République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Constantine 03
Salah Boubnider
Faculté de Médecine
Département de Médecine

Polycopié du Cours et Travaux Pratiques

Introduction à l'Analyse des Données Biomédicales

-Biostatistique-informatique-

Destiné aux Étudiants de 1ère Année de Médecine

Rédigé par : Dr. Abdennour Boulesnane

Année Universitaire : 2022-2023

Avant-propos

Cher(e)s Étudiant(e)s,

Il me fait plaisir de vous présenter ce polycopié dédié au module de Biostatistique-informatique pour les étudiants de première année de médecine. Dans ce module, vous trouverez une introduction complète à l'analyse de données biomédicales et son traitement statistique à l'aide d'Excel et SPSS. Le contenu de ce module est hautement pertinent pour les défis scientifiques actuels en médecine et en médecine computationnelle. L'utilisation efficace des données médicales et l'application de méthodes statistiques et computationnelles sont essentielles pour avancer dans le diagnostic et le traitement des maladies. C'est pourquoi, ce polycopié est conçu pour vous fournir une base solide en statistique et informatique, qui est essentielle pour la recherche en médecine, afin de prendre des décisions éclairées et de mener des recherches de qualité.

Le but de ce polycopié est de vous offrir un support pédagogique clair et complet pour faciliter votre apprentissage. Chaque chapitre commence par un résumé des points clés à retenir, suivi d'une explication détaillée des concepts abordés et de nombreux exemples pour vous aider à bien comprendre les notions présentées. Les travaux pratiques proposés vous permettront de mettre en pratique ces notions et d'acquérir les compétences nécessaires pour réussir dans notre discipline.

Dans les sept chapitres suivants, vous apprendrez à collecter vos données, à les prétraiter en utilisant différentes techniques, à effectuer des analyses statistiques avancées telles que l'analyse de relations entre variables, et à utiliser les distributions de probabilité avec SPSS. Chaque chapitre est accompagné d'exemples pratiques pour vous aider à mieux comprendre les concepts et les techniques abordés.

Nous espérons que ce polycopié vous offrira une expérience d'apprentissage enrichissante et stimulante. Nous vous souhaitons à tous une excellente lecture et une réussite éclatante dans vos études !

Cordialement,

Dr. Abdennour Boulesnane
Dernière mise à jour le : 19 février 2024

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Analyse des Données Biomédicales

1.1 Introduction

L'analyse des données biomédicales est importante, car elle permet aux chercheurs d'identifier les modèles de maladies, de développer des modèles prédictifs, de faciliter le développement de médicaments, de permettre une médecine personnalisée et d'intégrer des données provenant de plusieurs sources pour fournir une image plus complète de la santé des patients. Grâce à ces applications, l'analyse des données biomédicales joue un rôle crucial dans l'avancement de notre compréhension des maladies et dans le développement de nouveaux traitements et thérapies pour améliorer les résultats pour les patients.

1.2 Science des Données

1.2.1 Définition

La science des données est un domaine interdisciplinaire qui implique l'extraction, le traitement, l'analyse, la visualisation et l'interprétation d'ensembles de données volumineux et complexes. Il utilise une combinaison de méthodes statistiques et informatiques pour découvrir des modèles et des informations à partir de données qui peuvent éclairer la prise de décision et fournir un avantage concurrentiel.

La science des données implique une gamme de compétences, y compris l'analyse statistique, l'apprentissage automatique, la visualisation des données et la gestion des données. Cela implique souvent de travailler avec des ensembles de données volumineux et divers provenant de diverses sources, notamment les médias sociaux, les transactions de commerce électronique, les capteurs, les dossiers médicaux et les données financières.

L'objectif de la science des données est de transformer les données en informations exploitables qui peuvent éclairer la prise de décision, améliorer l'efficacité et fournir un avantage concurrentiel. Les scientifiques des données utilisent

une gamme d'outils et de techniques pour atteindre cet objectif, notamment le nettoyage et le prétraitement des données, l'analyse exploratoire des données, l'ingénierie des fonctionnalités, la sélection et la validation des modèles et la visualisation des données.

La science des données a un large éventail d'applications dans tous les secteurs, notamment la santé, la finance, le marketing, les médias sociaux et le commerce électronique. Parmi les exemples d'applications de science des données, citons la détection des fraudes, les systèmes de recommandation, le marketing personnalisé, la maintenance prédictive et le diagnostic médical.

1.2.2 Méthodologie de la Science des Données

La méthodologie de la science des données est une approche systématique de résolution de problèmes basés sur les données qui implique plusieurs étapes clés (voir Figure 1.1) :

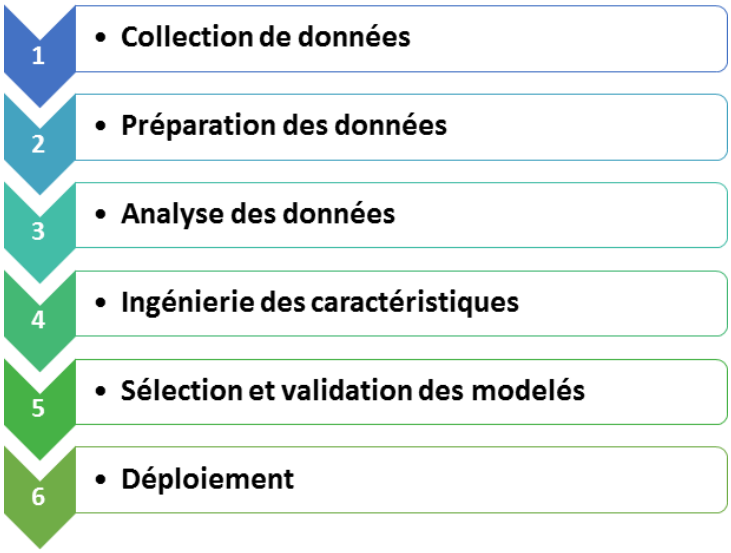


FIGURE 1.1 – Organigramme de la méthodologie de la science des données

1. **Collection de données** : Cela implique d'identifier les sources de données pertinentes, de collecter les données et de les stocker dans un format structuré.
2. **Préparation des données** : Cela implique le nettoyage, la transformation et le prétraitement des données pour les rendre aptes à l'analyse.
3. **L'analyse et L'analyse exploratoire des données** : L'analyse exploratoire des données (EDA) et l'analyse des données sont deux méthodes liées, mais distinctes d'analyse des données.
   L'EDA est une étape préliminaire dans l'analyse des données qui implique l'utilisation de diverses techniques statistiques et de visualisation pour explorer et comprendre les données. L'objectif de l'EDA est de découvrir des modèles, des tendances et des anomalies dans les données qui peuvent éclairer le processus d'analyse des données. L'EDA implique des techniques telles que les histogrammes, les boîtes à moustaches, les diagrammes de dispersion et les matrices de corrélation pour explorer

visuellement les données et identifier les principales caractéristiques et relations.
L'analyse des données, en revanche, est un processus plus formel et structuré d'analyse des données qui implique l'application de techniques statistiques et d'apprentissage automatique aux données. Le but de l'analyse des données est d'obtenir des informations et de faire des prédictions à partir des données en utilisant une méthodologie bien définie. L'analyse des données implique des techniques telles que les tests d'hypothèses, l'analyse de régression, la classification et le regroupement pour modéliser et analyser les données.
Bien que l'EDA et l'analyse des données soient des processus distincts, ils sont souvent utilisés ensemble dans le flux de travail d'analyse des données.

4. **Ingénierie des caractéristiques** : Cela implique de sélectionner et de transformer les caractéristiques ou variables pertinentes dans les données pour créer de nouvelles caractéristiques qui amélioreront les performances des modèles.
5. **Sélection et validation des modèles** : cela implique de sélectionner les modèles d'apprentissage automatique ou statistiques appropriés, de les entraîner sur les données et d'évaluer leurs performances à l'aide de métriques et de techniques de validation appropriées.
6. **Déploiement** : il s'agit de déployer le modèle dans un environnement de production et de l'intégrer dans le flux de travail de l'entreprise.

La méthodologie de la science des données est un processus itératif qui consiste à faire des allers-retours entre ces étapes jusqu'à ce qu'une solution satisfaisante soit trouvée. Cela nécessite une combinaison de compétences techniques, de connaissances du domaine et de créativité pour développer des solutions efficaces à des problèmes complexes axés sur les données.

1.3 Médecine Computationnelle

1.3.1 Définition

La médecine computationnelle est un domaine interdisciplinaire qui combine des principes de l'informatique, des mathématiques et de l'ingénierie avec la médecine pour développer des outils et des modèles informatiques pour résoudre des problèmes de santé. Cela implique l'utilisation d'approches basées sur les données, la modélisation informatique et la simulation pour améliorer le diagnostic médical, le traitement et les soins aux patients.

La médecine computationnelle vise à fournir des soins médicaux personnalisés et précis grâce à l'utilisation de modèles et de simulations basés sur les données. Ces modèles peuvent aider à identifier les facteurs de risque, à prédire la progression de la maladie et à optimiser les plans de traitement. Par exemple, les algorithmes d'apprentissage automatique peuvent être utilisés pour analyser des images médicales afin de détecter les premiers signes de cancer ou d'identifier des modèles dans de grands ensembles de données qui peuvent être utilisés pour développer des plans de traitement plus efficaces.

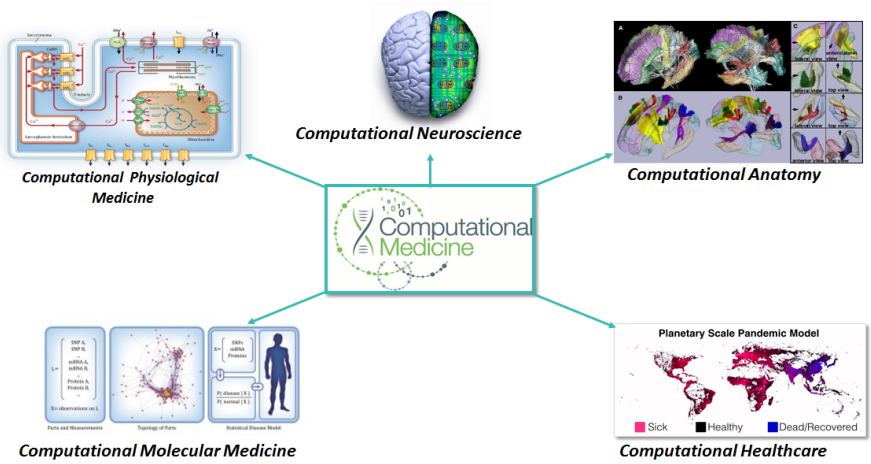


FIGURE 1.2 – Recherches en médecine computationnelle

Certaines des applications de la médecine computationnelle comprennent la découverte et le développement de médicaments, l'imagerie médicale et le diagnostic, la modélisation prédictive, la classification des maladies et la médecine personnalisée. La médecine computationnelle est un domaine en croissance rapide qui a le potentiel de transformer les soins de santé en fournissant des méthodes de diagnostic et de traitement plus précises et plus efficaces.

