

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Oran 2 Mohamed ben Ahmed

Faculté des sciences économiques, sciences de gestion et sciences
commerciales

Département des sciences économiques



Polycopie du module

Analyse des données

2^{ème} Année MASTER économie et gestion des entreprises

(3^{ème} Semestre)

Réalisé par :

BOUYACOUB Brahim

Maitre de conférences – A -

Département des sciences économiques

Année académique 2023-2024

Analyse des données

Description du cours

Le cours d'analyse des données est conçu pour fournir aux étudiants une compréhension approfondie des concepts, des méthodes et des techniques nécessaires pour collecter, analyser et interpréter des données dans divers domaines. Il met l'accent sur les compétences essentielles en statistiques et en science des données, en mettant en avant l'importance de l'analyse quantitative dans la prise de décisions éclairées. Le cours est généralement offert dans les programmes de gestion, d'économie ou de finance. Ce cours s'adresse aux étudiants de la deuxième année master, spécialité « économie et gestion des entreprise (EGE) ».

Objectifs du cours

Les objectifs du cours d'analyse des données varient en fonction du niveau d'enseignement et du programme académique, mais voici quelques objectifs généraux associés à un tel cours :

1. **Comprendre les Fondamentaux de l'Analyse des Données** : Acquérir une compréhension solide des concepts de base de l'analyse des données, y compris les types de données, les méthodes d'échantillonnage et les outils d'analyse.
2. **Maîtriser les Compétences en Statistiques** : Développer des compétences statistiques, notamment la capacité à résumer et à présenter des données, à effectuer des tests d'hypothèses et à interpréter les résultats.
3. **Savoir Collecter des Données de Manière Rigoureuse** : Apprendre à concevoir des enquêtes, à collecter des données de manière fiable et à évaluer la qualité des données.
4. **Utiliser des Outils Informatiques** : Familiarisation avec les logiciels et les outils d'analyse de données, tels que R, Python, Excel, SPSS, ou d'autres logiciels couramment utilisés.

5. **Appliquer des Méthodes d'Analyse des Données** : Être capable d'appliquer des méthodes d'analyse statistique et d'interpréter les résultats pour résoudre des problèmes concrets.
6. **Modélisation Statistique** : Comprendre comment construire des modèles statistiques pour expliquer des phénomènes complexes et prévoir des résultats futurs.
7. **Analyse de Séries Temporelles** : Apprendre à analyser des données temporelles, à détecter des tendances, des saisonnalités et à effectuer des prévisions.
8. **Savoir Prendre des Décisions Basées sur des Données** : Développer la capacité à utiliser l'analyse des données pour prendre des décisions éclairées dans des contextes professionnels.
9. **Travailler sur des Études de Cas Réelles** : Appliquer les compétences acquises à des études de cas réels dans divers domaines, ce qui renforce la compréhension pratique.
10. **Préparation à la Science des Données** : Pour les étudiants intéressés par la science des données, le cours peut servir de base pour des études plus avancées dans ce domaine.
11. **Développer des Compétences en Communication** : Apprendre à communiquer efficacement les résultats de l'analyse des données, que ce soit à l'écrit ou à l'oral, pour un public varié.
12. **Promouvoir la Pensée Critique** : Encourager les étudiants à évaluer de manière critique les méthodes d'analyse des données, à identifier les biais potentiels et à prendre des décisions éthiques.
13. **Développer des Compétences en Résolution de Problèmes** : Stimuler la résolution de problèmes en utilisant des approches analytiques.

The Data Analysis

The Data Analysis course is designed to provide students with a thorough understanding of the concepts, methods, and techniques required to collect, analyze, and interpret data in various fields. It emphasizes essential skills in statistics and data science, highlighting the importance of quantitative analysis in making informed decisions. The course is typically offered in management, economics, or finance programs. This course is intended for second-year master's students specializing in "Economics and Business Management (EGE)."

Course Objectives

The course aims to provide students with the skills and knowledge necessary to:

1. **Understand the Fundamentals of Data Analysis:** Acquire a solid understanding of the basic concepts of data analysis, including data types, sampling methods, and analysis tools.
2. **Master Statistical Skills:** Develop statistical skills, including the ability to summarize and present data, conduct hypothesis tests, and interpret results.
3. **Rigorously Collect Data:** Learn how to design surveys, collect data reliably, and assess data quality.
4. **Use Computer Tools:** Familiarize yourself with data analysis software and tools, such as R, Python, Excel, SPSS, or other commonly used software.
5. **Apply Data Analysis Methods:** Be able to apply statistical analysis methods and interpret results to solve practical problems.
6. **Statistical Modeling:** Understand how to build statistical models to explain complex phenomena and make future predictions.
7. **Time Series Analysis:** Learn how to analyze time-series data, detect trends, seasonality, and make forecasts.

8. **Make Data-Driven Decisions:** Develop the ability to use data analysis to make informed decisions in professional settings.
9. **Work on Real-World Case Studies:** Apply acquired skills to real-world case studies in various fields, enhancing practical understanding.
10. **Preparation for Data Science:** For students interested in data science, the course can serve as a foundation for more advanced studies in the field.
11. **Develop Communication Skills:** Learn to effectively communicate data analysis results, both in written and oral forms, to diverse audiences.
12. **Promote Critical Thinking:** Encourage students to critically evaluate data analysis methods, identify potential biases, and make ethical decisions.
13. **Problem-Solving Skills:** Foster problem-solving skills using analytical approaches.

تحليل البيانات

تم تصميم درس تحليل البيانات لتزويد الطلاب بفهم عميق للمفاهيم والأساليب والتقنيات اللازمة لجمع وتحليل وتفسير البيانات في مجموعة متنوعة من المجالات. تسلط الضوء على المهارات الأساسية في الإحصاءات وعلوم البيانات، مؤكدة أهمية التحليل الكمي في اتخاذ القرارات المستنيرة. عادةً ما تُقدّم الدورة في برامج الإدارة والاقتصاد والتمويل. تُوجه هذه الدورة إلى طلاب السنة الثانية للماجستير في تخصص "اقتصاد وتسيير المؤسسات." (EGE)

أهداف المقرر يهدف المقرر إلى تزويد الطلاب بالمهارات والمعرفة اللازمة للقيام بما يلي:

1. فهم أساسيات تحليل البيانات: اكتساب فهم قوي للمفاهيم الأساسية في تحليل البيانات، بما في ذلك أنواع البيانات، وأساليب العينات، وأدوات التحليل.
2. احتراف مهارات الإحصاءات: تطوير مهارات إحصائية، بما في ذلك القدرة على تلخيص وتقديم البيانات، وإجراء اختبارات فرض الفرضيات، وتفسير النتائج.
3. جمع البيانات بدقة: تعلم كيفية تصميم استطلاعات، وجمع البيانات بشكل موثوق، وتقييم جودة البيانات.
4. استخدام أدوات الحاسوب: التعرف على برمجيات وأدوات تحليل البيانات، مثل R، Python، Excel، SPSS، أو غيرها من البرمجيات المستخدمة بشكل شائع.
5. تطبيق أساليب تحليل البيانات: القدرة على تطبيق أساليب تحليل إحصائي وتفسير النتائج لحل مشكلات عملية.
6. نمذجة إحصائية: فهم كيفية بناء نماذج إحصائية لشرح الظواهر المعقدة والقيام بتوقعات مستقبلية.
7. تحليل السلاسل الزمنية: تعلم كيفية تحليل البيانات الزمنية، واكتشاف الاتجاهات والموسم، وإجراء التوقعات.
8. اتخاذ قرارات مستندة إلى البيانات: تطوير القدرة على استخدام تحليل البيانات لاتخاذ قرارات مستندة إلى البيانات في البيئات المهنية.
9. العمل على دراسات حالة من العالم الحقيقي: تطبيق المهارات المكتسبة على دراسات حالة من الواقع في مجموعة متنوعة من المجالات، مما يعزز الفهم العملي.

10. التحضير لعلوم البيانات: بالنسبة للطلاب المهتمين بعلوم البيانات، يمكن أن تكون الدورة أساسًا للدراسات المتقدمة في هذا المجال.

11. تطوير مهارات التواصل: تعلم كيفية التواصل بفعالية نتائج تحليل البيانات، سواء كان ذلك كتابيًا أو شفهيًا، لجمهور متنوع.

12. تعزيز التفكير النقدي: تشجيع الطلاب على تقييم أساليب تحليل البيانات بشكل نقدي، والتعرف على التحيزات المحتملة، واتخاذ قرارات أخلاقية.

13. تطوير مهارات حل المشكلات: تعزيز مهارات حل المشكلات باستخدام النهج التحليلية

Résumé du cours

Le cours d'analyse des données offre aux étudiants une compréhension approfondie des principes et des compétences essentielles nécessaires pour travailler avec les données. Les objectifs clés de ce cours incluent l'acquisition de compétences statistiques solides, la maîtrise des outils informatiques d'analyse de données tels que R et Python, et la capacité à collecter des données de manière rigoureuse.

Les étudiants apprendront à appliquer une variété de méthodes d'analyse statistique pour résoudre des problèmes concrets, à modéliser des phénomènes complexes, et à analyser des données temporelles pour identifier des tendances et réaliser des prévisions.

Ce cours favorise la prise de décisions basée sur les données, en mettant l'accent sur l'importance de l'analyse quantitative dans un contexte professionnel. Les étudiants auront également l'occasion de travailler sur des études de cas réels, renforçant ainsi leur compréhension pratique.

En fin de compte, ce cours prépare les étudiants à devenir des analystes de données compétents, capables de traiter efficacement les données dans une variété de domaines, tout en développant des compétences en communication, en pensée critique et en résolution de problèmes.

Ce cours est conçu pour les étudiants de la deuxième année en master, en particulier ceux se spécialisant en "économie et gestion des entreprises (EGE)." Il fournit une base solide pour ceux qui souhaitent utiliser l'analyse des données dans des contextes professionnels ou de recherche.

Avant-propos du module / matière :

Le cours "Analyse des données" a été conçu pour offrir une introduction complète aux principes, aux méthodes et aux compétences essentielles nécessaires pour travailler avec des données de manière efficace et significative.

Dans un monde de plus en plus axé sur les données, la capacité à collecter, analyser et interpréter des données est devenue une compétence essentielle dans une variété de domaines, qu'il s'agisse de la gestion, de l'économie, de la finance, de la science ou de la technologie. Ce module vise à vous équiper de ces compétences, que vous pourrez appliquer tout au long de votre carrière.

Au cours de ce module, vous explorerez les bases de l'analyse des données, y compris les types de données, les méthodes d'échantillonnage et les outils d'analyse statistique. Vous apprendrez à maîtriser des compétences statistiques importantes, à utiliser des logiciels d'analyse de données couramment utilisés, et à appliquer des méthodes d'analyse statistique pour résoudre des problèmes du monde réel.

Nous mettrons également l'accent sur la modélisation statistique, l'analyse de séries temporelles et la prise de décisions basées sur des données. Vous aurez l'occasion de travailler sur des études de cas réelles, ce qui renforcera votre compréhension pratique de l'analyse des données.

Nous encourageons également le développement de compétences en communication, de pensée critique et de résolution de problèmes, car ces compétences sont essentielles pour transformer l'analyse des données en actions concrètes.

Ce cours cible principalement les étudiants de la deuxième année master, en particulier ceux se spécialisant en "économie et gestion des entreprises (EGE)." Il fournit une solide fondation pour ceux qui cherchent à utiliser l'analyse des données dans un contexte professionnel ou de recherche.

Informations sur le cours

Faculté : Sciences économiques, commerciales et sciences de gestion

Département : Sciences économiques

Public cible : 2^{ème} année master, spécialité économie et gestion des entreprises.

Intitulé du cours : analyse des données

Crédit : 05

Coefficient : 02

Durée : 15 semaines

Enseignant : Dr. BOUYACOUB Brahim

Contact par mail : bouyacoub.brahim@gmail.com

Disponibilité :

- Au bureau : lundi, jeudi de 11h00 -12h00
- Par mail : Je m'engage à répondre par mail dans 48 heures qui suivent la réception du message.

Présentation du cours

L'analyse de données englobe un ensemble de méthodes descriptives visant à synthétiser et à représenter visuellement les informations essentielles présentes dans de vastes ensembles de données. Bien que cette technique ait des origines anciennes, remontant aux années 1930 avec les travaux de pionniers tels que Pearson, Spearman et Hotelling, elle a connu des développements significatifs dans les années 1960 et 1970, en grande partie grâce à l'avancée de l'informatique.

L'analyse de données se concentre sur l'exploration approfondie de tableaux de données, dans le but de mettre en évidence des relations pertinentes, se distinguant ainsi de l'analyse exploratoire des données. L'objectif sous-jacent de cette approche statistique est de révéler ces relations. Deux types fondamentaux de relations se distinguent : les relations d'équivalence et les relations d'ordre. En conséquence, il est possible de diviser une population en classes hiérarchisées en fonction de ces relations.

Ce cours est divisé en plusieurs unités d'apprentissage conçues pour vous doter de compétences essentielles en matière d'utilisation des statistiques et de l'analyse de données. De plus, il vous permettra d'acquérir des connaissances pratiques sur l'utilisation du logiciel SPSS pour explorer et résumer de manière significative les informations contenues dans des ensembles de données volumineux.

Contenu

Ce cours est scindé en 8 chapitres :

Chapitre 1 : Introduction aux Concepts Fondamentaux de l'Analyse de Données (Exploration des Méthodes Analytiques)

Chapitre 2 : Exploration des Variables Individuelles : Analyse Univariée (Approche des Techniques de Description Statistique)

Chapitre 3 : Analyse des Relations Croisées : Utilisation de Tableaux de Contingence et le Test du Chi-2 (Khi 2) (Exploration des Liens entre Variables)

Chapitre 4 : Visualisation des Données : Analyse d'un Nuage de Points, Recherche des Axes Principaux et Interprétation (Approche Géométrique en Statistique)

Chapitre 5 : Comparaison de Groupes : L'Analyse de la Variance ANOVA (Étude des Différences entre Groupes)

Chapitre 06 : Réduction de Dimension : Analyse en Composantes Principales (ACP)
(Simplification de la Représentation des Données)

Chapitre 07 : Classification des Données : Analyse Discriminante (Categorisation des Données)

Chapitre 08 : Mise en Pratique de l'Analyse de Données avec SPSS : Exemple Détaillé
d'Application

Table des matières

Analyse des données	2
Description du cours.....	2
Objectifs du cours.....	2
Résumé du cours	8
Avant-propos du module / matière :	9
Informations sur le cours	9
Chapitre 1 : Introduction aux Concepts Fondamentaux de l'Analyse de Données (Exploration des Méthodes Analytiques).....	14
Chapitre 2 : Exploration des Variables Individuelles : Analyse Univariée (Approche des Techniques de Description Statistique)	23
Chapitre 3 : Analyse des Relations Croisées : Utilisation de Tableaux de Contingence et le Test du Chi-2 (Khi 2) (Exploration des Liens entre Variables)	50
Chapitre 4 : Visualisation des Données : Analyse d'un Nuage de Points, Recherche des Axes Principaux et Interprétation (Approche Géométrique en Statistique)	57
Chapitre 5 : Comparaison de Groupes : L'Analyse de la Variance ANOVA (Étude des Différences entre Groupes)	66
Chapitre 06 : Réduction de Dimension : Analyse en Composantes Principales (ACP) (Simplification de la Représentation des Données).....	69
Chapitre 07 : Classification des Données : Analyse Discriminante (Categorisation des Données)	79
Chapitre 08 : Mise en Pratique de l'Analyse de Données avec SPSS : Exemple Détaillé d'Application	90
Conclusion.....	138
Référence bibliographie	138

Chapitre 1 :
Introduction aux Concepts
Fondamentaux de l'Analyse de Données
(Exploration des Méthodes Analytiques)

I. Introduction

L'analyse des données est une discipline largement applicable dans divers secteurs de l'activité humaine. Elle repose sur un ensemble de méthodes statistiques plus ou moins définies qui jouent un rôle essentiel dans la collecte, l'organisation, la synthèse, la présentation et l'examen de données.

L'objectif fondamental de cette démarche, comme souligné par Benzécri en 1982, est de permettre l'extrapolation de conclusions significatives à partir des données et de faciliter le processus de prise de décisions.

En d'autres termes, l'analyse des données agit comme un outil puissant pour extraire des informations utiles à partir de vastes ensembles de données, aidant ainsi les individus et les organisations à prendre des décisions éclairées. Cette discipline est un pilier fondamental de l'exploration et de l'exploitation des données dans divers contextes, qu'il s'agisse de la recherche scientifique, de la gestion d'entreprise, de la planification stratégique ou d'autres domaines.

II. C'est quoi l'analyse des données ?

L'analyse est une méthode qui se pose en opposition à la synthèse, car elle vise à comprendre un objet en le décomposant en ses constituants, comme le souligne Saporta en 2011. Cette démarche consiste en une étude minutieuse et précise, dont l'objectif est de disséquer un ensemble pour en extraire et éclairer les éléments qui le composent.

En d'autres termes, il s'agit de réaliser une analyse approfondie d'une situation ou d'un ensemble donné.

Au cœur de cette démarche analytique se trouvent les données. La donnée est l'élément fondamental et indispensable à tout raisonnement visant à extraire des informations nécessaires à la compréhension des phénomènes. Les données servent de matière première à l'analyse, permettant ainsi de décortiquer, d'interpréter et de donner du sens à ce qui peut sembler complexe. En somme, les données sont les éléments de base qui alimentent le processus analytique, contribuant ainsi à la clarification et à la compréhension des diverses composantes d'un ensemble.

III. Les types des données

Il existe principalement deux types de données : les données qualitatives et les données quantitatives, chacune ayant ses caractéristiques distinctes.

Les données quantitatives, de nature numérique, sont associées à des échelles de mesure d'intervalle ou de rapport, comme le souligne Guillaume Broc en 2018. Ces données se prêtent à des opérations mathématiques et sont exprimées sous forme de nombres. Par exemple, la taille

d'un objet, le prix d'un produit, le nombre d'articles vendus, ou les résultats chiffrés d'un test sont des exemples de données quantitatives. Elles sont particulièrement appropriées pour des analyses statistiques approfondies.

En revanche, les données qualitatives, également appelées données catégorielles, sont liées à des groupes ou des catégories. Elles correspondent à des échelles de mesure nominale, où les valeurs sont des noms ou des catégories, ou à des échelles de mesure ordonnée, qui reflètent une hiérarchie entre les catégories. Les données qualitatives décrivent la qualité d'un élément, comme la couleur, la texture, l'apparence d'un objet, ou encore la description d'une expérience. Contrairement aux données quantitatives, elles ne sont pas exprimées en chiffres.

En résumé, les données qualitatives se concentrent sur des caractéristiques qualitatives, tandis que les données quantitatives sont basées sur des valeurs numériques. Le choix entre l'utilisation de données qualitatives ou quantitatives dépendra du contexte de la collecte de données et des objectifs de l'analyse.

IV. Les étapes d'analyse des données

L'analyse des données est le processus qui consiste à examiner et à interpréter des données afin d'élaborer des réponses à des questions.

Les principales étapes du processus d'analyse consistent à cerner les sujets d'analyse, à déterminer la disponibilité de données appropriées, à décider des méthodes qu'il y a lieu d'utiliser pour répondre aux questions d'intérêt, à appliquer les méthodes et à évaluer, résumer et communiquer les résultats.

Exemple

- Une étude de cas (thème ou sujet étudié)
- Problématique
- Hypothèse
- Méthode et Analyse
- Résultat et interprétation

Application

1) Etude de cas :

La politique monétaire et la croissance économique

2) Problématique :

L'impact de la politique monétaire sur la croissance économique

3) Les hypothèses :

- H1 : la politique monétaire joue un rôle important sur la croissance économique.

- H2 : la politique monétaire joue un rôle modeste sur la croissance économique.

4) Les données de la politique monétaire :

- Le taux de la masse monétaire
- Le taux d'inflation
- Le taux de change
- Le taux d'intérêt

5) Les données de la croissance économique :

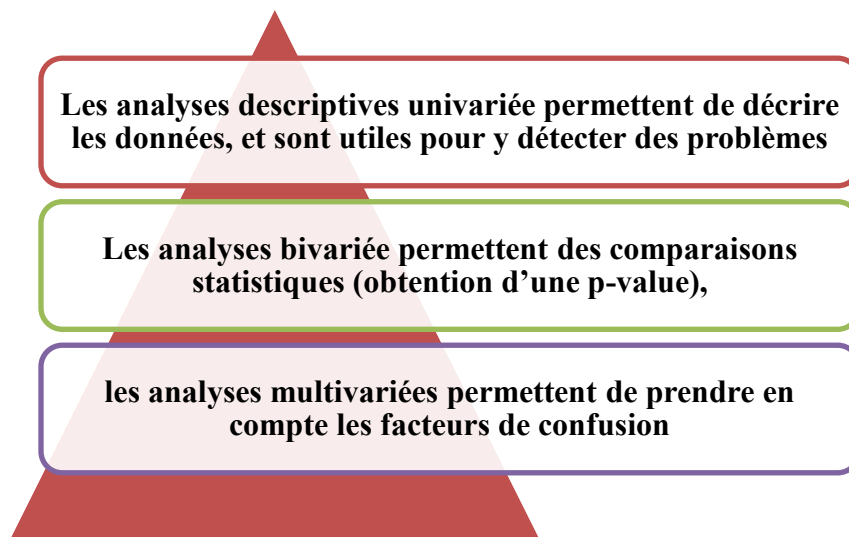
- Le produit intérieur brut (PIB)

6) Analyse :

Maintenant, il faut analyser les données, c'est-à-dire il faut vérifier l'hypothèse (soit confirmée l'hypothèse, soit infirmée).

V. Les types d'analyse des données

On distingue trois types d'analyses : les analyses univarié descriptives, les analyses bivarié et les analyses multivariées (ou multivariables).



- **Analyse univarié :** Les analyses descriptives permettent de décrire les données, et sont utiles pour y détecter des problèmes.
- **Analyse bivariée :** Étude statistique des relations pouvant exister entre deux variables en utilisant des tris croisés.
- **Analyse multivariée :** Étude statistique des relations pouvant exister entre plusieurs variables (méthodes explicatives). Elle peut aussi conduire à structurer les variables étudiées (méthodes descriptives).
 - Descriptive : par groupes de variation

- Explicative : entre groupes

Si les données ne sont relatives qu'à une seule variable, on parle de statistique descriptive univariée. Dans le cas où l'on s'intéresse à deux variables simultanément, on met en œuvre la statistique descriptive bivariée. Si l'ensemble de données provient de l'observation de plusieurs variables, on doit faire appel aux méthodes de la statistique descriptive multivariée.

VI. L'objectif des outils statistiques

- Des outils descriptifs : Pour décrire les données et les représenter graphiquement.
- Des outils de modélisation : Pour répondre aux questions et hypothèses afin de décider à partir des données.
- Pour expliquer certaines variables à partir d'autres variables.

VII. Exercices d'applications et solutions

M 1:

Question : Quelle est l'objectif principal de l'analyse de données ?

- A. Collecter autant de données que possible.
- B. Réduire la quantité de données collectées.
- C. Comprendre et interpréter les données.
- D. Stocker les données de manière sécurisée.

Solution : C. Comprendre et interpréter les données.

QCM 2:

Question : Quelles sont les deux principales catégories de données que l'on rencontre couramment en analyse de données ?

- A. Données de laboratoire et données de terrain.
- B. Données quantitatives et données qualitatives.
- C. Données primaires et données secondaires.
- D. Données brutes et données traitées.

Solution : B. Données quantitatives et données qualitatives.

QCM 3:

Question : Quelle échelle de mesure catégorise les données en groupes sans ordre particulier ?

- A. Échelle nominale.

- B. Échelle ordinale.
- C. Échelle d'intervalle.
- D. Échelle de ratio.

Solution : A. Échelle nominale.

QCM 4:

Question : Les données quantitatives sont exprimées en termes de :

- A. Catégories.
- B. Nombres.
- C. Couleurs.
- D. Textes.

Solution : B. Nombres.

QCM 5:

Question : L'analyse de données implique généralement :

- A. L'addition de nouvelles données.
- B. La suppression de données inutiles.
- C. L'organisation, la synthèse et l'interprétation des données.
- D. Le stockage des données brutes.

Solution : C. L'organisation, la synthèse et l'interprétation des données.

QCM 6:

Question : Qu'est-ce que l'échelle de ratio ?

- A. Une échelle qui classe les données en catégories sans ordre spécifique.
- B. Une échelle qui mesure des données sur une échelle continue avec un zéro absolu.
- C. Une échelle qui classe les données par ordre hiérarchique.
- D. Une échelle qui mesure des données sur une échelle continue sans zéro absolu.

Solution : B. Une échelle qui mesure des données sur une échelle continue avec un zéro absolu.

QCM 7:

Question : Quelle est la différence entre les données quantitatives et qualitatives ?

- A. Les données quantitatives sont exprimées en pourcentages, tandis que les données qualitatives sont exprimées en nombres.
- B. Les données quantitatives sont mesurées sur une échelle de ratio, tandis que les données qualitatives sont catégorielles.
- C. Les données quantitatives sont toujours subjectives, tandis que les données qualitatives sont objectives.
- D. Les données quantitatives concernent la qualité, tandis que les données qualitatives concernent la quantité.

Solution : B. Les données quantitatives sont mesurées sur une échelle de ratio, tandis que les données qualitatives sont catégorielles.

QCM 8:

Question : Quel est l'objectif principal de l'analyse de données ?

- A. Mesurer la variabilité des données.
- B. Stocker des données en toute sécurité.
- C. Comprendre, résumer et interpréter les données.
- D. Collecter des données brutes.

Solution : C. Comprendre, résumer et interpréter les données.

QCM 9:

Question : Quelle échelle de mesure catégorise les données en groupes avec un ordre spécifique ?

- A. Échelle nominale.
- B. Échelle ordinale.
- C. Échelle d'intervalle.
- D. Échelle de ratio.

Solution : B. Échelle ordinale.

QCM 10:

Question : Quelle est la principale différence entre les données qualitatives et les données quantitatives ?

- A. Les données qualitatives sont exprimées en pourcentages, tandis que les données quantitatives sont exprimées en nombres.
- B. Les données qualitatives sont basées sur des mesures, tandis que les données quantitatives sont basées sur des observations.
- C. Les données qualitatives sont catégorielles, tandis que les données quantitatives sont numériques.
- D. Les données qualitatives sont toujours objectives, tandis que les données quantitatives sont subjectives.

Solution : C. Les données qualitatives sont catégorielles, tandis que les données quantitatives sont numériques.

QCM 11:

Question : Quelle échelle de mesure est la plus élevée et permet l'utilisation des opérations mathématiques telles que l'addition et la multiplication ?

- A. Échelle nominale.
- B. Échelle ordinale.
- C. Échelle d'intervalle.
- D. Échelle de ratio.

Solution : D. Échelle de ratio.

QCM 12:

Question : Les données qualitatives sont principalement basées sur :

- A. Les chiffres.
- B. Les observations.
- C. Les calculs mathématiques.
- D. Les pourcentages.

Solution : B. Les observations.

QCM 13:

Question : Quel est le rôle principal de l'analyse des données ?

- A. Générer des données brutes.
- B. Faciliter la collecte de données.
- C. Extraire des informations significatives à partir de données existantes.
- D. Stocker des données sans les examiner.

Solution : C. Extraire des informations significatives à partir de données existantes.

QCM 14:

Question : Quelle échelle de mesure est la plus basse et ne permet que de classer les données en catégories sans ordre spécifique ?

- A. Échelle nominale.
- B. Échelle ordinale.
- C. Échelle d'intervalle.
- D. Échelle de ratio.

Solution : A. Échelle nominale.

QCM 15:

Question : Quelle est la caractéristique principale des données qualitatives ?

- A. Elles sont exprimées en nombres.
- B. Elles sont basées sur des observations.
- C. Elles mesurent des quantités.
- D. Elles sont catégorielles.

Solution : D. Elles sont catégorielles.

Chapitre 2 :
Exploration des Variables
Individuelles : Analyse Univariée
(Approche des Techniques de
Description Statistique)

I. Description de la variable qualitative

La description d'une variable qualitative consiste à présenter les effectifs, c'est-à-dire le nombre d'individus de l'échantillon pour chaque modalité de la variable, et les fréquences, c'est-à-dire le nombre de réponses associées aux modalités de la variable étudiée. En effet, dans de nombreux cas, le chargé d'étude cherche à répondre à une série de questions ne concernant qu'une seule et même variable.

II. Description de la variable quantitative

Plusieurs critères permettent de décrire une variable quantitative :

- les mesures de la tendance centrale : moyenne, médiane, mode.
- les mesures de la dispersion : étendue, variance, écart type, coefficient de variation.
- les représentations graphiques : histogrammes, diagramme en bâton.

III. Tableau statistique

Pour créer un tableau statistique il faut déterminer :

- 1) Effectif totale n : le nombre de toutes les valeurs prises par la variable.
- 2) Effectif n_i : nombre d'apparitions de la valeur x_i dans la population ou dans l'échantillon.

$$\sum_{i=1}^J n_i = n_1 + n_2 + \dots + n_J = n.$$

- 3) Fréquence f_i associée à la valeur x_i

$$\left\{ \begin{array}{l} f_i = \frac{n_i}{n}, \\ \sum_{i=1}^J f_i = f_1 + f_2 + \dots + f_J = 1. \end{array} \right.$$

- 4) Pourcentage p_i associé à la valeur x_i

$$\left\{ \begin{array}{l} p_i = 100 \times f_i \%, \\ \sum_{i=1}^J p_i = p_1 + p_2 + \dots + p_J = 100 \%. \end{array} \right.$$

5) Effectif cumulé Ni

$$\left\{ \begin{array}{l} N_1 = n_1, \\ N_2 = n_1 + n_2, \\ N_3 = n_1 + n_2 + n_3, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ N_J = n_1 + n_2 + \dots + n_J = n. \end{array} \right.$$

6) Fréquence cumulé Fi

$$\left\{ \begin{array}{l} F_1 = f_1, \\ F_2 = f_1 + f_2, \\ F_3 = f_1 + f_2 + f_3, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ F_J = f_1 + f_2 + \dots + f_J = 1. \end{array} \right.$$

Exemple 1 : Variable qualitative nominale

On note C : célibataire, M : marié, V : veuf, D : divorcé. On s'intéresse à la variable X= (état-civil) sur une population de n = 20 personnes. Considérons la série statistique Suivante:

M D M C C M C C C M C M V M V D C C M C

Tableau statistique :

x_i	n_i	f_i	$p_i\%$	N_i	F_i
C	9	0.45	45	9	0.45
M	7	0.35	35	16	0.75
V	2	0.10	10	18	0.85
D	2	0.10	10	20	1

Par exemple : le nombre d'apparition de la valeur $x_2 = M$ dans la série statistique est $n_2 = 7$, sa fréquence est $f_2 = n_2/n = 7/20 = 0,35$, son pourcentage $p_2 = 100 * f_2 = 100 * 0,35 = 35$

%, l'effectif cumulé $N_2 = n_1 + n_2 = 9 + 7 = 16$ et la fréquence cumulée $F_2 = f_1 + f_2 = 0,45 + 0,35 = 0,75$.

Exemple 2 : Variable qualitative ordinale

On interroge une population de $n = 50$ personnes sur leur dernier diplôme obtenu. On note : Sd : Sans diplôme, P : Primaire, Se : Secondaire, Su : Supérieur non-universitaire et U Universitaire.

Sd Sd Sd Sd P P P P P P P P P P Se Se Su
 Se Se Se Se Se Se Se Se Se Se Se Se Su Su Su
 Su Su Su Su U U U U U U U U U U Su

Tableau statistique

x_i	n_i	N_i	f_i	p_i	F_i
Sd	4	4	0.08	8	0.08
P	11	15	0.22	22	0.30
Se	14	29	0.28	28	0.58
Su	9	38	0.18	18	0.76
U	12	50	0.24	24	1

Exemple 3 : Variable quantitative discrète

Un quartier est composé d'une population de 50 ménages, et la variable X représente le nombre de personnes par ménage. Les valeurs de la variable sont :

1 1 1 1 1 2 2 2 2 2
 2 2 2 2 3 3 3 3 3 3
 3 3 3 3 3 3 3 3 3 4
 4 4 4 4 4 4 4 4 4 5
 5 5 5 5 5 6 6 6 8 8

Tableau statistique

x_i	n_i	N_i	f_i	F_i
1	5	5	0.10	0.10
2	9	14	0.18	0.28
3	15	29	0.30	0.58
4	10	39	0.20	0.78
5	6	45	0.12	0.90
6	3	48	0.06	0.96
8	2	50	0.04	1

IV. Les mesures de tendance :

Les mesures de tendance centrale permettent de résumer un ensemble de données relatives à une variable quantitative. Elles permettent de déterminer une valeur «typique» ou centrale autour de laquelle des données ont tendance à se rassembler.

Les indicateurs de tendance centrale sont :

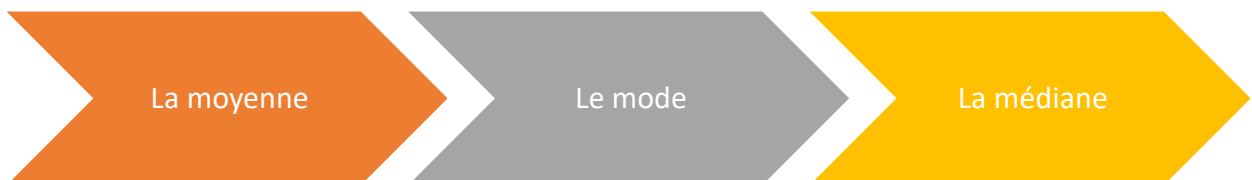
- La moyenne : Somme des valeurs de toutes les observations divisée par l'effectif.

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N} = \frac{\sum_{n=1}^N X_n}{N}$$

- La médiane : la valeur centrale d'une série statistique dont les valeurs observées ont été rangées dans l'ordre croissant, est la valeur qui partage la population étudiée en deux sous-ensembles de même effectif.

$$M = \begin{cases} X_{N/2} & \text{si } N \text{ est pair} \\ X_{\lfloor N/2 \rfloor + 1} & \text{si } N \text{ est impair} \end{cases}$$

- Le mode : Le mode représente la valeur présentant la plus grande fréquence d'occurrence. Il est défini comme la valeur la plus fréquente dans la série d'observation. (Cette valeur n'est pas nécessairement unique).



V. Les mesures de dispersion :

Comme le nom l'indique, les indicateurs de dispersions permettent de mesurer comment les données se «répartissent».

Les indicateurs de dispersions sont :

- L'écart type :

L'écart type est la mesure de la dispersion autour de la moyenne, exprimée dans la même unité que la variable. Il est défini comme la racine carrée de la variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N} \sum_{n=1}^N (X_n - \bar{X})^2}$$

- La variance :

La variance est la mesure de la dispersion autour de la moyenne, égale à la somme des carrés des écarts par rapport à la moyenne, divisée par le nombre d'observations moins un.

$$\sigma^2 = \frac{1}{N} \sum_{n=1}^N (X_n - \bar{X})^2$$

- L'étendue :

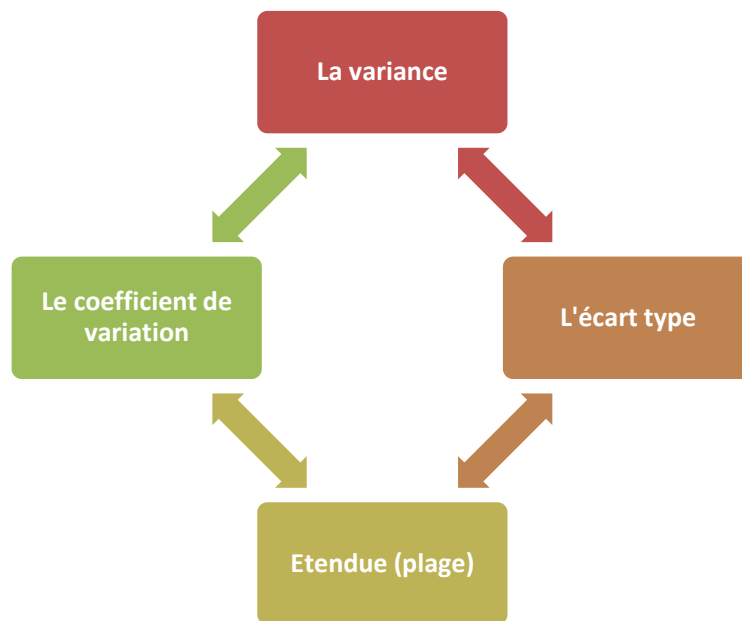
L'étendue d'une série statistique est l'écart entre sa plus grande valeur et sa plus petite. En d'autre terme, l'étendue c'est la différence entre la valeur la plus élevée et la valeur la plus bas.

$$e = \max X - \min X$$

- Le coefficient de variation :

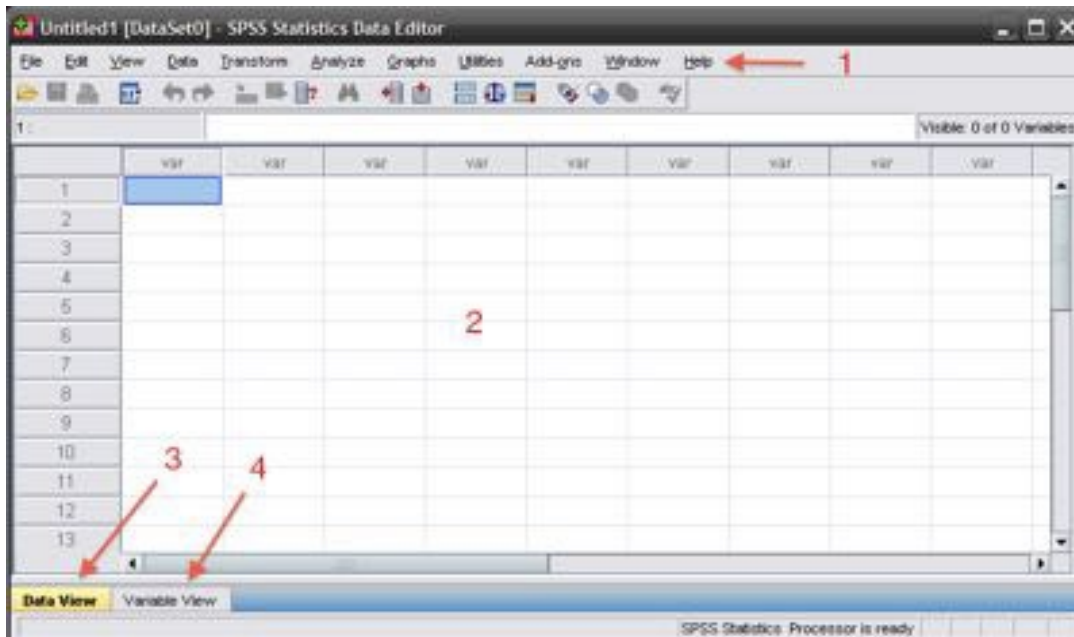
Le coefficient de variation est le rapport de l'écart type à la moyenne, exprimé en pourcentage. Son objet est de mesurer le degré de variation de la moyenne d'un échantillon à l'autre, lorsque ceux-ci sont issus de la même distribution.

