v=9MJk6dkrU7I

Thierry Ancelle. (2021, avril 15). *Graphiques*. https://www.youtube.com/watch?v=wj_TOPF1VXg

van den Berg, R. G. (2022, janvier 31). *Missing Values in SPSS - Quick Introduction*. https://www.spss-tutorials.com/spss-missing-values/

Veysseyre, R. (2014). *Aide-mémoire—Statistique et probabilités pour les ingénieurs—3ed* (3e édition). Dunod.

Wikipedia contributors. (2022). SPSS. In *Wikipedia, The Free Encyclopedia*. Wikimedia Foundation. https://en.wikipedia.org/w/index.php?title=SPSS&oldid=1063484669

Yergeau, E., & Poirier, M. (2021a). *SPSS à l'UdeS |Calculer ou Compute*. https://spss.espaceweb.usherbrooke.ca/calculer/

Yergeau, E., & Poirier, M. (2021b). *SPSS à l'UdeS Paramètres des variables* |. https://spss.espaceweb.usherbrooke.ca/parametres-des-variables/

Yergeau, E., & Poirier, M. (2021c). *SPSS à l'UdeS Sélection de cas*. https://spss.espaceweb.usherbrooke.ca/selection-de-cas/

Yergeau, E., & Poirier, M. (2021d). *SPSS à l'UdeS Tableau croisé*. https://spss.espaceweb.usherbrooke.ca/tableau-croise/

zedstatistics. (2019a, janvier 28). *What is skewness? A detailed explanation (with moments!)*. https://www.youtube.com/watch?v=_vDRKlTz7yo

zedstatistics. (2019b, janvier 30). *What is Kurtosis? (+ the « peakedness » controversy!)*. https://www.youtube.com/watch?v=TM033GCU-SY