1.3.2 Domaines d'application

La médecine computationnelle a un large éventail d'applications dans divers domaines, notamment :

1. **Imagerie médicale** : La médecine computationnelle peut être utilisée pour analyser des images médicales telles que des IRM, des tomodensitogrammes et des rayons X pour faciliter le diagnostic, la planification du traitement et la surveillance des maladies.
2. **Génomique et médecine personnalisée** : La médecine computationnelle peut être utilisée pour analyser les données génomiques afin d'identifier les mutations pathogènes, de prédire le risque de maladie et d'élaborer des plans de traitement personnalisés.
3. **Dossiers de santé électroniques** : La médecine computationnelle peut être utilisée pour analyser les dossiers de santé électroniques (DSE) afin d'identifier les modèles de maladie, de prédire les résultats de la maladie et d'élaborer des plans de traitement personnalisés.
4. **Découverte et développement de médicaments** : La médecine computationnelle peut être utilisée pour identifier de nouvelles cibles médicamenteuses, concevoir et optimiser des molécules médicamenteuses et prédire la toxicité des médicaments.
5. **Essais cliniques** : La médecine computationnelle peut être utilisée pour concevoir et analyser des essais cliniques afin d'améliorer leur efficience et leur efficacité.
6. **Santé publique** : la médecine computationnelle peut être utilisée pour analyser les données de santé publique afin d'identifier les épidémies, de

prévoir la propagation des maladies et d'élaborer des stratégies de prévention et de contrôle des maladies.

7. **Dispositifs médicaux et capteurs** : la médecine computationnelle peut être utilisée pour développer et optimiser des dispositifs médicaux et des capteurs pour le diagnostic, la surveillance et le traitement des maladies.
8. **Informatique de santé** : La médecine computationnelle peut être utilisée pour développer et analyser des systèmes informatiques de santé afin d'améliorer la prestation des soins de santé et les résultats pour les patients.

En résumé, la médecine computationnelle a diverses applications dans divers domaines, notamment l'imagerie médicale, la génomique et la médecine personnalisée, les dossiers de santé électroniques, la découverte et le développement de médicaments, les essais cliniques, la santé publique, les dispositifs et capteurs médicaux et l'informatique de la santé.

1.4 Science des données en médecine computationnelle

La science des données joue un rôle essentiel dans la médecine computationnelle. L'abondance de données médicales générées à partir de diverses sources, y compris les dossiers de santé électroniques, l'imagerie médicale, la génomique et les appareils portables, a créé un besoin d'approches axées sur les données pour analyser et interpréter ces données.

Les méthodes de la science des données, telles que l'apprentissage automatique, la modélisation statistique et l'exploration de données, peuvent être utilisées pour extraire des modèles et des informations à partir de données médicales qui peuvent éclairer le diagnostic médical, le traitement et les soins aux patients. Par exemple, les algorithmes d'apprentissage automatique peuvent être utilisés pour analyser des images médicales afin de détecter les premiers signes de cancer ou d'identifier des modèles dans de grands ensembles de données qui peuvent être utilisés pour développer des plans de traitement plus efficaces.

En médecine computationnelle, la science des données peut être utilisée pour développer des modèles prédictifs capables de prévoir la progression de la maladie ou de prédire les résultats pour les patients. La science des données peut également être utilisée pour identifier les facteurs de risque et personnaliser les plans de traitement en fonction de données spécifiques au patient, telles que les informations génétiques, les antécédents médicaux et les facteurs liés au mode de vie.

Dans l'ensemble, la science des données est un élément crucial de la médecine computationnelle, car elle permet aux chercheurs et aux cliniciens d'exploiter les vastes quantités de données médicales disponibles pour améliorer le diagnostic médical, le traitement et les soins aux patients.

1.4.1 Données Biomédicales

Les données biomédicales désignent tout type de données liées à la santé humaine et à la biologie. Il peut inclure un large éventail d'informations, telles que

les dossiers médicaux des patients, les données d'essais cliniques, les données génomiques, les données d'imagerie et de nombreux autres types de données liées à la santé.

Les données biomédicales sont généralement utilisées pour mieux comprendre les mécanismes de la maladie, identifier les traitements potentiels et améliorer les soins aux patients. Avec l'essor de l'analyse des mégadonnées et de l'intelligence artificielle, l'utilisation des données biomédicales suscite un intérêt croissant pour développer des modèles prédictifs et des approches de médecine personnalisée.

Cependant, les données biomédicales soulèvent également d'importantes préoccupations en matière d'éthique et de confidentialité, en particulier compte tenu de la nature sensible des données et du potentiel d'utilisation abusive. À ce titre, des réglementations strictes régissent la collecte, le stockage et l'utilisation des données biomédicales afin de garantir qu'elles sont traitées de manière responsable et éthique.

1.4.2 Outils d'analyse de données biomédicales

Il existe de nombreux outils disponibles pour analyser les données biomédicales, notamment :

1. **Progiciels statistiques** : les logiciels statistiques tels que R, SAS, Python et **SPSS** sont couramment utilisés pour analyser les données biomédicales, y compris les essais cliniques et les études épidémiologiques.
2. **Outils de visualisation de données** : des outils tels que Tableau, MATLAB et **Excel** peuvent être utilisés pour créer des visualisations de données biomédicales, telles que des graphiques et des diagrammes, afin d'identifier des modèles et des tendances.
3. **Outils d'apprentissage automatique et d'intelligence artificielle** : ces outils, tels que TensorFlow et Keras, peuvent être utilisés pour développer des modèles prédictifs et identifier des modèles dans de grands ensembles de données biomédicales.
4. **Outils d'analyse génomique** : Des outils tels que GATK, SAMtools et Picard peuvent être utilisés pour analyser les données génomiques, y compris les données de séquençage de l'ADN et les données d'expression génique.
5. **Outils d'analyse d'imagerie** : des logiciels tels que ImageJ et OsiriX peuvent être utilisés pour analyser les données d'imagerie médicale, y compris les tomodensitogrammes, les IRM et les radiographies.
6. **Outils d'analyse de réseau** : Ces outils, tels que Cytoscape et Gephi, peuvent être utilisés pour analyser des réseaux biologiques complexes et identifier les relations entre les différents composants.
7. **Outils d'exploration de texte et de traitement du langage naturel** : ces outils, tels que Pubmed et MetaMap, peuvent être utilisés pour extraire des informations de la littérature biomédicale et des dossiers de santé électroniques.

Ce ne sont là que quelques exemples des nombreux outils disponibles pour analyser les données biomédicales. Le choix de l'outil dépendra des besoins spécifiques du chercheur et de la nature des données analysées.

1.4.3 Excel

Excel est un progiciel de tableur couramment utilisé pour l'analyse et la gestion de données dans de nombreux domaines, y compris la recherche biomédicale. Bien qu'Excel puisse être utile pour certains types d'analyse de données biomédicales, ce n'est peut-être pas le progiciel le plus approprié pour tous les types de données biomédicales.

Une limitation d'Excel est qu'il a une limite maximale sur le nombre de lignes et de colonnes pouvant être traitées, ce qui peut être un problème pour les grands ensembles de données. De plus, Excel ne dispose pas de fonctionnalités intégrées pour l'analyse statistique avancée ou la visualisation des données, ce qui peut limiter son utilité pour certains types d'analyse de données biomédicales.

Cependant, Excel peut être utile pour les tâches de gestion de données de base telles que la saisie de données, le tri, le filtrage et les calculs de base. Il peut également être utile pour générer des tableaux et des graphiques simples pour visualiser les données. De plus, Excel peut être utilisé pour suivre des données expérimentales ou créer des feuilles de calcul de base pour calculer des mesures statistiques de base telles que la moyenne, la médiane et l'écart type.

1.4.4 Statistical Package for the Social Sciences : SPSS

SPSS signifie Statistical Package for the Social Sciences, qui est un progiciel utilisé pour l'analyse statistique dans les sciences sociales, y compris la psychologie, la sociologie et d'autres domaines connexes. C'est un outil populaire pour analyser les données dans les études de recherche et il est largement utilisé dans les milieux universitaires et commerciaux.

SPSS offre une interface conviviale pour effectuer des analyses statistiques, permettant aux chercheurs de saisir facilement des données, d'exécuter des analyses et de générer des tableaux et des graphiques pour présenter les résultats. Certains des tests statistiques pouvant être effectués à l'aide de SPSS comprennent les tests t, l'ANOVA, l'analyse de régression, l'analyse factorielle et l'analyse par grappes.

SPSS fournit également une large gamme d'outils de gestion et de transformation des données, tels que le tri, la fusion et le recodage des données, qui peuvent être utiles pour préparer les données à analyser.

En outre, SPSS peut être utile pour certains types d'analyse de données biomédicales, telles que l'analyse de données d'enquête ou l'exécution de tests statistiques sur des données d'essais cliniques. Il fournit une interface conviviale pour la saisie et l'analyse des données, et peut générer des tableaux et des graphiques pour présenter les résultats.

Dans l'ensemble, SPSS est un outil puissant et polyvalent pour l'analyse statistique, en particulier pour ceux qui travaillent dans les sciences des données. Il est conçu pour être facile à utiliser, même pour ceux qui ont des connaissances statistiques limitées, et peut produire des résultats précis et fiables.

1.5 Conclusion

Dans l'analyse de données biomédicales, Excel est souvent utilisé pour des analyses simples et la visualisation de données de base, tandis que SPSS est utilisé pour des analyses statistiques plus complexes, telles que l'analyse de régression, l'ANOVA et l'analyse factorielle. Ces deux outils sont importants pour les chercheurs biomédicaux, car ils offrent un moyen flexible et convivial d'analyser et de présenter des données.