CV=l'écart type/ la moyenne *100



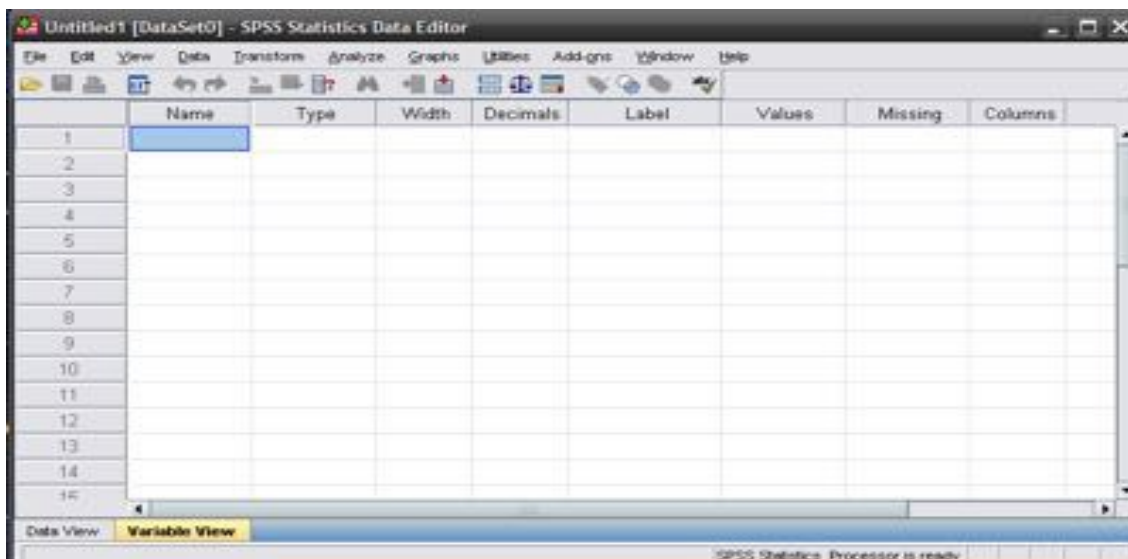
VI. Présentation du logiciel SPSS

1. Fenêtre principale du logiciel SPSS



Elle se compose de plusieurs parties :

1. La barre des menus et des boutons de commande
2. La fenêtre principale de SPSS pour l'entrée et le traitement des données
3. Ici nous sommes sur la fenêtre « Data View », c'est-à-dire la fenêtre des données
4. Si vous cliquez sur « Variable View », vous vous trouvez alors sur la fenêtre de résumé des variables :



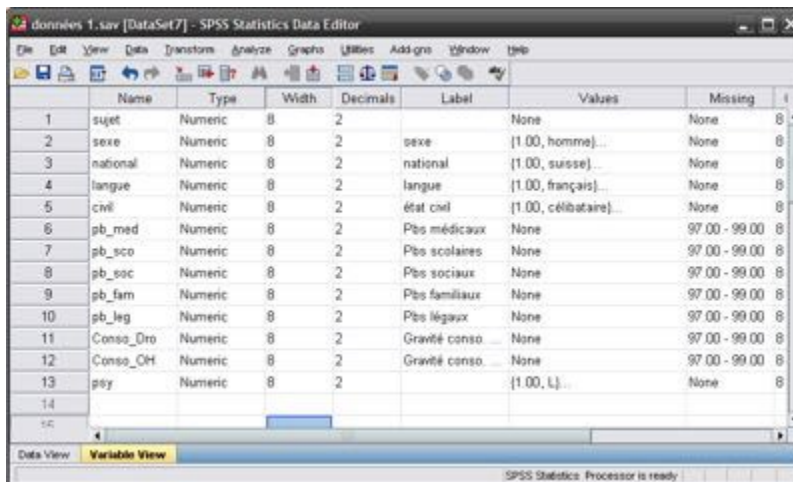
Pour la saisie des données, il faut dans un premier temps définir le nom des variables que l'on utilise.

2. Définition d'une variable et de ses propriétés

Cette opération s'effectue dans la fenêtre Variable View. Vous pouvez passer d'une fenêtre à l'autre en cliquant sur les onglets correspondants dans la barre en bas à gauche de la fenêtre.

Sous la colonne Name on indique le nom de la variable (nom sans espaces et sans accents), par exemple âge pour l'âge des participant-e-s. On appuie ensuite sur Enter pour valider cette entrée.

Des propriétés par défaut s'inscrivent alors sur la ligne qui concerne cette variable :



	Name	Type	Width	Decimals	Label	Values	Missing	
1	sujet	Numeric	8	2		None	None	8
2	sexe	Numeric	8	2	sexe	[1.00, homme]...	None	8
3	national	Numeric	8	2	national	[1.00, suisse]...	None	8
4	langue	Numeric	8	2	langue	[1.00, français]...	None	8
5	civil	Numeric	8	2	état civil	[1.00, célibataire]...	None	8
6	pb_med	Numeric	8	2	Pts médicaux	None	97.00 - 99.00	8
7	pb_sco	Numeric	8	2	Pts scolaires	None	97.00 - 99.00	8
8	pb_soc	Numeric	8	2	Pts sociaux	None	97.00 - 99.00	8
9	pb_fam	Numeric	8	2	Pts familiaux	None	97.00 - 99.00	8
10	pb_leg	Numeric	8	2	Pts légaux	None	97.00 - 99.00	8
11	Conso_Dro	Numeric	8	2	Gravité conso	None	97.00 - 99.00	8
12	Conso_OH	Numeric	8	2	Gravité conso	None	97.00 - 99.00	8
13	psy	Numeric	8	2		[1.00, L]...	None	8
14								

3. Enregistrement des données

La première fois que vous enregistrez vos données (même principe que dans Word) : On enregistre les données en exécutant l'option **Save As** du menu **File**. Dans un premier temps, on choisit l'emplacement où l'on veut enregistrer le fichier (**Look in :**).

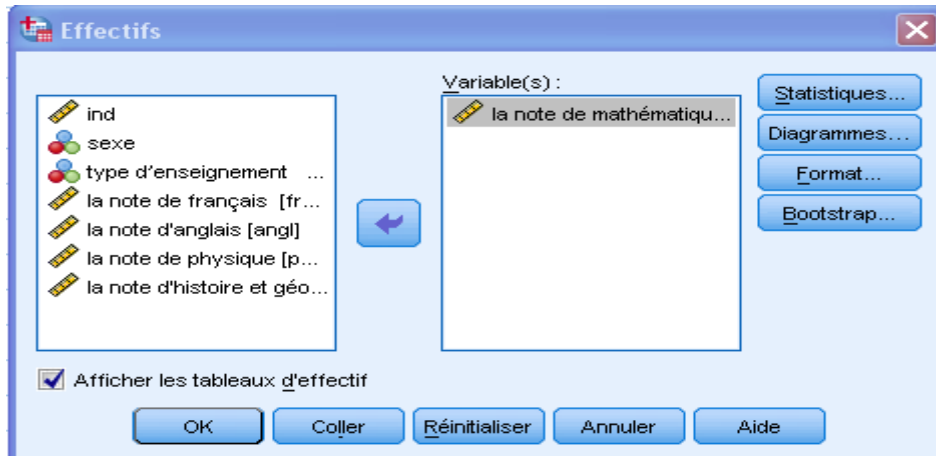
Dans un deuxième temps, on entre un nom dans la fenêtre **File name:** ; ensuite on clique sur le bouton **Save**. L'extension d'un fichier de données SPSS est « **.sav** » et cette extension s'inscrit automatiquement à la suite du nom de votre fichier. Si vous voyez un fichier du type « nom.sav » sachez que c'est un fichier de données SPSS. Il est important de vérifier de bien avoir enregistré le fichier « **.sav** » depuis la fenêtre du fichier de données SPSS, et non seulement le fichier de résultats (Output) dont l'extension est « **.spv** ».

Pour ajouter des données à un fichier déjà existant (même principe que dans Word) : lorsque votre document est ouvert, cliquez simplement sur **Save** (pas **Save as**).

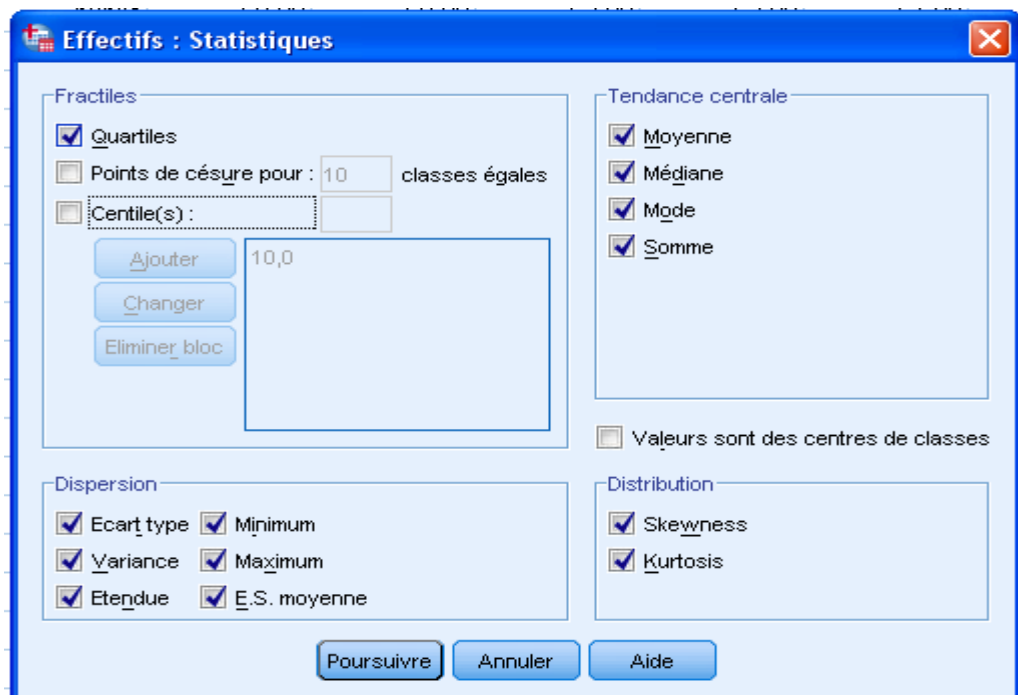
VII. Application SPSS

Exemple 1 :

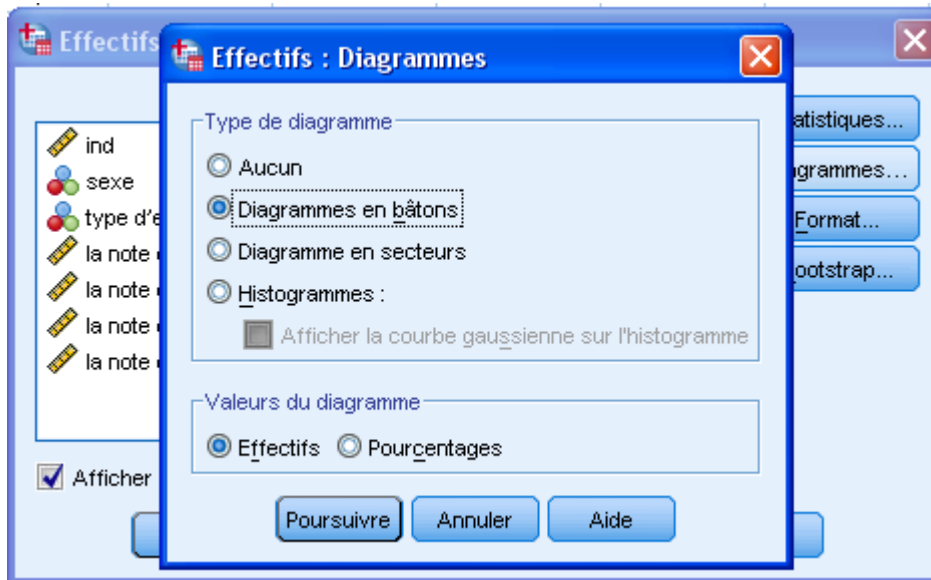
- Ouvrir le fichier SPSS et analyser la variable la note de mathématiques.
- Utiliser la commande : **Analyse + Statistiques descriptives + Fréquences**
- Sélectionner la variable pour laquelle on souhaite connaître les caractéristiques statistiques, puis la déplacer à l'aide de la case flèche.



- Cliquer sur **statistiques** et sélectionner les éléments désirés :
 1. Pour une variable nominale : mode, distribution des fréquences, minimum, maximum ;
 2. Pour une variable ordinale : mode, distribution des fréquences, minimum, maximum, médiane ;
 3. Pour une variable métrique : écart-type, moyenne, minimum, maximum



- Cliquer sur **Poursuivre**.
- Cliquer sur **Diagrammes** et sélectionner les éléments désirés :
 1. diagramme en bâtons : variable discrète ;
 2. graphique en secteur : caractère qualitatif ;
 3. histogrammes : variable continue.



On obtient le tableau des résultats suivants :

→ Effectifs

[Ensemble_de_données1] E:\ENSA2010\enseignement\GC\TP\Etudiant.sav

Statistiques

la note de mathématiques

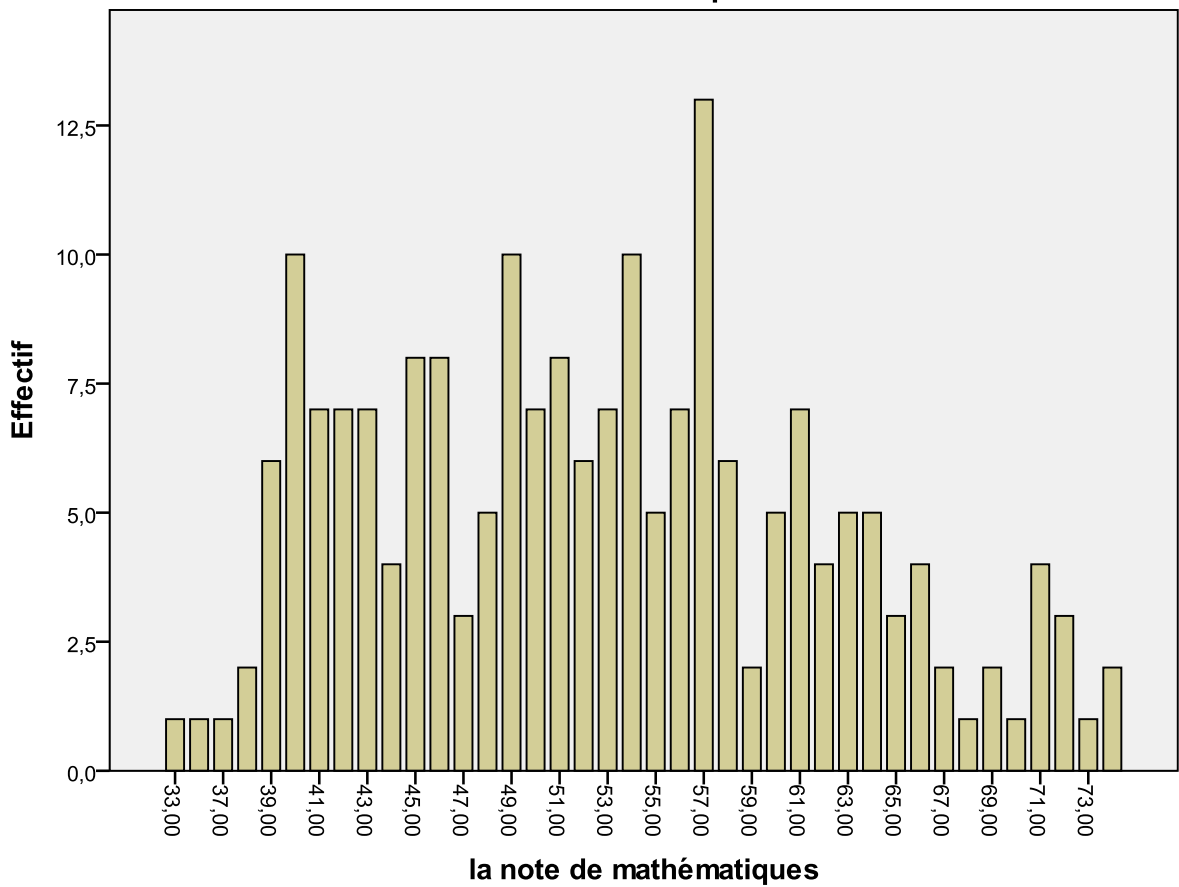
N	Valide	200
	Manquante	0
Moyenne		52,6450
Erreur std. de la moyenne		,66245
Médiane		52,0000
Mode		57,00
Ecart-type		9,36845
Variance		87,768
Asymétrie		,287
Erreur std. d'asymétrie		,172
Aplatissement		-,649
Erreur std. d'aplatissement		,342
Intervalle		42,00
Minimum		33,00
Maximum		75,00
Somme		10529,00
Centiles	25	45,0000
	50	52,0000
	75	59,0000

la note de mathématiques

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	33,00	1	,5	,5	,5
	35,00	1	,5	,5	1,0
	37,00	1	,5	,5	1,5
	38,00	2	1,0	1,0	2,5
	39,00	6	3,0	3,0	5,5
	40,00	10	5,0	5,0	10,5
	41,00	7	3,5	3,5	14,0
	42,00	7	3,5	3,5	17,5
	43,00	7	3,5	3,5	21,0
	44,00	4	2,0	2,0	23,0
	45,00	8	4,0	4,0	27,0
	46,00	8	4,0	4,0	31,0
	47,00	3	1,5	1,5	32,5
	48,00	5	2,5	2,5	35,0
	49,00	10	5,0	5,0	40,0
	50,00	7	3,5	3,5	43,5
	51,00	8	4,0	4,0	47,5
	52,00	6	3,0	3,0	50,5
	53,00	7	3,5	3,5	54,0
	54,00	10	5,0	5,0	59,0
	55,00	5	2,5	2,5	61,5
	56,00	7	3,5	3,5	65,0
	57,00	13	6,5	6,5	71,5
	58,00	6	3,0	3,0	74,5
	59,00	2	1,0	1,0	75,5
	60,00	5	2,5	2,5	78,0
	61,00	7	3,5	3,5	81,5
	62,00	4	2,0	2,0	83,5
	63,00	5	2,5	2,5	86,0

64,00	5	2,5	2,5	88,5
65,00	3	1,5	1,5	90,0
66,00	4	2,0	2,0	92,0
67,00	2	1,0	1,0	93,0
68,00	1	,5	,5	93,5
69,00	2	1,0	1,0	94,5
70,00	1	,5	,5	95,0
71,00	4	2,0	2,0	97,0
72,00	3	1,5	1,5	98,5
73,00	1	,5	,5	99,0
75,00	2	1,0	1,0	100,0
Total	200	100,0	100,0	

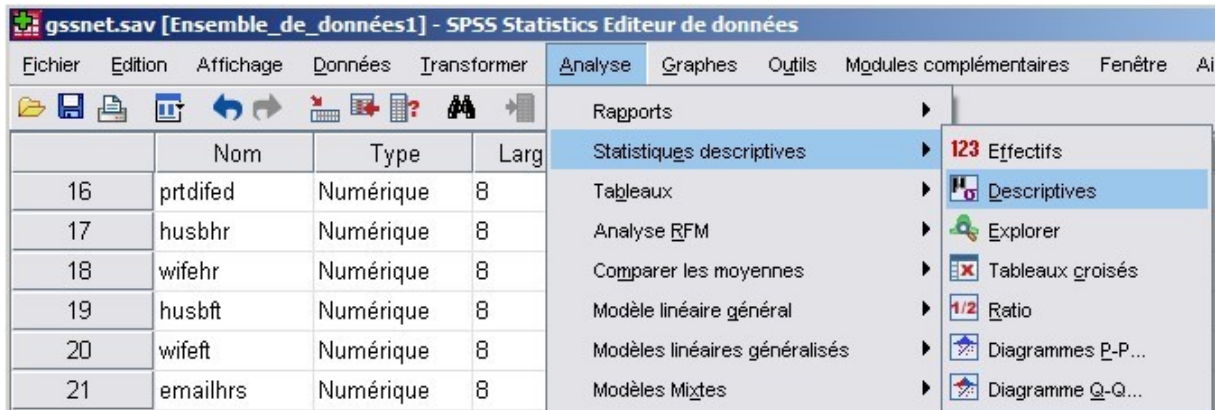
la note de mathématiques



Exemple 2 :

Etape 1 :

Pour réaliser une analyse descriptive, cliquez sur Analyse dans la barre d'outils. Choisissez Statistiques descriptives dans le menu déroulant, puis choisissez Descriptives dans le second menu déroulant.



Ensuite, vous sélectionnez la ou les variable(s) choisie(s) et vous la (ou les) transférez dans la boîte de variable de droite à l'aide de la flèche.



Lorsque vous cliquez sur le bouton Options dans la première boîte de dialogue, vous obtenez cette nouvelle boîte. Celle-ci permet d'ajouter des informations supplémentaires à celles données par défaut dans les tableaux de résultats. En plus de la moyenne, de l'écart-type, de l'étendue et des valeurs minimales et maximales, les indices suivants sont disponibles en option.



Vous pouvez alors choisir quelles informations vous désirez obtenir par rapport à la dispersion, la distribution et l'ordre de présentation des informations.

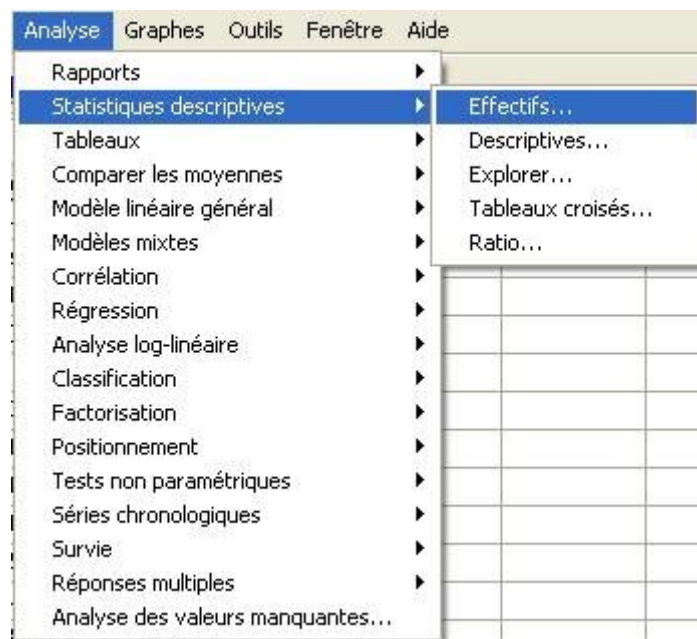
Etape 2

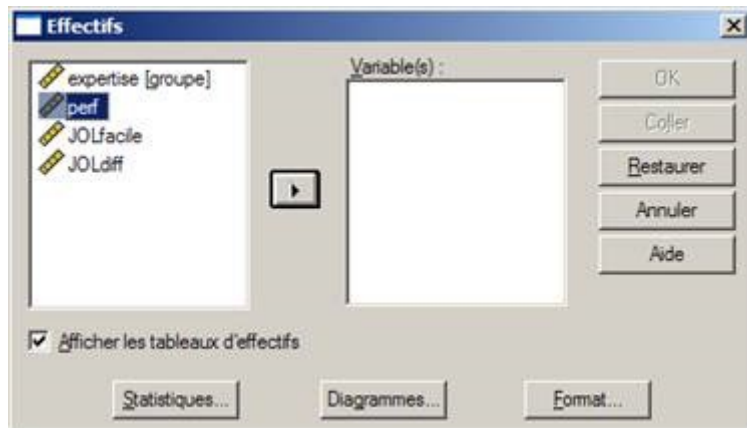
Il existe plusieurs manières de calculer une moyenne, un écart-type et autres paramètres descriptifs sur SPSS.

Pour calculer la moyenne d'une variable quantitative, une des procédures est la suivante :

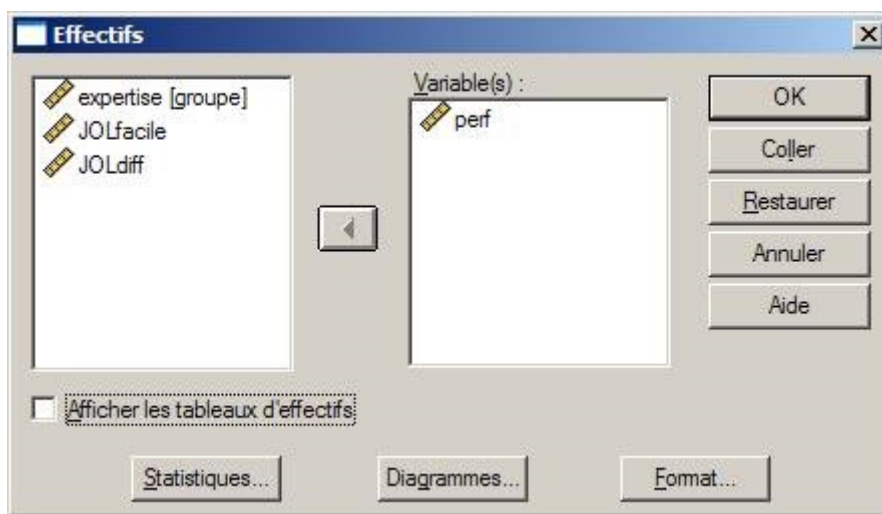
Choisir dans le menu déroulant : Analyse, Statistiques Descriptives, Effectifs....

Pour ouvrir la boîte de dialogue Effectifs

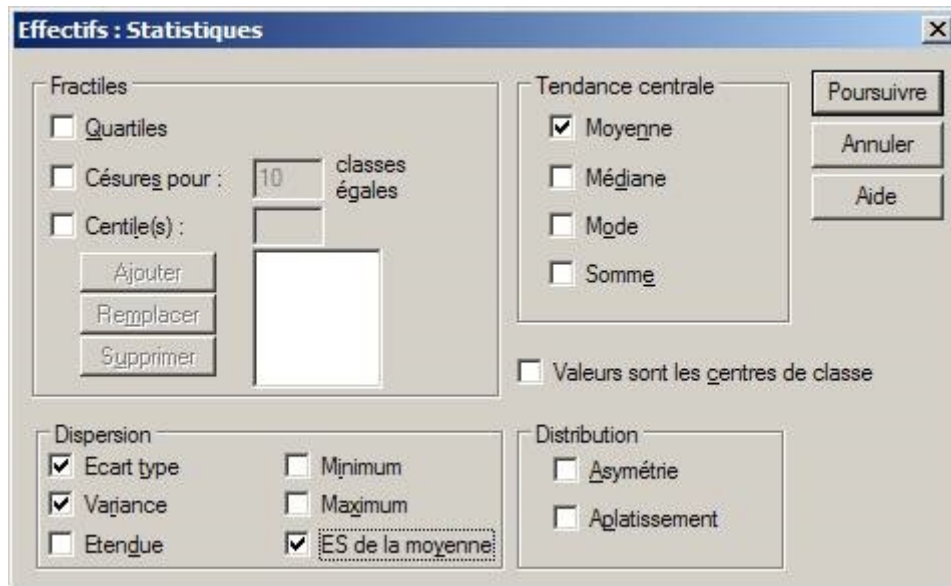


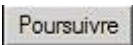



Pour calculer la moyenne de la variable « Performance » notée *Perf* dans le fichier de données SPSS, surlignez *Perf*, et cliquez sur ► pour les transférer dans la zone Variable(s).



Cliquez ensuite sur **Statistiques...** pour obtenir la boîte de dialogue Effectifs : statistiques (Figure 3) et cochez Moyenne, Ecart-type, Variance, ES (Erreur standard) de la moyenne.



Cliquez ensuite sur  . Vous revenez alors à la boîte de dialogue Effectifs et ensuite cliquez sur .

Etape 3

Tableau de statistiques descriptives

Statistiques		
Age of respondent		
N	Valide	1417
	Manquant	2
Moyenne		46,56
Erreur standard de la moyenne		,460
Médiane		44,00
Mode		43
Ecart type		17,330
Variance		300,341
Plage		71
Minimum		18
Maximum		89
Somme		65979

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.

Etape 4 :

Dans l'exemple, 1 419 participants ont donné leur âge et que 1 417 ont répondu à la question portant sur le nombre d'années de scolarité.

La colonne « Intervalle » (étendue) ou « plage » indique l'écart entre 18 et 89 ans, soit la différence entre la valeur minimale et maximale.

Ce tableau vous indique que la performance moyenne est égale à 46,56 ; l'erreur standard de la moyenne est égale à ,46 ; l'écart-type est égal à 17,33 et la variance à 300,34.

VIII. La représentation graphique des données :

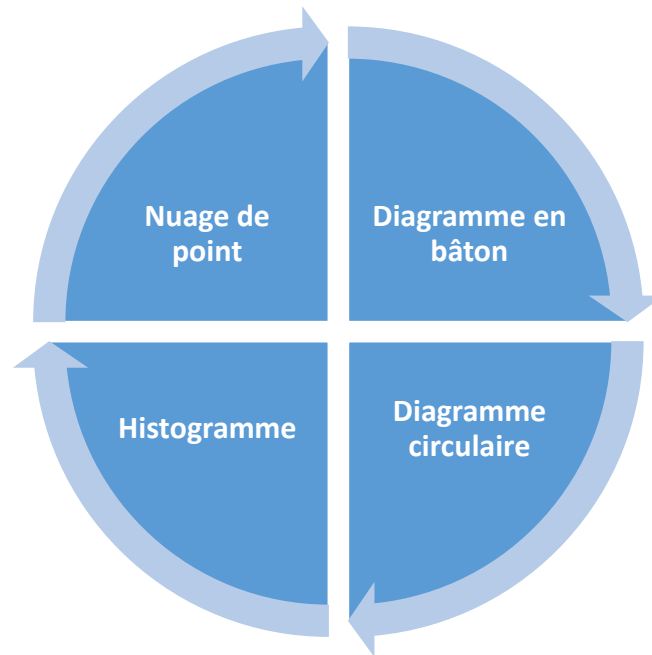


Diagramme en bâton

Les diagrammes en bâtons sont utilisés pour représenter des séries à variable discrète.

Exemple :

Le nombre d'employer	20	40	15	5
l'âge d'employer	25	30	35	40

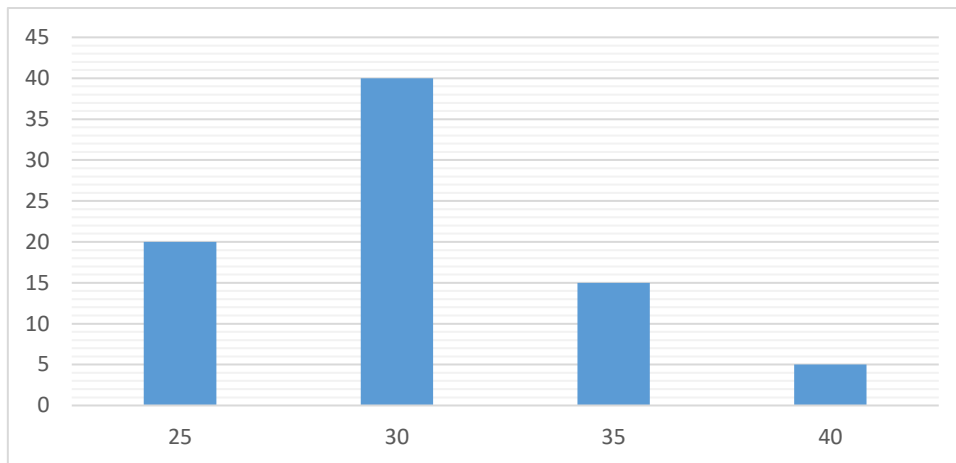
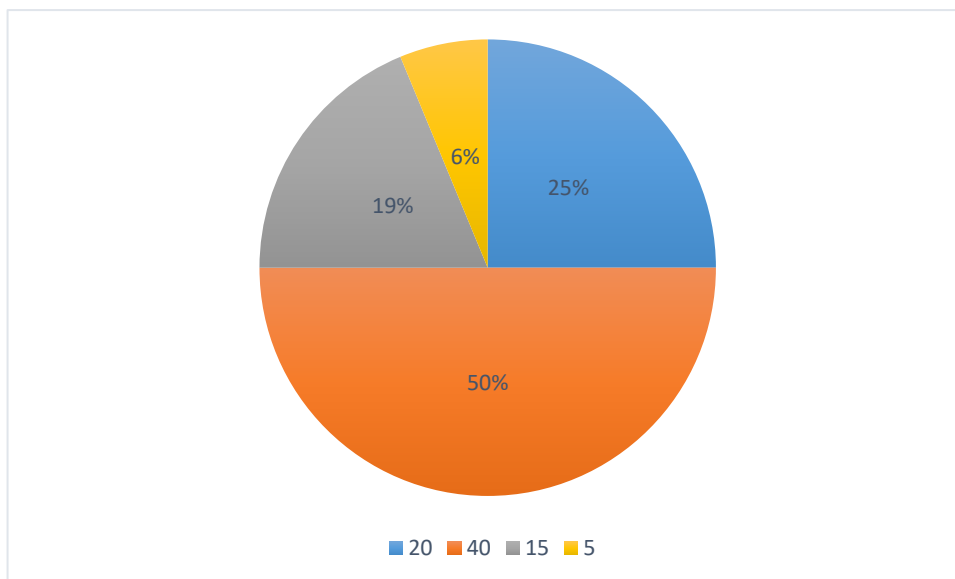


Diagramme circulaire

Un diagramme circulaire admet pour support un disque découpé en secteurs dont les aires sont proportionnelles aux pourcentages des différents constituants de la population statistique.

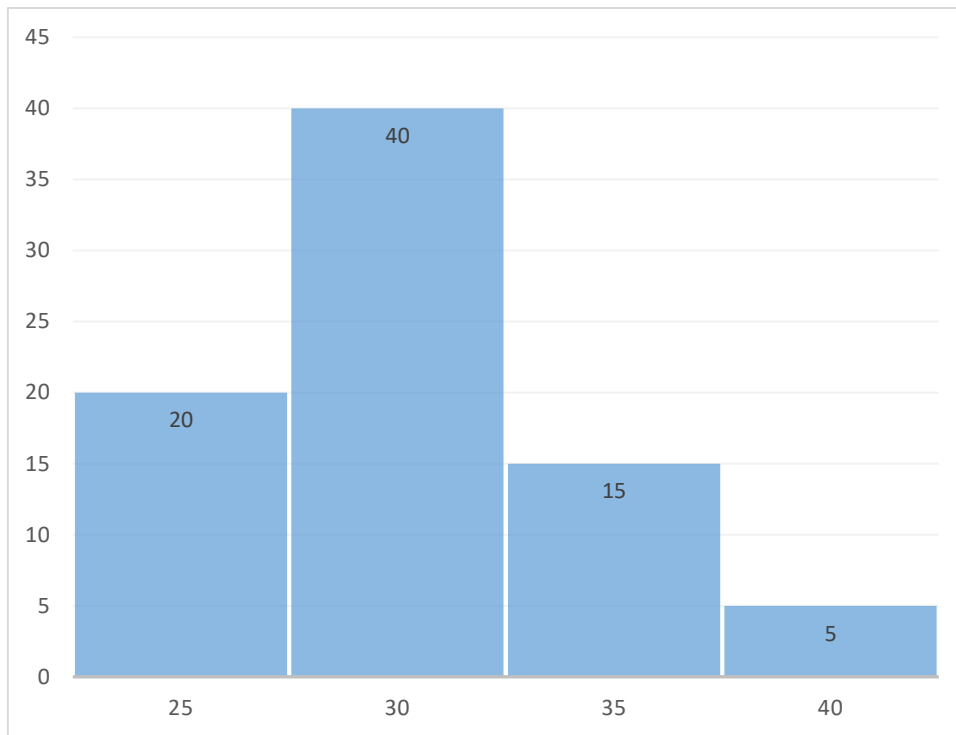
Exemple :



Histogramme

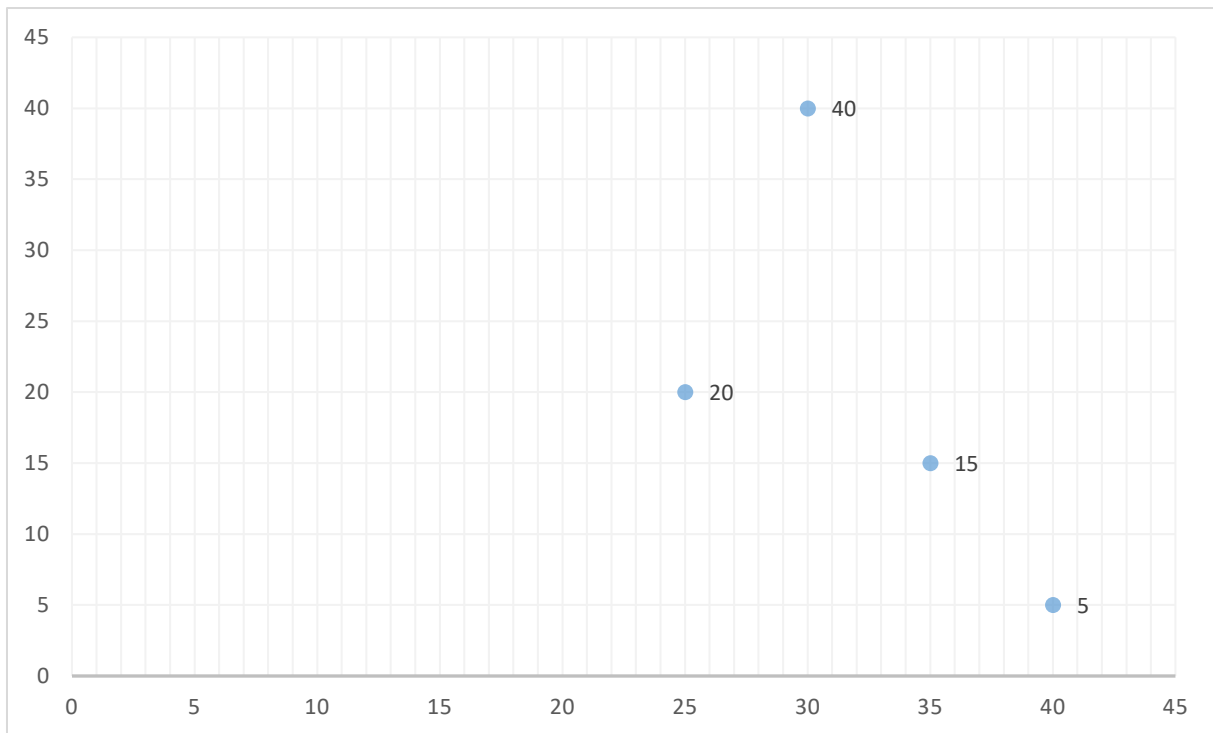
L'histogramme est un moyen rapide pour étudier la répartition d'une variable.

Exemple :



Nuage de point :

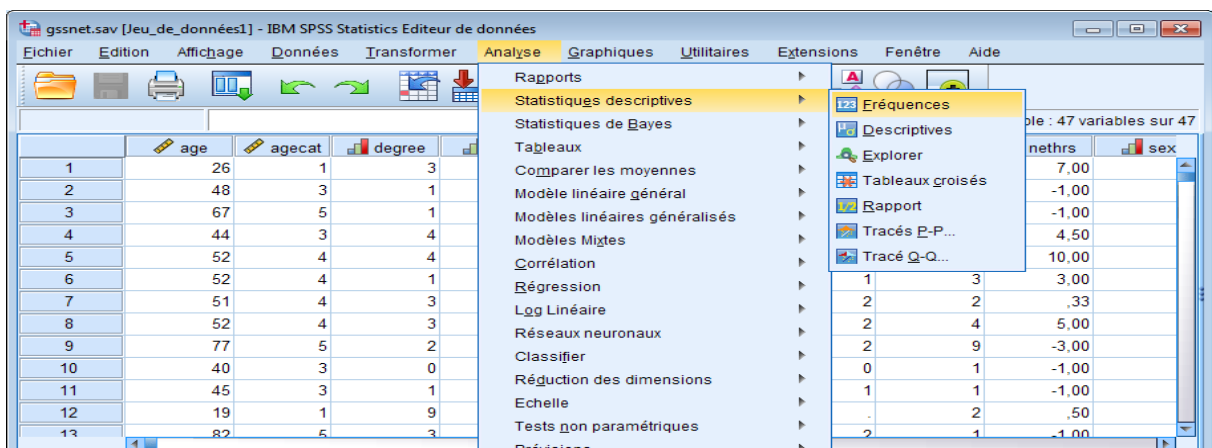
Un nuage de points est une représentation de données dépendant de plusieurs variables.