Chapitre 2

Collection de Données avec Excel

2.1 Introduction

La collecte de données est un élément essentiel de la recherche biomédicale, car elle contribue à générer les informations et les connaissances nécessaires pour faire progresser notre compréhension des maladies et développer de nouveaux traitements et thérapies. Sans collecte de données, il serait difficile de faire progresser la recherche biomédicale et d'améliorer les résultats pour les patients.

Dans la recherche biomédicale, Excel est couramment utilisé pour la collecte de données. Excel est un outil polyvalent qui permet aux utilisateurs de créer des feuilles de calcul personnalisées pour stocker et organiser les données de manière structurée. Il offre également des fonctionnalités telles que la validation des données, des formules et des modèles qui peuvent aider à garantir l'exactitude des données, à rationaliser l'analyse des données et à standardiser la collecte des données.

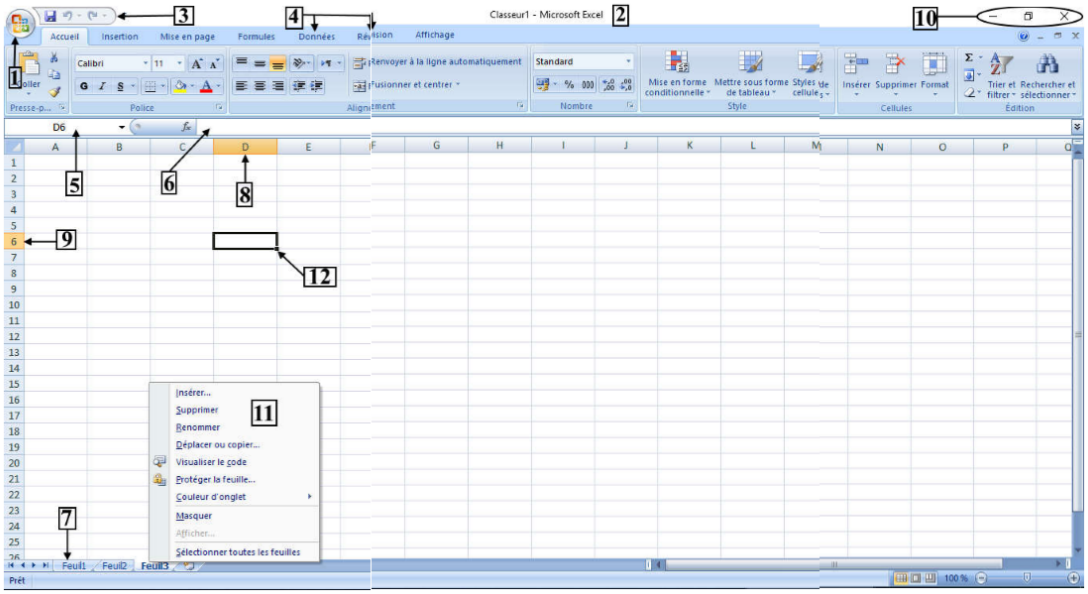
2.2 Démarrer EXCEL

Il existe différentes méthodes de démarrage d'Excel :

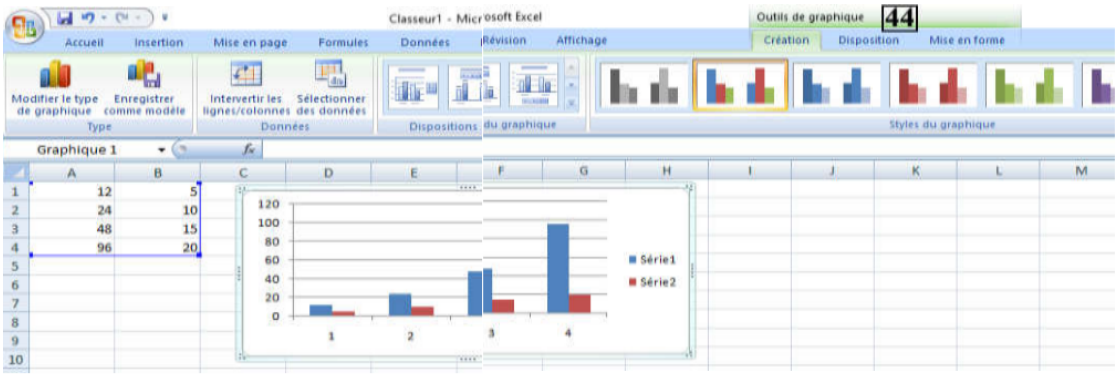
- Menu démarrer : Bouton démarrer ➜ tous les programmes ➜ Microsoft Office ➜ Microsoft Office Excel 2007
- Raccourci : Pour faciliter le lancement d'Excel, il est préférable de créer un raccourci sur le Bureau. Ensuite, double-cliquer sur ce raccourci pour démarrer Excel.

2.3 Terminologie

Excel, Classeur, Feuille de calcul, colonne, ligne, cellule, référence cellule (voir Figure 2.1).



(a)



(b)

FIGURE 2.1 – Environnement Excel

- Un fichier Excel est nommé Classeur. Il peut contenir plusieurs feuilles de calcul **(7)**. Par défaut, il contient trois feuilles de calcul : Feuil1, Feuil2 et Feuil3. Chaque feuille de calcul est constituée de colonnes **(8)** et des lignes **(9)**. L'intersection de ces colonnes et lignes donne des cellules.
- Chaque colonne est référencée par une, deux ou trois lettres (A, B, C,..., MC, MD,...XFC, XFD) (16384 colonnes). Les lignes sont numérotées (1, 2, ..., 1048576). La référence d'une cellule est obtenue par la combinaison de la référence de sa colonne et le numéro de sa ligne (sans espace entre les deux). L'intersection entre la 4e colonne (D) et la 6e ligne donne la cellule D6.
- Par défaut, les cellules sont vides, mais elles peuvent contenir de valeurs.

2.4 Fenêtre EXCEL

■ Bouton office **(1)** : Ce bouton est situé dans le coin supérieur à gauche de la fenêtre Excel. Il ouvre le menu office qui contient deux volets. Le volet droit affiche la liste des classeurs récemment utilisés. Le volet gauche affiche la liste de commandes souvent utilisées : "Nouveau", "Ouvrir", "Enregistrer", "Imprimer", etc. (voir Figure 2.1).

■ Barre de titre **(2)** : Elle indique le nom du classeur en cours, suivi du nom de l'application utilisée (Microsoft Excel). À droite, il existe les trois boutons **(10)** : réduire, niveau inférieur (agrandir) et fermer.

■ Barre d'outils accès rapide **(3)** : Elle contient les boutons de commandes fréquemment utilisés (vous pouvez les utiliser sans passer par les onglets). Par défaut, elle affiche les trois boutons (enregistrer, annuler et de répéter). Mais on peut la personnaliser en y ajoutant d'autres outils tels que nouveau, ouvrir ou impression rapide, etc.

■ Ruban : il s'affiche sous la barre de titre. Le ruban est organisé hiérarchiquement ; il comporte plusieurs onglets **(4)** et **(44)**. Ils peuvent être des onglets fixes **(4)** : "Accueil, Insertion, Mise en page, Formules, Données, Révision et Affichage" ou des onglets contextuels **(44)**. Ces derniers s'affichent lorsque vous sélectionnez un objet spécifique. Les onglets contextuels affichés dépendent de l'objet qui est sélectionné. Par exemple : la sélection d'un graphique fait apparaitre les trois onglets contextuels Outils de graphique **(44)** : création, disposition, mise en forme. Ensuite, chaque onglet est divisé en plusieurs groupes qui comportent des boutons de commande et des galeries. Par exemple : l'onglet Accueil comporte les groupes : "Presse-papiers, Police, Alignement, Nombre, Style, Cellules et Édition". Remarque : vous pouvez réduire le ruban (masquer les groupes) : double-cliquez sur l'onglet actif (**[Ctrl] + [F1]**) :

- Pour réafficher les groupes temporairement, faire un simple clic sur un onglet.
- Pour restaurer le ruban, double-cliquez de nouveau sur un onglet.

■ Zone de nom **(5)** : elle affiche l'adresse ou le nom de la cellule active (actuellement sélectionnée) ou de la sélection. Une cellule est active à la fois. Elle est entourée d'une bordure épaisse.

■ Barre de formule **(6)** : elle affiche le contenu de a cellule active. Elle peut être utilisée pour la saisie ou la modification du contenu de cette cellule.

■ Poignée de la recopie **(12)** : le petit carré noir dans le coin inférieur droit d'une cellule (lorsque vous pointez sur la poignée de recopie, le pointeur se transforme en croix noire).

2.5 Fichier XLS ou XLSX ?

Voir la Table 2.1.

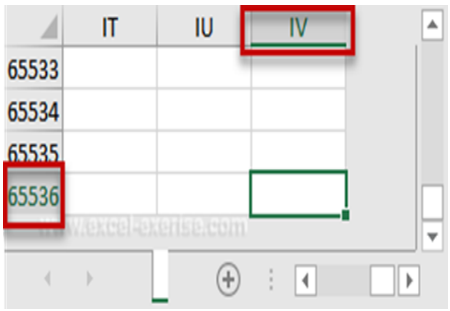
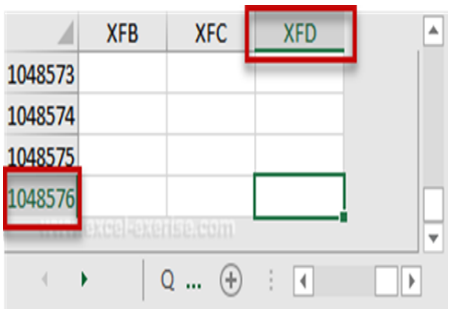
Excel 97-2003	**Excel 2007**
Dans un classeur xls, les limites sont de 65 536 lignes et 256 colonnes, ce qui correspond à la colonne IV.	Maintenant, avec un classeur xlsx, les limites sont de 1 048 576 lignes et 16 384 colonnes, qui est la colonne XFD.
	

TABLE 2.1 – XLS vs XLSX

2.6 Gestion des Classeurs

2.6.1 Créer classeur

Pour créer un nouveau classeur **[Ctrl]+[N]** : Bouton Office ⇨ Nouveau ⇨ Nouveau Classeur Excel ⇨ Créer.