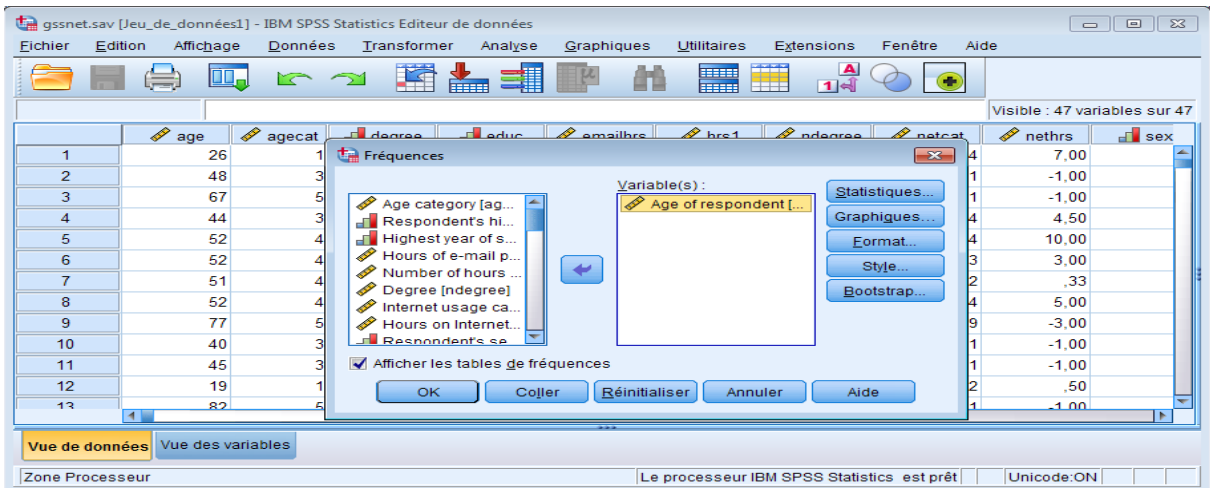
IX. Analyse des données avec SPSS : Représentation graphique

Méthode 1 :

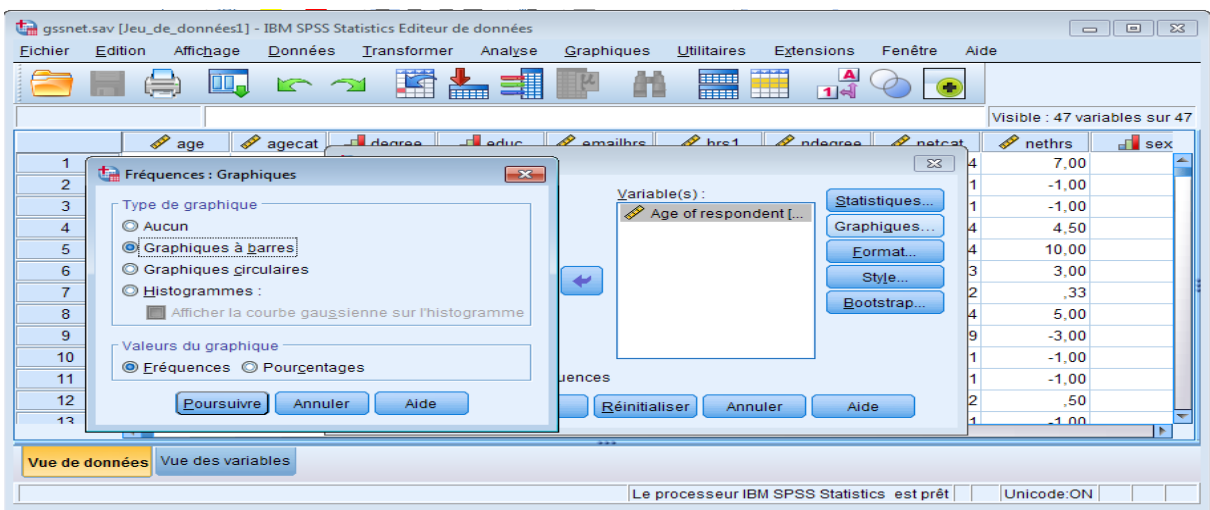
Pour réaliser la représentation graphique, cliquez sur Analyse dans la barre d'outils. Choisissez Statistiques descriptives dans le menu déroulant, puis choisissez fréquences dans le second menu déroulant.



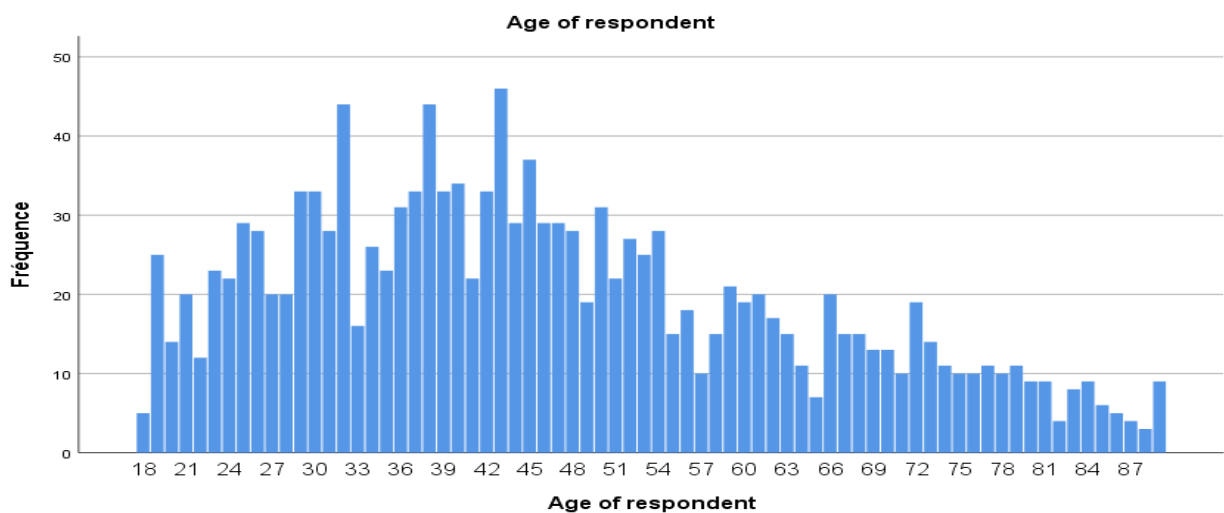
Choisir graphiques



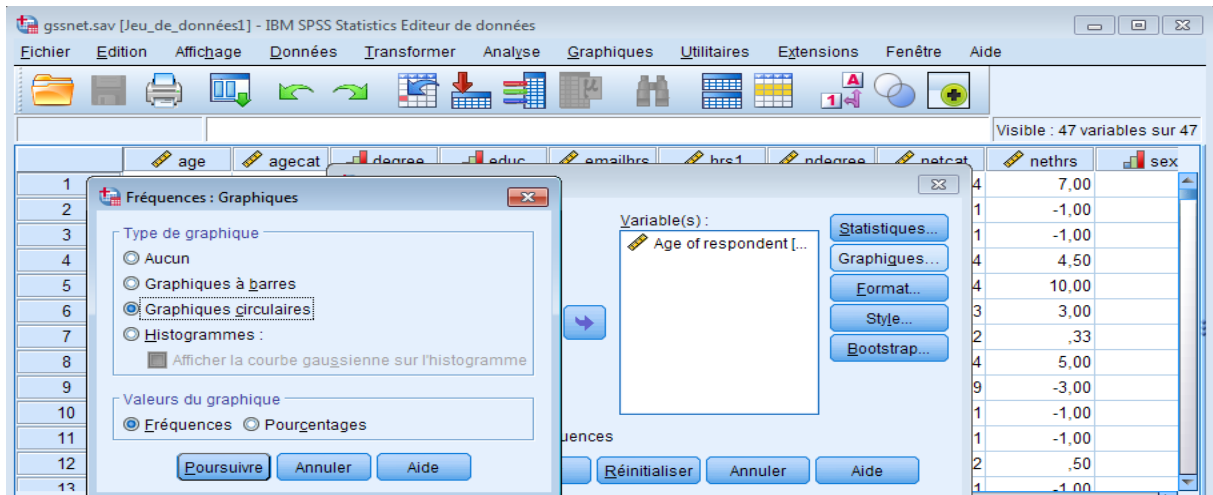
Choisir graphiques à barres



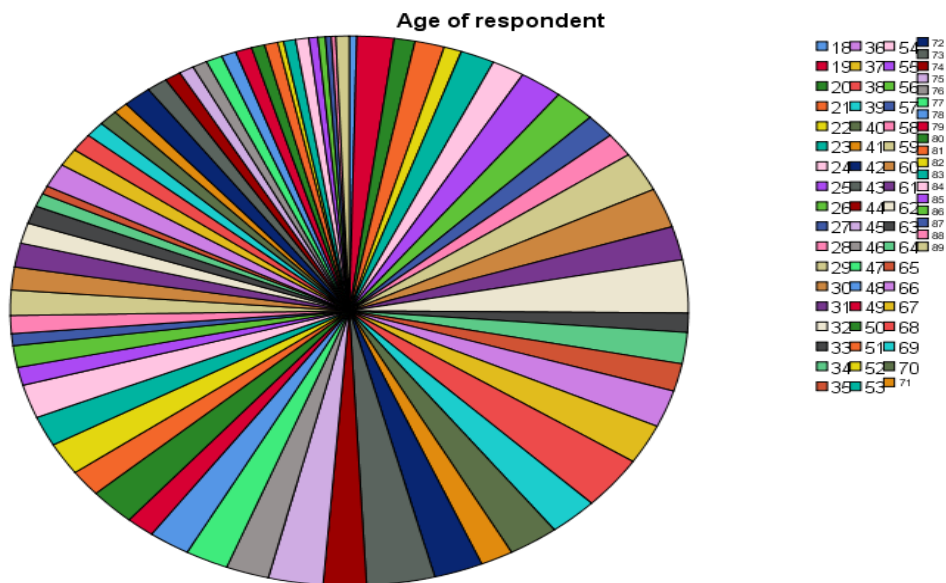
Résultat : graphiques à barres



Choisir graphiques circulaires

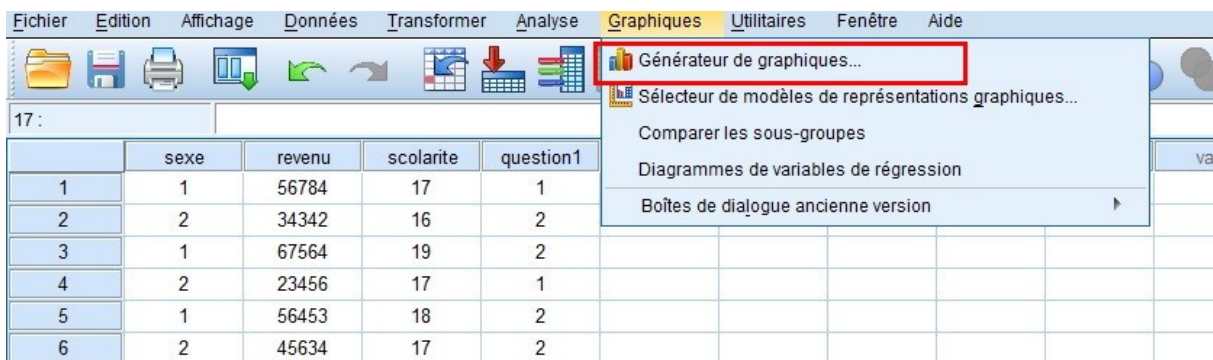


Résultat : graphiques circulaires



Méthode 2 :

Choisir le menu GRAPHIQUES.



Une petite fenêtre s'ouvre : confirmez l'utilisation du générateur de graphiques en cliquant au bas de la page sur OK.

Fichier Edition Affichage Données Transformer Analyse Graphiques Utilitaires Fenê

17 :

	sexe	revenu	scolarite	question1	var	var	v
1	1	56784	17	1			
2	2	34342	16	2			
3	1	67564	19	2			
17	1	52567	19	1			
18	2	25678	17	2			
19	2	57567	17	2			
20	1	34567	17	2			
21	1	43567	18	2			
22	1	43234	18	2			
23	1	67887	18	1			

Générateur de graphiques

Avant d'utiliser cette boîte de dialogue, le niveau de mesure doit être défini correctement pour chaque variable de votre graphique. De plus, si votre graphique contient des variables catégorielles, les libellés de valeur doivent être définis pour chaque catégorie.

Cliquez sur OK pour définir votre graphique.

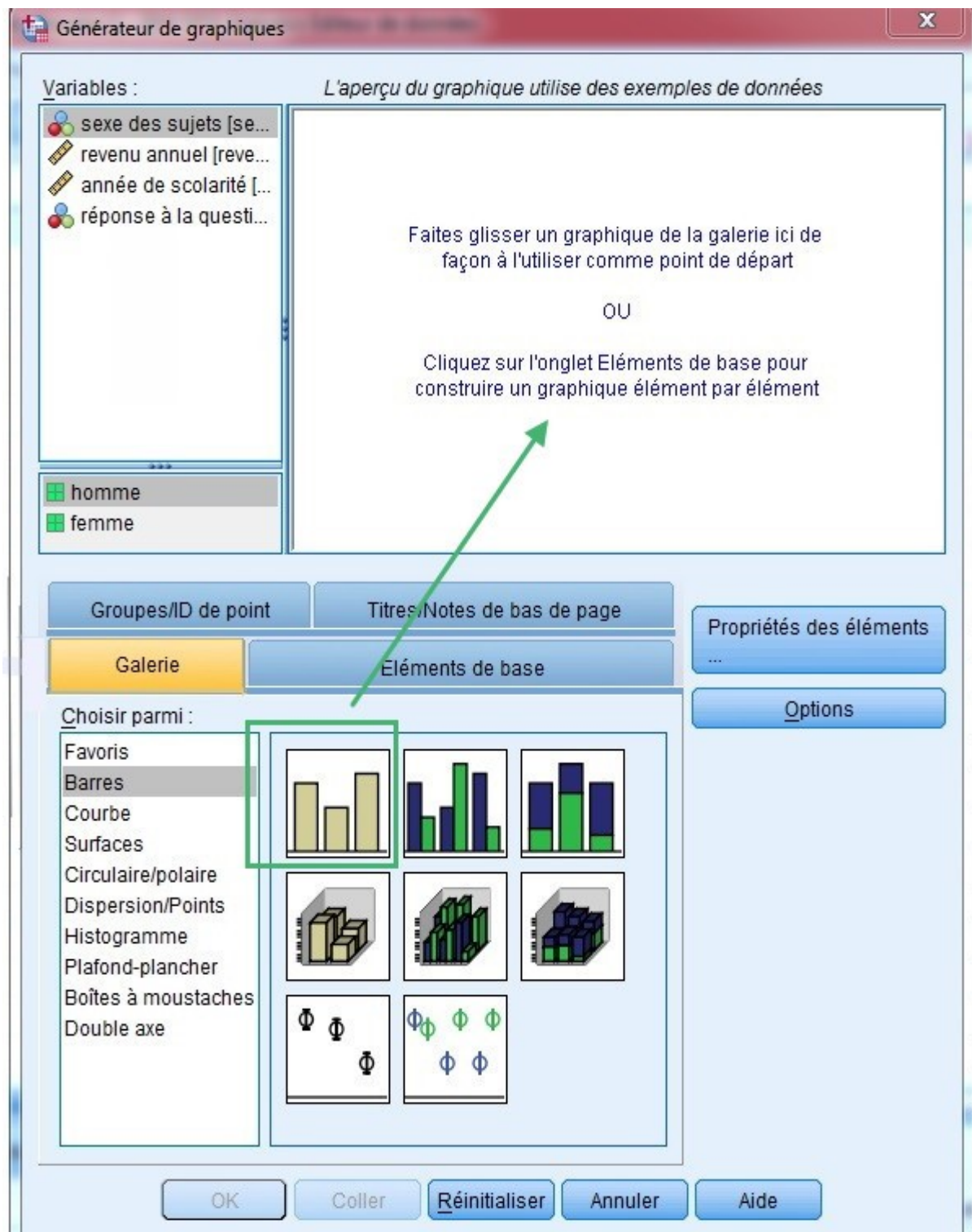
Cliquez sur Définir les propriétés de variable pour définir le niveau de mesure ou les libellés de valeur pour les variables de graphique.

Ne plus afficher cette boîte de dialogue

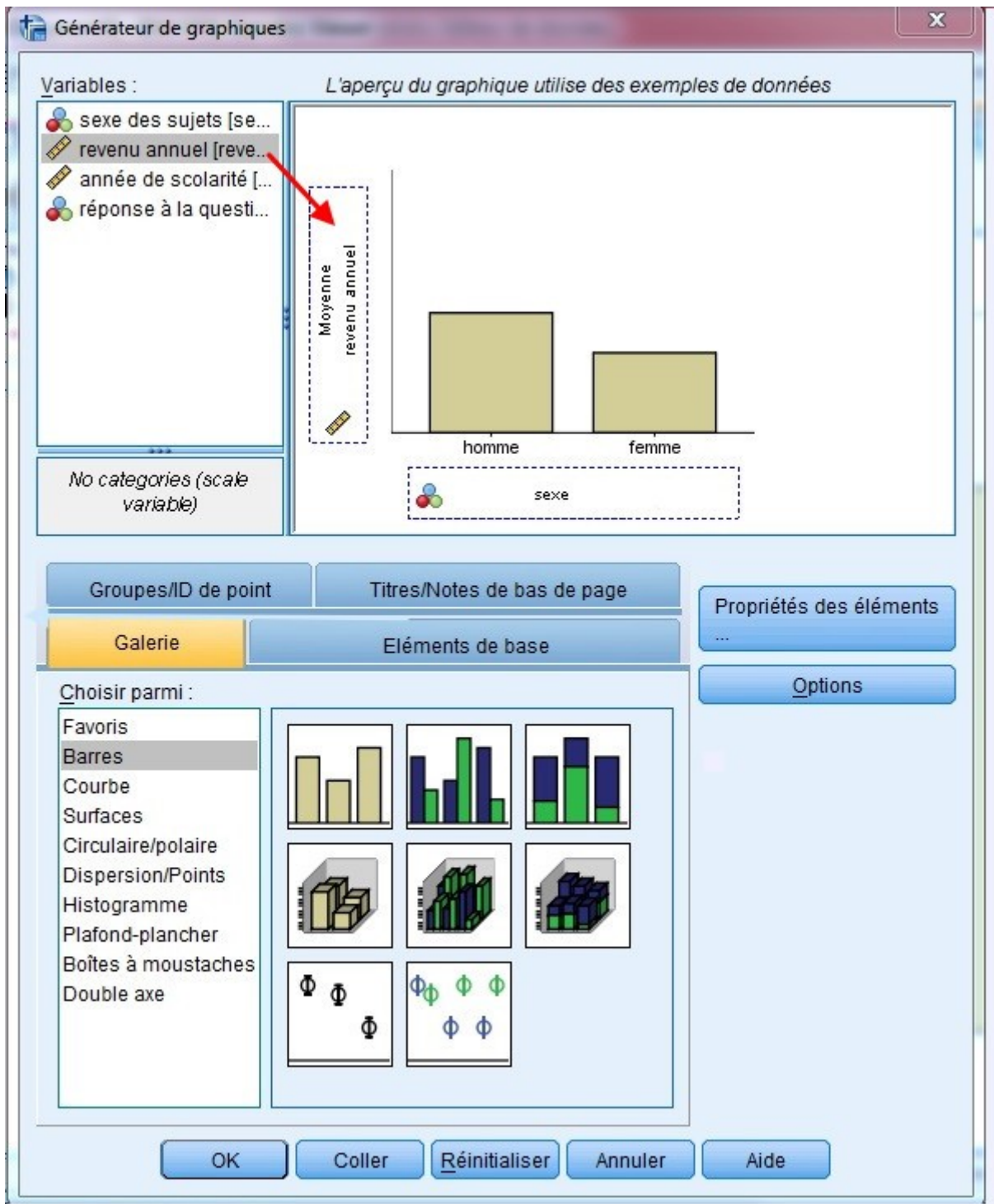
OK Définir les propriétés de variables

1

Vue de données Vue des variables



Ensuite, glissez vos variables X [Sexe] et Y [Revenu] dans les rectangles de la fenêtre du haut, comme suit :



Puis cliquez sur OK pour confirmer vos choix.

Chapitre 3 :
Analyse des Relations
Croisées : Utilisation de
Tableaux de Contingence et
le Test du Chi-2 (Khi 2)
(Exploration des Liens entre
Variables)

I. Analyse bivariée

L'objectif de cette partie est d'étudier deux variables différents X et Y et de rechercher s'il existe un lien entre ces deux variables. Chacune des deux variables peut être, soit quantitative, soit qualitative.

1. Deux variables qualitatives (tableau de contingence).
2. Une variable quantitative et une variable qualitative (boite à moustache).
3. Deux variables quantitatives (régression linéaire).

II. Deux variables qualitatives (tableau de contingence).

- La variable qualitative nominale est une information non mesurable. Elle présente des catégories que l'on nomme avec un nom codé qu'on appelle modalité.
- La variable peut avoir plusieurs modalités codées. Par exemple : la variable couleur avec plusieurs modalités (1-rouge, 2-vert, 3-bleu, 4-...), variable âge plusieurs modalités, et variable logement avec plusieurs modalités (1- logement individuel, 2- logement semi collectif, 3- logement collectif).

III. Tableau de contingence (tableau croisé)

- tableau croisé (appelé aussi tableau de contingence) examine la relation entre deux variables catégorielles.
- C'est un arrangement dans lequel les données sont classées selon deux variables catégorielles. Les catégories d'une variable apparaissent dans les lignes et les catégories de l'autre variable apparaissent dans les colonnes.

IV. Hypothèses (H_0) et (H_1)

L'hypothèse nulle (H_0) et l'hypothèse alternative (H_1) du test Chi-2 peuvent être exprimées de la manière suivante :

- H_0 : « (Variable 1) n'est pas associée à (Variable 2) » (**Absence de relation entre les 2 variables catégorielles**)
- H_1 : «(Variable 1) est associée à (Variable 2) » (**Existence de relation entre les 2 variables**).

V. Le principe du test Khi 2

- analyse bi-variée consiste à déterminer s'il existe une association entre deux variables qualitatives nominales.
- Déceler une éventuelle relation d'indépendance ou d'influence d'une variable sur une autre.

- Le Chi-deux est une analyse dite non-paramétrique, pas de prémisses des paramètres de la distribution de la variable (moyenne, écart-type et normalité).
- Significativité globale de la relation : d.d.l. = (l - 1) * (c - 1)
- Si $\alpha > 5\%$ la différence n'est pas significative au seuil de 5%
- Si $\alpha < \text{ou} = 5\%$ la différence est significative au seuil de 5%

Exemple :

On s'intéresse à une éventuelle relation entre la variable X= (A, B) de n = 200 et la variable Y = (C, D, E).

Tableau 1. Effectifs observés nij

X/Y	C	D	E	total
A	n11 = 10	n12 = 50	n13 = 20	n1. = 80
B	n21 = 20	n22 = 60	n23 = 40	n2. = 120
total	N .1 = 30	n.2 = 110	n.3 = 60	n = 200

Les nombres n1. ; n2. Et n.1 ; n.2 ; n.3 sont appelés effectifs marginaux.

Tableau 1. Fréquence observés nij

X/Y	C	D	E	total
A	f11 = 0,05	f12 = 0,25	f13 = 0,10	f1. = 0,40
B	f21 = 0,10	f22 = 0,30	f23 = 0,20	f2. = 0,60
total	f.1 = 0,15	f.2 = 0,55	f.3 = 0,30	1

Les nombres f1. ; f2. et f.1 ; f.2 ; f.3 sont appelées fréquences marginales.

VI. Tableau des effectifs théoriques

A partir des effectifs marginaux Li et Cj, on peut calculer les effectifs attendus lorsque X et Y sont indépendants.

Ces effectifs théoriques, notés eij , sont donnés par la formule

$$e_{ij} = \frac{L_i C_j}{n}.$$

Tableau 2. Effectifs théoriques eij

X/Y	C	D	E	total
A	n11 = 80*30/200	n12 = 80*110/200	n13 = 80*60/200	n1. = 80
B	n21 = 120*30/200	n22 = 120*110/200	n23 = 120*60/200	n2. = 120

total

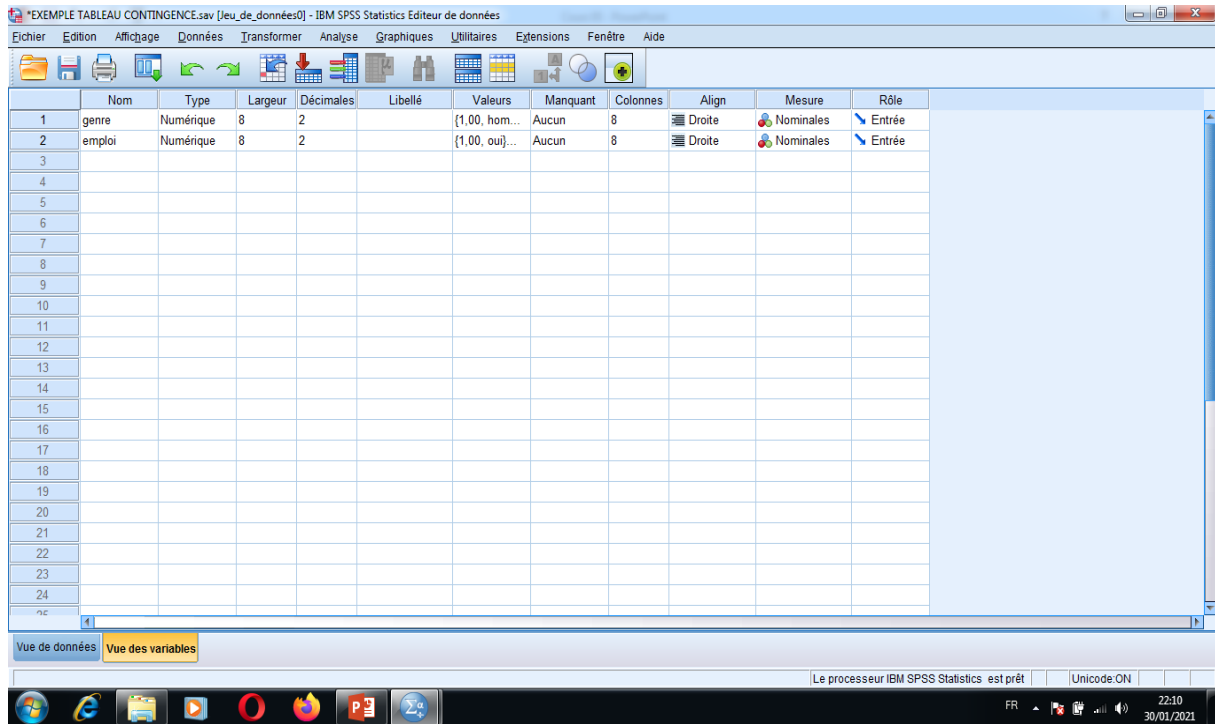
N .1 = 30

n.2 = 110

n.3 = 60

n = 200

VII. Application logiciel SPSS



Récapitulatif de traitement des observations

	Observations					
	Valide		Manquant		Total	
	N	Pourcentage	N	Pourcentage	N	Pourcentage
genre * emploi	50	100,0%	0	0,0%	50	100,0%

Tableau croisé genre * emploi

			emploi		Total
			oui	non	
genre	homme	Effectif	15	11	26
		% du total	30,0%	22,0%	52,0%
	femme	Effectif	9	15	24
		% du total	18,0%	30,0%	48,0%
Total		Effectif	24	26	50
		% du total	48,0%	52,0%	100,0%

Tests du khi-carré

	Valeur	ddl	Signification asymptotique (bilatérale)	Sig. exacte (bilatérale)	Sig. exacte (unilatérale)
khi-carré de Pearson	2,039 ^a	1	,153		
Correction pour continuité ^b	1,310	1	,252		
Rapport de vraisemblance	2,054	1	,152		
Test exact de Fisher				,171	,126
Association linéaire par linéaire	1,998	1	,158		
N d'observations valides	50				

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 11,52.

b. Calculée uniquement pour une table 2x2

VIII. Interprétation des résultats

- Signification du khi-deux de Person est SUPERIEUR à 0,05 (Seuil de signification).
- On rejette H1 (dépendance) et on accepte H0 (Indépendance)

Interprétation : il n'existe pas d'une dépendance entre la variable du genre et celle de l'employabilité.

Exercice 1 :

Pour cibler la clientèle d'un nouveau produit de consommation, une entreprise fait un sondage auprès de 321 personnes. L'intérêt dans le produit est noté par "aucun intérêt", "un intérêt mineur" ou un "intérêt important". La situation familiale (au moins un enfant à charge :

oui ou non) est notée également. On cherche à vérifier si l'intérêt dans le produit dépend de la situation familiale. Les résultats sont les suivants :

<i>Enfant</i>	<i>aucun</i>	<i>mineur</i>	<i>important</i>
<i>oui</i>	10	12	3
<i>non</i>	7	38	9

On a donc 79 personnes qui répondent. On veut vérifier s'il y a un lien entre les deux mesures au niveau 5%.

Questions :

1. Déterminer les hypothèses du modèle bi-varie (tableau de contingence).
2. Calculer les effectifs marginaux, les fréquences observées n_{ij} et les fréquences marginaux.
3. calculer les effectifs théoriques.

Après utilisation du logiciel SPSS, interpréter les résultats suivants :

Tests du khi-carré

	Valeur	ddl	Signification asymptotique (bilatérale)
khi-carré de Pearson	6,162 ^a	2	,046
Rapport de vraisemblance	5,839	2	,054
Association linéaire par linéaire	4,020	1	,045
N d'observations valides	79		

a. 1 cellules (16,7%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 3,80.

Exercice 2 :

Deux variables qualitatives mesurées simultanément sur une population.

P : candidats.

- Variable X : Résultat au test d'aptitude, qualitative à $l = 2$ modalités.
- Variable Y : Résultat au test psychologique, qualitative à $c = 2$ modalités.
- On dispose d'un échantillon de taille $n = 120$

Voici le tableau des effectifs observés n_{ij}

X \ Y	introverti	extraverti	total ligne L_i
apte	14	34	48
inapte	31	41	72
total colonne C_j	45	75	$n = 120$

Travail à faire :

- 1) Quel sont les variables de ce modèle ?
- 2) Quel est le type de ce modèle ? expliquer.
- 3) Donner les principales étapes de ce modèle.
- 4) Déterminer les hypothèses du modèle.
- 5) Calculer les fréquences observées n_{ij} et marginaux.
- 6) calcule les effectifs théoriques et marginaux.
- 7) Calculer de degré de liberté DDL.
- 8) Après utilisation du logiciel SPSS, interpréter les résultats suivants :

Tests du khi-carré

	Valeur	ddl	Signification asymptotique (bilatérale)	Sig. exacte (bilatérale)	Sig. exacte (unilatérale)
khi-carré de Pearson	2,370 ^a	1	,124		
Correction pour continuité ^b	1,815	1	,178		
Rapport de vraisemblance	2,406	1	,121		
Test exact de Fisher				,178	,088
Association linéaire par linéaire	2,351	1	,125		
N d'observations valides	120				

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 18,00.

b. Calculée uniquement pour une table 2x2

**Chapitre 4 : Visualisation des
Données : Analyse d'un Nuage de
Points, Recherche des Axes
Principaux et Interprétation
(Approche Géométrique en
Statistique)**

I. Définition d'une série statistique

Lorsqu'on étudie deux caractères statistiques sur une population donnée, on obtient une série statistique double.

On note souvent les valeurs prises par le premier caractère x_1, x_2, \dots, x_N et celles prises par le second y_1, y_2, \dots, y_N . Les valeurs prises par cette série sont alors les couples $(x_1; y_1), (x_2; y_2), \dots, (x_N; y_N)$.

II. Qu'est-ce qu'un nuage de points ?

En statistiques, un nuage de points est une représentation de données dépendant de plusieurs variables. Il permet de mettre en évidence le degré de corrélation entre au moins deux variables liées.

En d'autre terme, un nuage de points ou diagramme de dispersion est une représentation graphique dans un repère du plan d'une série statistique à deux variables X et Y.

Chaque individu i est représenté par un point M_i de coordonnées $(x_i; y_i)$ où x_i et y_i sont les valeurs respectives des variables X et Y prises par l'individu i .

III. Point moyen du nuage

On appelle point moyen $G(x; y)$ le point dont les coordonnées sont les moyennes des valeurs x_i et y_i de la série.

Exemple d'application :

Le tableau suivant donne le prix de vente d'un article en fonction de la quantité commandée.

Quantité (x_i)	50	100	150	200	250	300
Prix (y_i)	2,8	2,4	2,3	2	1,65	1,6

Questions :

- Calculer les coordonnées du point moyen G du nuage de points correspondant.

Solutions :

Il faut tout d'abord calculer :

1. Calculer la moyenne des valeurs X_i .
2. Calculer la moyenne des valeurs Y_i .
3. Le point G_a pour coordonnées (X, Y) .

La moyenne des quantités commandées est :

$$\bar{x} = \frac{50 + 100 + 150 + 200 + 250 + 300}{6} = 175.$$

La moyenne des prix est :

$$\bar{y} = \frac{2,8 + 2,4 + 2,3 + 2 + 1,65 + 1,6}{6} = 2,125.$$

Le point G a pour coordonnées (175 ; 2,125).

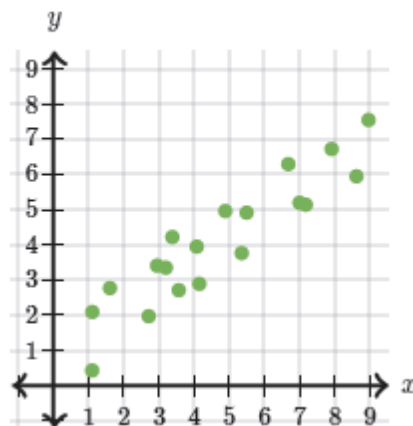
IV. La corrélation des variables x et y

Chaque individu i est représenté par un point M_i de coordonnées $(x_i ; y_i)$ où x_i et y_i sont les valeurs respectives des variables X et Y prises par l'individu i.

L'allure du nuage de points révèle s'il existe une liaison ou non entre les deux variables quantitatives. En général, les variables sont, dans une certaine mesure, dépendantes l'une de l'autre : elles sont en corrélation. La liaison entre elles est dite relative.

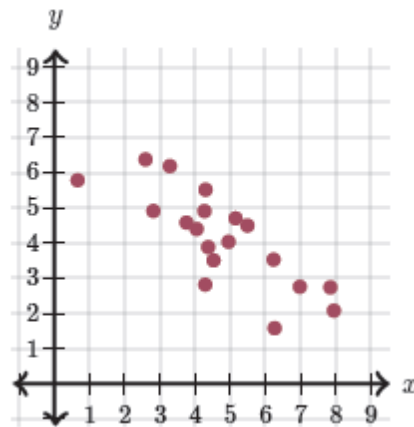
La corrélation est positive quand x et y varient dans le même sens. Ici, lorsque les valeurs de x augmentent, les valeurs de y augmentent.

Corrélation positive



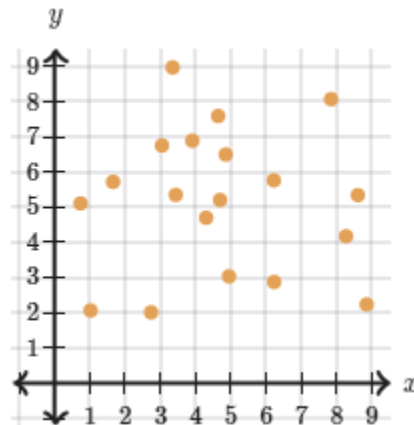
La corrélation est négative quand x et y varient en sens contraire. Ici, lorsque les valeurs de x augmentent, les valeurs de y diminuent.

Corrélation négative



Il n'y a pas de corrélation entre les variables s'il n'existe pas de liaison entre elles. Elles sont indépendantes.

Pas de corrélation

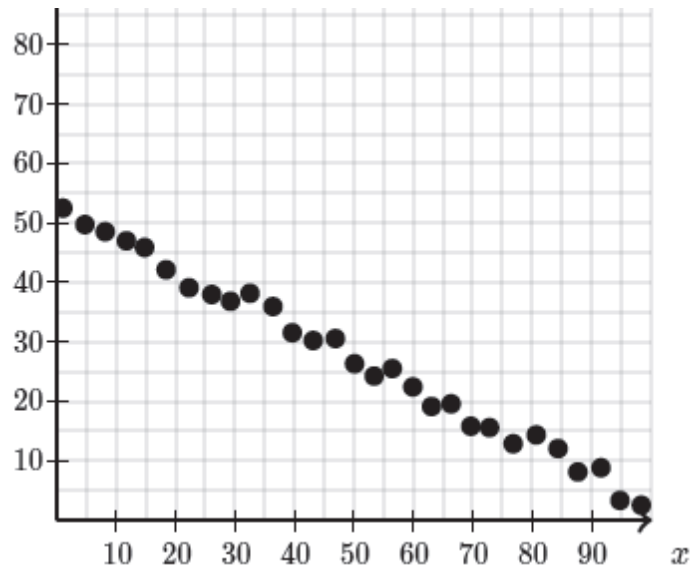


Exercices 1 :

D'après le nuage de points, que peut-on en conclure sur la relation entre les deux variables ?

Choisir une réponse :

- 1) La relation entre les deux variables est une relation linéaire positive
- 2) La relation entre les deux variables est une relation linéaire négative
- 3) La relation entre les deux variables est une relation non linéaire
- 4) Il n'y a pas de relation entre les deux variables



La réponse juste est la réponse 2.

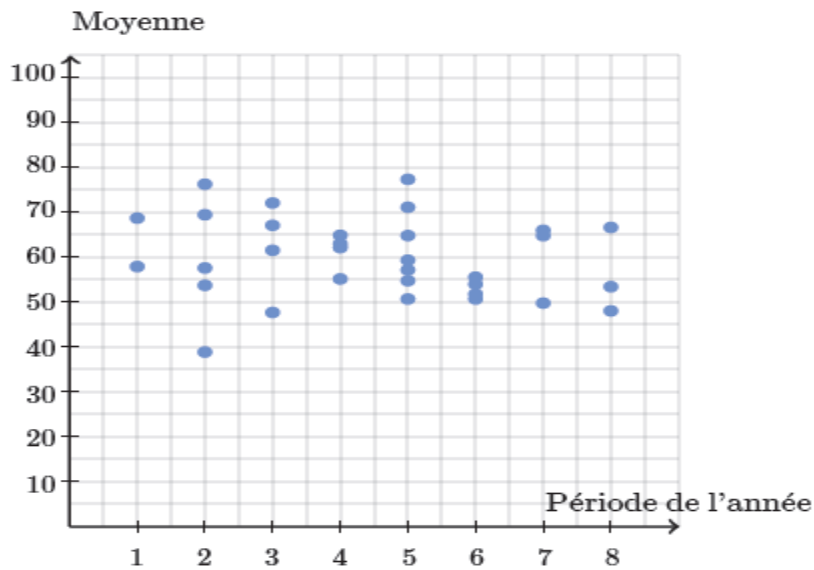
Exercice 2 :

Alexandra voulait savoir s'il y avait un lien entre la date d'un contrôle et la moyenne de la classe à ce contrôle. Elle a récapitulé pour tous les contrôles de l'année précédente, la moyenne de la classe et la période de l'année où il a été donné. Pour plus de lisibilité, elle a ramené les notes sur 202020 à des notes sur 100100100 et elle a obtenu ce nuage de points.

Quelle est la phrase qui décrit le mieux la relation suggérée ?

Choisir une réponse :

- 1) La relation entre les deux variables est une relation linéaire positive
- 2) La relation entre les deux variables est une relation linéaire négative
- 3) La relation entre les deux variables est une relation non linéaire
- 4) Il n'y a pas de relation entre les deux variables



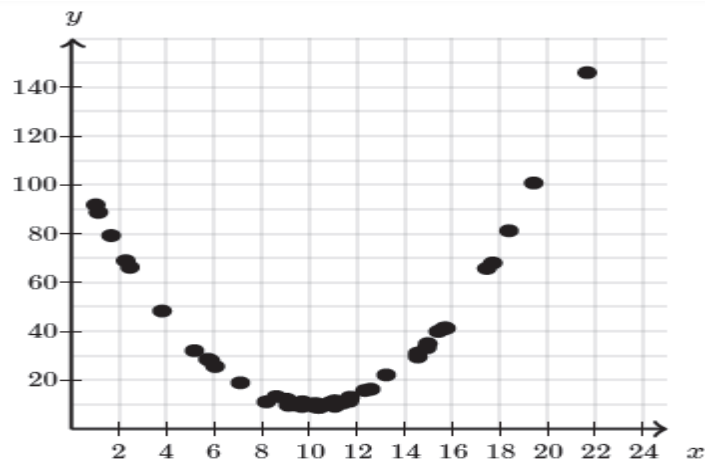
La réponse juste est la réponse 4.