2.6.2 Ouvrir un classeur

Pour ouvrir un classeur depuis Excel **[Ctrl]+[O]** : Bouton Office ⇨ Ouvrir

2.6.3 Enregistrer un classeur

Excel offre plusieurs choix pour enregistrer un nouveau classeur (le créer physiquement) et le conserver sur votre disque dur :

1. Bouton Office ⇨ Enregistrer (**[ctrl]+[S]**)
2. Cliquer sur la disquette de la barre d'outils accès rapide (**[ctrl]+[S]**)
3. Bouton Office ⇨ Enregistrer sous ([**F12**]).

Lorsque vous utilisez une de ces méthodes d'enregistrement (1, 2 ou 3) pour la première fois, la boite de dialogue "enregistrer sous" s'affiche, dans laquelle vous spécifiez :

- ✓ Le nom du fichier
- ✓ Le disque et le dossier de sauvegarde
- ✓ Le type de fichier : le format de fichier souhaité pour l'enregistrement. Par défaut, les classeurs sont enregistrés au format (.xlsx) (Excel 2007)

2.7 Cellule Active Et Plage De Cellules

■ La cellule active est celle dans laquelle la saisie sera enregistrée. Elle se distingue par une bordure plus marquée. Par défaut, A1 est la cellule active à l'ouverture du classeur. L'adresse (ou le nom) et le contenu de la cellule active sont affichés dans la zone de nom et la barre de formule respectivement.

■ Tout rectangle de cellules est appelé "plage de cellules", ou "plage". Pour désigner une plage de cellules, il est courant d'utiliser la référence de la première cellule en haut à gauche suivie d'un double-point et de la référence de la dernière cellule en bas à droite.

Exemple :
● A1 :B3 fait référence aux cellules : A1, B1, A2, B2, A3, B3
● C1 :E3 fait référence aux cellules : C1, D1, E1, C2, D2, E2, C3, D3, E3
● A1 :A8 fait référence aux cellules : A1, A2, A3, A4, A5, A6, A7, A8
● A1 :E1 fait référence aux cellules : A1, B1, C1, D1, E1

2.8 Saisie des Données

Avant saisir de données dans une cellule, il faut la sélectionnée.

❖ Texte

■ Le texte est automatiquement aligné à gauche.
■ Le texte ne passe pas à la ligne même s'il est long et dépasse la largeur de la colonne.
■ Pour passer à la ligne suivante : **[alt]+[entrée]** au sein d'une cellule
■ Si le texte commence par " + ", " - " ou " = ", Excel affiche un message d'erreur "($\#NOM?$)" car il interprète ce texte comme une formule. Pour éviter cela, ajoutez une apostrophe « ' » avant le texte (exemple : '+Médecine).
■ Ajouter une apostrophe « ' » avant un nombre pour que ce dernier soit interprété comme un texte (exemple : '2019 dans la cellule A2).

❖ Nombre

■ Le nombre est automatiquement aligné à droite.
■ Pour saisir un nombre négatif, il faut le précéder du signe " - " ou le mettre entre parenthèses.
■ L'écriture 32e9 signifie $32 * 10^9$ c.-à-d. 32 suivis de neuf zéros.

❖ Date

■ La date est automatiquement alignée à droite.
■ Pour entrer une date dans une cellule, vous pouvez la saisir sous la forme : jj/mm/aaaa ou jj-mm-aaaa.
■ Pour saisir la date d'aujourd'hui, tapez **[ctrl]+[;]** (cette date est constante).
■ Pour qu'une date soit interprétée comme un texte, faites-la précéder par une apostrophe « ' » (exemple : '2019) (voir Figure 2.2).

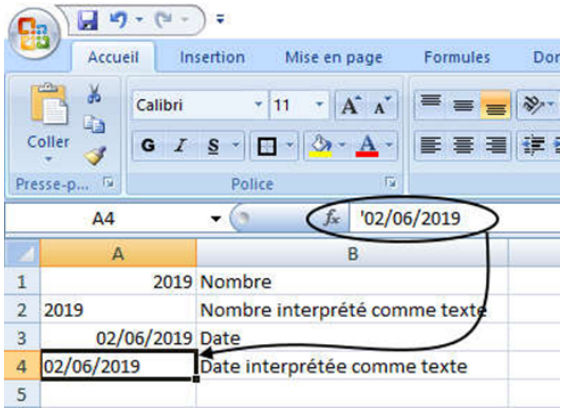


FIGURE 2.2 – Utilisation de l'apostrophe

2.9 Séries de Données

➤ Taper Patient, Date, dans les cellules A1, B1 respectivement.
➤ Sélectionner la cellule A1, faire glisser la poignée de recopie sur quelque cellules (vers le bas) (voir Figure 2.3).
➤ Pour les dates vous avez la possibilité d'incrémentation par jour mois...

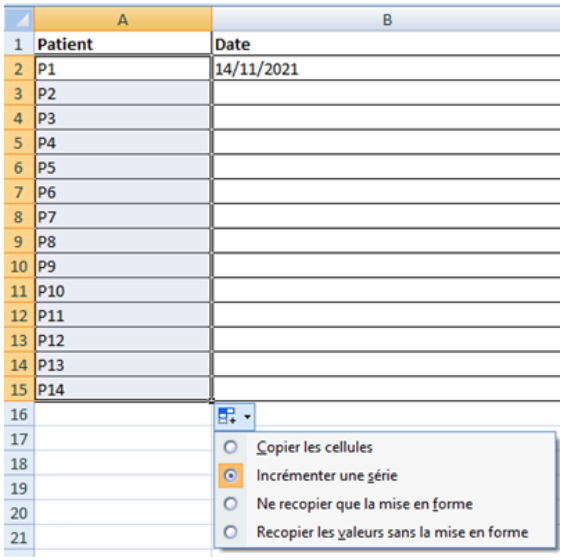


FIGURE 2.3 – Glissement de cellule Excel

2.10 Validation des Variables Statistiques

Les variables sont utilisées pour décrire les individus d'une population. Chaque colonne correspond à une variable.
■ Une variable a un nom : "ID", "Age", "poids" ...
■ Une variable a une valeur à un certain moment : MED01 , 22 ans, 59 kg (chaque ligne représente un individu, unité statistique ...)

TABLE 2.2 – Variables quantitatives et qualitatives

Types de variables			
Quantitative (Numérique)		**Qualitative (Catégorielle)**	
Continue	**Discrète**	**Ordinale**	**Nominale**
Se compose de valeurs numériques qui peuvent être mesurées, mais pas comptées (infinie).	Se compose de valeurs numériques qui peuvent être mesurées et comptées (finie).	Se compose de texte ou d'étiquettes qui ont un ordre logique.	Se compose de texte ou d'étiquettes sans ordre logique.
Par exemple. Poids {56,06 kg, 87 kg}	Par exemple. Nombre d'enfants {0, 1, 2, 3,..., 10}	Par exemple. Taille de la tumeur {petite, moyenne, grande}	Par exemple. Sexe {Homme, Femme}

2.10.1 Variables Quantitatives

■ En utilisant la fenêtre "Validation des données", il est possible de limiter une saisie en imposant des restrictions telles que la saisie d'un nombre entier.
Exemple : pour la variable age : Pour restreindre la saisie à des nombres entiers supérieurs à 0, commencez par sélectionner les cellules concernées. (avant la saisie de ces valeurs) dans notre exemple C2 :C6 (l'étape S1). Ensuite **Données** → **Outils de données** → **Validation des données** (S2 et S3). Choisir Autoriser : **nombre entier** (S4). Ensuite Données : **supérieur à** et taper 0 dans source (S5) (voir Figure 2.4).

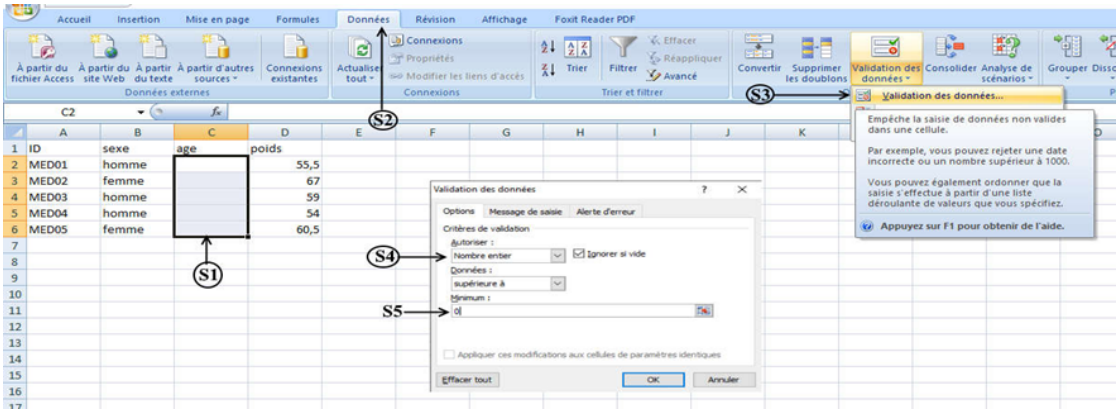


FIGURE 2.4 – Validation des données quantitatives

Notes :
■ Pour les variables quantitatives continues, nous choisissons **Décimal** dans

la liste "Autoriser".
■ Dans la liste "Données", vous avez plusieurs choix comme : **supérieur à**, **inférieure à**, **comprise entre**..., etc.

2.10.2 Variables Qualitatives

■ Sélectionner les cellules où vous allez saisir le sexe (l'étape E1 dans la figure suivante) (dans notre exemple B2 :B6) (voir Figure 2.5).

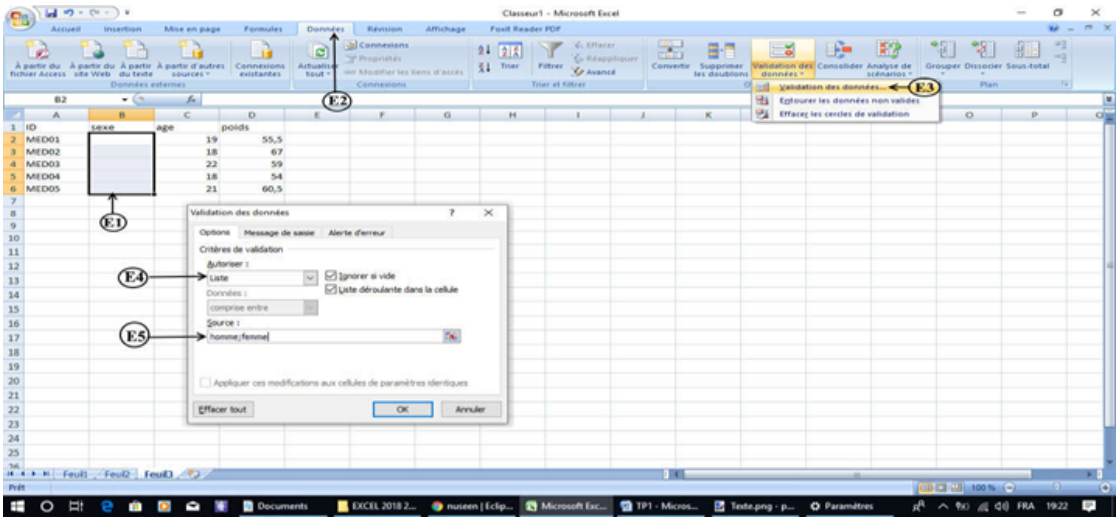


FIGURE 2.5 – Validation des données qualitatives

Ensuite, choisir autoriser "liste" (E4) dans : **Données→validation des données→validation de données** (E2 et E3). Taper les éléments de la liste (**homme ; femme**) dans le champ source (E5) (séparés par point-virgule **;**).

Le résultat sera :

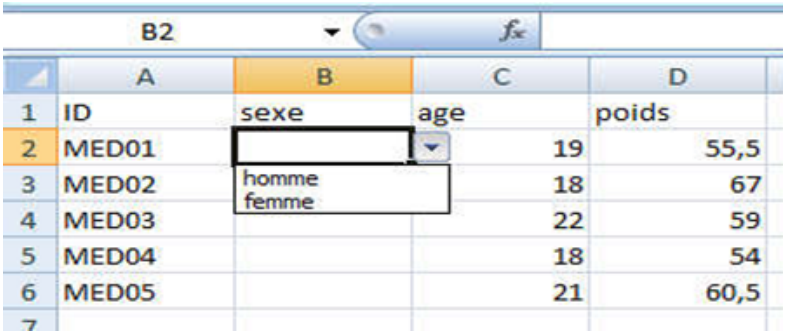


FIGURE 2.6 – Résultat de la validation des données qualitatives

Note :
■ Veuillez trouver le fichier de données Excel de ce chapitre à partir de ce lien : `https://aboulesnane.net/wp-content/datafiles/DATAEXCEL2.xlsx`

2.11 Conclusion

Excel est un outil populaire pour la collecte de données dans la recherche biomédicale. Pour collecter des données efficacement avec Excel, il est important de planifier votre feuille de calcul, d'utiliser la validation des données pour garantir l'exactitude, d'utiliser des formules pour rationaliser l'analyse et de sécuriser vos données. Par conséquent, Excel peut être un outil puissant pour organiser et analyser les données dans la recherche biomédicale.

Chapitre 3

Organisation des Données Collectées dans SPSS

3.1 Introduction

SPSS (Statistical Package for the Social Sciences) est un outil logiciel puissant largement utilisé pour organiser, analyser et visualiser des données dans la recherche sociale et biomédicale. Il offre une gamme d'options personnalisables pour l'organisation et la gestion des données, de puissants outils statistiques pour l'analyse des données et une documentation claire et transparente pour la reproductibilité.

3.2 L'Environnement SPSS

L'environnement SPSS se compose de plusieurs composants, comme le montre l'image ci-dessous (Figure 3.1) :

1. La barre de titre : affiche le nom du fichier actuel et de l'application.
2. La barre de menus : Cette barre donne accès à différentes commandes qui sont regroupées selon leur fonction. SPSS a un certain nombre d'options de menu situées en haut de l'écran (comme tout autre programme informatique). Ouvrez SPSS et sélectionnez chacune des options de menu une par une.
   - **Le menu 'Fichier' (raccourci Alt + F)** : Essentiellement, ce menu vous permet d'ouvrir des fichiers existants, d'en créer de nouveaux et d'imprimer ou d'enregistrer tout ce sur quoi vous travaillez. Les listes Données récemment utilisées et Fichiers récemment utilisés sont utiles, car elles vous permettent d'accéder rapidement aux fichiers que vous avez récemment ouverts ou sur lesquels vous avez travaillé.
   - **Le menu 'Edition' (raccourci Alt + E)** : Ce menu devrait vous être familier si vous avez déjà utilisé des traitements de texte. Annuler et Rétablir peuvent aider à rectifier les erreurs que vous faites. Couper, Copier et Coller vous permettent de déplacer des blocs de nombres

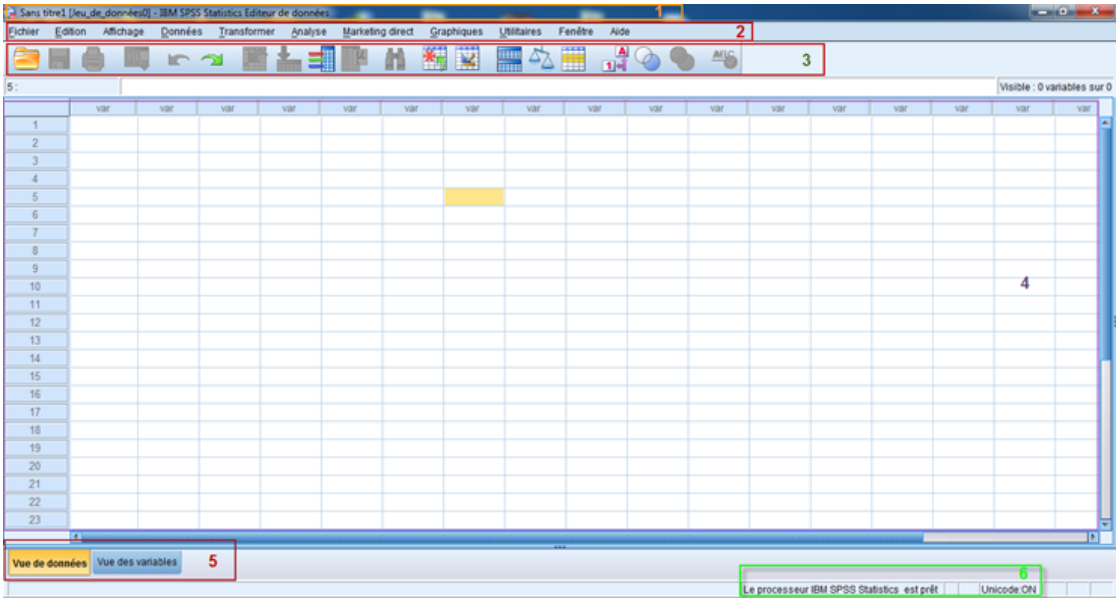


FIGURE 3.1 – Environnement SPSS

d'une zone de la feuille de calcul à une autre. Chercher... et Aller à l'observation... vous permet de localiser un score de données ou un participant particulier, ce qui est très pratique lorsque vous traitez un grand nombre de données.

- **Le menu 'Affichage' (raccourci Alt + V)** : Le menu Affichage traite des aspects visuels de la feuille de calcul, en particulier : quelles barres d'outils sont affichées, quelles polices sont utilisées, pour voir les lignes de la grille sur la feuille de calcul ou les étiquettes de valeur sont affichées pour vos variables..., etc.
- **Le menu 'Données' (raccourci Alt + D)** : Ce menu vous permet d'organiser votre fichier de données. Il est peu probable que vous utilisiez initialement la plupart des options de ce menu ; cependant, quelques-unes des options peuvent être utiles. Par exemple, vous pouvez identifier certaines erreurs potentielles commises lors de la saisie de données en signalant d'éventuelles entrées de données en double à l'aide de l'outil **Identifier les observations dupliquées**.
- **Le menu 'Transformer' (raccourci Alt + T)** : Ce menu vous permet de manipuler vos variables.
- **Le menu 'Analyser' (raccourci Alt + A)** : C'est le menu que vous utiliserez probablement le plus et dont vous aurez initialement besoin : Statistiques descriptives, Comparer les moyennes, Modèle linéaire général, Corrélation et Régression,..., etc.
- **Le menu 'Marketing direct' (raccourci Alt + M)** : C'est plus pour les entreprises qui souhaitent réaliser des études de marché. Vous n'aurez pas besoin d'utiliser ce menu !
- **Le menu 'Graphiques' (raccourci Alt + G)** : Ce menu vous permet de présenter les données sous forme graphique, ce qui vous aidera à mieux comprendre vos données. Il existe plusieurs façons de créer

des graphiques dans SPSS, mais c'est un bon point de départ.

- **Le menu 'Utilitaires' (raccourci Alt + U)** : En pratique, il est utile pour créer des analyses personnalisées et automatisées..., mais n'hésitez pas à l'ignorer pour l'instant !
- **Le menu 'Fenêtre' (raccourci Alt + W)** : Ce menu vous permet d'accéder rapidement à d'autres fenêtres qui pourraient être masquées.
- **Le menu 'Aide' (raccourci Alt + H)** : Ce menu peut être très utile, car il vous offre de l'aide et des informations à la fois sur le système du programme lui-même et sur les tests statistiques qu'il propose.

3. La barre d'outils : fournit des raccourcis vers des commandes de menu couramment utilisés.
4. La fenêtre de l'éditeur de données : Le nom en tête de chaque colonne est le nom de la variable, c'est-à-dire le nom que vous utiliserez pour faire référence à une variable, tandis que, chaque ligne représente une observation (un cas).
5. Les onglets : vue de données et vue des variables : vue de données est l'endroit où nous inspectons nos données réelles et la vue des variables est l'endroit où nous voyons des informations supplémentaires sur nos données.
6. La barre de statut : en bas de chaque fenêtre, SPSS fournit plusieurs informations telles que : l'état de la commande, l'état du filtre, etc.

3.3 Manipulation de Données dans SPSS

■ Les données SPSS ont trois composants principaux : les observations, les variables et les métadonnées. Lorsque vous recevez des données, vous aurez rarement un problème avec les observations, occasionnellement un problème avec les variables, mais presque toujours un problème avec les métadonnées.