Exercice 3 :

D'après le nuage de points, que peut-on en conclure sur la relation entre les deux variables ?

Choisir une réponse :

- 1) La relation entre les deux variables est une relation linéaire positive
- 2) La relation entre les deux variables est une relation linéaire négative
- 3) La relation entre les deux variables est une relation non linéaire
- 4) Il n'y a pas de relation entre les deux variables



La réponse juste est la réponse 3.

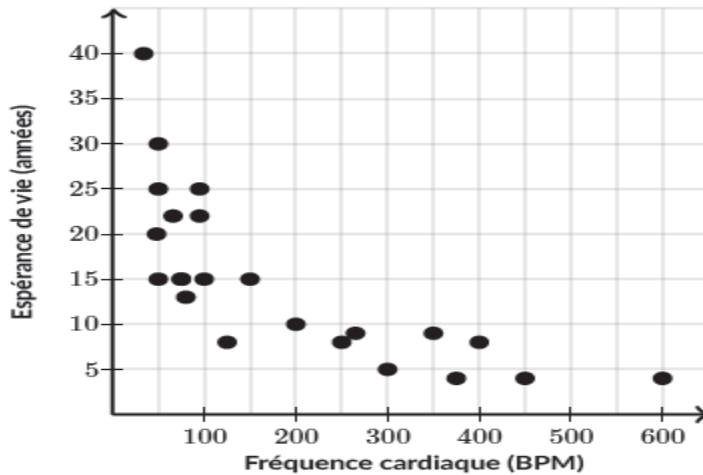
Exercice 4 :

Le graphique ci-dessous présente la relation entre le rythme cardiaque moyen (en battements par minute) et l'espérance de vie (en années) de différents mammifères.

Quelle est la phrase qui décrit le mieux la relation suggérée ?

Choisir une réponse :

- 1) La relation entre les deux variables est une relation linéaire positive
- 2) La relation entre les deux variables est une relation linéaire négative
- 3) La relation entre les deux variables est une relation non linéaire
- 4) Il n'y a pas de relation entre les deux variables



La réponse juste est la réponse 3.

Exercice :

Actifs occupés dans le primaire (en %) et part du primaire dans le PIB (en %) dans 12 pays de l'OCDE :

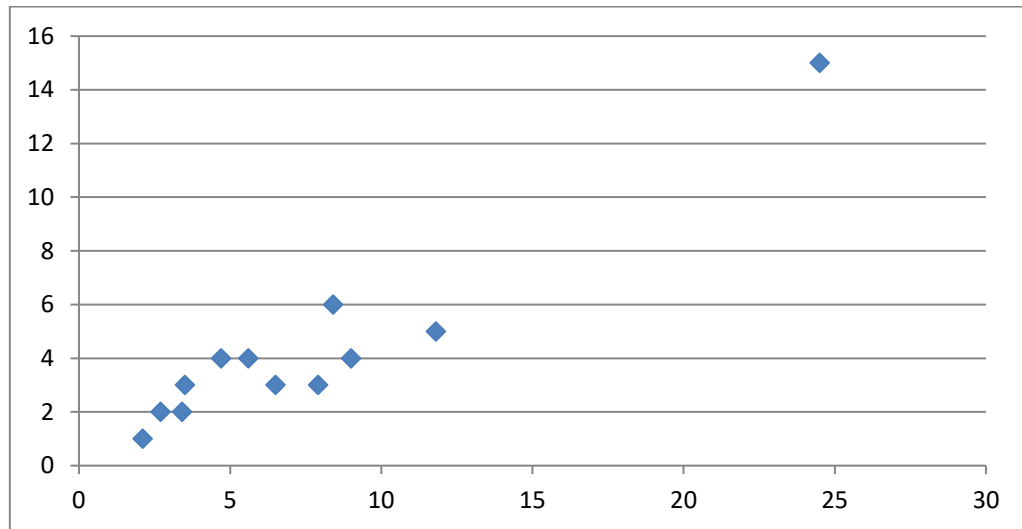
Pays	Actifs	Part
Allemagne	3,4	2
Autriche	7,9	3
Belgique	2,7	2
Danemark	5,6	4
Espagne	11,8	5
Finlande	8,4	6
Grèce	24,5	15
Italie	9	4
Royaume-Uni	2,1	1
Suède	3,5	3
Norvège	6,5	3
Pays-Bas	4,7	4

Travail à faire :

1. Représenter le nuage des observations avec le centre de gravité.
2. Calculer le coefficient de corrélation linéaire ; interpréter le résultat
3. Calculer le coefficient de corrélation linéaire en enlevant des données le point singulier ; commenter

Solution :

1. Représenter graphiquement cette série par un nuage de points.



X barre	7,50833
Y barre	4,33333

M a pour coordonner **G (7,50, 4,33)**

2. Calculer le coefficient de corrélation linéaire ; interpréter le résultat

covariance (x,y)	19,3056
variance (x)	34,0141
variance (y)	12,0556
écart type (x)	5,83216
écart type (y)	3,47211

Le coefficient de corrélation	0,953365024
-------------------------------	-------------

On est tenté de dire que r est proche de 1, donc il y a une corrélation linéaire positive entre X et Y

3. Calculer le coefficient de corrélation linéaire en enlevant des données le point singulier ;
commenter

Quand on a un point isolée, il faut reconstruire le nuage de point et refaire les calculs, car il peut modifier la vision du nuage

Pays	Actifs	Part
Allemagne	3,4	2
Autriche	7,9	3
Belgique	2,7	2
Danemark	5,6	4
Espagne	11,8	5
Finlande	8,4	6
Italie	9	4
Royaume- Uni	2,1	1
Suède	3,5	3
Norvège	6,5	3
Pays-Bas	4,7	4
X barre	5,96363636	
Y barre	3,36363636	

G(5,96, 3,363)

covariance (x,y)	3,08595041
variance (x)	8,47322314
variance (y)	1,8677686
écart type (x)	2,91088013
écart type (y)	1,36666331

Le coefficient de corrélation 0,7757166

On est tenté de dire que r est proche de 1, donc il y a une corrélation linéaire positive entre X et Y, on peut conclure cela quand r est supérieur à 0,7.

Chapitre 5 :
Comparaison de Groupes :
L'Analyse de la Variance
ANOVA (Étude des
Différences entre Groupes)

I. Introduction

- L'analyse de la variance ANOVA permet d'étudier l'effet d'une variable qualitative (nombre fini de valeurs) sur une (ou des) variable quantitative.
- L'analyse de variance entre dans le cadre général du modèle linéaire, où une variable quantitative est expliquée par une variable qualitative.
- L'objectif essentiel est de comparer les moyennes empiriques de la variable quantitative observées pour les variables qualitatives (facteurs).
- Ce type d'analyse teste si les moyennes des échantillons sont égales.

II. Définition du Test d'Anova à un Facteur

- TEST ANOVA permet de tester si la variable qualitative influence la distribution de la variable quantitative.
- L'analyse de la variance ANOVA permet de répondre aux questions du genre :
 1. La productivité varie elle selon le niveau d'étude ?
 2. Le niveau en langue étrangères affecte il les notes des étudiants ?
- Les variables dépendantes sont les variables quantitatives dont on veut tester la distribution
 - Le facteur c'est la variable qualitative pour laquelle on veut vérifier si elle impacte la distribution des variables dépendantes.

III. Les Hypothèses de l'ANOVA

- H_0 : La variable qualitative n'exerce aucun effet sur la variable quantitative.
- H_1 : La variable qualitative exerce un effet sur la variable quantitative.
- L'hypothèse nulle est que la variable qualitative n'influence pas la variable quantitative.
- Donc pour la rejeter (le facteur influence la variable dépendante) $\text{sig} < 0,05$

IV. La Variance Expliquée R^2

- La variance expliquée R^2 doit être calculée à partir du tableau de l'ANOVA
- $R^2 = \text{racine}(\text{somme des carrés "inter-groupes"} / \text{somme des carrés "total"})$
- Interprétation : La taille de l'effet est considérée (Cohen 1988):
 1. Faible à 0,10
 2. Modérée à 0,25
 3. Forte à 0,40

V. Exemple d'application

- Q1 : Vous possédez quel diplôme :
 Licence LMD Master LMD
- Q2 – Quel est votre salaire en net (DA) :.....

VI. Résultats et interprétation

Avant d'examiner les résultats de l'ANOVA, il importe de vérifier la prémisse d'égalité des variances avec le test de Levene.

Test d'homogénéité des variances

hrs1

Statistique de Levene	ddl1	ddl2	Signification
1,179	4	899	,319

- H0 : Les variances sont égales
- H1 : Les variances ne sont pas égales

On accepte hypothèse H0 car $0,319 > 0,05$

ANOVA à 1 facteur					
SALAIRE					
	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Inter-groupes	7,763E+9	1	7,763E+9	450,937	,000
Intra-groupes	1,429E+9	83	17215025,7		
Total	9,192E+9	84			

- Il existe une différence de salaire entre les salariés avec le diplôme de Licence et ceux avec le diplôme de Master.
- Donc on accepte H1 : La moyenne des salaires diffère selon le diplôme ($< 0,05$) et on rejette l'hypothèse H0 : la moyenne des salaires est semblable selon le diplôme.

Chapitre 06 :
Réduction de Dimension :
Analyse en Composantes
Principales (ACP)
(Simplification de la
Représentation des Données)

I. Introduction

- Analyse en composantes principales (ACP) est une analyse statistique descriptive multivariée.
- Analyse en composantes principales (ACP) est appliqué à un ensemble de variables initiales qu'on veut réduire en quelques facteurs ou composantes ou axes.
- ACP représente graphiquement les données par rapport à ces facteurs sous forme d'axes. Ces représentations graphiques sont du type nuage de points.

II. Objectifs

- L'Analyse en Composante Principale (ACP) fait partie des analyses descriptives multi-variées. Le but de cette analyse est de résumer le maximum d'informations possibles en en perdant le moins possible pour : Faciliter l'interprétation d'un grand nombre de données initiales, Donner plus de sens aux données réduites
- L'ACP permet donc de réduire des tableaux de grandes tailles en un petit nombre de variables (2 ou 3 généralement) tout en conservant un maximum d'information. Les variables de départ sont dites 'métriques'.
- L'ACP permet donc de réduire les variables initiales en un petit nombre de variables tout en conservant un maximum d'information.
- Dans ce type d'analyse, il n'y a pas de variable dépendante ou indépendante préalablement identifiée. Aussi, aucune vérification de l'hypothèse nulle n'est exigée.
- Dans le monde de l'ACP les données sont appelées inertie.

III. Conditions d'utilisation de l'ACP

- Utilisation de variables quantitatives continues et les variables qualitatives ordinales.
- Relation linéaire entre les variables.
- Corrélations inter items : On doit s'assurer qu'il existe des corrélations minimales entre les items ou les variables qui feront l'objet de l'analyse.

IV. Procédure de l'ACP

- 1) Formuler le problème.
- 2) Lancer l'ACP sur SPSS.

- 3) Calcul la matrice de corrélations et vérifier si les données sont-elles factorisables (Test KMO, Test Bartlett).
- 4) Extraire les facteurs et déterminer leur nombre
- 5) Interpréter les facteurs (Les Items à retenir dans chaque facteur).

V. Test KMO

Mesure de l'adéquation de l'échantillonnage (KMO) (La mesure Kaiser-Meyer-Olkin) : Cette mesure donne un aperçu global de la qualité des corrélations inter-items.

L'indice KMO varie entre 0 et 1 et donne une information complémentaire à l'examen de la matrice de corrélation. Son interprétation va comme suit :

- 0,80 et plus Excellent
- 0,70 et plus Bien
- 0,60 et plus Médiocre
- 0,50 et plus Misérable
- Moins de 0,50 Inacceptable

VI. Test Bartlett

Test de Bartlett de la sphéricité : Ce test doit être très significatif < 0.05 , Entre 0.05 et 0.10 acceptable et au-dessus de 0.10, rejeté.

- Ce test permet de dire si toutes les variables sont indépendantes (H0).
- Pour la validité de l'ACP on doit rejeter l'hypothèse nulle
- La probabilité de ce test doit être proche de 0
- Une valeur inférieure à 0,05 est acceptable

VII. Le nombre de facteurs à retenir

Concernant le nombre de facteurs à retenir, trois règles peuvent être suivies :

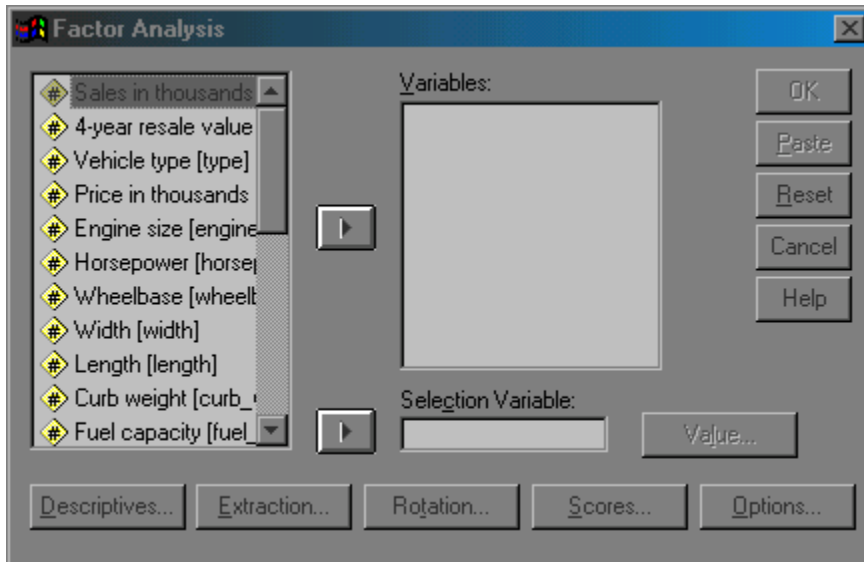
1. 1ere règle : la règle de Kaiser qui veut qu'on ne retienne que les facteurs aux valeurs propres supérieures à 1.
2. 2eme règle : on choisit le nombre d'axe en fonction de la variance expliquée que l'on veut obtenir (restitution minimale d'information) si par exemple veut que le modèle restitue au moins 70% de l'information on ajoutera le nombre de facteurs nécessaire à l'obtention d'une variance expliquée égale ou supérieure à 70% .
3. 3eme méthode : le « Scree-test » ou test du coude. En observant le graphique des valeurs propres, le nombre de facteurs serait égale au nombre de points ayant une valeur se trouvant à gauche du point d'inflexion.

VIII. Les Items à retenir dans chaque facteur

Le poids minimal pour associer un Item à un facteur est donc de 0,30. A cette règle ajoutant celle proposée par Roussel et Wacheux (2005,) proposent l'élimination des items ayant une contribution (poids) supérieurs à 0,3 sur plusieurs facteurs (lignes).

IX. Application sur SPSS

Aller dans Analyze > Data Reduction > Factor... La boîte de dialogue suivante apparaît alors :



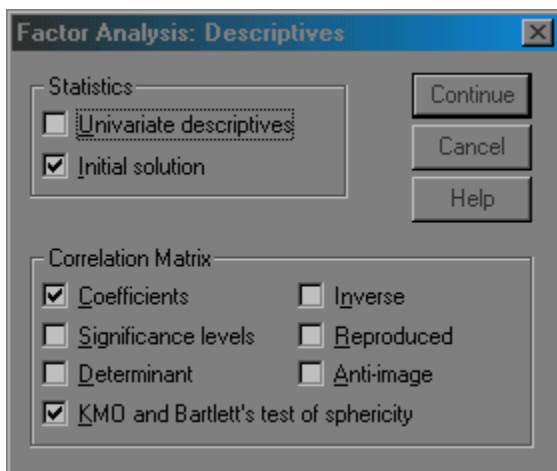
On choisit les variables qui nous paraissent les mieux adaptées à l'analyse en les sélectionnant dans la partie de droite puis en cliquant sur la flèche qui pointe vers la droite.

Cinq boîtes de dialogue d'options s'offrent maintenant à nous : 1. Descriptives... 2.

Extraction... 3. Rotation... 4. Scores... 5. Options... que nous allons maintenant examiner une à une.

« Descriptives... »

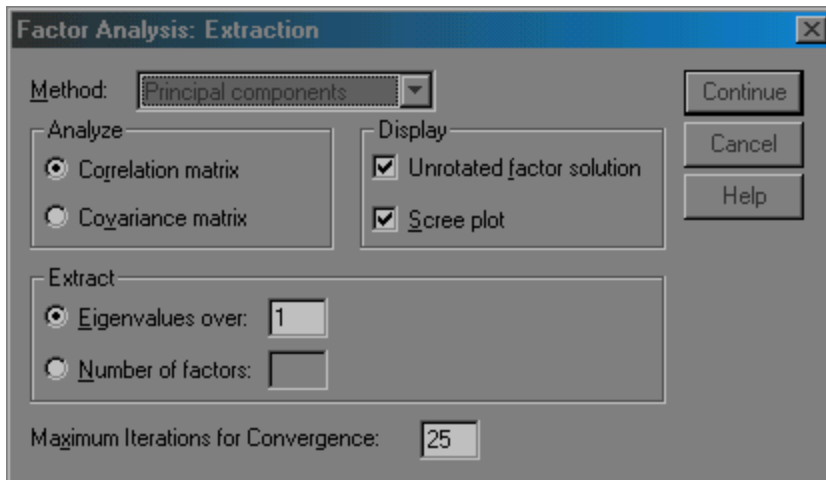
La boîte de dialogue « Factor Analysis : Descriptives » apparaît.



Dans « Correlation Matrix », cliquer sur « Coefficients » et « KMO and Bartlett's test of sphericity ».

« Extraction... »

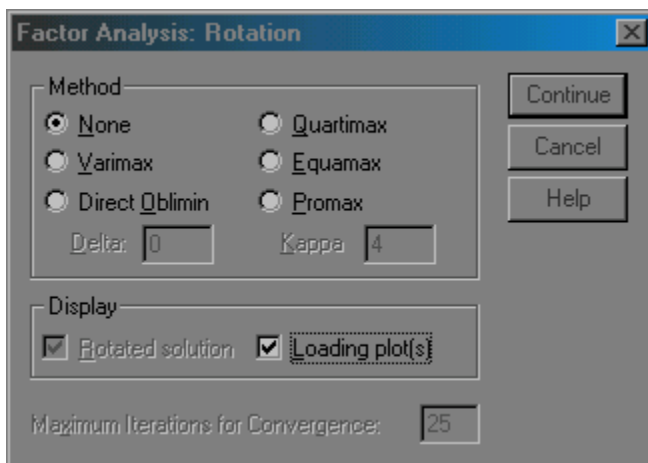
La boîte de dialogue « Factor Analysis : Extraction » apparaît.



Cliquer sur « Scree Plot » (Graphique des valeurs propres). Ne pas toucher aux autres options.

« Rotation... »

La boîte de dialogue « Factor Analysis : Rotation » apparaît.

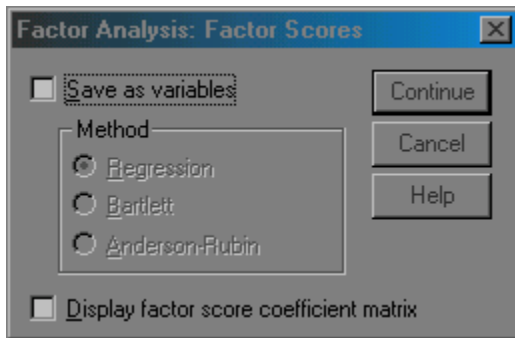


Pour l'instant, il ne faut rien toucher dans « Method ». L'option « Varimax » pourra être choisie si les résultats ne sont pas suffisants dans un premier temps.

Par contre, cocher l'option « Loading plot(s) » (Carte(s) factorielle(s)). Cette option permet d'avoir une représentation des différents axes.

« Scores... »

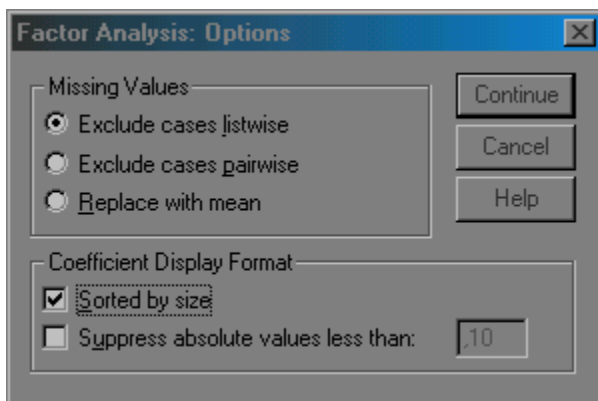
La boîte de dialogue « Factor Analysis : Factor Scores » apparaît.



Pour l'instant, il ne faut toucher à rien. L'option « Save as variables » (enregistrer dans des variables) permettra d'attribuer à chaque individu ses coordonnées factorielles une fois l'analyse terminée.

« Options... »

La boîte de dialogue « Factor Analysis : Options » apparaît.



Choisir l'option « Sorted by size » (Classement des variables par taille) dans Affichage des projections.

Analyse des résultats

Analyser les résultats d'une ACP, c'est répondre à trois questions :

1. Les données sont-elles factorisables ?
2. Combien de facteurs retenir ?
3. Comment interpréter les résultats ?

1. Les données sont-elles factorisables ?

Pour répondre à cette question, dans un premier temps, il convient d'observer la matrice des corrélations (« Correlation Matrix »). Si plusieurs variables sont corrélées (> 0.5), la factorisation est possible. Si non, la factorisation n'a pas de sens et n'est donc pas conseillée.

Correlation Matrix

	Vehicle type	Price in thousands	Engine size	Horsepower	Wheelbase	Width	Length	Curb weight	Fuel capacity	
Correlation	Vehicle type	1,000	-,040	,268	,013	,391	,251	,141	,524	,599
	Price in thousands	-,040	1,000	,623	,838	,106	,323	,150	,526	,424
	Engine size	,268	,623	1,000	,836	,470	,688	,537	,760	,667
	Horsepower	,013	,838	,836	1,000	,283	,536	,387	,610	,504
	Wheelbase	,391	,106	,470	,283	1,000	,682	,840	,651	,654
	Width	,251	,323	,688	,536	,682	1,000	,709	,721	,658
	Length	,141	,150	,537	,387	,840	,709	1,000	,627	,565
	Curb weight	,524	,526	,760	,610	,651	,721	,627	1,000	,864
	Fuel capacity	,599	,424	,667	,504	,654	,658	,565	,864	1,000

Dans notre exemple, plusieurs variables sont corréllées entre elles :

Dans un deuxième temps, il faut observer l'indice de KMO (Kaiser-Meyer-Olkin) qui doit tendre vers 1. si ce n'est pas le cas, la factorisation n'est pas conseillée. Pour juger de l'indice de KMO, on peut utiliser l'échelle suivante :

- 0,50 et moins est misérable
- entre 0,60 et 0,70, c'est médiocre
- entre 0,70 et 0,80 c'est moyen
- entre 0,80 et 0,90 c'est méritoire
- et plus 0,9 c'est merveilleux.

Enfin, on utilise le test de sphéricité de Bartlett. : si la signification (Sig.) tend vers 0.000, c'est très significatif, inférieur à 0.05 significatif, entre 0.05 et 0.10 acceptable et au dessus de 0.10, on rejette.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,810
Bartlett's Test of Sphericity	Approx. Chi-Square	1212,128
	df	28
	Sig.	,000

Si l'ACP satisfait à au moins deux de ces trois conditions, on peut continuer.

2. Combien de facteurs retenir ?

Trois règles sont applicables :

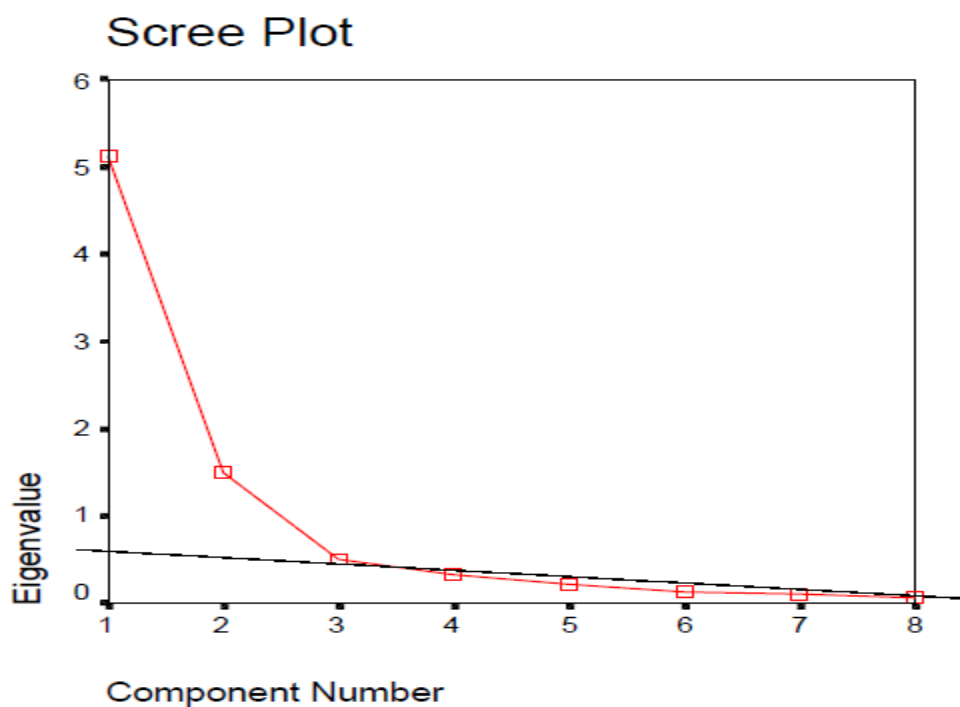
- 1ere règle : la règle de Kaiser qui veut qu'on ne retienne que les facteurs aux valeurs propres supérieures à 1.
- 2eme règle : on choisit le nombre d'axe en fonction de la restitution minimale d'information que l'on souhaite. Par exemple, on veut que le modèle restitue au moins 80% de l'information. Pour ces deux premières règles, on examine le tableau « Total Variance Explained ».

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,121	64,008	64,008	5,121	64,008	64,008
2	1,510	18,874	82,882	1,510	18,874	82,882
3	,496	6,205	89,087			
4	,328	4,100	93,187			
5	,223	2,793	95,980			
6	,141	1,757	97,736			
7	,115	1,433	99,169			
8	6,645E-02	,831	100,000			

Extraction Method: Principal Component Analysis.

3eme méthode : le « Scree-test » ou test du coude. On observe le graphique des valeurs propres et on ne retient que les valeurs qui se trouvent à gauche du point d'inflexion. Graphiquement, on part des composants qui apportent le moins d'information (qui se trouvent à droite), on relie par une droite les points presque alignés et on ne retient que les axes qui sont au-dessus de cette ligne.



Dans notre exemple, nous ne retenons que les deux premiers axes.

3. Interprétation des résultats

C'est la phase la plus délicate de l'analyse. On donne un sens à un axe grâce à une recherche lexicale (ou recherche de mots) à partir des coordonnées des variables et des individus. Ce sont les éléments extrêmes qui concourent à l'élaboration des axes.

Component Matrix^a

	Component	
	1	2
Curb weight	,912	-2,57E-02
Engine size	,878	,265
Fuel capacity	,847	-,109
Width	,843	-,221
Horsepower	,772	,554
Length	,760	-,487
Wheelbase	,742	-,569
Price in thousands	,606	,715

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Dans notre exemple, ce sont les variables « Curb Weight » et « Engine Size » qui concourent le plus à la construction de l'axe 1. Si la recherche lexicale à partir des variables ne donne rien, il faut alors donner un sens à l'axe en s'appuyant là aussi aux individus qui ont les coordonnées extrêmes.

Chapitre 07 :
Classification des Données :
Analyse Discriminante
(Categorisation des Données)

Définition

L'analyse discriminante est une technique d'analyse des données connue sous l'abréviation AD. Elle est utilisée dans le cadre de la modélisation d'une variable qualitative Y à K catégories (modalités) dite variable à expliquer (ou variable endogène ou encore variable à prédire) (J-P.Benzécri, 1982),

En d'autre terme, l'analyse discriminante est une technique statistique qui vise à décrire, expliquer et prédire l'appartenance à des groupes prédéfinis (classes, modalités de la variable à prédire...) d'un ensemble d'observations (individus, exemples...) à partir d'une série de variables prédictives (descripteurs, variables exogènes...) (Saporta, 2006).

Objectif

- Prédire une variable qualitative à k classes à l'aide de p prédicteurs.
- Ensemble des méthodes utilisées pour prédire une variable qualitative à k catégories à l'aide de p prédicteurs (S.AMBAPOUR., 2003).
- Le but de l'analyse discriminante est d'étudier les relations entre une variable qualitative et un ensemble de variables explicatives quantitatives. C'est une méthode utilisée notamment par les banques pour le scoring.
- Déterminer les variables explicatives les plus discriminantes vis à vis des classes déterminées
- Déterminer à quel groupe appartient un individu à partir de ses caractéristiques.
- Mais surtout à valider une classification ou à faire un choix entre plusieurs classifications pour savoir laquelle est la plus pertinente. L'analyse discriminante intervient donc a posteriori d'une classification.
- Deux conditions sont à remplir :
 1. Les variables explicatives doivent être métriques.

2. Elles ne doivent pas être trop corrélées entre elles. Cela se vérifie par l'observation des corrélations entre les variables. Si c'est le cas, on peut passer par une analyse factorielle qui permet de réduire les données à quelques axes. Ces axes sont, par propriété, non corrélés entre eux.

Les méthodes utilisées

1. Méthodes géométriques : recherche des meilleures fonctions discriminantes.
 - Linéaires
 - Non linéaires
2. Méthodes probabilistes : estimation directe des probabilités d'appartenance aux groupes définis par y .
 - Paramétrique
 - Semi paramétrique
 - Non paramétrique

En analyse discriminante, on considère trois types de matrice de variances-covariances et donc trois types de corrélations.



Principe de l'analyse discriminante

- Les centres de gravité de chaque sous-nuage appartenant à une seule classe sont éloignés (J-M.ROMEDER, 1973).
- Les sous-nuages appartenant à une seule classe sont les plus homogènes possibles autour de ces centres de gravité (F.CAILLIEZ et J-P.PAGES., 1976).
- Pour se faire il faut maximiser les variances intergroupes (entre les groupes) et minimiser la variance intra-groupe (à l'intérieur des groupes) (L.Lebart, 1977).

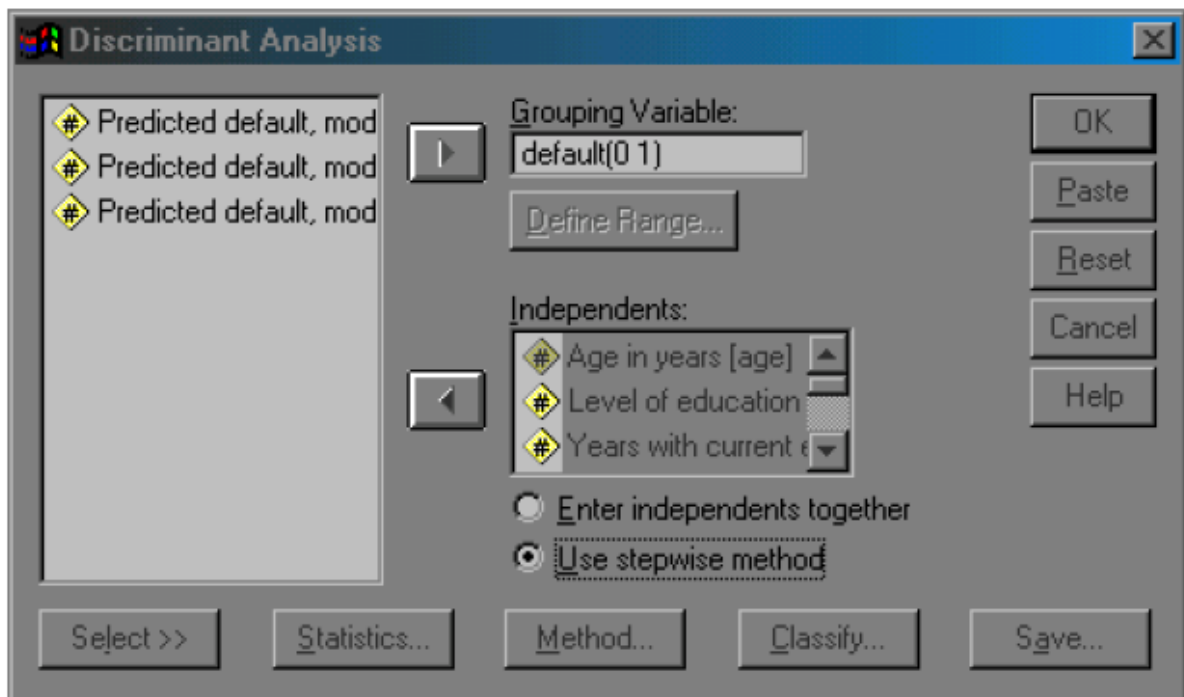
Évaluation

L'évaluation se situe à deux niveaux (Bardos, 2001) :

1. évaluer le pouvoir discriminant d'un axe factoriel ;
2. évaluer le pouvoir discriminant d'un ensemble d'axes factoriels.

Application de l'analyse discriminante sur SPSS

Aller dans Analyser > Classifier > Discriminant... La boîte de dialogue suivante apparaît alors :



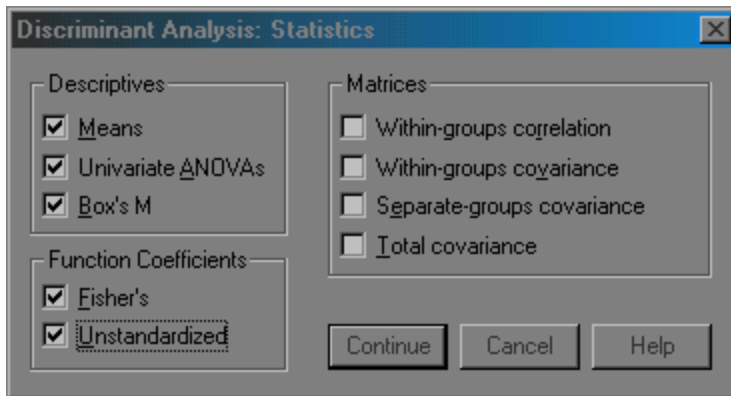
Dans « Grouping Variable » (i.e. les critères de regroupement), il faut indiquer la variable à expliquer en la sélectionnant dans la partie de droite puis en cliquant sur la flèche qui pointe vers la droite. SPSS demande alors de définir l'intervalle, c'est-à-dire les différentes modalités que la variable peut prendre.

Dans « Independents » (i.e. les variables explicatives), il faut indiquer les variables métriques que l'on souhaite intégrer à l'analyse. Il est important de choisir « Use stepwise method » (i.e. la méthode pas à pas).

Trois options s'offrent alors à nous : « Statistics... », « Method... » et « Classify... ». On ne touchera pas aux différentes options de « Méthod... »

1. Statistics...

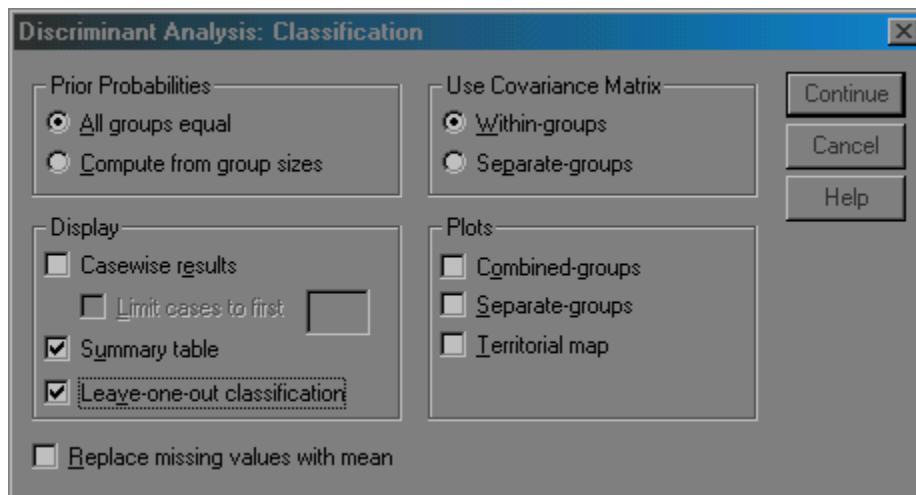
La boîte de dialogue « Discriminant Analysis : Statistics » apparaît.



Dans la boîte qui apparaît, il convient de cocher « Means » (moyennes), « Univariate ANOVAs » (ANOVA à 1 facteur) et « Box's M » (Test de Box) dans « Descriptives » et « Fischer's » ainsi que « Unstandardized » dans « Function Coefficients ».

2. Classify...

La boîte de dialogue « Discriminant Analysis : Classification » apparaît.



Dans la boîte qui apparaît, il convient de cocher « Summary Table » (option qui permet l'affichage de la matrice de confusion) et « Leave-one-out classification » dans « Display ».

Analyse des résultats

Une analyse discriminante se déroule en 3 étapes :

1. On vérifie l'existence de différences entre les groupes.
2. On valide l'étude.
3. On vérifie le pouvoir discriminant des axes.
4. On juge la qualité de la représentation du modèle.

La 3ème étape peut être passée dans la plupart des cas.

1. Vérification de l'existence de différences entre les sous-groupes.

On vérifie s'il existe bien des différences entre les groupes grâce à trois indicateurs : la moyenne ou la variance, le test du F et le Lambda de Wilks. Ils s'interprètent de la façon suivante :

	En cas d'influence	En absence d'influence
Moyenne ou variance	Différence	Similitude
Test du F	F élevé Sig F tend vers 0,000	F faible SIG F \geq 0,01 ou 0,05
Lambda de Wilks	\leq 0,90	Tend vers 1

Cette première analyse permet de déterminer quelles sont les variables qui sont les plus discriminantes entre les groupes.

Les moyennes et écart-types s'observent dans le tableau « Group Statistics ». Les variables « Years with current employes », « Years at current adress », « Debt to income ration » et « Credit card debt » dans l'exemple ci-dessous semblent être les variables les plus discriminantes.

Group Statistics

Previously defaulted		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
No	Age in years	35,5145	7,70774	517	517,000
	Level of education	1,6596	,90443	517	517,000
	Years with current employer	9,5087	6,66374	517	517,000
	Years at current address	8,9458	7,00062	517	517,000
	Household income in thousands	47,1547	34,22015	517	517,000
	Debt to income ratio (x100)	8,6793	5,61520	517	517,000
	Credit card debt in thousands	1,2455	1,42231	517	517,000
	Other debt in thousands	2,7734	2,81394	517	517,000
Yes	Age in years	33,0109	8,51759	183	183,000
	Level of education	1,9016	,97279	183	183,000
	Years with current employer	5,2240	5,54295	183	183,000
	Years at current address	6,3934	5,92521	183	183,000
	Household income in thousands	41,2131	43,11553	183	183,000
	Debt to income ratio (x100)	14,7279	7,90280	183	183,000
	Credit card debt in thousands	2,4239	3,23252	183	183,000
	Other debt in thousands	3,8628	4,26368	183	183,000
Total	Age in years	34,8600	7,99734	700	700,000
	Level of education	1,7229	,92821	700	700,000
	Years with current employer	8,3886	6,65804	700	700,000
	Years at current address	8,2786	6,82488	700	700,000
	Household income in thousands	45,6014	36,81423	700	700,000
	Debt to income ratio (x100)	10,2606	6,82723	700	700,000
	Credit card debt in thousands	1,5536	2,11720	700	700,000
	Other debt in thousands	3,0582	3,28755	700	700,000

Le test du F et du Lambda de Wilks s'observe dans le tableau « Tests of Equality of Group Means ».

L'examen du F dans notre exemple nous confirme que ce sont bien les variables « Years at current address », « Credit card debt in thousands », « Years with current employer », et « Debt to income ratio (x100) » qui sont les plus discriminantes.

De plus, d'après le test du Lambda de Wilks, seule la variable « Debt to income ratio (x100) » semble avoir une influence.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Age in years	,981	13,482	1	698	,000
Level of education	,987	9,301	1	698	,002
Years with current employer	,920	60,759	1	698	,000
Years at current address	,973	19,402	1	698	,000
Household income in thousands	,995	3,533	1	698	,061
Debt to income ratio (x100)	,848	124,889	1	698	,000
Credit card debt in thousands	,940	44,472	1	698	,000
Other debt in thousands	,979	15,142	1	698	,000

2. Vérification de la validité de l'étude.

On estime la validité d'une analyse discriminante à partir de indicateurs :

- Le test de Box.
- La corrélation globale.
- Le Lambda de Wilks.

On observe le test de Box grâce au tableau « Test Results ».

Test Results

Box's M		364,962
F	Approx.	36,182
	df1	10
	df2	552413,8
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

Le M doit être le plus élevé possible. La significativité du test de F doit tendre vers 0. S'il est supérieur à 0,05, l'analyse n'est pas valide.

La corrélation globale se mesure quant à elle se retrouve dans le tableau « Eigenvalues » (Valeurs propres).

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,395 ^a	100,0	100,0	,532

a. First 1 canonical discriminant functions were used in the analysis.

On observe notamment la colonne « Canonical Correlation » (Corrélation Canonique).

Plus elle est proche de 1, meilleur est le modèle.

Le Lambda de Wilks s'observe quant à lui dans le tableau « Wilks' Lambda ».

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,717	231,524	4	,000

Plus la valeur du Lambda de Wilks (deuxième colonne) est faible, plus le modèle est bon.

On observe également sa significativité : plus elle est tend vers 0, meilleur, plus le modèle est bon.

3. Estimation des coefficients de la fonction discriminante.

On observe le pouvoir discriminant des axes grâce au tableau « Canonical Discriminant Function Coefficients ».

Canonical Discriminant Function Coefficients

	Function 1
Years with current employer	-,120
Years at current address	-,037
Debt to income ratio (x100)	,075
Credit card debt in thousands	,312
(Constant)	,058

Unstandardized coefficients

Ce tableau permet d'obtenir la fonction discriminante. Dans notre exemple, la fonction est égale à :

$$0,058 - 0,12*(\text{Years with current employer}) - 0,037*(\text{Years at current adress}) + 0,075*(\text{Debet to income ratio}) + 0,312*(\text{Credit card ddebt in thousands})$$

4. Qualité de la représentation.

on observe la qualité de la représentation : on s'assure que la fonction discriminante classe bien les individus en sous-groupes. Pour cela, on analyse la matrice de confusion qui regroupe les individus bien classés et les mal classés :

Groupes prévus (ou théoriques)

	Groupes réels (ou observés)	Groupes prévus (ou théoriques)	Total
		Groupe 1	Groupe 2
Groupes réels (ou observés)	Groupe 1	22	4
	Groupe 2	4	18
	Total	26	22
	Total	26	22
	Total	48	48

Ainsi, dans notre exemple, 22 éléments du groupe 1 ont été bien reclassés grâce à la fonction discriminante et 4 l'ont mal été. De même, pour le groupe 2, 4 individus ont été mal reclassés et 18 bien reclassés. Au total, c'est donc 40 individus (22 + 18) qui ont été correctement reclassés soit 83% de réussite ($40 / 48 = 83\%$).

Sous SPSS, la matrice de confusion s'observe dans le tableau « Classification Results ».

Classification Results^{b,c}

			Predicted Group Membership		Total
			No	Yes	
Original	Count	Previously defaulted			
		No	391	126	517
		Yes	42	141	183
		Ungrouped cases	96	54	150
	%	No	75,6	24,4	100,0
		Yes	23,0	77,0	100,0
Ungrouped cases		64,0	36,0	100,0	
Cross-validated ^a	Count	No	391	126	517
		Yes	43	140	183
	%	No	75,6	24,4	100,0
		Yes	23,5	76,5	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 76,0% of original grouped cases correctly classified.

c. 75,9% of cross-validated grouped cases correctly classified.

La note (b.) nous indique le pouvoir de reclassement de la fonction discriminante, ici 76,0%. On peut retrouver ce chiffre en additionnant les observations bien reclassées (ici 398 et 138 soit un total de 536) et en les divisant par le nombre total d'observations classées (dans le cas présent 700 soit 517 + 183)

Il existe une dernière étape qui consiste à observer les mal-classés et savoir si c'est dû à un atypisme ou à une défaillance de la fonction discriminante. S'il s'agit d'un atypisme, il convient de les enlever et de recommencer l'étude.

Chapitre 08 :
Mise en Pratique de
l'Analyse de Données avec
SPSS : Exemple Détaillé
d'Application

Application d'analyse des données avec SPSS

Le logiciel SPSS pour Windows constitue un système de traitement de données permettant, à partir de fichiers SPSS ou à partir d'autres types de fichiers (Excel, dBase, FoxPro, MS Access) de générer divers tableaux, graphiques et diagrammes ou encore d'effectuer divers traitements statistiques comme le dépouillement de données, le calcul de diverses mesures de tendance centrale et de dispersion, la construction de tableaux croisés, l'exécution de divers tests statistiques paramétriques et non paramétriques, l'élaboration d'études de régression, de corrélation et d'analyse de variance, l'analyse de séries chronologiques et de divers modèles prévisionnels, le tracé de cartes de contrôle pour la maîtrise statistique des procédés.

The image shows a screenshot of the SPSS (PASW Statistics) software interface. The main window, titled "02456defsupimptu2.sav [Ensemble de données1] - PASW Statistics Éditeur de données", displays a list of variables in a table. Three other windows are overlaid on top of the main window, each with an arrow pointing to it from a label on the right:

- Fenêtre des données:** Points to the main data editor window.
- Fenêtre des résultats:** Points to the "Résultats1 [Document1] - PASW Statistics Viewer" window, which shows the output of a command.
- Fenêtre des syntaxes:** Points to the "Syntaxe1 - PASW Statistics Éditeur de syntaxe" window, which shows the command being executed.

The variable list in the main window includes:

Nom	Type	Largeur	Décimales	Etiquette	Valeurs	Manquant	Colonnes	Align	Mesure
1 REG	Numérique	2	0	DRAHRH	(1,BOUCL...	Aucun	8	Droite	Ordinales
2 PROV									Ordinales
3 COM									helle
4 VILL									helle
5 MEN									helle
6 NUMPAR									helle
7 ESPECE									Ordinales
8 LOCA									helle
9 nbespec									helle
10 SUPERFICIE									helle
11 SUPESP									helle
12 SUPESPcor									helle
13 SUPOPT									helle
14 SUPOPTPA...									helle
15 INDEX									helle
16 strate									helle
17 NUMCAH									helle
18 NUMRESP									helle
19 D24A1M									helle
20 D24A2M									helle
21 D24A3M									helle
22 D24V1									helle
23 D24V1A1									helle
24 D24V1A2									helle
25 D24V1A3									helle

The "Résultats1" window shows the following output:

```
GET
FILE='C:\Documents and Settings\Mamounata\Bureau\ARBORIC
DATASET NAME Ensemble de données1 WINDOW=FRONT.
```

The "Syntaxe1" window shows the following command:

```
DATASET ACTIVATE Ensemble de données1.
IF (NUMPAR > 2) supesp=SUPERFICIE / NBRPIED.
VARIABLE LABELS supesp "Superficie réelle de l'espèce".
EXECUTE.
```

Éditeur de données

L'éditeur de données fournit une méthode pratique, semblable à celle d'un tableur, permettant de créer et de modifier des fichiers de données. La fenêtre de l'éditeur de données s'ouvre automatiquement lorsque vous lancez une session.

L'éditeur de données permet d'afficher les données de deux façons :

- Affichage des données. Affiche les valeurs réelles des données ou les étiquettes de valeurs définies.
- Affichage des variables. Affiche les informations de définition des variables, à savoir les étiquettes de valeurs et de variables définies, le type des données (par exemple, chaîne, date ou valeur numérique), le niveau de mesure (nominale, ordinale ou échelle) et les valeurs utilisateur manquantes.

Dans les deux affichages, vous pouvez ajouter, modifier et supprimer les informations contenues dans le fichier de données.

Affichage des données

Un grand nombre des fonctions de l'affichage des données sont similaires à celles que proposent les tableurs. Il y a toutefois des différences importantes :

- Les lignes sont des observations. Chaque ligne représente une observation. Par exemple, chaque répondant d'un questionnaire est considéré comme étant une observation.
- Les colonnes sont des variables. Chaque colonne représente une variable ou une caractéristique étant mesurée. Par exemple, chaque élément ou élément d'un questionnaire est une variable.
- Les cellules contiennent des valeurs. Chaque cellule contient une seule valeur pour une variable et pour une observation. La cellule correspond au point d'intersection de l'observation et de la variable. Les cellules ne contiennent que des valeurs de données. A la différence des tableurs, les cellules de l'éditeur de données ne peuvent pas contenir de formules.
- Le fichier de données est rectangulaire. La taille du fichier de données est déterminée par le nombre d'observations et de variables. Vous pouvez entrer des données dans n'importe quelle cellule. Si vous entrez des données dans une cellule en dehors des limites du fichier de données défini, le rectangle de données est agrandi pour inclure toutes les lignes et/ou colonnes nécessaires entre cette cellule et les limites du fichier. Il n'y a pas de cellule « vide » à l'intérieur des limites du fichier de données. En ce qui concerne les variables numériques, les cellules à

blanc sont converties en valeurs manquantes par défaut. En ce qui concerne les variables chaîne, un blanc est considéré comme une valeur valide.

Viewer

Les résultats sont affichés dans le Viewer. Vous pouvez utiliser le Viewer pour :

- Parcourir les résultats
- Afficher ou masquer les tableaux et diagrammes sélectionnés.
- Modifier l'ordre d'affichage des résultats en déplaçant les éléments sélectionnés.
- Déplacer des éléments entre le Viewer et d'autres applications.

Le Viewer est divisé en deux panneaux :

- Le panneau gauche contient la légende du contenu du résultat.
- Le panneau droit contient les tableaux statistiques, les diagrammes et les textes.

Vous pouvez cliquer sur un élément affiché dans la légende pour accéder directement au tableau ou au diagramme correspondant. Vous pouvez cliquer et faire glisser le bord droit de la fenêtre de légende pour modifier la largeur de la fenêtre.

Entrée / Sortie des Données dans SPSS

L'affichage des variables présente les descriptions des attributs de chaque variable du fichier de données. Dans l'Affichage des variables :

- Les lignes sont des variables.
- Les colonnes sont des attributs de variable.

Oui	100	2 000.00	...		
Non	200	3 000.00	...		
Oui	300	4 000.00	...		

Vous pouvez ajouter ou supprimer des variables et modifier les attributs de ces dernières, y compris les attributs suivants :

- Nom de variable
- Le type de données
- Le nombre de chiffres ou de caractères
- Le nombre de décimales
- Les étiquettes descriptives de variables et de valeurs.

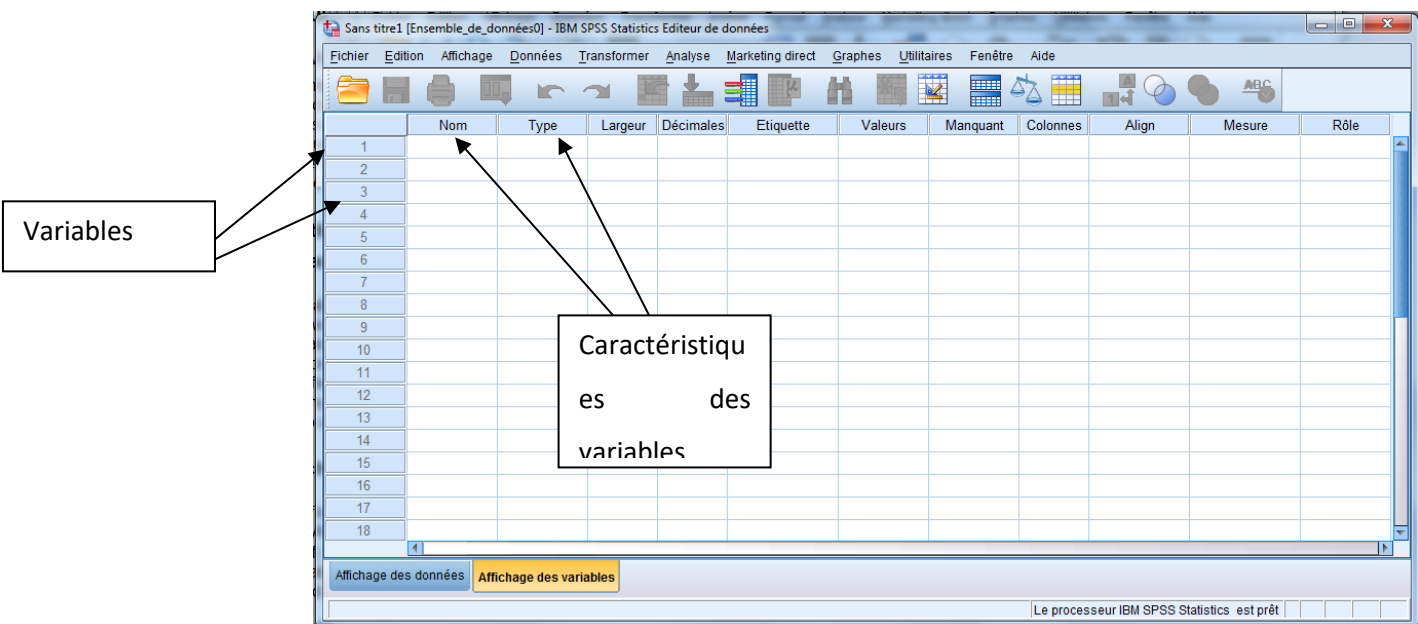
- Les valeurs manquantes définies par l'utilisateur
- Largeur des colonnes
- Le niveau de mesure

Tous ces attributs sont enregistrés lorsque vous sauvegardez le fichier de données.

Vous pouvez définir les propriétés des variables dans Affichage des variables, mais vous disposez également de deux autres méthodes pour ce faire :

- L'assistant Copier des propriétés de données permet d'utiliser un fichier de données IBM® SPSS® Statistics externe ou un autre ensemble de données disponible dans la session en cours comme modèle pour définir les propriétés de fichier et de variable dans l'ensemble de données actif. Vous pouvez également utiliser des variables de l'ensemble de données actif comme modèle pour d'autres variables de l'ensemble de données actif. L'option Copier des propriétés de données est disponible dans le menu Données de la fenêtre de l'éditeur de données..
- L'option Définir les propriétés de variable (également disponible dans le menu Données de la fenêtre de l'éditeur de données) permet d'analyser vos données et de répertorier toutes les valeurs de données uniques pour les variables sélectionnées et d'identifier les valeurs non étiquetées, et fournit une fonction d'étiquetage automatique. Cette méthode est particulièrement utile pour les variables qualitatives qui utilisent des codes numériques pour représenter les modalités. Par exemple, 0 = Masculin, 1 = Féminin.

Pour la définition des variables vous utiliser la vue variables « Affichage des Variables » accessible via Menu Affichage->Variables (CTRL+T)



Les caractéristiques des variables :

Nom :

le nom d'une variable sert d'identificateur pour cette dernière, les noms de variables n'excèdent généralement pas 8 caractères, les noms doivent :

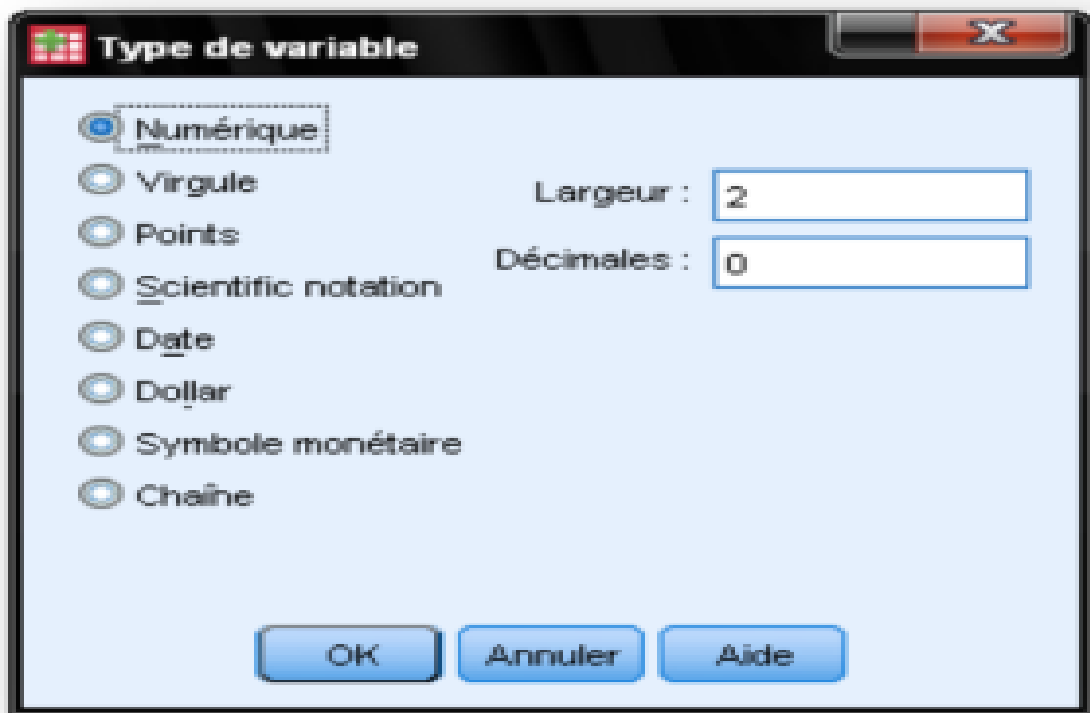
Les règles suivantes s'appliquent pour les noms des variables :

- Chaque nom de variable doit être unique ; aucune duplication n'est admise.
- Les noms de variable peuvent contenir jusqu'à 64 octets, le premier caractère étant une lettre ou l'un des caractères suivants : @, # ou \$. Les caractères suivants peuvent être une combinaison de lettres, de chiffres, un point (.) et des caractères autres que ceux de ponctuation. En mode page de code, soixante-quatre octets correspondent à 64 caractères dans les langues sur un octet (anglais, français, allemand, espagnol, italien, hébreu, russe, grec, arabe et thaï par exemple) et à 32 caractères dans les langues sur deux octets (japonais, chinois et coréen par exemple). De nombreux caractères qui n'occupent qu'un seul octet en mode page de code en occupent au moins deux en mode Unicode. Par exemple, é ne représente qu'un seul octet en mode page de code, mais en occupe deux au format Unicode. Ainsi, résumé est égal à six octets dans un fichier page de code et à huit en mode Unicode.

Remarque : Les lettres incluent tout caractère autre que ceux de ponctuation utilisé dans l'écriture de mots courants dans les langues prises en charge dans le jeu de caractères de la plateforme.

- Les noms de variable ne doivent pas contenir d'espaces.
- Le caractère # au début du nom de la variable désigne une variable temporaire. Vous ne pouvez créer des variables temporaires qu'avec une syntaxe de commande. Vous ne pouvez pas entrer le signe # comme premier caractère d'une variable dans une boîte de dialogue de création de variables.
- Le symbole \$ en début de nom indique que la variable est une variable système.. Vous ne pouvez pas utiliser le symbole \$ comme premier caractère d'une variable définie par l'utilisateur.
- Le point, le trait de soulignement et les caractères \$, # et @ peuvent être utilisés dans les noms de variable. Par exemple, A_.\$@#1 est un nom de variable valide.
- Evitez les noms de variable se terminant par un point car celui-ci peut être interprété comme un caractère de fin de commande. Vous ne pouvez créer des variables se terminant par un point que dans une syntaxe de commande. Vous ne pouvez pas créer de variables se terminant par un point dans une boîte de dialogue de création de variables.

- Evitez d'utiliser des noms de variable se terminant par des traits de soulignement, étant donné que ceux-ci peuvent entrer en conflit avec des noms de variable automatiquement créés par les commandes et les procédures.
- Les mots-clés réservés ne peuvent pas être utilisés pour les noms de variables : Les mots-clés réservés sont ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO et WITH.
- Les noms de variables peuvent être définis par n'importe quelle combinaison de majuscules et de minuscules. La casse est respectée pour des raisons d'affichage.
- Lorsque des noms longs de variable occupent plusieurs lignes au niveau du résultat, les sauts de ligne sont segmentés au niveau des traits de soulignement, des virgules et des passages de minuscule à majuscule.



Type :

La plupart des données manipulées avec SPSS vont être de simples nombres, alors que d'autres nécessitent soit un formatage spécifique (monétaire), ou un traitement spécifique (Date & Heure).

Les types de données disponibles sont les suivants :

Numérique. Variable dont les valeurs sont des nombres. Les valeurs sont affichées en format numérique standard. L'éditeur de données accepte les valeurs numériques au format standard ou sous forme de notation scientifique.

Virgule. Variable numérique dont les valeurs sont affichées avec des virgules toutes les trois positions, le point servant de séparateur décimal. L'outil Data Editor accepte les valeurs numériques pour les variables de virgule avec ou sans virgule ou sous forme de notation scientifique. Les valeurs ne peuvent pas contenir de virgule à droite de l'indicateur décimal.

Point. Variable numérique dont les valeurs sont affichées avec des points toutes les trois positions, la virgule servant de séparateur décimal. L'outil Data Editor accepte les valeurs numériques pour les variables de point avec ou sans point ou sous forme de notation scientifique. Les valeurs ne peuvent pas contenir de point à droite de l'indicateur décimal.

Notation scientifique. Variable numérique dont les valeurs sont affichées avec un E intégré et un exposant de puissance dix avec signe. L'éditeur de données accepte des valeurs numériques pour les variables de notation scientifique avec ou sans exposant. L'exposant peut être précédé d'un E ou d'un D avec ou sans signe, ou seulement d'un signe. Par exemple, 123, 1.23E2, 1.23D2, 1.23E+2 et même 1.23+2.

Date. Variable numérique dont les valeurs sont affichées dans l'un des formats de date ou d'heure possibles. Sélectionnez un format dans la liste. Vous pouvez entrer des dates avec, comme séparateur, des barres obliques, des traits d'union, des points, des virgules ou des espaces. La valeur du siècle pour les années à 2 chiffres est déterminée par les paramètres Options (accessibles depuis le menu Edition, sélectionnez Options, puis cliquez sur l'onglet Données).

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.

Symbole monétaire : Variable numérique dont les valeurs sont affichées dans l'un des formats monétaires personnalisés que vous avez définis dans l'onglet Devise de la boîte de dialogue Options. Les caractères de symbole monétaire définis ne sont pas utilisables lors de la saisie de données mais sont affichés dans l'éditeur de données.

Chaîne. Variable dont les valeurs ne sont pas numériques et ne sont donc pas utilisées pour les calculs. Ces valeurs peuvent contenir n'importe quel caractère, dans la limite de la longueur définie. Les majuscules et les minuscules sont différenciées. Ce type de variable est aussi connu sous le nom de variable alphanumérique.

Format numérique restreint. 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.

Largeur :

Le nombre de caractères utilisés pour l’affichage des valeurs de la variable

Décimales :

Le nombre de caractères utilisés pour l’affichage des valeurs décimales

Etiquette :

Vous pouvez attribuer des étiquettes de variables descriptives dont le nombre de caractères ne dépasse pas 256 (128 caractères pour les langages sur deux octets). Les étiquettes de variable peuvent contenir des espaces et des caractères réservés qui ne sont pas autorisés dans les noms de variable.

Valeurs :

Vous pouvez affecter des étiquettes descriptives de valeur pour chaque valeur d'une variable. Ce processus se révèle particulièrement utile si votre fichier de données utilise des codes numériques pour représenter des modalités non numériques (par exemple, les codes 1 et 2 pour homme et femme).

Manquant :

L'option Valeurs manquantes permet de définir les valeurs de données spécifiées comme valeurs manquantes spécifiées par l'utilisateur. Par exemple, vous pouvez faire la distinction entre les données manquantes parce qu'une personne interrogée a refusé de répondre et les données manquantes parce que la question ne s'appliquait pas au répondant. Les valeurs des données définies comme valeurs utilisateur manquantes sont repérées par un indicateur en vue d'un traitement spécial et sont exclues de la plupart des calculs.

- Les spécifications des valeurs manquantes de type utilisateur sont enregistrées avec le fichier de données. Vous n'avez pas besoin de redéfinir des valeurs manquantes de type utilisateur à chaque ouverture de fichier de données.
- Vous pouvez entrer jusqu'à trois valeurs manquantes de votre choix, un intervalle de valeurs manquantes ou un intervalle plus une valeur de votre choix.
- Les intervalles ne peuvent être spécifiés que pour des valeurs numériques.
- Toutes les valeurs de chaîne, y compris les valeurs nulles ou vides, sont considérées comme des valeurs valides à moins que vous ne les définissiez comme manquantes.

- Les valeurs manquantes pour les variables chaîne ne peuvent pas dépasser huit octets. (Il n'y a pas de limite pour la largeur définie de la variable chaîne, mais les valeurs manquantes définies ne peuvent pas dépasser huit octets)
- Pour définir des valeurs nulles ou vides comme manquantes pour une variable chaîne, entrez un seul espace dans l'un des champs sous la sélection Valeurs manquantes discrètes.

Colonnes :

Vous pouvez spécifier le nombre de caractères définissant la largeur des colonnes. Vous pouvez également modifier la largeur des colonnes dans Affichage des données en cliquant et en tirant les bords des colonnes.

- La largeur des colonnes des polices proportionnelles est basée sur la largeur moyenne des caractères. En fonction des caractères utilisés dans la valeur, un nombre plus ou moins important de caractères peut être affiché dans la largeur spécifiée.
- La largeur des colonnes affecte seulement l'affichage des valeurs dans l'éditeur de données. Modifier la largeur de la colonne ne change pas la largeur définie d'une variable.

Align :

L'alignement contrôle l'affichage des valeurs des données et/ou des étiquettes de valeur dans Affichage des données. L'alignement par défaut est à droite pour les variables numériques et à gauche pour les variables chaînes. Ce paramètre n'affecte que l'affichage d'Affichage des données.

Mesure :

Vous pouvez spécifier un niveau de mesure d'échelle (données numériques sur un intervalle ou une échelle de rapport), ordinal ou nominal Les données nominales et ordinales peuvent être des chaînes de caractères (alphanumériques) ou numériques.

- **Nominal.** Une variable peut être traitée comme étant nominale si ses valeurs représentent des modalités sans classement intrinsèque (par exemple, le service de la société dans lequel travaille un employé). La région, le code postal ou l'appartenance religieuse sont des exemples de variables nominales.
- **Ordinal.** Une variable peut être traitée comme étant ordinale si ses valeurs représentent des modalités associées à un classement intrinsèque (par exemple, des niveaux de satisfaction allant de Très mécontent à Très satisfait). Exemples de variable ordinale : des scores d'attitude représentant le degré de satisfaction ou de confiance, et des scores de classement des préférences.

• **Echelle.** Une variable peut être traitée comme une variable d'échelle (continue) si ses valeurs représentent des modalités ordonnées avec une mesure significative, de sorte que les comparaisons de distance entre les valeurs soient adéquates. L'âge en années et le revenu en milliers de dollars sont des exemples de variable d'échelle.

Remarque : Pour les variables chaîne ordinales, l'ordre alphabétique des valeurs chaîne est supposé refléter l'ordre des modalités. Par exemple, pour une variable chaîne comportant des valeurs Faible, Moyen, Elevé, l'ordre des modalités 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.

Pour les nouvelles variables numériques créées avec des transformations, des données provenant de sources extérieures, et des fichiers de données IBM® SPSS® Statistics créés avant la version 8, le niveau de mesure par défaut est déterminé par les conditions du tableau suivant. Les conditions sont évaluées dans l'ordre répertorié dans le tableau. Le niveau de mesure pour la première condition qui correspond aux données est appliqué.

Saisie de données

Dans Affichage des données, vous pouvez entrer les données directement dans l'éditeur de données. Vous pouvez entrer des données dans n'importe quel ordre. Vous pouvez entrer des données par observation ou par variable, pour des zones sélectionnées ou des cellules individuelles.

- La cellule active est mise en surbrillance.
- Le nom de variable et le numéro de ligne de la cellule active sont affichés dans le coin supérieur gauche de l'éditeur de données.
- Lorsque vous sélectionnez une cellule et lorsque vous entrez une valeur de données, la valeur est affichée dans l'éditeur de cellules en haut de l'éditeur de données.
- Les valeurs de données ne sont pas enregistrées tant que vous n'avez pas appuyé sur Entrée ou que vous n'avez pas sélectionné une autre cellule.
- Pour entrer autre chose que des données numériques simples, vous devez d'abord définir le type de variable.

Si vous entrez une valeur dans une colonne vide, l'éditeur de données crée automatiquement une nouvelle variable et affecte un nom de variable.

Pour entrer des données numériques

- ▶ Sélectionnez une cellule dans Affichage des données.

▶ Entrez la valeur des données. (La valeur est affichée dans l'éditeur de cellules dans la partie supérieure de Data Editor.)

▶ Appuyez sur Entrée ou sélectionnez une autre cellule pour enregistrer la valeur.

Pour entrer des données non-numériques

▶ Double-cliquez sur un nom de variable dans la zone supérieure de la colonne dans Affichage des données ou cliquez sur l'onglet Affichage des variables.

▶ Cliquez sur le bouton de la cellule Type de la variable.

▶ Sélectionnez le type de données dans la boîte de dialogue Type de variable.

▶ Cliquez sur OK.

▶ Double-cliquez sur le numéro de ligne ou cliquez sur l'onglet Affichage des données.

▶ Pour la variable que vous venez de définir, entrez les données dans la colonne.

1- Lecture des données à partir des fichiers

Fichiers de données

Les fichiers de données se présentent sous une grande diversité de formats et ce logiciel a été conçu pour traiter nombre d'entre eux, dont :

- Feuilles de calcul créées sous Excel et Lotus
- Tableaux de bases de données issus de plusieurs sources de bases de données, notamment Oracle, SQLServer, Access, dBASE, etc.
- Fichiers texte délimités par des tabulations et autres types de fichiers texte simples
- les fichiers de données au format IBM® SPSS® Statistics créés avec d'autres systèmes d'exploitation ;
- Fichiers de données SYSTAT
- Fichiers de données SAS
- Fichiers de données Stata
- Rapports de liste et packages de données Business Intelligence IBM® Cognos®

Pour ouvrir des fichiers de données

▶ A partir des menus, sélectionnez :

Fichier > Ouvrir > Données

▶ Dans la boîte de dialogue Ouvrir données, sélectionnez le fichier à ouvrir.

▶ Cliquez sur Ouvrir.

Sinon, vous pouvez :

- Définir automatiquement la largeur de chaque variable chaîne à la valeur observée la plus longue pour cette variable, en utilisant Minimize string widths based on observed values. Ceci

est très utile lors de la lecture de fichiers de données de page de code en mode Unicode. Consultez la section [Options générales](#) pour plus d'informations.

- Lire les noms de variable sur la première ligne des fichiers de feuilles de calcul.
- spécifier une plage de cellules à lire dans le cas de fichiers de feuilles de calcul.
- Spécifier une feuille de calcul à lire dans un fichier Excel (version Excel 95 ou supérieure).

Types de fichier de données

SPSS Statistics. Ouvre les fichiers de données enregistrés au format IBM® SPSS® Statistics ainsi que le produit DOS SPSS/PC+.

SPSS Statistics Compressé. Ouvre les fichiers de données enregistrés au format SPSS Statistics compressé.

SPSS/PC+ : Ouvre les fichiers de données SPSS/PC+. Cette option n'est disponible que sous les systèmes d'exploitation Windows.

SYSTAT. Ouvre les fichiers de données SYSTAT.

SPSS Statistics Portable. Ouvre les fichiers de données sauvegardés en format portable. Sauvegarder un fichier en format portable prend bien plus de temps que de sauvegarder le fichier en format SPSS Statistics.

Excel : Ouvre les fichiers Excel.

Lotus 1-2-3. Ouvre les fichiers de données enregistrés au format 1-2-3 pour les versions 3.0 et 2.0, ainsi que pour la version 1A de Lotus.

SYLK : Ouvre les fichiers de données sauvegardés en format SYLK (lien symbolique), format utilisé par quelques applications de tableurs.

dBASE : Ouvre des fichiers de format dBASE pour dBASE IV, dBASE III ou III PLUS, ou pour dBASE II. Chaque observation est un enregistrement. Les étiquettes de variable et de valeurs ainsi que les spécifications de valeurs manquantes sont perdues lorsque vous sauvegardez un fichier dans ce format.

SAS. SAS versions 6–9 et fichiers de transfert SAS. A l'aide de la syntaxe de commande, vous pouvez également lire les étiquettes des valeurs depuis un fichier de catalogue au format SAS.

Stata. Version Stata 4–8.

age	marital	address	income	inccat	car	carca
55	1	12	72	3	36.2	3
56	0	29	153	4	76.9	3
28	1	9	28	2	13.7	1
24	1	4	26	2	12.5	1
25	0	2	23	1	11.3	1
45	1	9	76	4	37.2	3
42	0	19	40	2	19.8	2
35	0	15	57	3	28.2	2
46	0	26	24	1	12.2	1
34	1	0	89	4	46.1	3
55	1	17	72	3	35.5	3
28	0	3	24	1	11.8	1
31	1	9	40	2	21.3	2
42	0	8	137	4	68.9	3
35	0	8	70	3	34.1	3
52	1	24	159	4	78.9	3
21	1	1	37	2	18.6	2
32	0	0	28	2	13.7	1
42	0	9	109	4	54.7	3

Les fichiers délimités par une virgule ou une tabulation se rapportent aux lignes de données utilisant des virgules ou des tabulations pour indiquer chaque variable. Dans cet exemple, les données sont délimitées par des tabulations

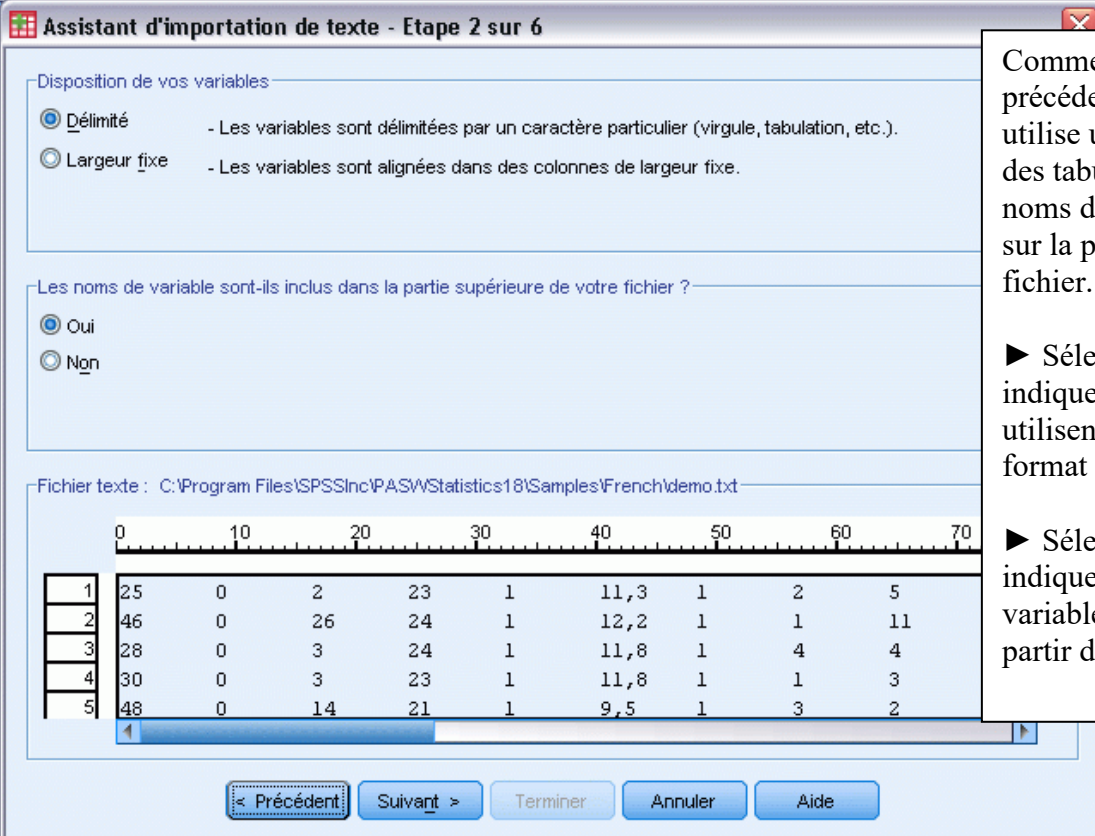
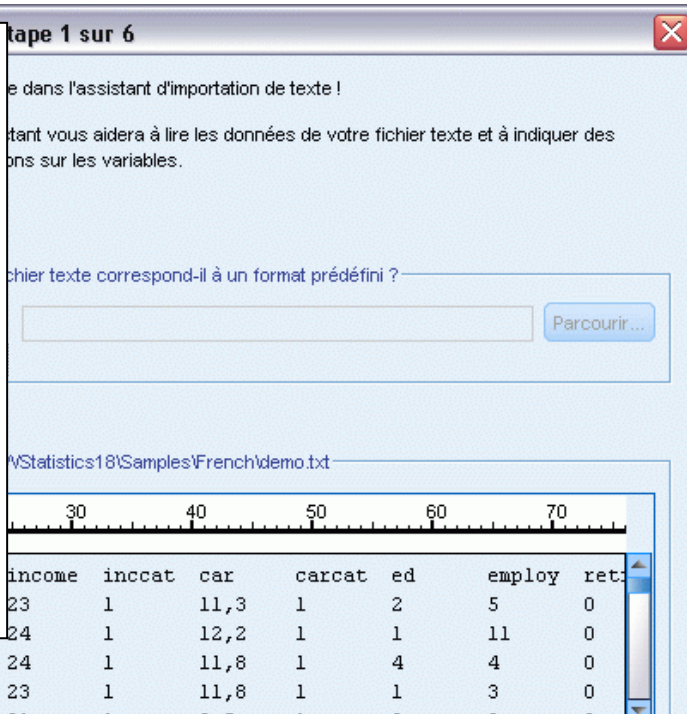
Fichier	Edition	Affichage	Données	Transformer	Analyse	Graphes	Outils	Fenêtre	Aide
Nouveau									
Ouvrir									
Capturer base de données									
Lire les données texte...									
<hr/>									
Fermer									Ctrl+F4
Enregistrer									Ctrl+S
Enregistrer sous...									
Enregistrer toutes les données									
Marquer le fichier comme étant en lecture seule									
<hr/>									
Renommer l'ensemble de données...									
Afficher des informations sur les fichiers de données									
Données cache...									
Arrêter processeur									Ctrl+,
Changer serveur...									
<hr/>									
Aperçu avant impression									
Imprimer...									Ctrl+P
<hr/>									
Données récemment utilisées									
Fichiers récemment utilisés									
<hr/>									
Quitter									

A partir des menus, sélectionnez :
Fichier > Lire les données texte...
Sélectionnez Texte (*.txt) comme type de fichier à afficher.

L'Assistant d'importation de texte vous guide tout au long du processus permettant de définir le mode d'impression du fichier texte indiqué.

► A l'étape 1, vous pourrez sélectionner un format prédéfini ou créer un format dans l'Assistant. Sélectionnez Non pour indiquer qu'un nouveau format doit être créé.

► Cliquez sur Suivant pour



Comme indiqué précédemment, ce fichier utilise un format délimité par des tabulations. En outre, les noms de variable sont définis sur la première ligne de ce fichier.

► Sélectionnez Délimité pour indiquer que les données utilisent une structure de format délimité.

► Sélectionnez Oui pour indiquer que les noms de variable doivent être lus à partir du début du fichier.

Assistant d'importation de texte - Délimité, étape 3 sur 6

► Saisissez 2 dans la section supérieure de la boîte de dialogue suivante pour indiquer que la première ligne de données commence sur la deuxième ligne du fichier texte.

► Cliquez sur Suivant pour continuer.

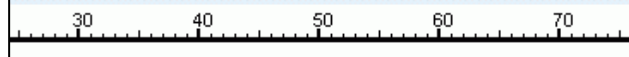
► Conservez les valeurs par défaut des autres champs de cette boîte de dialogue et cliquez sur Suivant pour continuer.

2

Présente une observation : 29

Importer ?