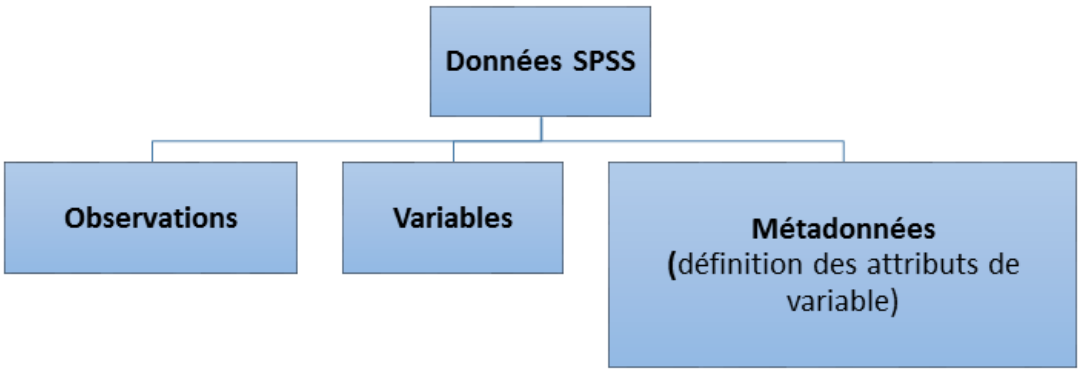


FIGURE 3.2 – Données SPSS

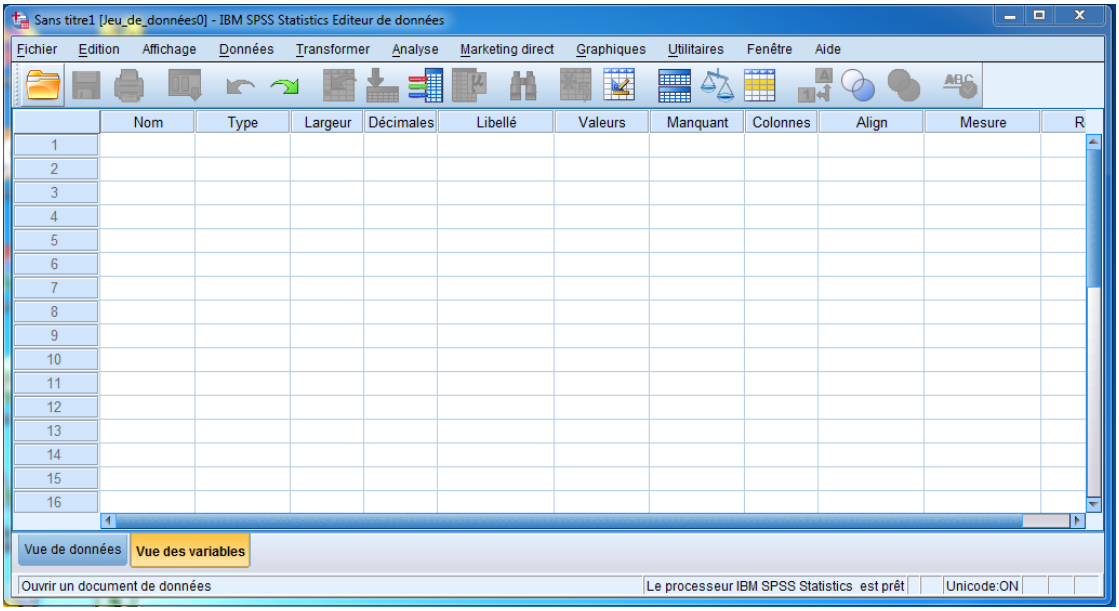


FIGURE 3.3 – Vue des variables

■ SPSS peut lire des données à partir de divers formats, notamment des bases de données, des fichiers texte, Microsoft Excel, CSV.... etc. Vous pouvez également taper directement dans SPSS ; et, vous pouvez même coller les données copiées dans SPSS.

3.3.1 Définition des métadonnées

■ Dans SPSS, les données sont organisées sous forme d'observations (lignes) et chaque observation est constituée d'un ensemble de variables (colonnes). Tout d'abord, vous définissez les caractéristiques des variables qui composent une observation, puis vous entrez les données dans les variables qui composent le contenu des observations.

■ Pour saisir des données dans SPSS, utilisez l'onglet «Vue des variables ». Comme vous pouvez le voir dans la figure ci-dessous (voir Figure 3.3), les attributs de la variable (tels que le nom, le type et la largeur) sont définis en haut de la fenêtre. Tout ce que vous avez à faire est d'entrer quelque chose dans chaque colonne pour chaque variable.

■ Les 11 caractéristiques sont les seuls nécessaires pour spécifier complètement tous les attributs d'une variable. Lorsque vous ajoutez une nouvelle variable, vous constaterez que des valeurs par défaut raisonnables apparaissent pour la plupart des caractéristiques. Les 11 caractéristiques d'une variable sont :

1. **Nom** : Cliquez simplement sur la cellule et saisissez un court descriptif, tel que : Age, revenu, sexe, patient.... Bien que vous puissiez saisir des noms plus longs ici, il est recommandé de les maintenir courts, car ils seront utilisés dans des listes nommées ainsi que pour des balises d'identification sur des graphiques de données et autres formats où l'es-

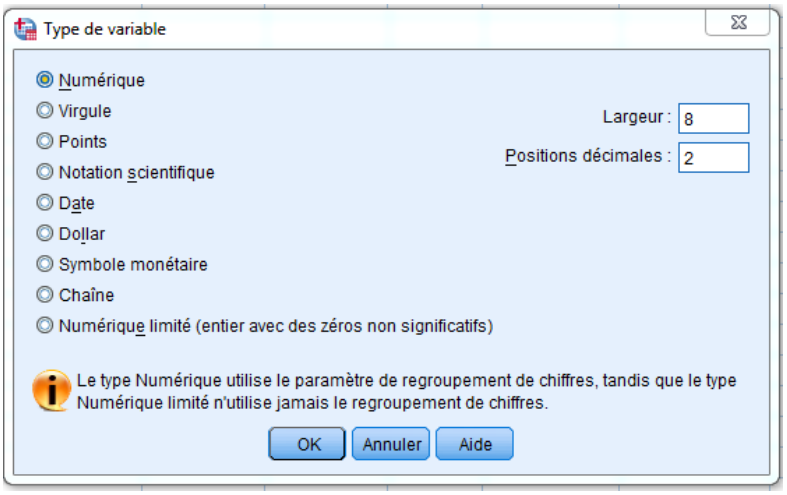


FIGURE 3.4 – Type de variable

pace peut être limité. Les espaces ne sont pas autorisés dans les noms de variables !

2. **Type** : La plupart des données que vous saisirez ne seront que des nombres normaux. Cependant, les données telles que la devise doivent être affichées dans un format spécial, et les données telles que les dates nécessitent des procédures de calcul spéciales. Pour ce type de données, il vous suffit de spécifier le type dont vous disposez et SPSS s'occupe des détails pour vous. Pour afficher la boîte de dialogue illustrée à la Figure 3.4, sélectionnez une cellule dans la colonne Type, puis cliquez sur le bouton représenté par trois points qui apparaît.
3. **Largeur** : La colonne largeur dans la définition d'une variable détermine le nombre de caractères utilisés pour afficher la valeur. Si la valeur à afficher n'est pas assez grande pour remplir l'espace, la sortie sera complétée par des blancs. S'il est plus grand que vous ne l'avez spécifié, il sera reformaté pour s'adapter ou des astérisques seront affichés.
4. **Décimales** : La colonne Décimales contient le nombre de chiffres qui apparaissent à droite de la virgule décimale lorsque la valeur apparaît à l'écran.
5. **Libellé** : Le nom et le libellé ont le même objectif fondamental : ce sont des descripteurs qui identifient la variable. La différence est que le nom est l'identifiant court et l'étiquette est le long. Vous pouvez également ignorer la définition du libellé. Si vous n'avez pas d'étiquette définie pour une variable, SPSS utilisera le nom de variable que vous avez défini pour tout.

6. **Valeurs** : La colonne Valeurs est l'emplacement où vous pouvez assigner des libellés à toutes les valeurs possibles d'une variable. Pour afficher la boîte de dialogue illustrée à la Figure 3.5, sélectionnez une cellule dans la colonne Valeurs, puis cliquez sur le bouton représenté par trois points qui apparaît.

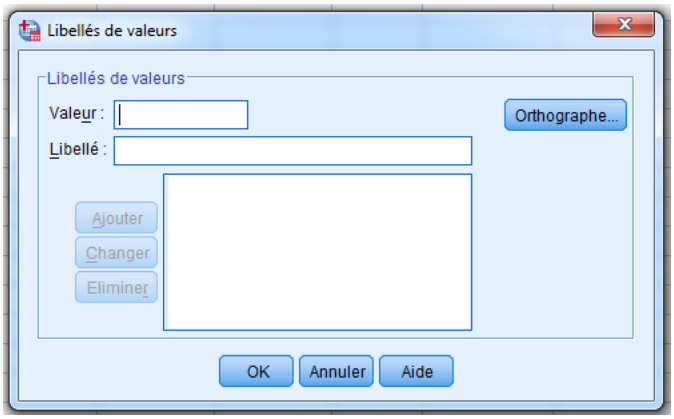


FIGURE 3.5 – Libellés de valeurs

Normalement, vous faites une seule entrée pour chaque valeur possible qu'une variable peut prendre. Mais, par exemple, pour une variable nommée Sexe, vous pouvez affecter la valeur 1 à l'étiquette Homme et 2 à l'étiquette Femme. Si vous définissez des libellés, votre sortie peut afficher des étiquettes au lieu de valeurs.

Pour définir un libellé pour une valeur, procédez comme suit :

(a) Dans la zone Valeur, entrez la valeur.

(b) Dans la zone Libellé, entrez un libellé.

(c) Cliquez sur le bouton Ajouter. (La valeur et l'étiquette apparaissent dans le grand bloc de texte.)

(d) Pour modifier ou supprimer une définition, sélectionnez-la simplement dans le bloc de texte, apportez vos modifications, puis cliquez sur le bouton Changer.

(e) Répétez les étapes 1 à 4 si nécessaire.

(f) Pour enregistrer les étiquettes de valeur et fermer la boîte de dialogue, cliquez sur OK.

7. **Manquant** : Vous pouvez spécifier des codes pour les données manquantes. Pour afficher la boîte de dialogue illustrée à la Figure 3.6, sélectionnez une cellule dans la colonne Manquant. Un bouton représenté par trois points apparaîtra, cliquez dessus pour ouvrir la boîte de dialogue.