Précision (approximatif) : 10 %



23	1	11,3	1	2	5	0
24	1	12,2	1	1	11	0
24	1	11,8	1	4	4	0
23	1	11,8	1	1	3	0

Assistant d'importation de texte - Délimité, étape 4 sur 6

L'aperçu des données de l'étape 4 vous offre un moyen rapide de vérifier que vos données ont été lues correctement.

► Sélectionnez Tab et désélectionnez les autres options.

► Cliquez sur Suivant pour continuer.

Qu'est-ce qu'un qualificateur de texte ?

- Aucun
- Apostrophe
- Guillemet
- Autre :

	inccat	car	carcat	ed
	1	11,3	1	2
	1	12,2	1	1
	1	11,8	1	4
	1	11,8	1	1
	1	9,5	1	3
	1	8,5	1	4
	1	10	1	3
	1	11,4	1	3
	1	10,5	1	4

< Précédent

Suivant >

Terminer

Annuler

Aide

Etant donné que des noms de variable peuvent avoir été tronqués pour des raisons de formatage, cette boîte de dialogue vous permet de modifier les noms superflus.

Vous pouvez également définir les types de données dans cette boîte de dialogue. Par exemple, nous pouvons supposer que la variable de revenus doit contenir une certaine somme en dollars

Pour modifier un type de données :

- Sélectionnez Dollar dans la liste déroulante Format des données.
- Cliquez sur Suivant pour continuer.

Assistant d'importation de texte - Etape 5 sur 6

Spécifications pour les variables sélectionnées dans l'aperçu de données

Nom de variable : Nom d'origine : age

Format de données :

Aperçu des données

age	marital	address	income	inccat	car	carcat	ed
25	0	2	23	1	11,3	1	2
46	0	26	24	1	12,2	1	1
28	0	3	24	1	11,8	1	4
30	0	3	23	1	11,8	1	1
48	0	14	21	1	9,5	1	3

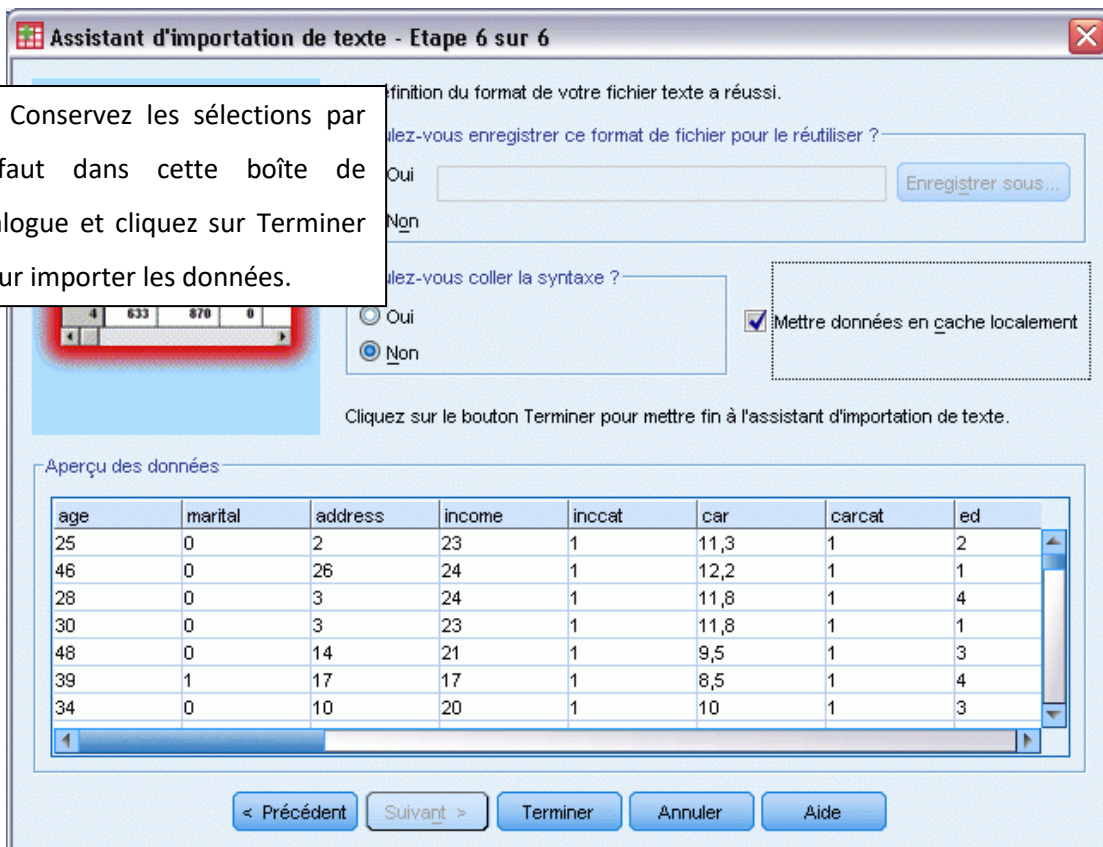
Annuler Aide

sur 6

Aperçu des données

age	marital	address	income	inccat	car	carcat	ed
25	0	2	23	1	11,3	1	2
46	0	26	24	1	12,2	1	1
28	0	3	24	1	11,8	1	4
30	0	3	23	1	11,8	1	1
48	0	14	21	1	9,5	1	3

< Précédent Suivant > Terminer Annuler Aide



Analyses statistiques

- 1- Exécution d'une analyse
- 2- Des Analyses simples

Examen des statistiques récapitulatives pour chaque variable

Ce chapitre traite des mesures récapitulatives simples et de la façon dont le niveau de mesure d'une variable influence le type de statistiques devant être utilisé. Nous utiliserons le fichier de données demo.sav.

Niveau de mesure

Différentes mesures récapitulatives sont adaptées à différents types de données, selon le niveau de mesure :

demo.sav [Ensemble_de_données1] - Editeur de données

Fichier Edition Affichage Données Transforme Analyse Graphes Outils Modules complémentaires Fenêtre Aide

20 : age 40 Visible : 29 variables sur 29

	age	situatio	address	revenu	inccat	car
13	31	1	12	40,00	2,00	21,30
14	42	0	29	137,00	4,00	68,90
15	35	0	8	70,00	3,00	34,10
16	52	1	24	159,00	4,00	78,90
17	21	1	1	37,00	2,00	18,60
18	32	0	0	28,00	2,00	13,70
19	42	0	9	109,00	4,00	54,70
20	40	1	12	117,00	4,00	58,30
21	30	0	3	23,00	1,00	11,80
22	48	0	14	21,00	1,00	9,50

Affichage des données Affichage des variables

Processeur prêt

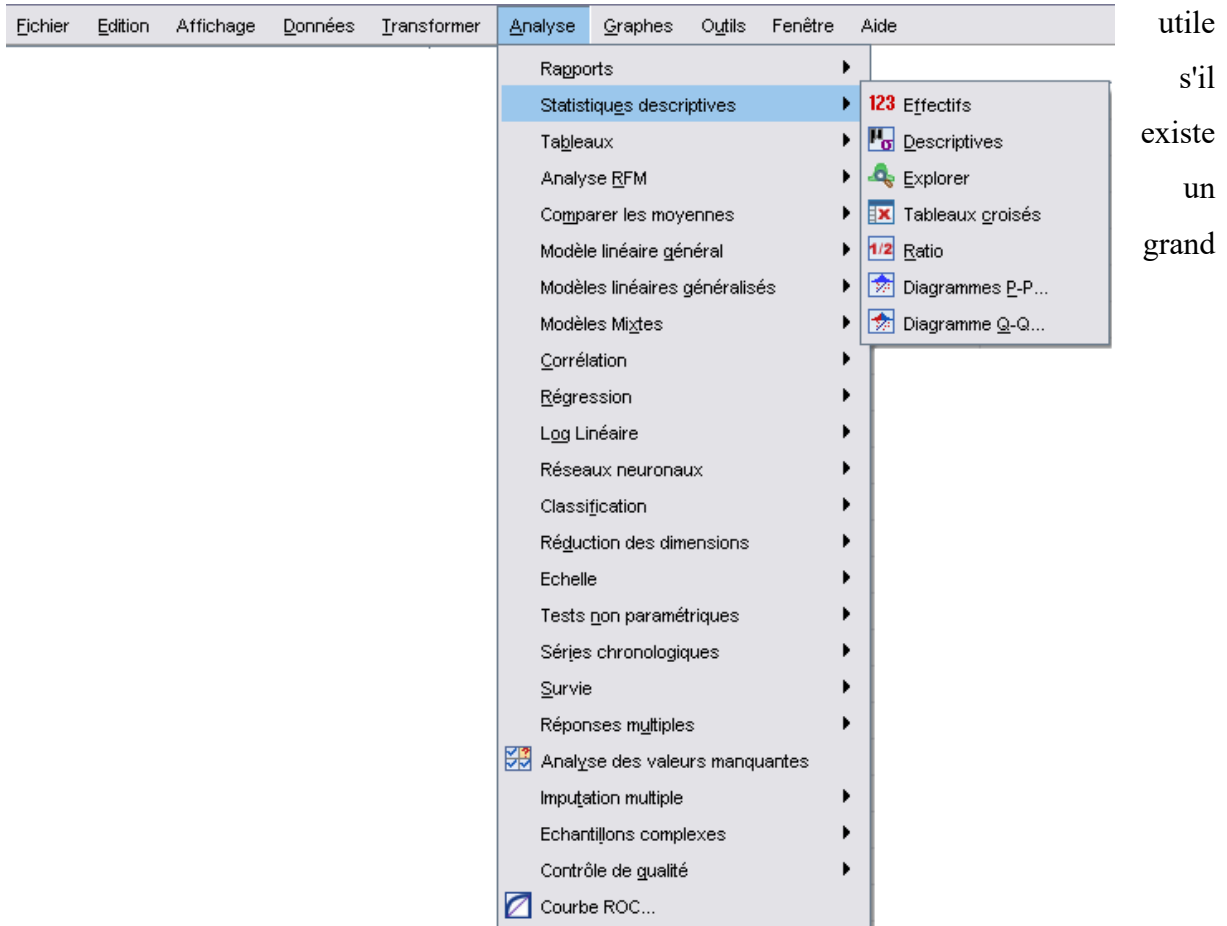
Qualitatives : Données ayant un nombre limité de valeurs ou de modalités distinctes (par exemple, sexe ou situation de famille). Elles sont parfois également qualifiées de données qualitatives. Les variables qualitatives peuvent être des données chaîne (alphanumérique) ou des variables numériques qui utilisent des codes chiffrés pour représenter les modalités (par exemple, 0 = Célibataire et 1 = Marié). Il existe deux types essentiels de données qualitatives :.....voir (1)

- **Nominal**. Données qualitatives dont les modalités n'ont aucun ordre inhérent. Par exemple, une modalité d'emploi de type ventes n'est pas supérieure ou inférieure à une modalité d'emploi de type marketing ou étude.
- **Ordinal**. Données qualitatives dont les modalités possèdent un ordre significatif, mais pour lesquelles il n'existe aucune distance mesurable entre les modalités. Par exemple, les valeurs élevée, moyenne et faible doivent être classées dans un certain ordre, mais il est impossible de calculer la « distance » entre ces valeurs.

Echelle. Données mesurées sur une échelle d'intervalle ou de rapport, où les valeurs de données indiquent à la fois l'ordre des valeurs et la distance qui les sépare. Par exemple, un salaire de 58 160 € est supérieur à un salaire de 42 212 € et la distance entre les deux valeurs est de 15 948 €. Ces données sont aussi appelées données quantitatives ou données continues. ... Voir (2)

Mesures récapitulatives pour données qualitatives

Pour les données qualitatives, la mesure récapitulative la plus courante est le nombre ou le pourcentage d'observations dans chaque modalité. Le mode est la modalité ayant le plus grand nombre d'observations. Pour les données ordinales, la médiane (valeur au-dessus ou au-dessous de laquelle se trouve la moitié des observations) peut également être une mesure récapitulative



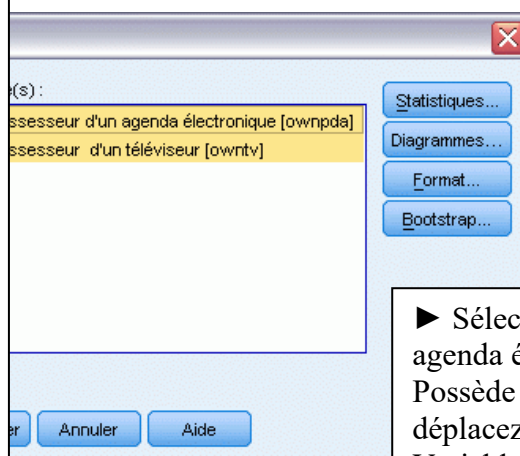
nombre de modalités.

La procédure Fréquences produit des tableaux de fréquences qui affichent le nombre et le pourcentage d'observations pour chaque valeur observée d'une variable.

► A partir des menus, sélectionnez :

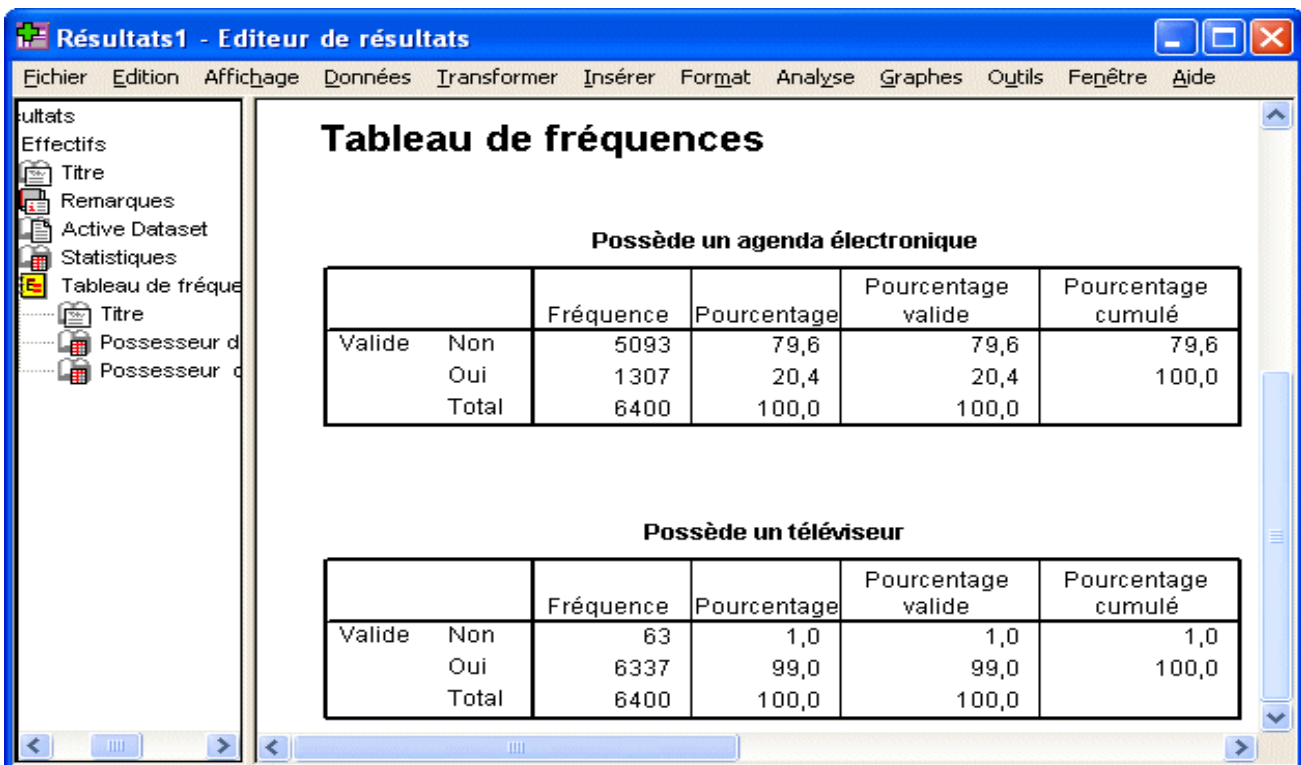
Analyse > Statistiques descriptives > Effectifs...

Remarque : Cette fonction nécessite l'option Statistiques de base.



► Sélectionnez Possède un agenda électronique [pda] et Possède un téléviseur [tv], et déplacez-les vers la liste Variable(s).

► Cliquez sur OK pour exécuter la procédure.



Les tableaux de fréquences apparaissent dans la fenêtre du Viewer. Les tableaux d'effectifs révèlent que seuls 20,4 % des personnes possèdent un agenda électronique, mais que la quasi-totalité possèdent une télévision (99,0 %). Ces informations ne semblent pas vraiment pertinentes, mais il peut être intéressant d'en savoir plus sur le petit groupe de personnes qui ne possèdent pas de télévision.

Vous pouvez afficher graphiquement les informations dans un tableau de fréquences avec un diagramme en bâtons ou un diagramme en secteurs.

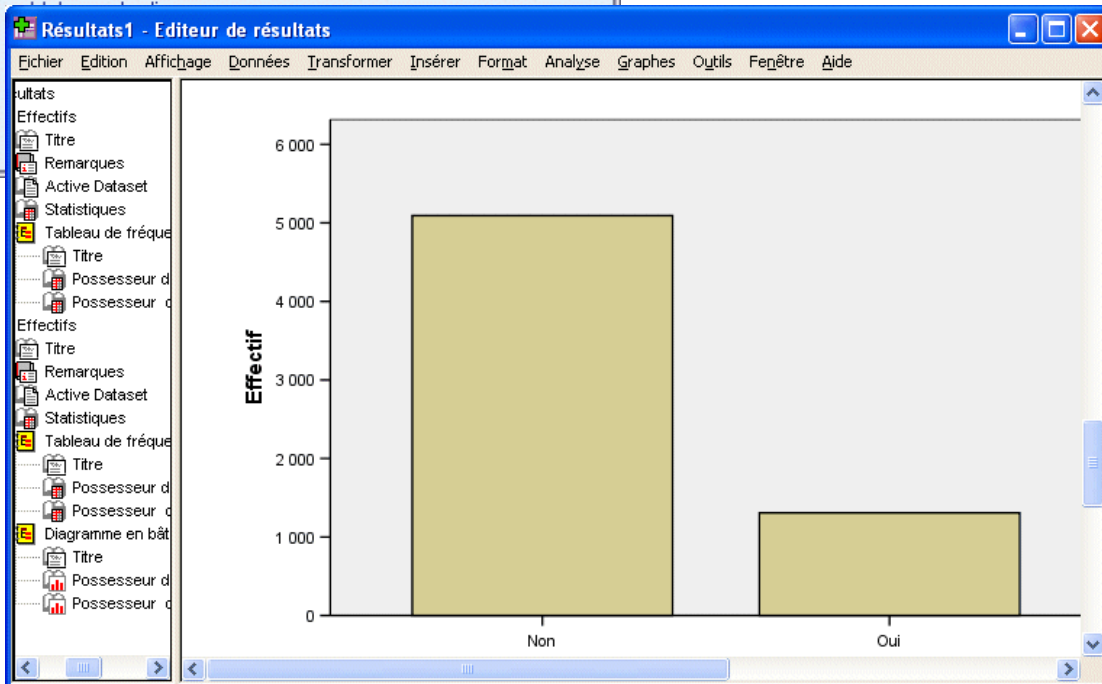
► Ouvrez à nouveau la boîte de dialogue Fréquences. (Les deux variables doivent toujours être sélectionnées.) (Vous pouvez utiliser le bouton Rappeler boîte de dialogue de la barre d'outils pour revenir rapidement aux dernières procédures utilisées.)





- ▶ Cliquez sur Diagrammes.
- ▶ Cliquez sur Diagrammes en bâtons, puis sur Poursuivre.
- ▶ Cliquez sur OK dans la boîte de dialogue principale pour exécuter la procédure.

Outre les tableaux de



fréquences, les mêmes informations sont à présent affichées sous forme de diagrammes en bâtons ; vous pouvez ainsi voir que la plupart des personnes n'ont pas d'agenda électronique alors que la quasi-totalité d'entre elles possèdent une télévision.

Mesures récapitulatives pour variables d'échelle

De nombreuses mesures récapitulatives sont disponibles pour les variables d'échelle, dont :

- Mesures de la tendance centrale. Les mesures les plus courantes de la tendance centrale sont la moyenne (moyenne arithmétique) et la médiane (valeur au-dessus ou au-dessous de laquelle se trouve la moitié des observations).

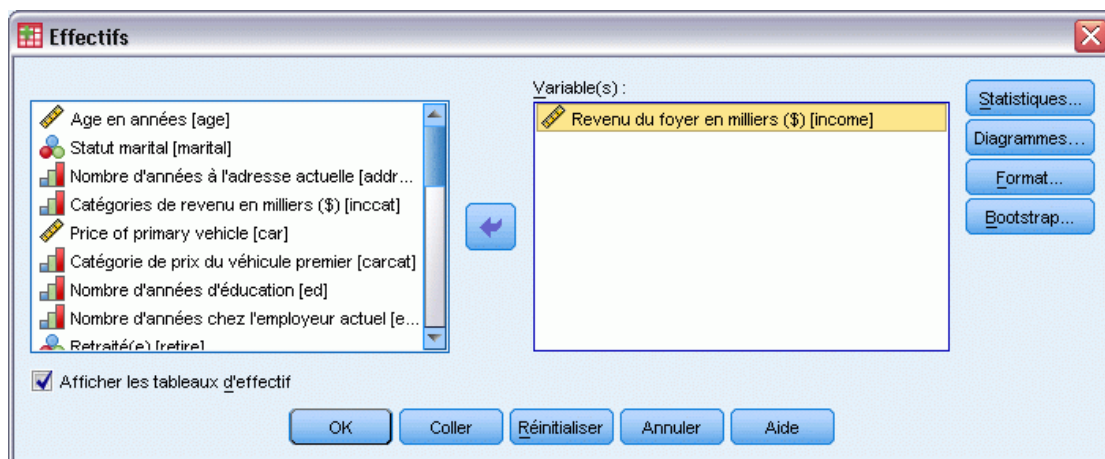
- Mesures de la dispersion. Les statistiques qui mesurent la quantité de variation ou de dispersion dans les données comprennent l'écart-type, minimal et maximal.

The screenshot shows the 'Statistiques' window in SPSS. The table displays the following data:

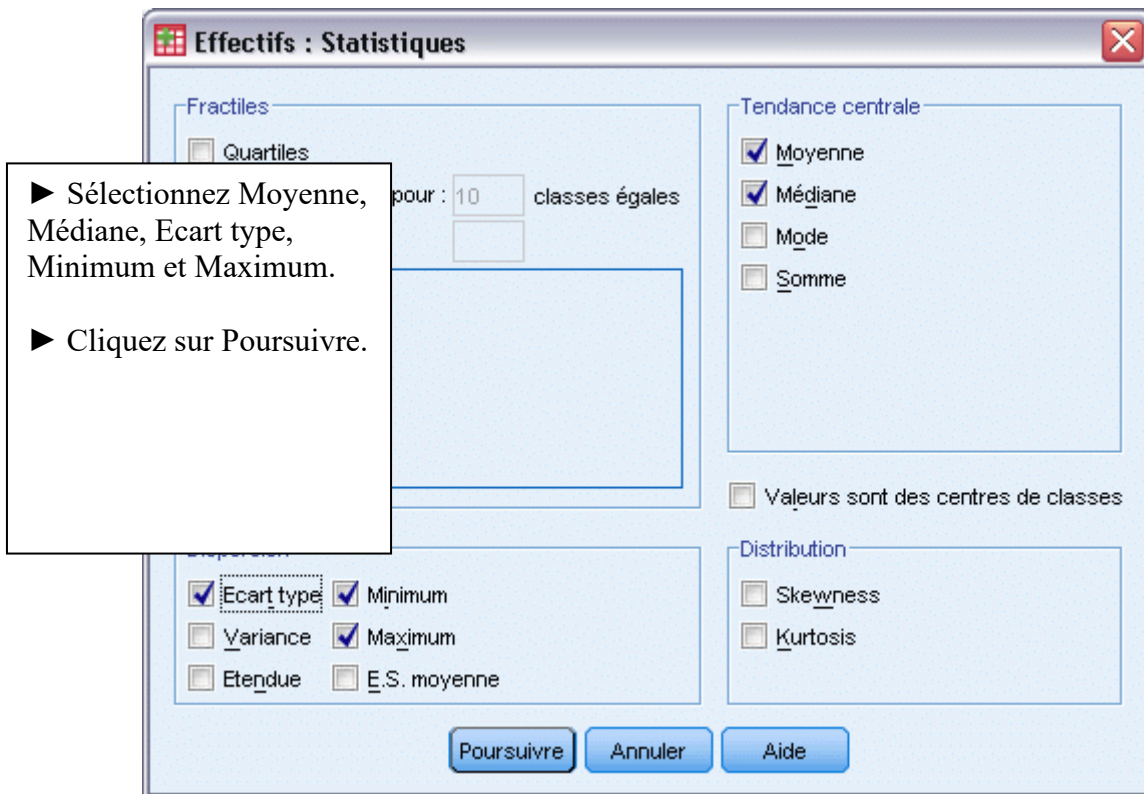
Statistiques		
Revenu du ménage en milliers (revenu)		
N	Valide	6400
	Manquante	0
Moyenne		69,4748
Médiane		45,0000
Ecart-type		78,71856
Minimum		9,00
Maximum		1116,00

Ouvrez à nouveau la boîte de dialogue Fréquences.

- ▶ Cliquez sur Réinitialiser pour effacer les paramètres précédents.
 - ▶ Sélectionnez la variable Revenu du ménage en milliers [revenu] et déplacez-la dans la liste Variable(s).
- Variable(s).



Cliquez sur Statistiques.

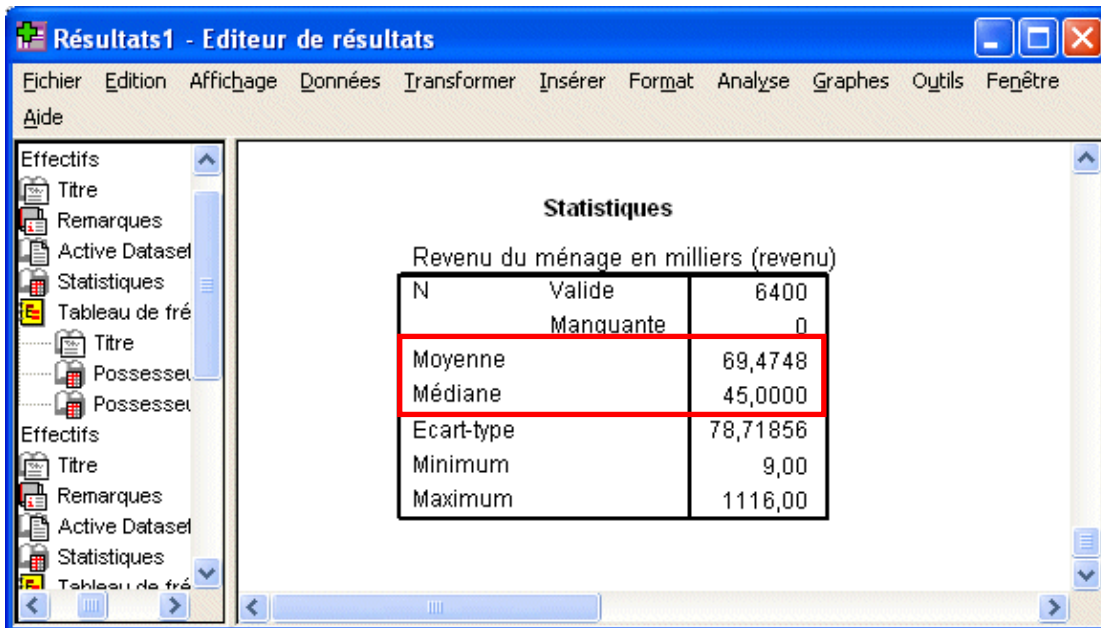


Désélectionnez Afficher les tableaux d'effectif dans la boîte de dialogue principale Effectifs. (En général, les tableaux de fréquences ne sont pas très utiles pour les variables d'échelle car il peut exister presque autant de valeurs distinctes que d'observations dans le fichier de données.)

► Cliquez sur OK pour exécuter la procédure.

Le tableau statistique de fréquences est affiché dans la fenêtre du Viewer.

Dans cet exemple, la différence entre la moyenne et la médiane est importante. La moyenne est plus importante que la médiane de quasiment 25 000, ce qui indique que les valeurs ne sont pas distribuées normalement. Vous pouvez vérifier visuellement la distribution grâce à un histogramme.

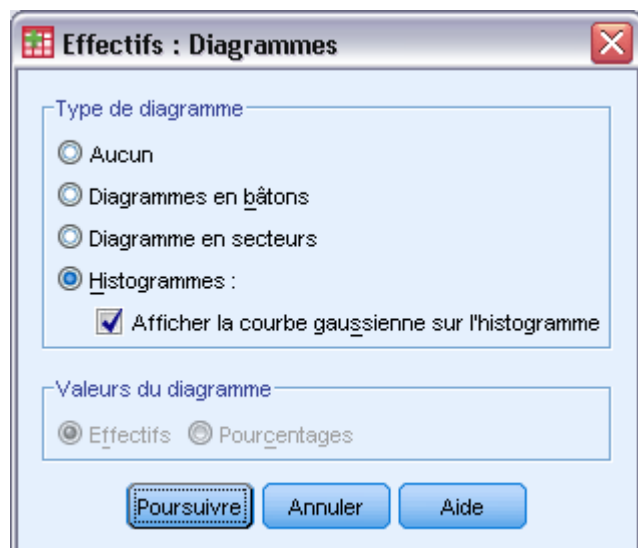


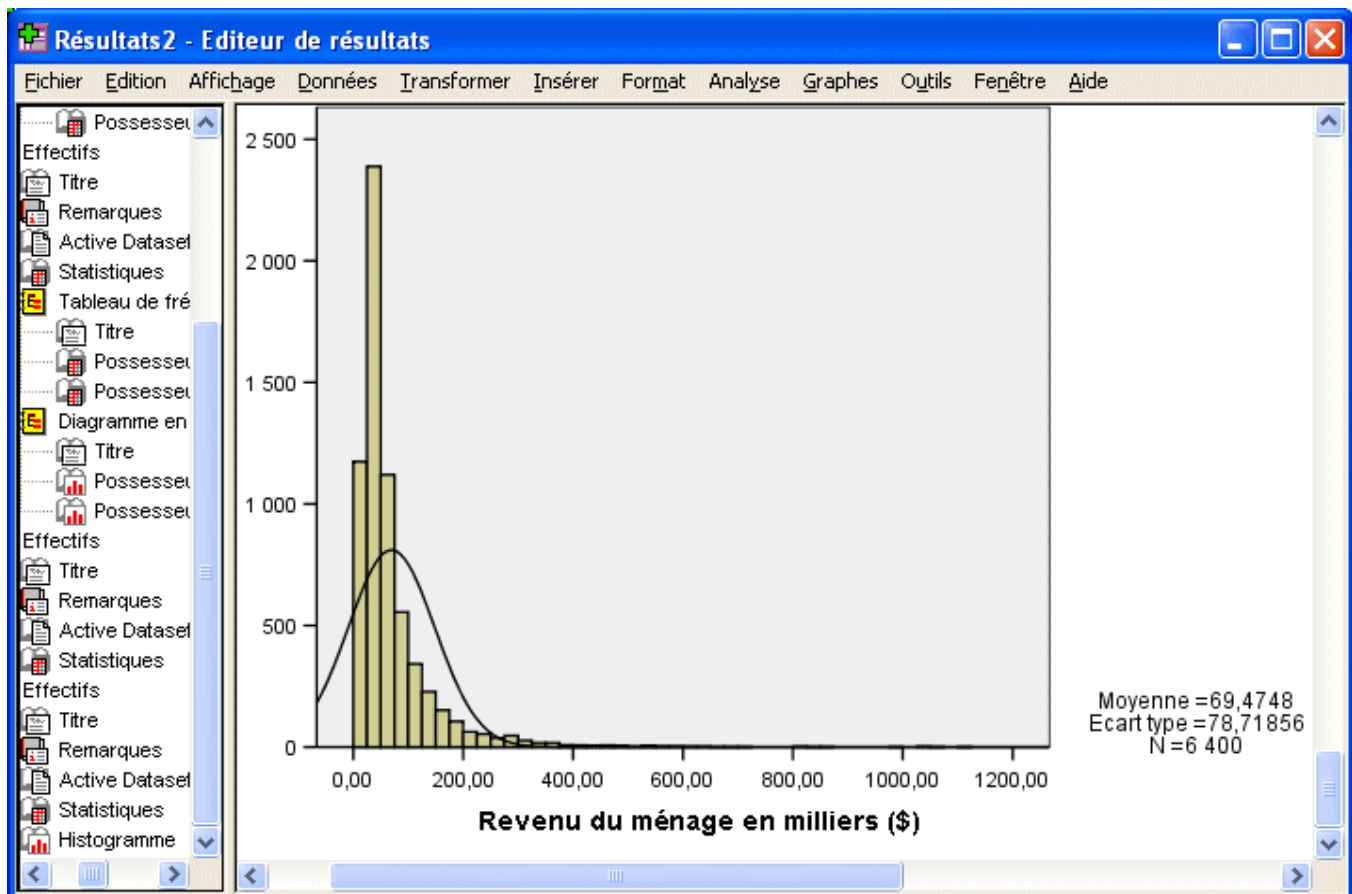
► Ouvrez à nouveau la boîte de dialogue Fréquences.

► Cliquez sur Diagrammes.

► Cliquez sur Histogrammes et Avec courbe gaussienne.

► Cliquez sur Poursuivre, puis sur OK dans la boîte de dialogue principale pour exécuter la procédure.





La grande majorité des observations est regroupée au bas de l'échelle, la plupart se trouvant au-dessous de 100 000. Quelques observations, cependant, se trouvent dans l'intervalle 500 000 et au-delà (elles sont si peu nombreuses que vous devez modifier l'histogramme pour les voir). Ces valeurs très élevées pour quelques observations seulement ont un effet significatif sur la moyenne mais peu d'effet, voire aucun, sur la médiane ; cela signifie que, dans cet exemple, la médiane est un meilleur indicateur de la tendance centrale.

Graphs

1- Les Fondamentaux

Vous pouvez créer et modifier des types de diagramme divers et variés. Dans les exemples suivants, nous allons créer et modifier trois types de diagramme fréquemment utilisés :

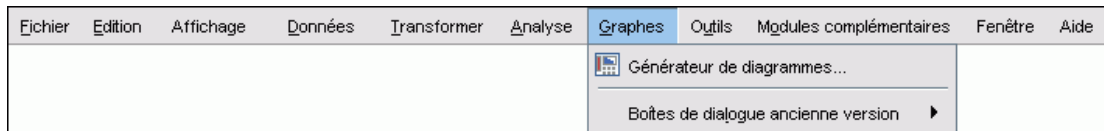
- Diagramme en bâtons simples
- Diagramme en secteurs
- Diagramme de dispersion contenant des groupes

2- Quelques types de graphs

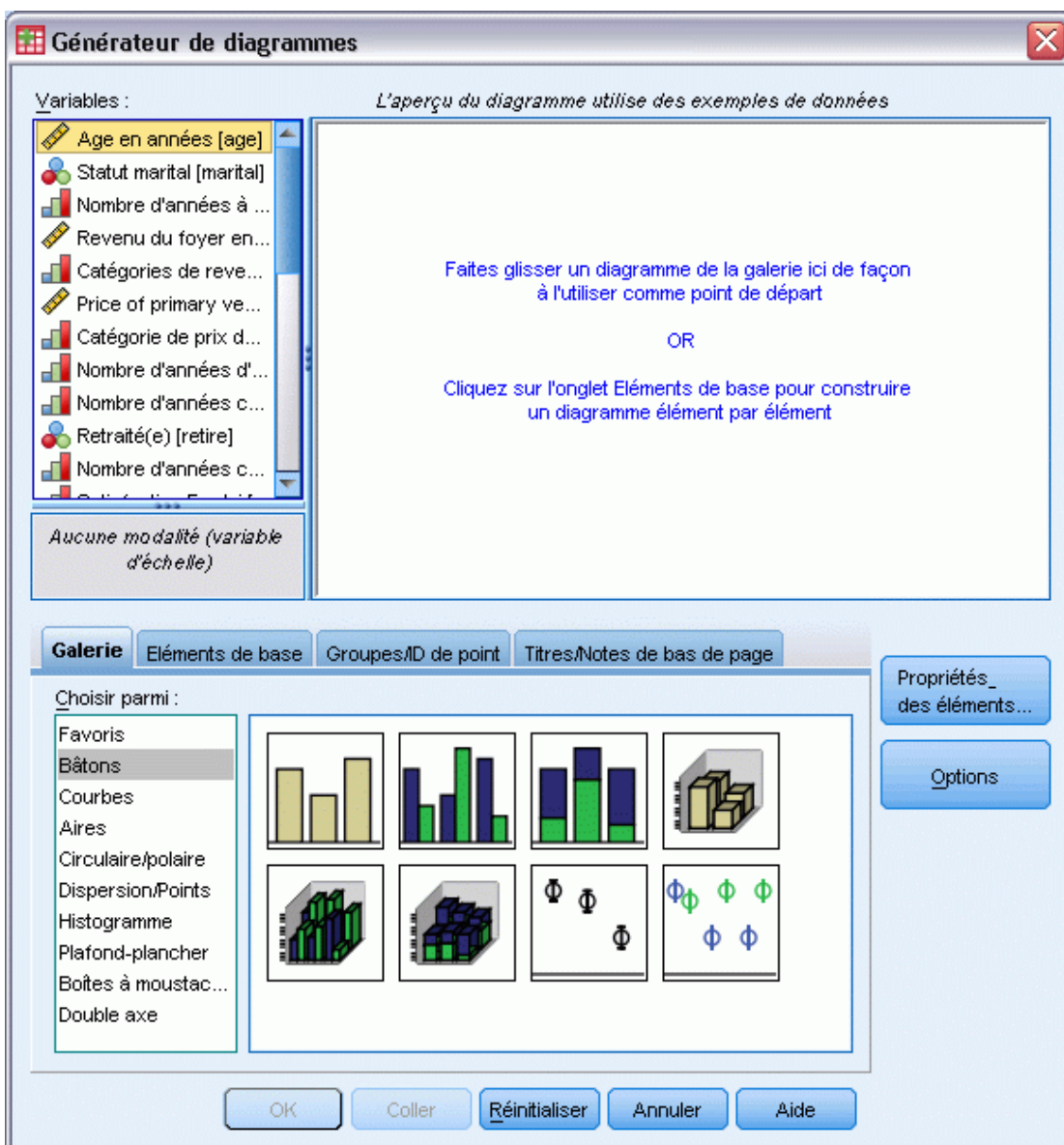
Pour illustrer les notions de base de la création de diagrammes, nous allons créer un diagramme en bâtons du revenu moyen pour plusieurs niveaux de satisfaction professionnelle. Cet exemple utilise le fichier de données demo.sav.

► A partir des menus, sélectionnez :

Graphiques > Générateur de diagrammes...



La
boîte
de



dialogue Générateur de diagrammes est une fenêtre interactive qui vous permet d'obtenir l'aperçu d'un diagramme avant que vous ne le génériez.