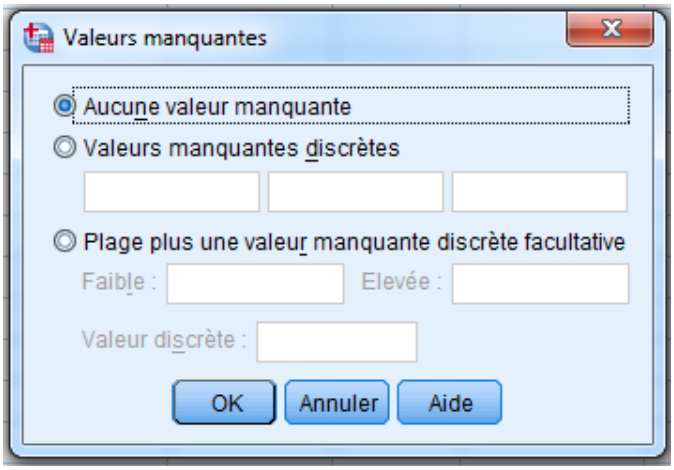


FIGURE 3.6 – Valeurs manquantes

Par exemple, supposons que vous saisissiez des réponses à des questions, et que l'une des questions soit : " Combien d'enfants avez-vous ?" La réponse normale à cette question est un nombre, vous définissez donc le type de variable comme un nombre. Vous pouvez définir –99 comme la valeur saisie lorsque la réponse est « Je ne me souviens pas » et –98 peut être utilisé lorsque la réponse est « Je ne peux pas dire ».

8. **Colonnes** : Dans l'attribut Colonnes, vous pouvez spécifier la largeur de la colonne à utiliser pour saisir les données.
9. **Align** : La colonne Align détermine la position des données dans son espace alloué.
10. **Mesure** : Votre valeur pour l'attribut Measure spécifie le niveau de mesure de votre variable. Voici le niveau des options de mesure dans SPSS :
    - **Nominal** : Une valeur qui spécifie une catégorie ou un type de chose. Vous pouvez avoir 0 pour Désapprouver et 1 pour Approuver. Ou vous pouvez utiliser 1 pour signifier Femme et 2 pour signifier Homme.
    - **Ordinal** : Une valeur qui spécifie la position (ordre) de quelque chose dans une liste. Par exemple, premier, deuxième et troisième sont des nombres ordinaux.
    - **Échelle** : Un nombre qui spécifie une magnitude. L'échelle peut être la distance, le poids, l'âge ou un décompte de quelque chose.
11. **Rôle** : Vous n'avez pas besoin de vous soucier de la colonne Rôle pour le moment.

3.4 Saisie et affichage des éléments de données dans l'onglet «Vue de données»

Après avoir défini les variables, vous pouvez commencer à saisir les données. Cliquez sur l'onglet **Vue de données** de la fenêtre de l'éditeur de données. En haut des colonnes, vous voyez les noms des variables. La saisie de données dans l'une de ces cellules est simple : il vous suffit de cliquer sur la cellule et de commencer à taper.

3.5 Sauvegarde des données SPSS

Tout ce que vous avez à faire est de choisir **Fichier ➜ Enregistrer ou Enregistrer sous, (Ctrl + S)** sélectionnez votre type de fichier, puis entrez un nom de fichier. The SPSS Statistics File Format is «**.sav**» (voir Figure 3.7).

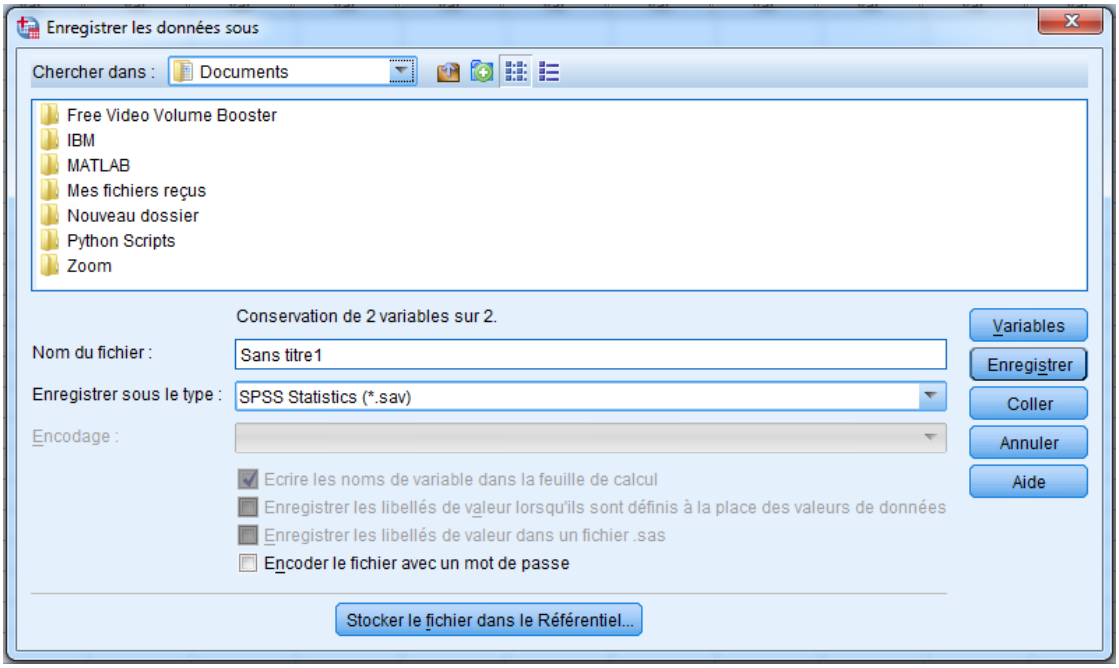


FIGURE 3.7 – Enregistrer les données dans SPSS

Vous avez le choix entre de nombreux types de fichiers : formats de texte brut, formats de feuille de calcul Excel, formats Lotus, formats dBase, formats SAS, format SYLK, format Portable et 18 formats Stata.

3.6 Ouverture de fichiers de données SPSS

Pour ouvrier un fichier de données, choisissez **Fichier ➜ Ouvrir ➜ Données ou (Ctrl + O)** et sélectionnez le fichier à charger. Lorsque vous le faites, les noms de variables et les données sont chargés dans SPSS.

3.7 Transfert de données d'un fichier Excel vers SPSS

Pour ouvrir votre fichier Excel dans SPSS :

1. **Fichier ➜ Ouvrir ➜ Données (Ctrl + O)**, dans le menu SPSS.
2. Sélectionnez le type de fichier que vous souhaitez ouvrir, Excel *.xls *.xlsx, *.xlsm .

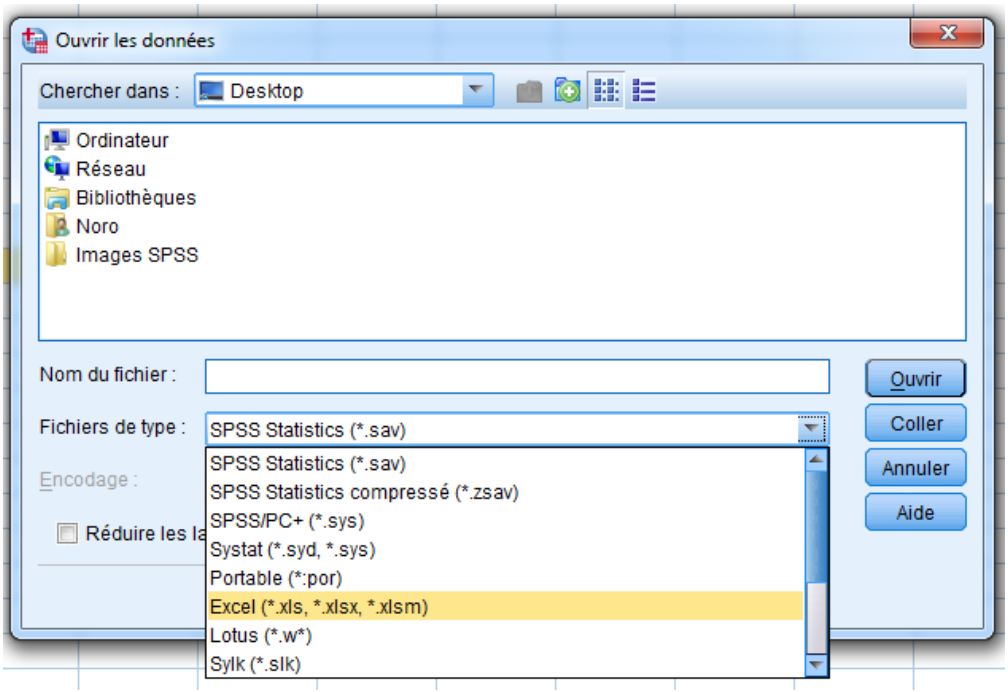


FIGURE 3.8 – Ouvrir les données dans SPSS

3. Dans la boîte de dialogue "Ouvrir les données", sélectionnez le fichier que vous souhaitez ouvrir (voir Figure 3.8)..
4. Cliquez sur **Ouvrir**.
5. La boîte de dialogue suivante apparaît (Figure 3.9).

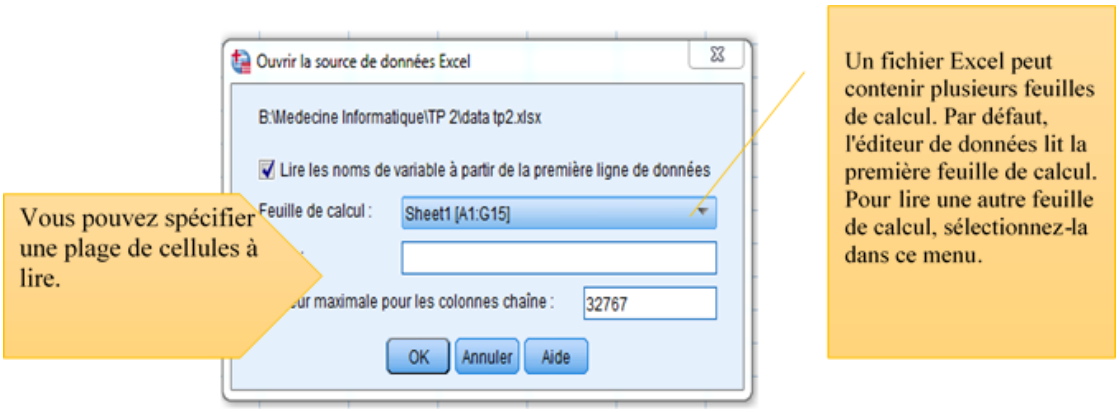


FIGURE 3.9 – Ouvrir la source de données Excel

6. Cliquez sur **OK**.

Note :
■ Veuillez trouver le fichier de données SPSS de ce chapitre à partir de ce lien :
`https://aboulesnane.net/wp-content/datafiles/DataSPSS3.sav`

3.8 Conclusion

L'organisation des données collectées dans SPSS est importante pour une gestion précise et efficace des données, la reproductibilité des résultats de recherche, la sécurité et la confidentialité des données de recherche et la collaboration entre les chercheurs. SPSS offre une interface conviviale et des options personnalisables pour organiser les données, telles que le recodage, les transformations de données et le nettoyage des données. SPSS permet une gestion des données rapide et efficace, comme la possibilité d'importer et d'exporter des données à partir de différentes sources, de fusionner des ensembles de données et de nettoyer des données. Dans l'ensemble, l'organisation des données collectées dans SPSS est essentielle pour mener des recherches fiables et valides en sciences sociales et biomédicales.