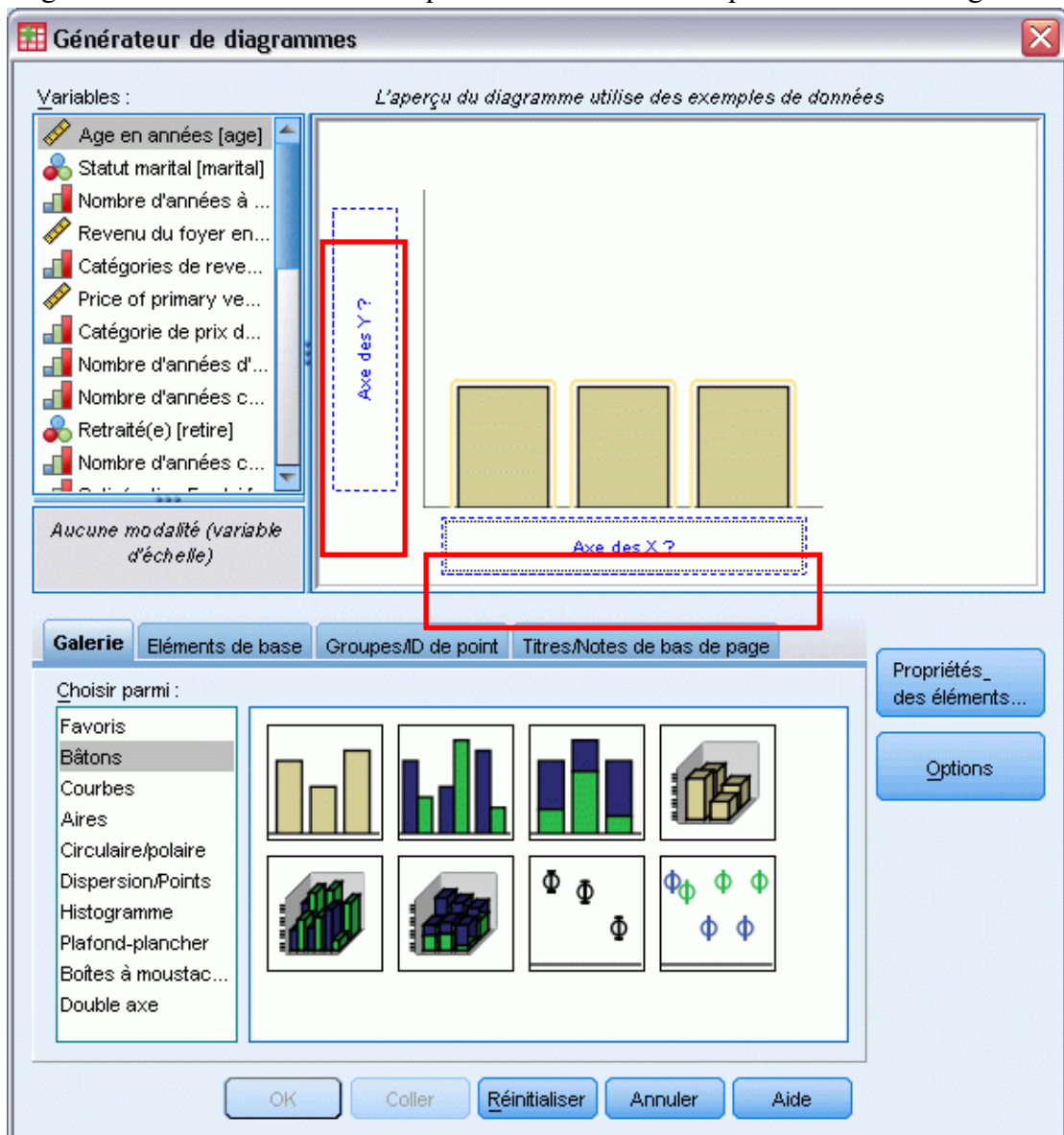
► Cliquez sur l'onglet Galerie s'il n'est pas sélectionné.

La galerie inclut plusieurs diagrammes différents prédéfinis, qui sont organisés par type de diagramme. L'onglet Eléments de base fournit également des éléments de base (comme les axes et les éléments graphiques) pour créer des diagrammes en partant de zéro, mais il est plus facile d'utiliser la galerie.

► Cliquez sur Bâton s'il n'est pas sélectionné.

Les icônes représentant les diagrammes en bâtons disponibles dans la galerie apparaissent dans la boîte de dialogue. Les images doivent fournir suffisamment d'informations pour identifier le type de diagramme spécifique.

► Faites glisser l'icône du diagramme en bâtons simples sur le « canevas », qui est en fait la zone étendue au-dessus de la galerie. Le Générateur de diagrammes affiche un aperçu du diagramme sur le canevas. Notez que les données utilisées pour dessiner le diagramme ne sont



pas

vos données en cours. Il s'agit d'exemples.

Bien qu'il y ait un diagramme sur le canevas, il n'est pas complet car il n'y a aucune variable ou statistique pour contrôler la hauteur des bâtons et pour spécifier la catégorie de variable correspondant à chaque bâton. Vous ne pouvez pas avoir de diagramme sans variable et statistique.

Vous pouvez ajouter des variables en les glissant de la liste Variables qui se trouve à gauche du canevas.

Lorsque vous glissez les variables, les cibles sont les « zone de déplacement » du canevas. Certaines zones de déplacement requièrent une variable, d'autres non.

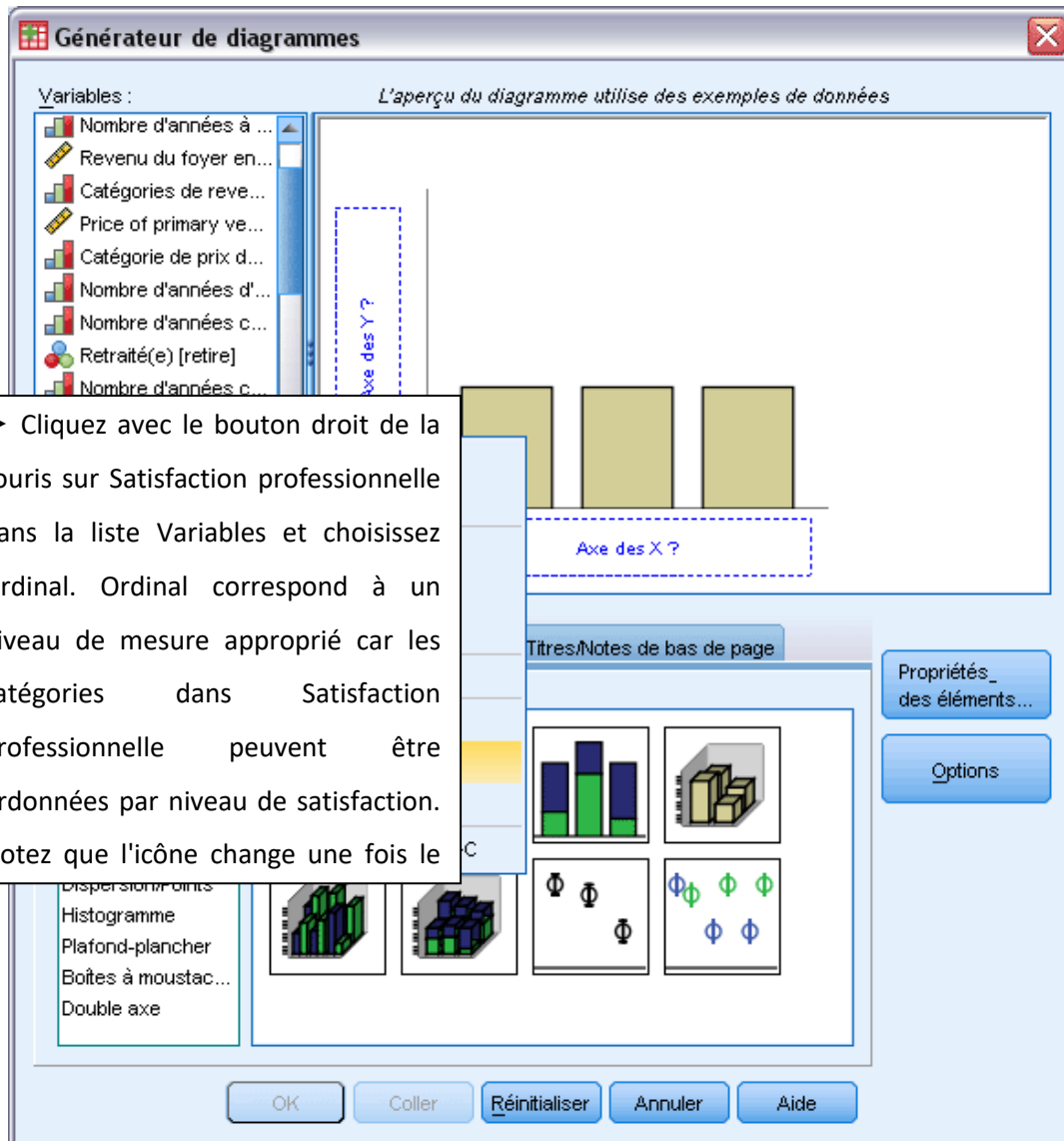
En fonction du type de diagramme que vous créez, vous aurez peut-être besoin d'une variable dans la zone de déplacement de l'axe y. Par exemple, lorsque vous voulez afficher une statistique récapitulative d'une autre variable (comme une moyenne de salaire), vous avez besoin d'une variable dans la zone de déplacement de l'axe y. Les diagrammes de dispersion requièrent également une variable dans l'axe y. Dans ce cas, la zone de déplacement identifie la variable dépendante.

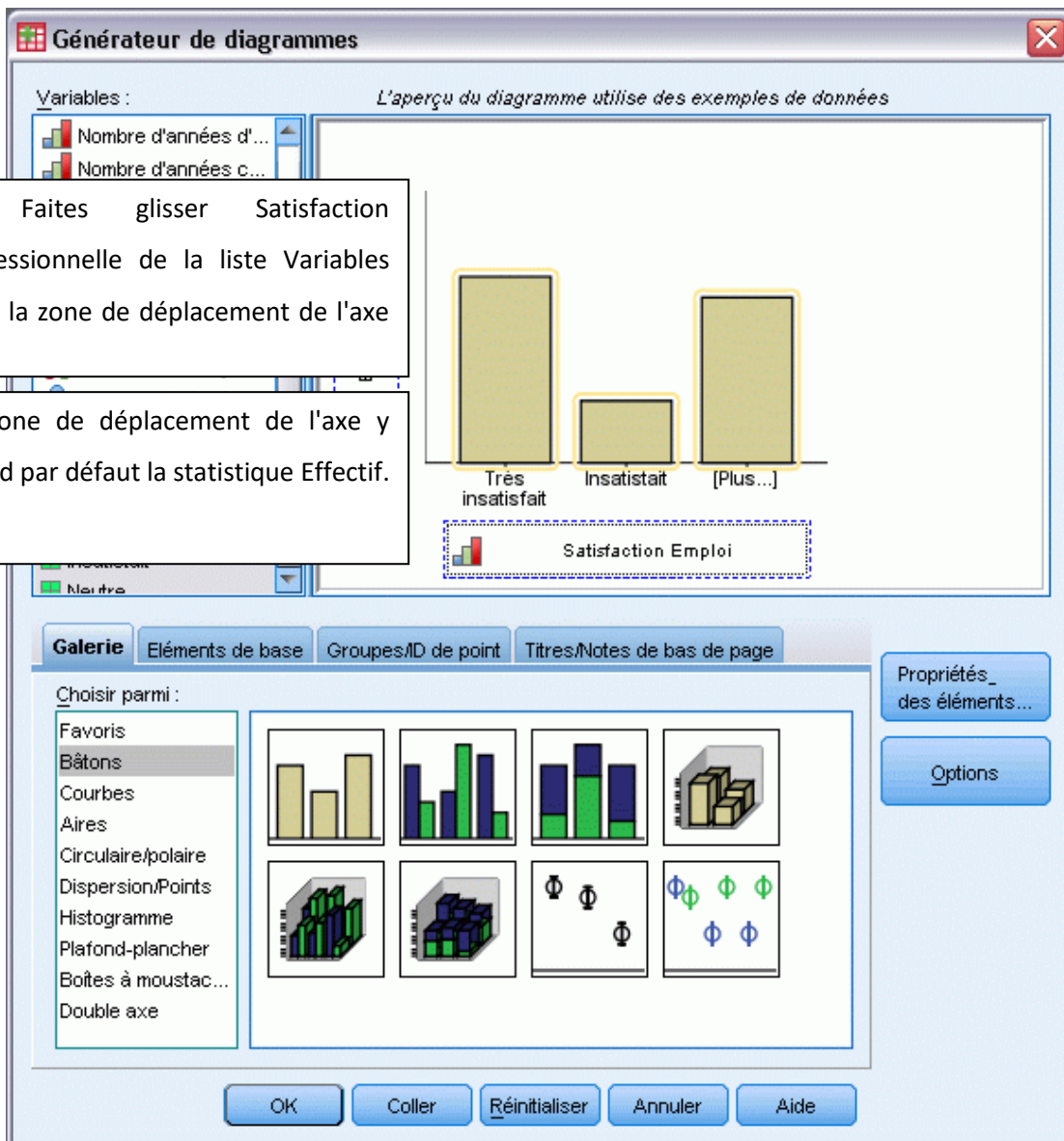
La zone de déplacement de l'axe x est obligatoire. La variable de cette zone de déplacement contrôle l'apparition des bâtons sur l'axe x.

Vous allez créer un diagramme qui affiche des bâtons relatifs au salaire moyen de chaque catégorie de satisfaction professionnelle, les deux zones de déplacement sont donc nécessaires. Il y aura une variable qualitative sur l'axe x et une variable d'échelle sur l'axe y pour calculer la moyenne.

Le niveau de mesure d'une variable est important dans le Générateur de diagrammes. Vous allez utiliser la variable Satisfaction professionnelle de l'axe x. Cependant, l'icône (qui ressemble à une règle) à côté de la variable indique que son niveau de mesure est défini en tant que variable d'échelle. Pour créer le diagramme correct, vous devez utiliser un niveau de mesure qualitatif. Plutôt que de revenir et de modifier le niveau de mesure dans l'Affichage des variables, vous

pouvez le modifier temporairement dans le Générateur de diagrammes.



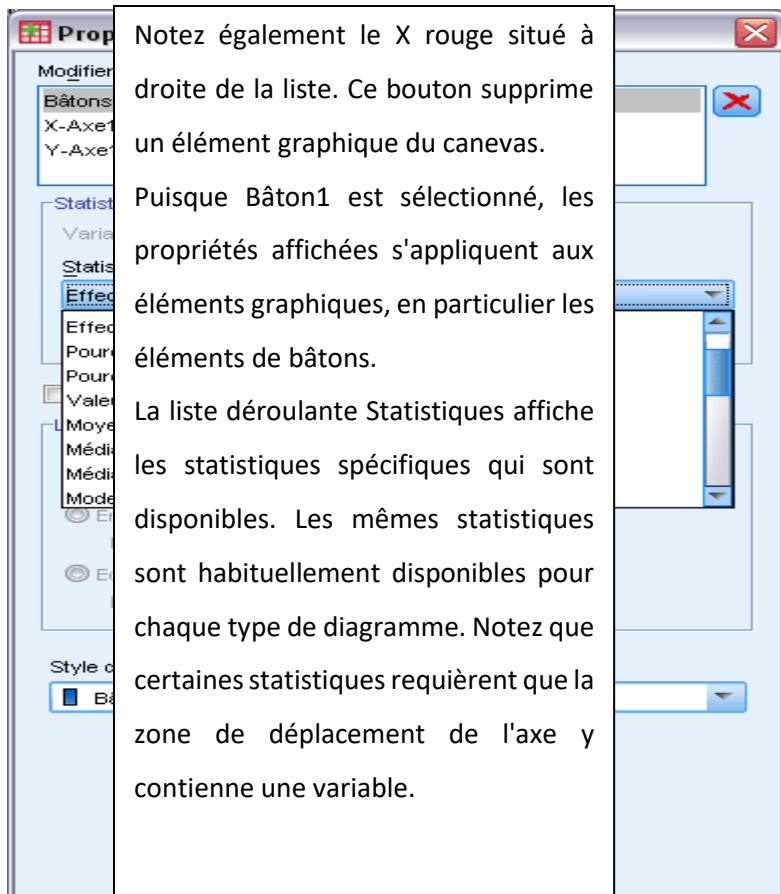


Si vous souhaitez utiliser une autre statistique (comme pourcentage ou moyenne),

vous pouvez facilement en changer. Vous n'utiliserez aucune de ces statistiques dans cet exemple, mais nous allons revoir le processus au cas où vous devez changer cette statistique plus tard.

► Cliquez sur Propriété des éléments pour afficher la fenêtre Propriété des éléments.

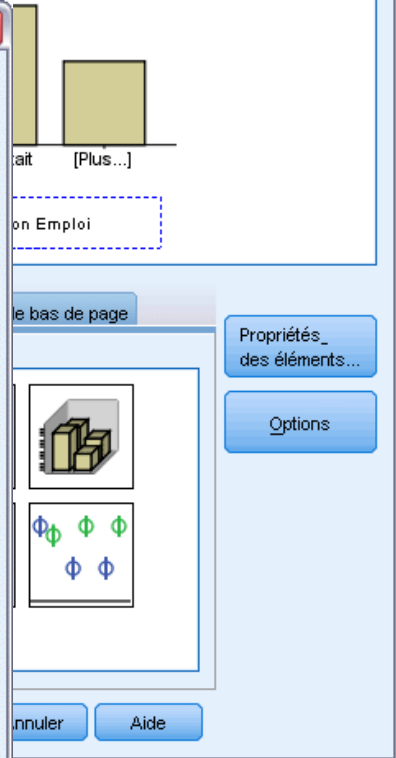
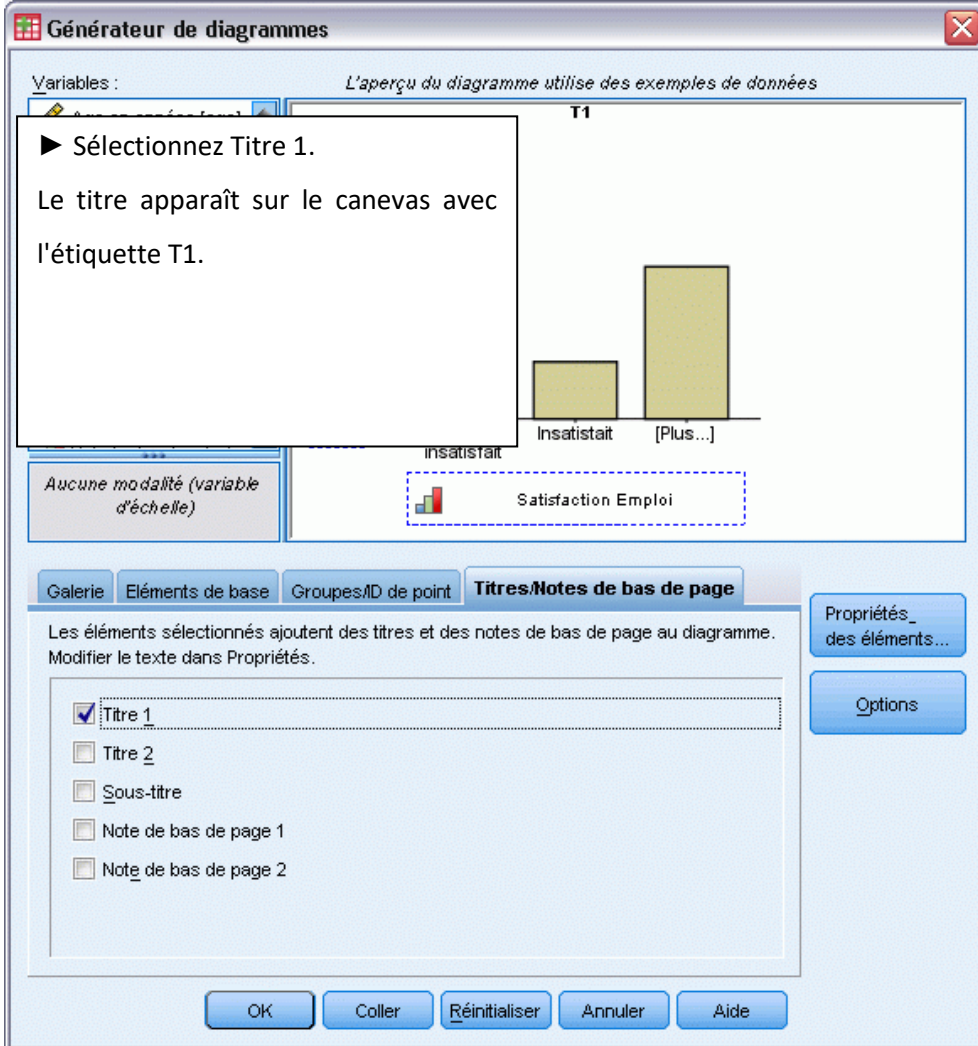
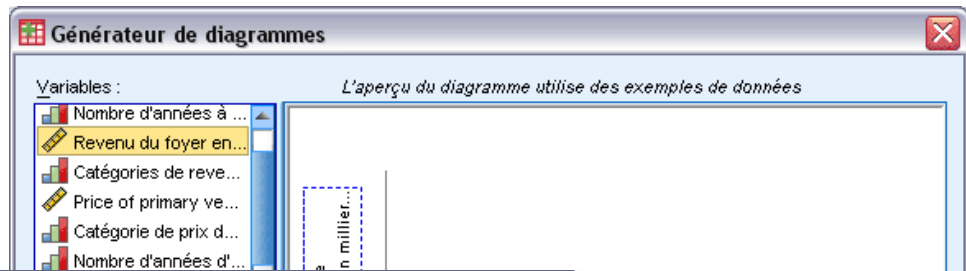
La fenêtre Propriété des éléments vous permet de modifier les propriétés des différents éléments du diagramme. Ces éléments comprennent les éléments graphiques (comme les bâtons du diagramme) et les axes du diagramme. Sélectionnez un des éléments dans Modifier les propriétés de liste pour modifier les propriétés associées à cet élément.



Revenez à la boîte de dialogue Générateur de diagrammes et faites glisser Revenu du ménage en milliers de la liste Variables vers la zone de déplacement de l'axe y.

Puisque la variable sur l'axe y est sous forme d'échelle et que la variable de l'axe x est qualitative (ordinal est un type de niveau de mesure qualitatif), la zone de déplacement de l'axe y prend par défaut la statistique Moyenne. Il s'agit des variables et des statistiques souhaitées, il n'y a donc aucun besoin de modifier les propriétés de l'élément.

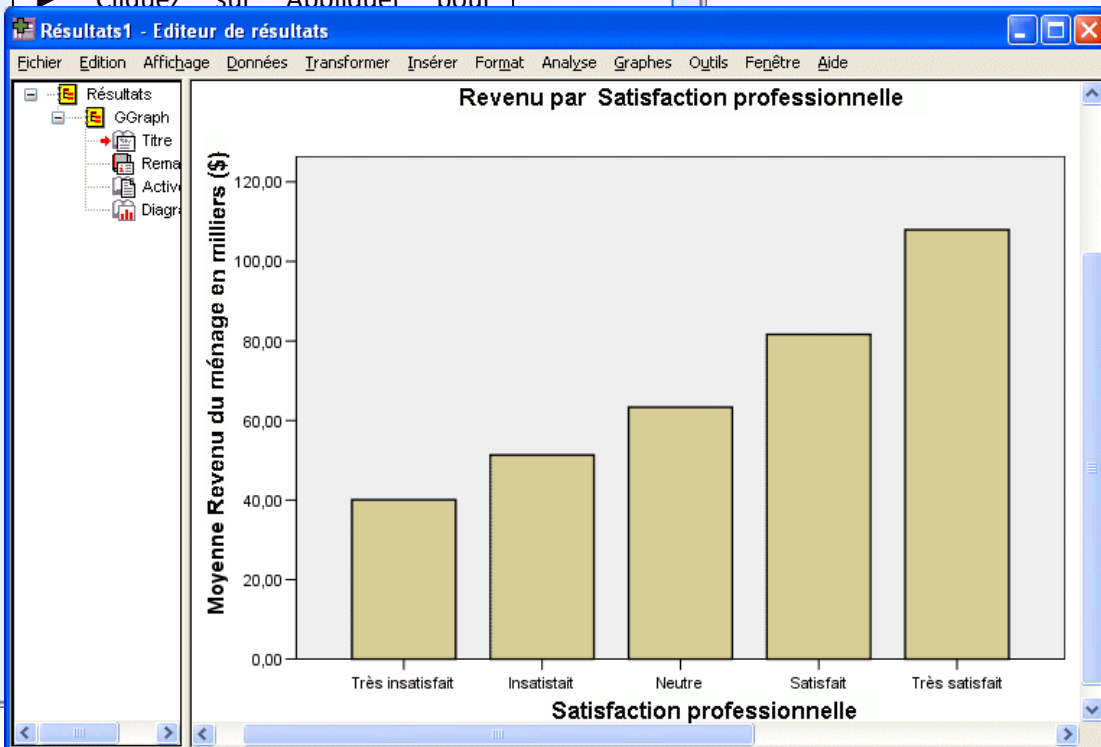
Vous pouvez également ajouter des titres et des notes de bas de page au



► Dans la fenêtre Propriété des éléments, sélectionnez Titre 1 dans Modifier les propriétés de liste.

► Dans la zone de texte Contenu, saisissez Revenu par satisfaction professionnelle. Il s'agit du texte que le titre affichera.

► Cliquez sur Appliquer pour



Le diagramme en bâtons indique que les plus satisfaits de satisfaction professionnelle ont à avoir des revenus plus

Modification de diagrammes – Notions de base

Vous pouvez modifier les diagrammes de différentes façons. Pour l'exemple de diagramme en bâtons créé, vous allez effectuer les tâches suivantes :

- Changement de couleur.
- Formatage des valeurs des étiquettes de graduation.
- Modification du texte.

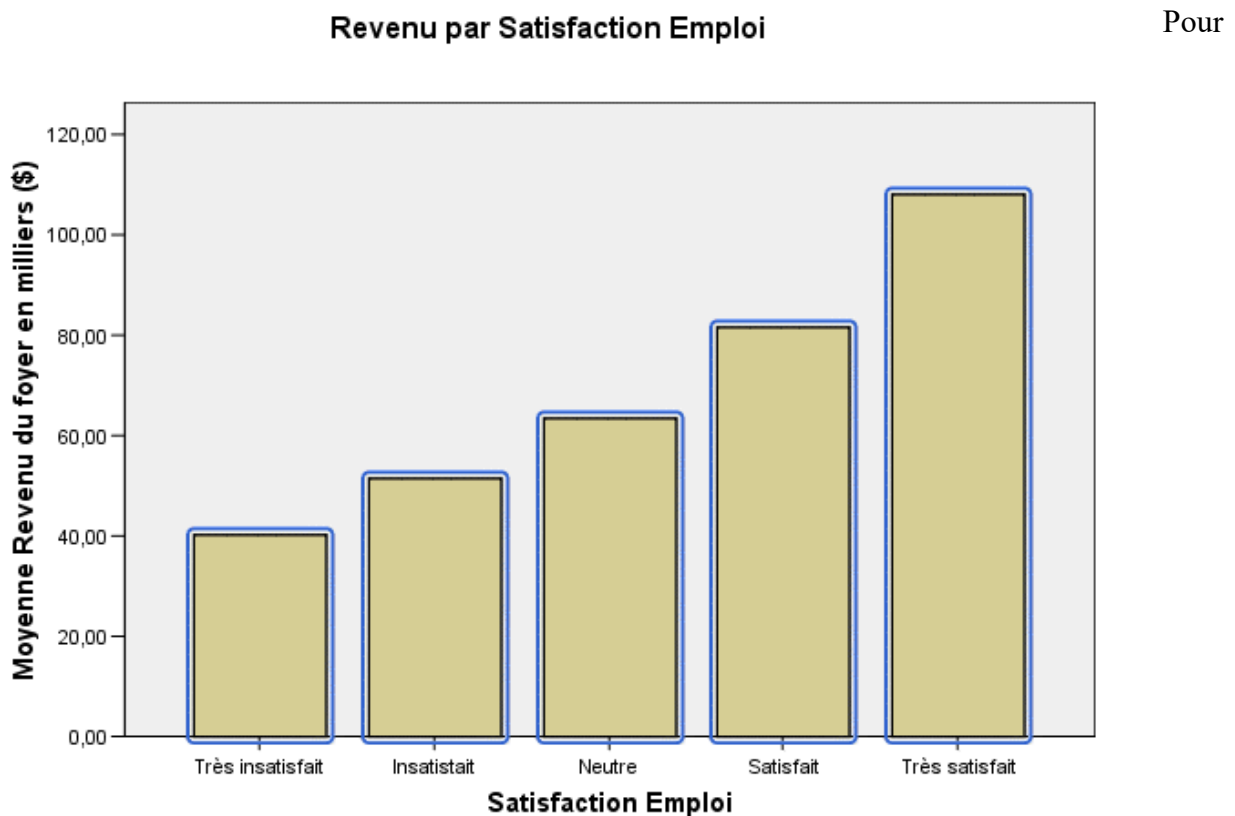
- Affichage des étiquettes des valeurs de données.
- Utilisation de modèles de diagramme.

Pour modifier le diagramme, ouvrez-le dans l'éditeur de diagrammes.

► Double-cliquez sur le diagramme en bâtons pour l'ouvrir dans l'éditeur de diagrammes.

Pour modifier un élément de diagramme, sélectionnez-le.

► Cliquez sur l'un des bâtons. Les rectangles contenant les bâtons signifient qu'ils sont sélectionnés.



sélectionner un élément d'un groupe, faites défiler l'affichage pour cliquer dessus lorsque le groupe est sélectionné.

► Cliquez sur le bâton correspondant aux répondants très satisfaits. Seul ce bâton reste sélectionné.

Pour sélectionner un autre bâton, il vous suffit de cliquer dessus.

► Cliquez sur le bâton correspondant aux répondants relativement satisfaits. Désormais, seul ce bâton est sélectionné.

Pour sélectionner plusieurs éléments, cliquez dessus tout en maintenant la touche Ctrl enfoncée.

► Si le bâton correspondant aux répondants relativement satisfaits est sélectionné, appuyez sur la touche Ctrl et cliquez sur le bâton des répondants très satisfaits. Ces deux bâtons sont désormais sélectionnés.

Cet exemple illustre une règle générale de hiérarchisation vers le bas applicable aux diagrammes simples :

- Si aucun des éléments graphiques n'est sélectionné, cliquez sur l'un d'eux afin de les sélectionner tous.
- Si tous les éléments graphiques sont sélectionnés, cliquez sur l'un d'eux afin que la sélection porte sur lui seul. Pour sélectionner un autre élément graphique, vous pouvez cliquer dessus. Pour sélectionner plusieurs éléments graphiques, cliquez dessus tout en maintenant la touche Ctrl enfoncée.

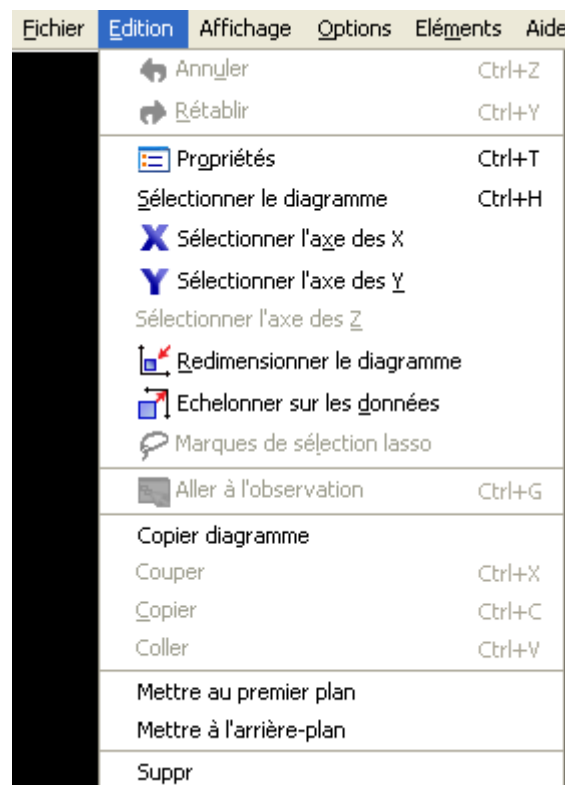
Remarque : Cette procédure est légèrement différente pour les diagrammes regroupés. Les diagrammes regroupés sont traités dans l'exemple de diagramme de dispersion.

► Pour désélectionner tous les éléments, appuyez sur la touche Echap.

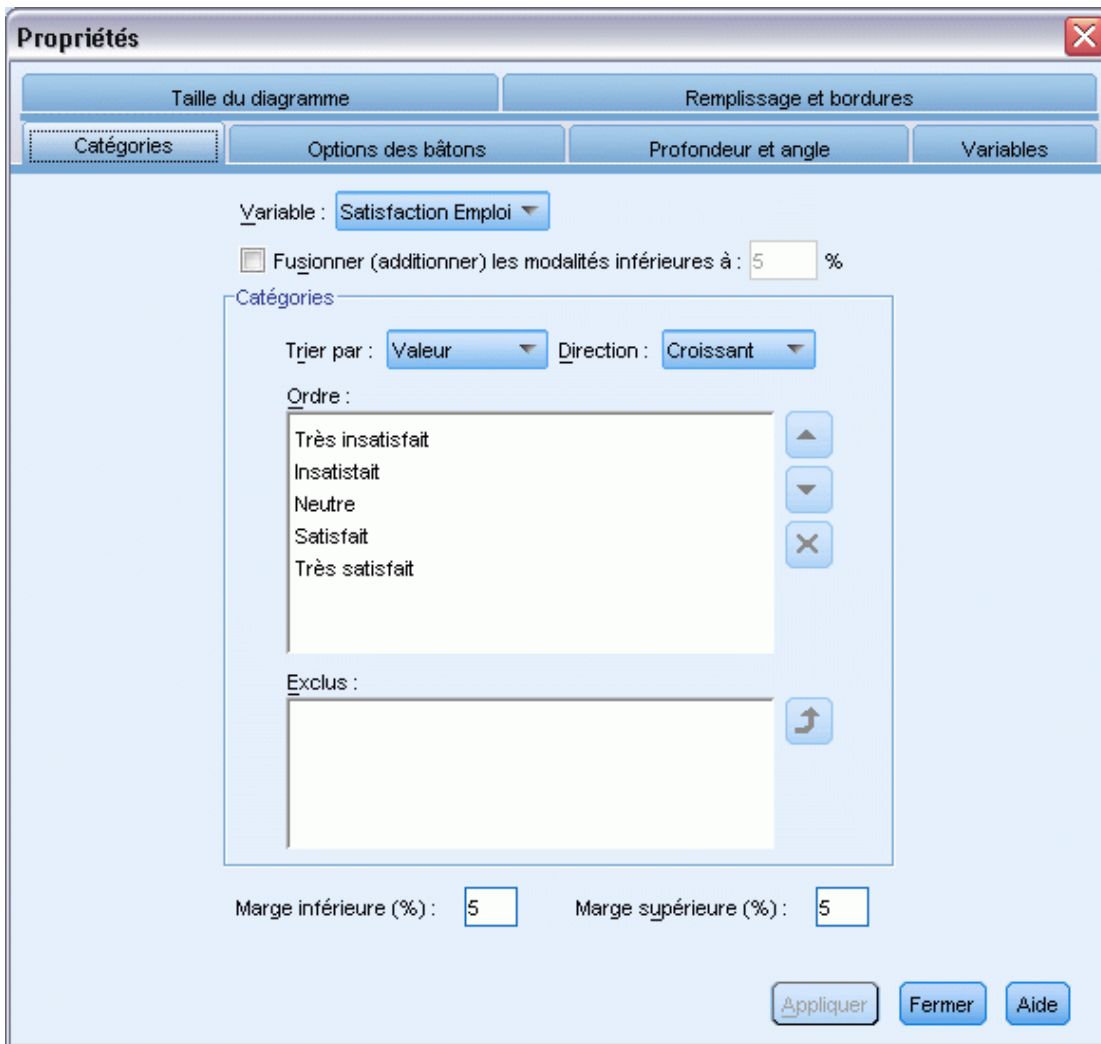
► Cliquez sur un bâton pour resélectionner tous les bâtons.

► A partir des menus de l'éditeur de diagrammes, sélectionnez :

Edition > Propriétés



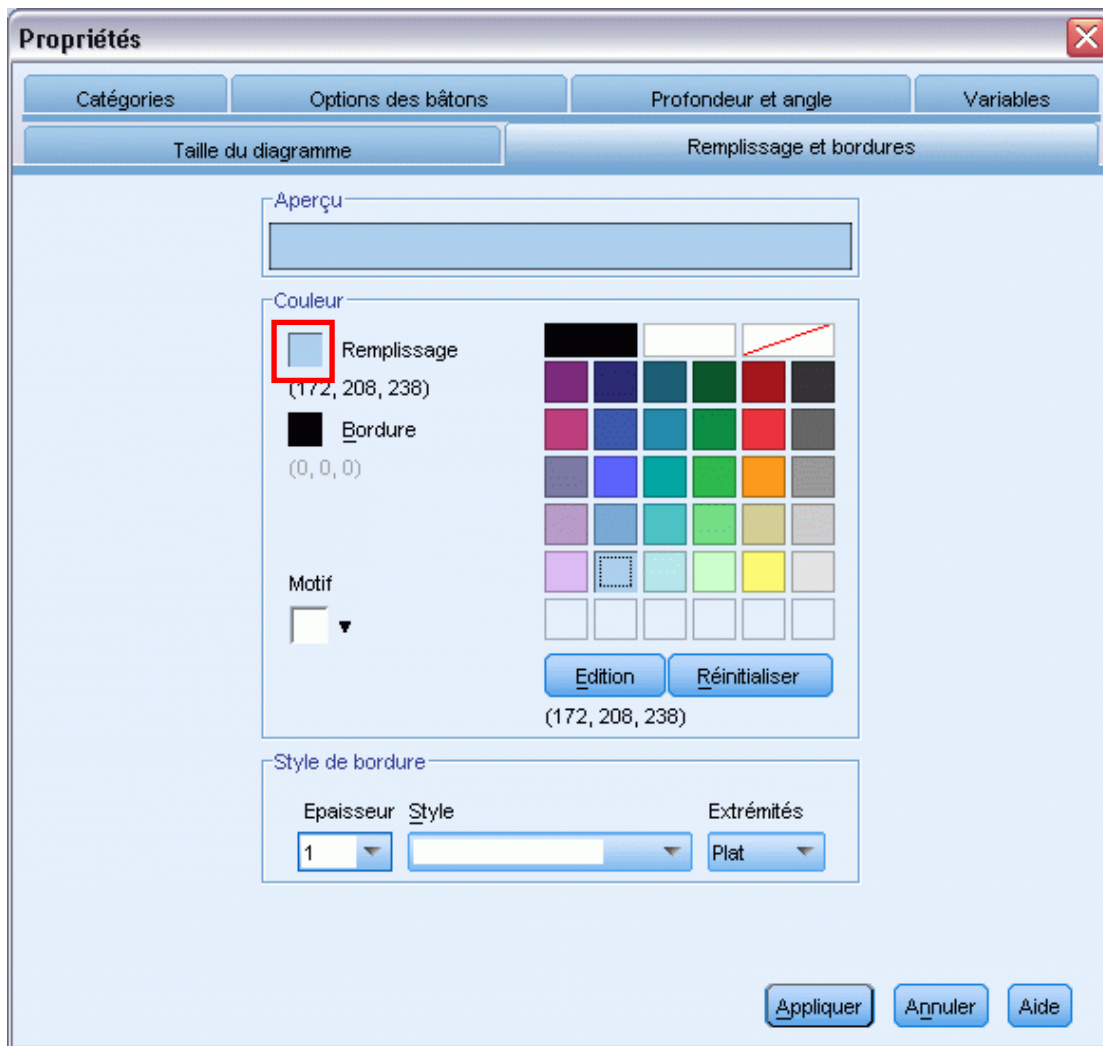
Ce menu affiche la fenêtre Propriétés, qui contient les onglets s'appliquant aux bâtons sélectionnés. Ces onglets varient selon l'élément de diagramme que vous sélectionnez dans l'éditeur de diagrammes. Par exemple, si vous aviez sélectionné un cadre de texte au lieu de bâtons, plusieurs onglets apparaissent dans la fenêtre Propriétés. Ces onglets vous permettent d'effectuer la plupart des modifications de diagramme.



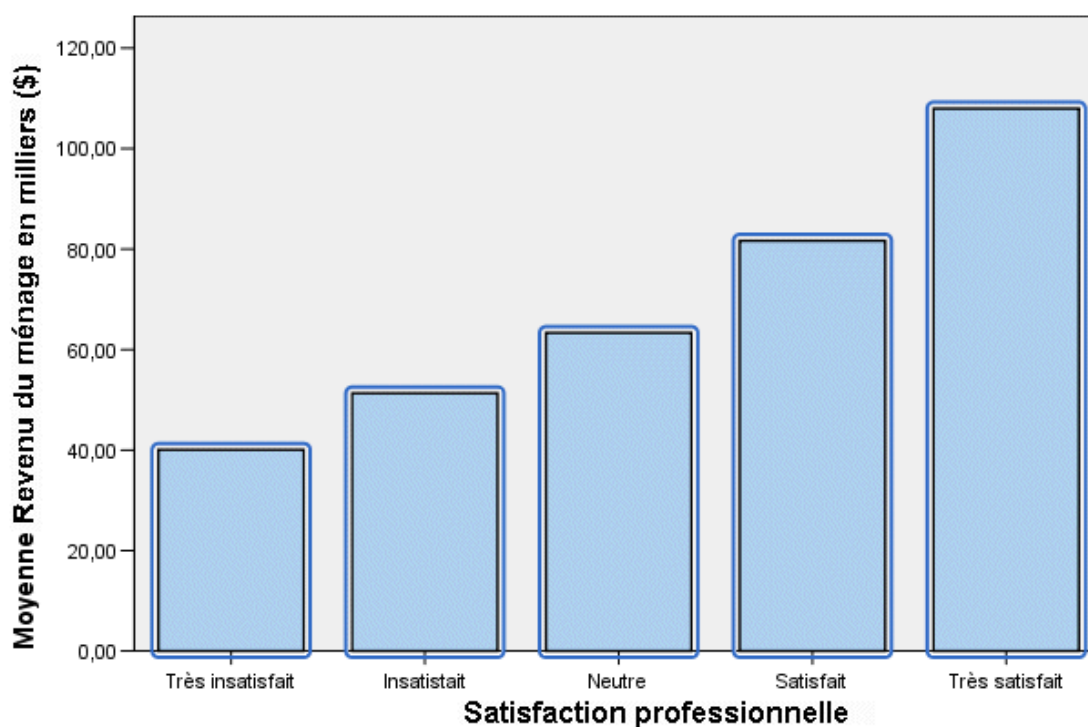
Vous allez d'abord changer la couleur des bâtons. En outre, vous

spécifiez les attributs de couleur des éléments graphiques (à l'exception des courbes et des marques) dans l'onglet de remplissage et de bordures.

- ▶ Cliquez sur l'onglet Remplissage et bordures.
- ▶ Cliquez sur l'échantillon en regard de l'option de remplissage pour indiquer que vous souhaitez modifier la couleur de remplissage des bâtons. Les numéros figurant sous l'échantillon font référence aux composantes rouge, verte et bleue de la couleur actuelle.



Revenu par Satisfaction professionnelle



Notez que les valeurs sur l'axe des y sont exprimées en milliers. Pour améliorer la présentation du diagramme et faciliter son interprétation, nous allons modifier le format numérique des étiquettes de graduation, puis le titre de l'axe en conséquence.

- ▶ Sélectionnez les étiquettes de graduation sur l'axe des y en cliquant dessus.
- ▶ Pour rouvrir la fenêtre Propriétés (si vous l'avez fermée précédemment), sélectionnez les options suivantes :

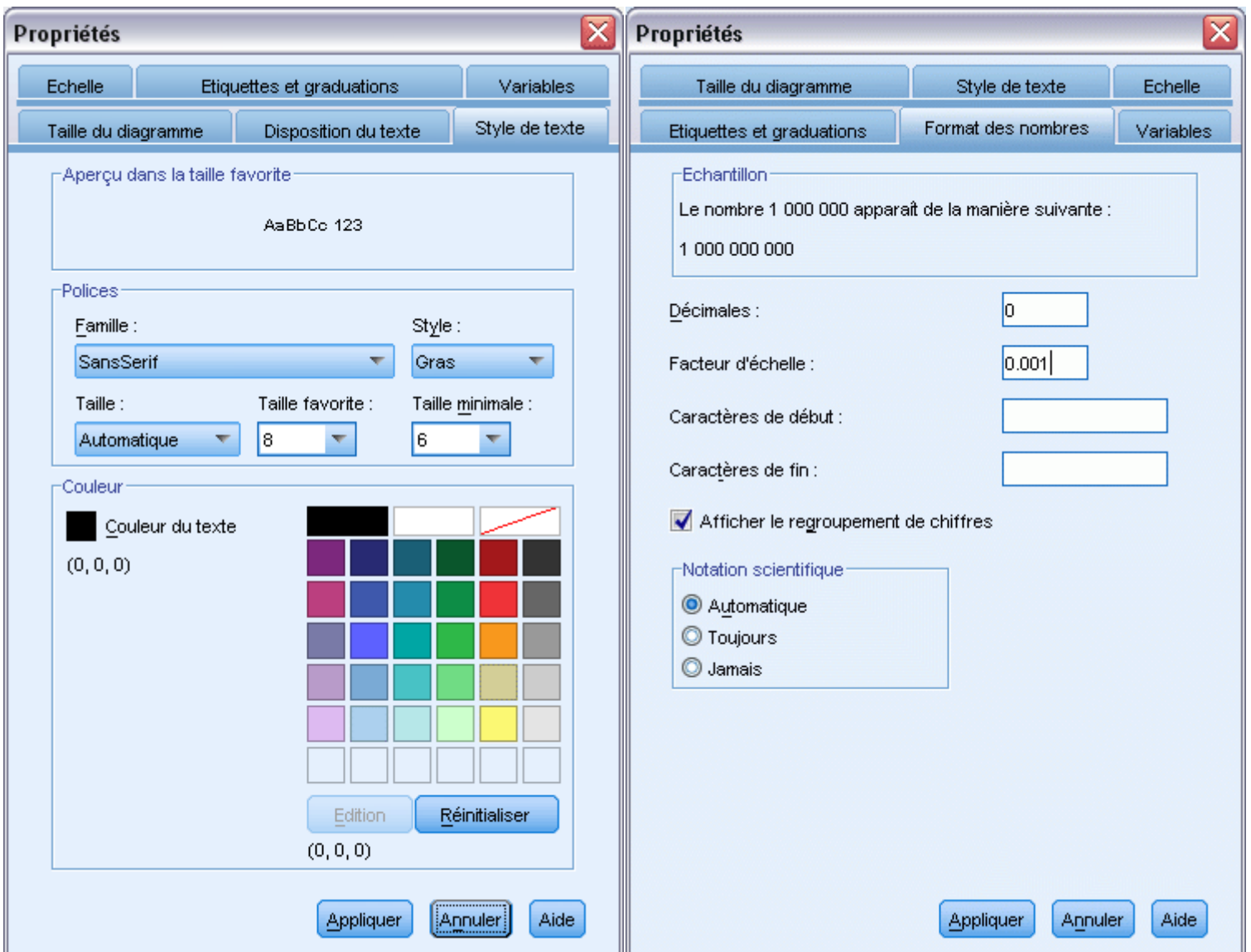
Edition > Propriétés

Remarque : A ce stade, nous supposons que la fenêtre Propriétés est ouverte. Si vous avez fermé la fenêtre Propriétés, suivez l'étape précédente pour la rouvrir. Il est également possible d'utiliser le raccourci clavier Ctrl+T afin de rouvrir cette fenêtre.

Plusieurs onglets sont disponibles maintenant que les étiquettes de graduation sont sélectionnées, et non les bâtons

- ▶ Cliquez sur l'onglet Format numérique.

► Si vous ne souhaitez pas afficher les décimales sur les étiquettes de graduation, entrez 0 dans la zone de texte Décimales.

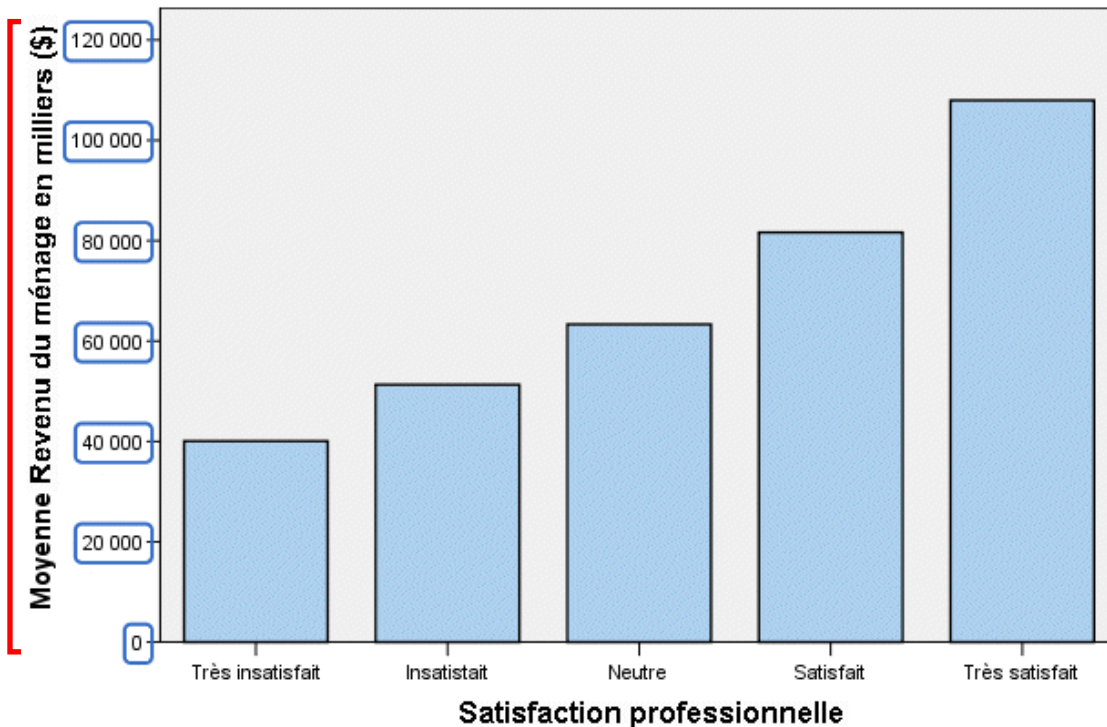


► Saisissez 0,001 dans la zone de texte Facteur d'échelle. Le facteur d'échelle correspond au diviseur de la valeur affichée, utilisé par l'éditeur de diagrammes. Etant donné que 0,001 représente une fraction, employer ce diviseur incrémente de 1 000 les valeurs des étiquettes de graduation. Par conséquent, elles ne sont plus exprimées en milliers et leur mise à l'échelle est annulée.

► Sélectionnez Afficher le regroupement de chiffres. La fonction de regroupement des chiffres utilise un caractère spécial (fourni par les paramètres régionaux de votre ordinateur) pour marquer la position des milliers dans la valeur concernée.

Les

Revenu par Satisfaction professionnelle



étiquettes de graduation reflètent ce nouveau formatage numérique : Aucune décimale n'apparaît, les valeurs ne sont plus mises à l'échelle et les milliers sont signalés par un caractère spécial.

Remarque : Vous n'avez pas besoin d'ouvrir la fenêtre Propriétés pour modifier le texte. Vous pouvez modifier du texte directement à l'intérieur des diagrammes.

- Cliquez sur le titre de l'axe des y pour le sélectionner.
- Cliquez à nouveau dessus afin d'activer le mode d'édition. Une fois en mode d'édition, l'éditeur de diagrammes positionne horizontalement le texte après rotation. Il affiche également un curseur en forme de barre rouge qui clignote (non illustré dans cet exemple).

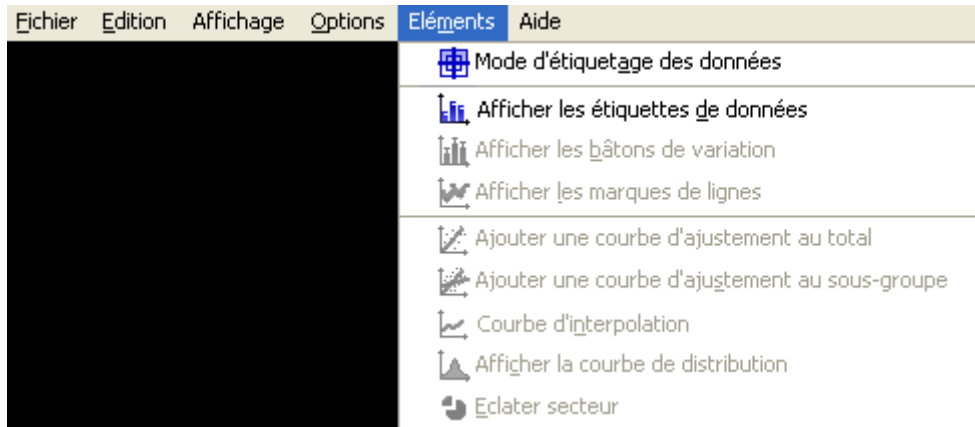
► Supprimez le texte suivant :
en milliers

► Appuyez sur Entrée pour désactiver le mode d'édition et mettre à jour le titre de l'axe. Ce titre décrit désormais avec précision le contenu des étiquettes de graduation.

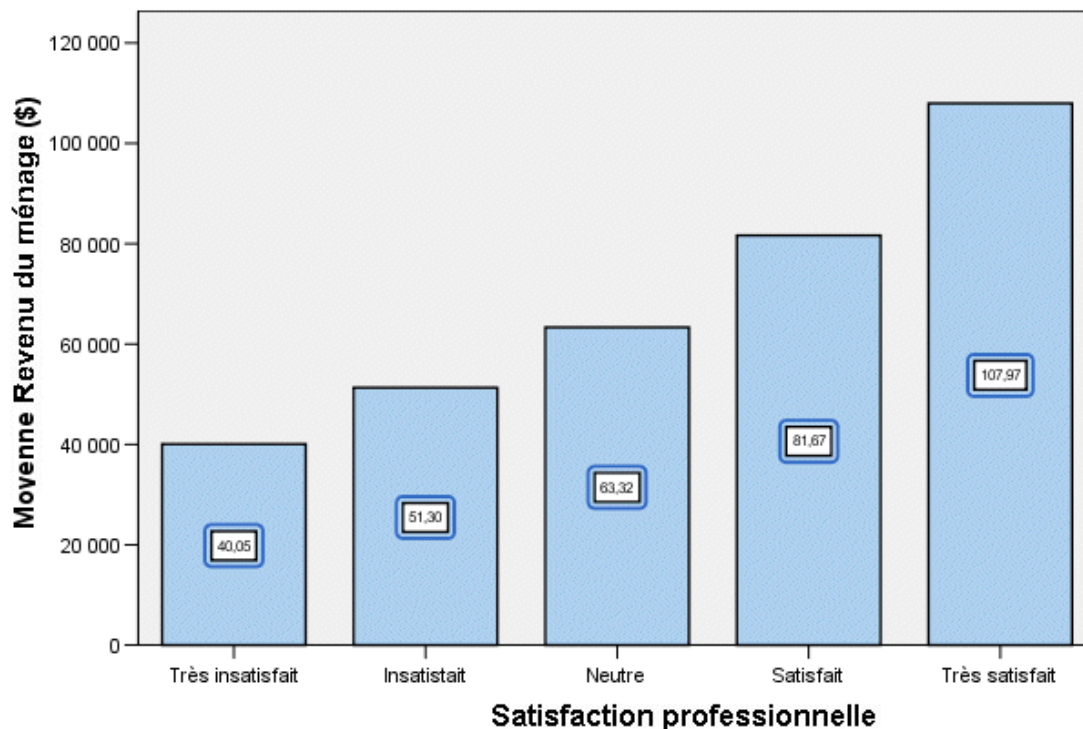
Une autre tâche courante consiste à afficher les valeurs précises associées aux éléments graphiques (sous forme de bâtons dans cet exemple). Ces valeurs apparaissent dans les étiquettes de données.

► A partir des menus de l'éditeur de diagrammes, sélectionnez :

Eléments > Afficher les étiquettes de données



Revenu par Satisfaction professionnelle



Chaque bâton du diagramme indique désormais le revenu moyen exact du ménage. Notez que les unités sont en milliers. Par conséquent, vous pouvez réutiliser l'onglet Format numérique pour modifier le facteur d'échelle.

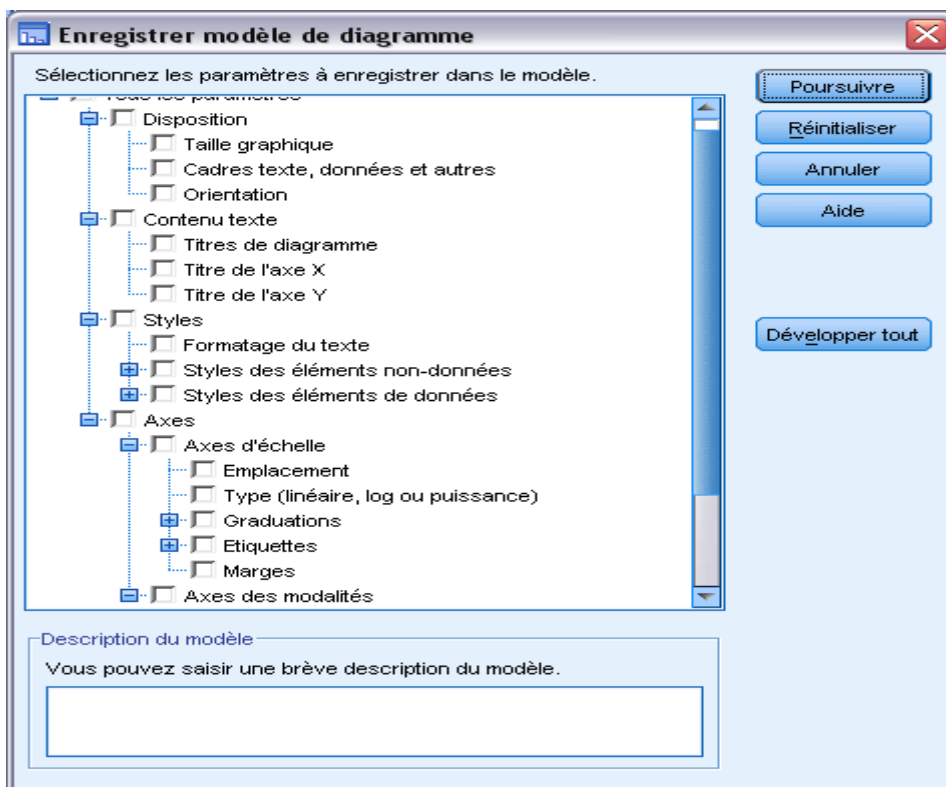
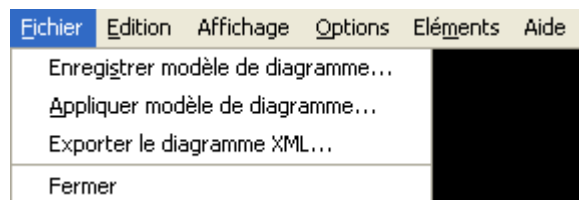
Si vous modifiez régulièrement vos diagrammes, vous pouvez utiliser un modèle de diagramme pour réduire le temps nécessaire à leur création et à leur modification. Un modèle de diagramme enregistre les attributs d'un diagramme spécifique. Vous pouvez par la suite appliquer ce modèle lorsque vous créez ou modifiez un diagramme.

Nous allons enregistrer le diagramme actuel en tant que modèle, puis appliquer ce modèle lors de la création d'un nouveau diagramme

► A partir des menus, sélectionnez :

Fichier > Enregistrer modèle de diagramme...

La boîte de dialogue Enregistrer modèle de diagramme vous permet de spécifier les attributs de diagramme à ajouter au modèle.

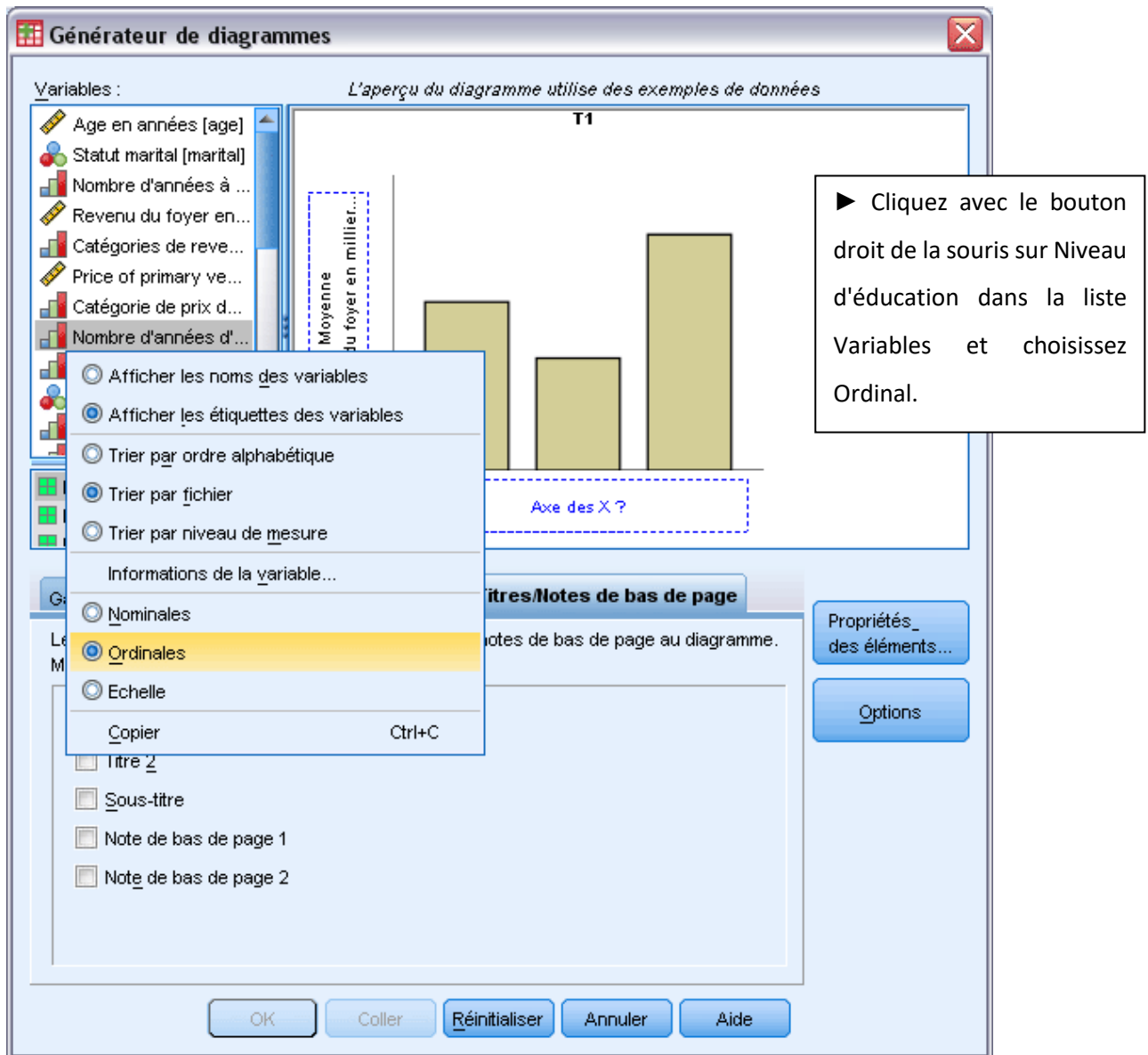


Vous pouvez appliquer le modèle quand vous créez un diagramme ou dans l'éditeur de

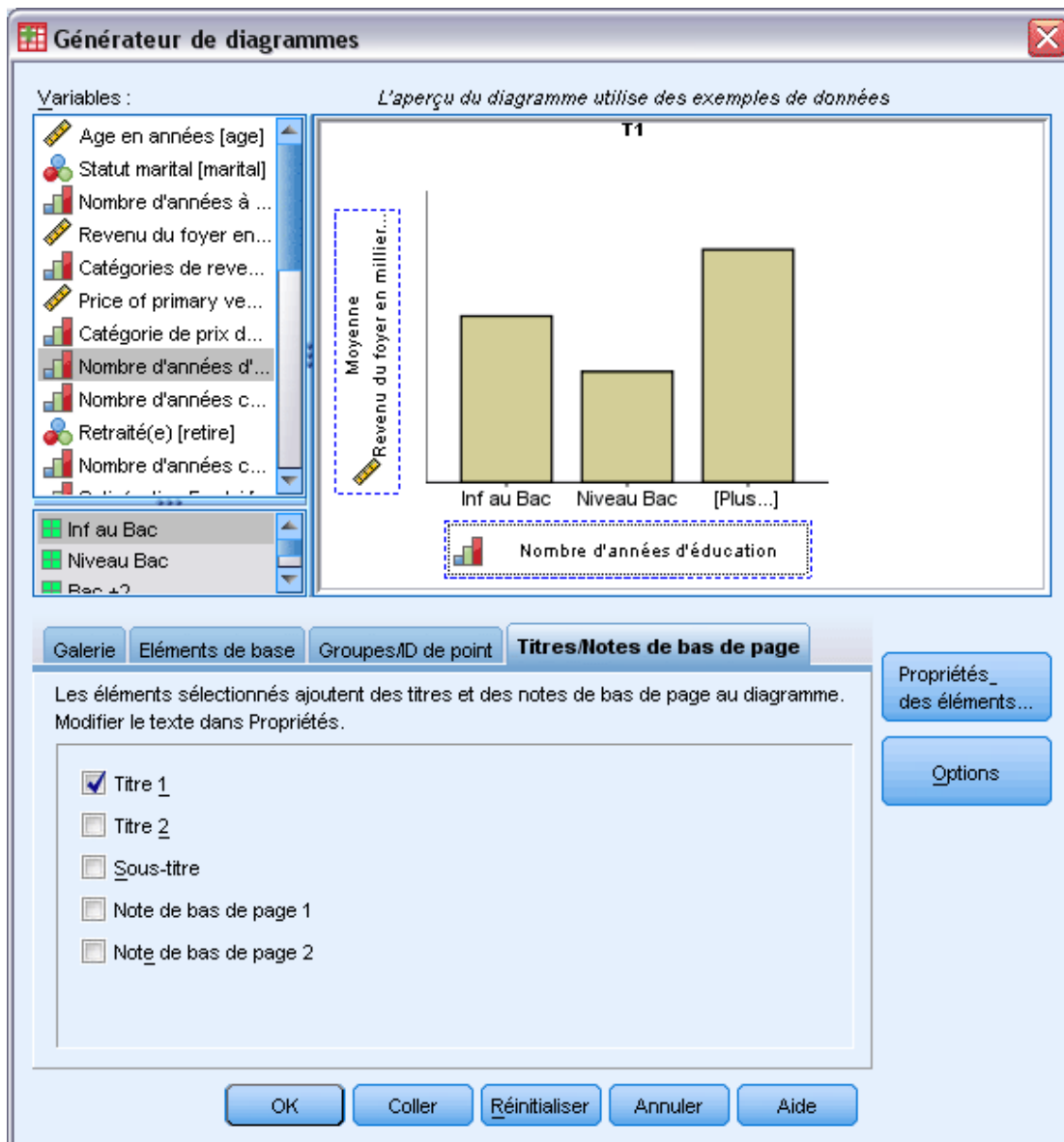
diagrammes. Dans l'exemple suivant, nous allons l'appliquer durant la création d'un diagramme.

La boîte de dialogue Générateur de diagrammes « mémorise » les variables entrées lors de la création du premier diagramme. Toutefois, dans le cas présent, vous allez créer un diagramme légèrement différent pour voir comment l'application d'un modèle formate un diagramme.

► Supprimez Satisfaction professionnelle de l'axe x en le glissant de la zone de déplacement vers la liste Variables. Vous pouvez également cliquer sur la zone de déplacement et appuyer sur Supprimer.



► Faites glisser Niveau d'éducation de la liste Variables vers la zone de déplacement de l'axe x.

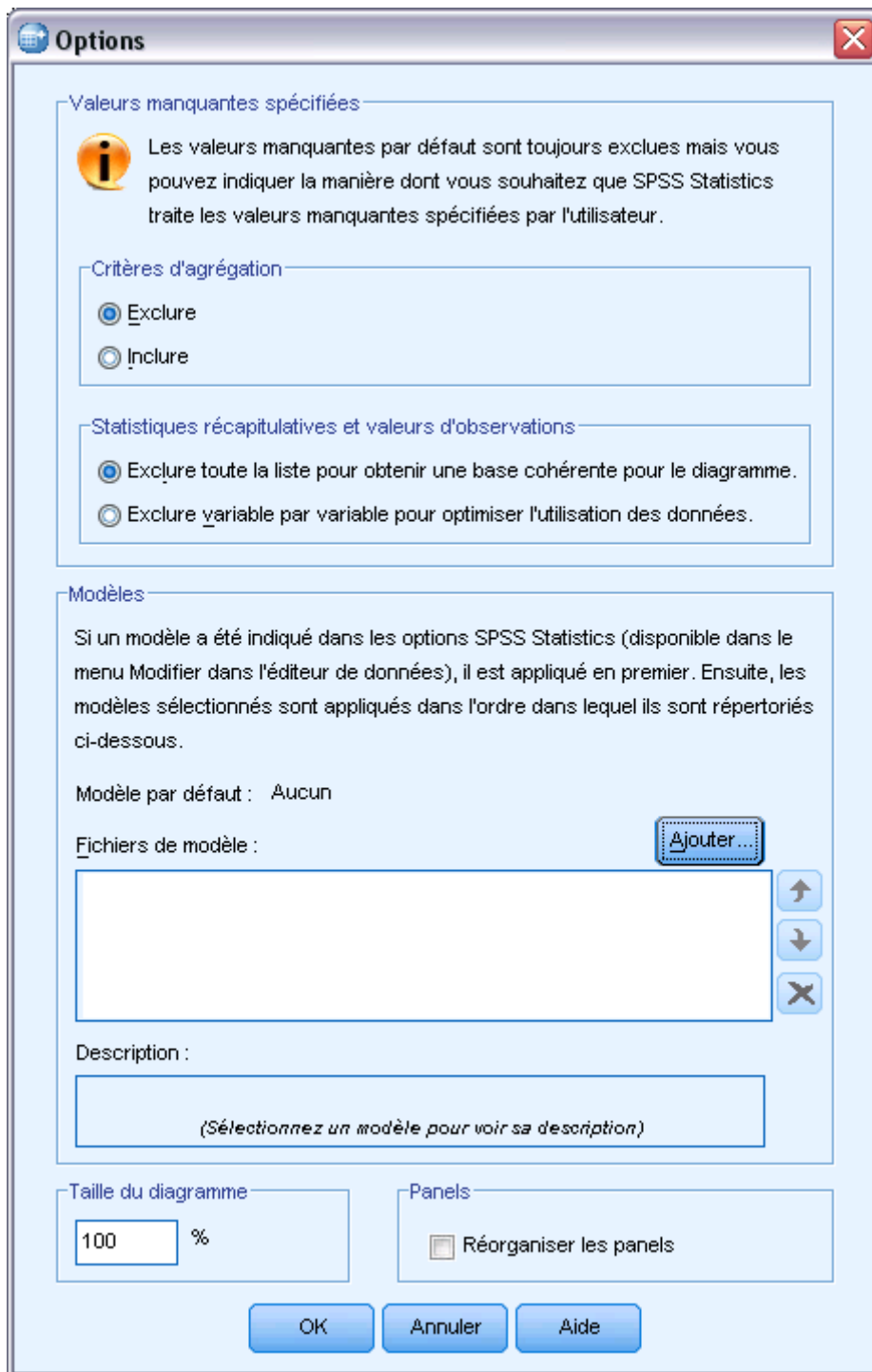


Puisque le titre n'est plus pertinent, nous allons le supprimer.

► Dans l'onglet Titres/Notes de bas de page, désélectionnez Titre 1.

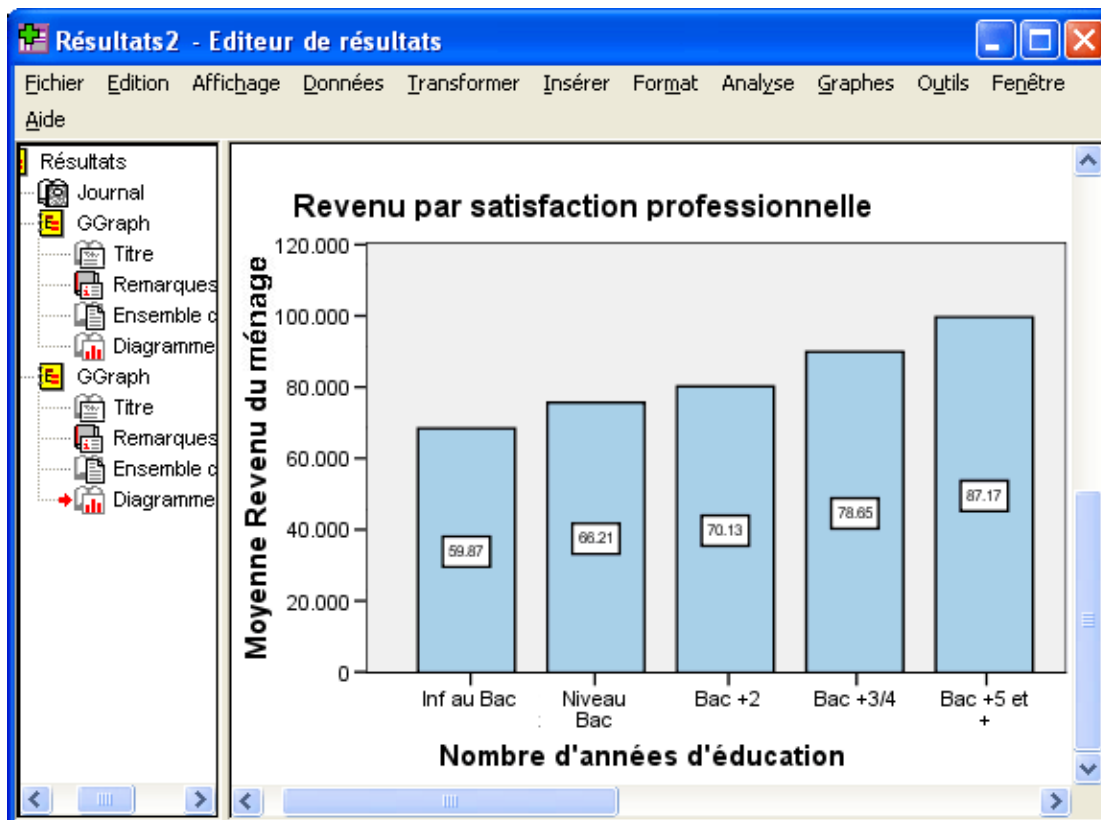
Nous allons maintenant indiquer le modèle à appliquer au nouveau diagramme.

► Cliquez sur Options



- ▶ Dans le groupe Modèles de la boîte de dialogue Options, cliquez sur Ajouter.
- ▶ Dans la boîte de dialogue Trouver les fichiers du modèle, repérez le fichier du modèle préalablement enregistré à l'aide de la boîte de dialogue Enregistrer modèle de diagramme.
- ▶ Sélectionnez ce fichier et cliquez sur Ouvrir.
- ▶ Cliquez sur OK pour fermer la boîte de dialogue Options.

Exemple



Le formatage du nouveau diagramme correspond à celui du diagramme que vous avez créé et modifié précédemment. Même si les variables sur l'axe des x sont différentes, les diagrammes, eux, sont semblables. Notez que le titre du diagramme précédent a été conservé dans le modèle, même si vous avez supprimé le titre dans le Générateur de diagrammes.

Si vous souhaitez appliquer des modèles après avoir créé un diagramme, vous pouvez le faire dans Chart Editor (choisissez l'option Appliquer modèle de diagramme du menu Fichier).

Conclusion

L'analyse des données est une discipline fondamentale qui joue un rôle essentiel dans la compréhension, l'interprétation et la prise de décisions basées sur des données. Au cours de ce cours, nous avons exploré un large éventail de concepts, techniques et méthodes utilisés pour analyser et tirer des informations significatives à partir de données. Voici quelques points clés que nous pouvons retenir :

1. Collecte de Données : L'analyse des données commence par la collecte de données pertinentes. Une collecte de données précise et appropriée est cruciale pour garantir que les résultats de l'analyse sont significatifs.
2. Nettoyage et Prétraitement : Les données brutes nécessitent souvent un nettoyage et un prétraitement pour éliminer les valeurs aberrantes, les valeurs manquantes et les erreurs, afin de garantir la qualité des données.
3. Exploration des Données : L'exploration des données, y compris la visualisation, permet de découvrir des tendances, des modèles, des corrélations et des informations cachées dans les données.
4. Statistiques Descriptives : Les statistiques descriptives, telles que la moyenne, la médiane, l'écart type et les quartiles, fournissent des résumés numériques des données.
5. Inferential Statistics : Les statistiques inférentielles permettent de tirer des conclusions générales à partir d'un échantillon de données en utilisant des tests d'hypothèses et des intervalles de confiance.
6. Modélisation Statistique : La modélisation statistique comprend des techniques telles que la régression linéaire, la régression logistique, l'ANOVA et l'analyse de séries temporelles, qui permettent de modéliser les relations entre les variables.
7. Apprentissage Automatique : L'apprentissage automatique est une sous-discipline de l'analyse des données qui se concentre sur la construction de modèles prédictifs à partir des données, en utilisant des algorithmes d'apprentissage supervisé et non supervisé.
8. Visualisation des Données : La visualisation des données est un outil puissant pour communiquer les résultats de l'analyse de données de manière claire et efficace.
9. Éthique des Données : L'analyse des données doit être menée de manière éthique, en respectant la confidentialité, la sécurité et la légalité des données.
10. Applications dans Divers Domaines : L'analyse des données est utilisée dans divers domaines, notamment la science des données, la recherche, les affaires, la médecine, la finance, la météorologie, l'économie, etc.

En fin de compte, l'analyse des données est une discipline en constante évolution, car de nouvelles techniques et technologies émergent continuellement. Ce cours vous a fourni des bases solides pour explorer davantage ce domaine passionnant et pour appliquer vos connaissances à des problèmes réels. L'analyse des données est un outil puissant pour découvrir des informations, prendre des décisions éclairées et résoudre des problèmes complexes, et elle continuera à jouer un rôle essentiel dans un large éventail de domaines professionnels.

Référence bibliographique

1. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis*. Cengage Learning.
2. Everitt, B. S., & Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer.
3. Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson.
4. Rencher, A. C., & Christensen, W. F. (2012). *Methods of Multivariate Analysis*. Wiley.
5. Stevens, J. P. (2012). *Applied Multivariate Statistics for the Social Sciences*. Routledge.
6. Tabachnick, B. G., & Fidell, L. S. (2019). *Using Multivariate Statistics*. Pearson.
7. Everitt, B. S. (2009). *The Cambridge Dictionary of Statistics*. Cambridge University Press.
8. Sheskin, D. J. (2010). *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press.
9. Wilcox, R. R. (2017). *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press.
10. Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.
11. Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer.
12. Todorov, V., & Filzmoser, P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, 32(3), 1-47.
13. Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern Classification*. Wiley.
14. Johnson, R. A., & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall.
15. Flury, B. (1997). *A First Course in Multivariate Statistics*. Springer.
16. Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. Wiley.
17. Greenacre, M. J. (2017). *Correspondence Analysis in Practice*. CRC Press.
18. Agresti, A., & Finlay, B. (1997). *Statistical Methods for the Social Sciences*. Upper Saddle River, Prentice Hall.
19. Bachelard, G. (1967). *La formation de l'esprit scientifique*. Paris, Vrin.
20. Baker, L.-R. (1995). *Explaining Attitudes*. Cambridge, Cambridge University Press.

21. Baillargeon, G. (2004). Méthodes statistiques avec application en gestion, production, marketing, relations industrielles et comptables. Trois-Rivières, SMG.
22. Benzecri, J.-P., et al. (1976). L'analyse des données. Tome I : La taxinomie. Tome II: L'analyse des correspondances. Paris, Dunod.
23. Besnier, J.-M. (1996). Les théories de la connaissance. Paris, Flammarion, coll. « Dominos », no 105.
24. Besson, J.-L. (1992). « Les statistiques : vraies ou fausses ». Autrement, no 5.
25. Boudon, R. (1971). Les mathématiques en sociologie. Paris, Presses universitaires de France.
26. Boudon, R. (1990). L'art de se persuader. Paris, Fayard.
27. Kennedy, P. (2003). A Guide to Econometrics. Cambridge, MA, MIT Press.
28. Kim, J.O., & Mueller, C. (1978). Introduction to Factor Analysis. Beverly Hills, Sage University Paper no 13.
29. Kinnear, P., & Gray, C. (2000). SPSS for Windows Made Simple. Hove (R.-U.), Psychology Press.
30. Reimann, C., & Filzmoser, P. (2000). Normal and Lognormal Data Transformation Methods for Geochemical Data. Environmental Science & Technology, 34(20), 4283-4290.