Chapitre 4

Prétraitement des Données

4.1 Introduction

Le prétraitement des données est une étape cruciale de l'analyse des données qui implique le nettoyage, la transformation et la préparation des données brutes pour l'analyse. Il contribue à améliorer la qualité, la précision et la compatibilité des données avec les techniques d'analyse et les outils logiciels. Le prétraitement des données peut également réduire le temps et les efforts requis pour l'analyse des données en rationalisant le processus de nettoyage et de transformation des données, et peut aider à créer des visualisations et des résumés des données qui facilitent la compréhension et l'interprétation des données.

4.2 Navigation dans le SPSS Visualiseur (Viewer)

Lorsque vous exécutez une analyse, produisez un graphique ou même l'enregistrement d'un fichier, **la fenêtre SPSS Statistics Viewer** apparaît automatiquement pour afficher ce que vous avez créé (Figure 4.1).

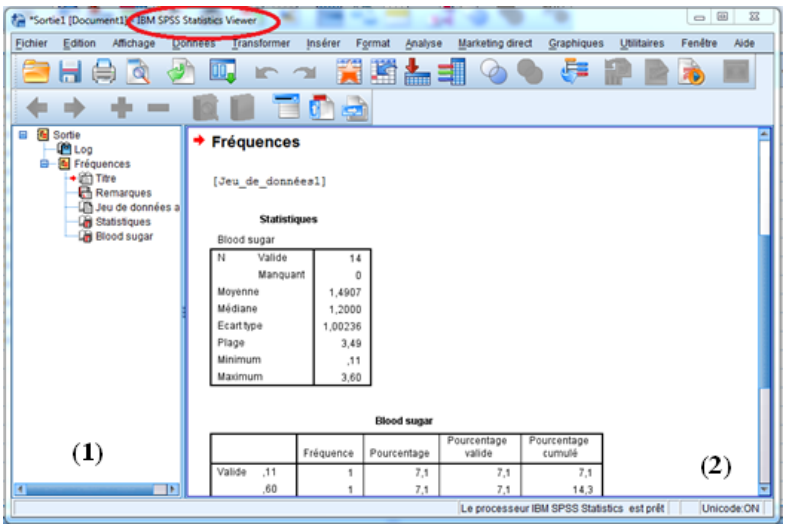


FIGURE 4.1 – IBM SPSS statistics viewer

■ Le volet de plan (1), sur la gauche, contient un aperçu de toutes les informations stockées dans le Viewer.
■ Le volet de contenu (2), à droite, contient les tableaux statistiques, les graphiques et les textes.
■ Pour masquer un objet dans le volet de contenu, il faut :

- ✓ Sélectionnez l'objet (table, graphe..) depuis le volet de plan ou le volet de contenu.
- ✓ Cliquez sur l'icône du livre dans la barre d'outils  .
- ✓ Pour afficher à nouveau cet objet, appuyez sur l'icône 

Masquer des éléments sans les supprimer permet à l'utilisateur de se concentrer plus facilement sur les résultats qui l'intéressent tout en conservant tous les résultats.

4.3 Jouer avec les données dans SPSS

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P1	53,00	1,20	1	1
2	P2	-99,00	1,50	1	3
3	P3	88,00	1,40	2	2
4	P4	45,00	,60	1	4
5	P5	90,00	2,40	2	2
6	P6	175,00	3,60	2	1

FIGURE 4.2 – Fichier de données SPSS

1. Télécharger le fichier de données SPSS appelé « DataSPSS4.sav » sur : `https://aboulesnane.net/wp-content/datafiles/DataSPSS4.sav`
2. Les données contiennent cinq variables nommées : Patient, Weight, Bloodsugar, Sexe, et Bloodgroup (voir la Figure 4.2).
   (a) La Variable « **Patient** » est une variable de type Chaîne.
   (b) La Variable « **Weight** » est une variable quantitative continue de type Numérique. (Pour la variable Weight, on représente les valeurs manquantes par le nombre -99 (voir Figure 3.6)).
   (c) La Variable « **Bloodsugar** » est une variable quantitative continue de type Numérique.
   (d) Les valeurs possibles pour la variable qualitative « **Sexe** » : 1=Homme et 2=Femme.
   (e) Les valeurs possibles pour la variable qualitative « **Bloodgroup** » : 1=AB , 2=A, 3=B et 4=O.

4.4 Remplacement des valeurs manquantes

Afin de remplacer les valeurs manquantes par des valeurs acceptables, nous utiliserons des paramètres de position centrale tels que la moyenne, la médiane, le mode...etc. Par exemple, dans la variable **Weight** de nos données,

nous avons une valeur manquante représentée par -99. Afin de ne pas perturber la distribution de nos données, nous pouvons remplacer cette valeur par une valeur moyenne de la série. Pour cela :

1. Choisissez **Transformer ➜ Remplacer les valeurs manquantes** : la boîte de dialogue **Remplacer les valeurs**...apparaît.
2. Passez la variable **Weight** vers la zone «**Nouvelles variables**».
3. Dans le menu **Méthode**, (voir Figure 4.3), vous pouvez sélectionner la meilleure méthode utilisée pour remplacer les valeurs manquantes (Interpolation linéaire, Paramètres de position centrale...). Dans notre cas, nous utiliserons **Moyenne série**.

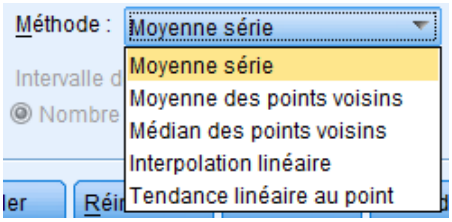


FIGURE 4.3 – Menu Méthode

4. Cliquez sur **OK** (voir la Figure 4.4).

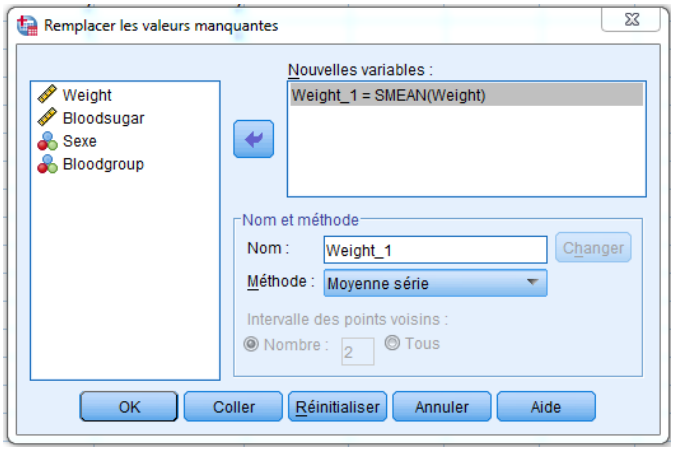


FIGURE 4.4 – Remplacement les valeurs manquantes

En conséquence, une nouvelle colonne apparaît avec le nom « **Weight_1** » où la valeur -99 est remplacée par la valeur moyenne 90,2 (voir la Figure 4.5).

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup	Weight_1
1	P1	53,0	1,2	1	1	53,0
2	P2	-99,0	1,5	1	3	90,2
3	P3	88,0	1,4	2	2	88,0
4	P4	45,0	,6	1	4	45,0
5	P5	90,0	2,4	2	2	90,0
6	P6	175,0	3,6	2	1	175,0

FIGURE 4.5 – Résultat après le remplacement

4.5 Trier les observations

1. Choisissez **Données ➜ Trier les observations** : la boîte de dialogue **Trier les observations** s'affiche (Figure 4.6).

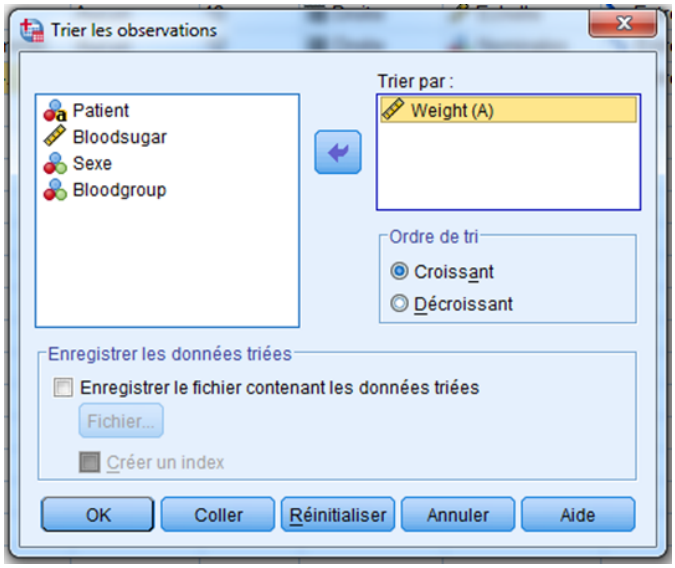


FIGURE 4.6 – Trier les observations

2. Sélectionnez la variable **Weight**.
3. Cliquez sur **OK** pour trier les données.

■ L'ordre dans lequel les données sont affichées n'affecte jamais l'analyse. Vous triez les données uniquement pour mieux voir les informations dans l'éditeur de données.
■ Il est possible de trier les données soit par ordre croissant, soit par ordre décroissant. Notez que par défaut, les données sont triées par ordre **croissant**.
■ Le tri peut être basé sur un ou plusieurs critères.

Note :
■ Pour annuler le tri :
✓ Nous ne pouvons pas annuler le tri des données.
✓ Une astuce consiste à trier à nouveau les données en fonction de la variable « **Patient** » (répétez la même procédure vue ci-dessus juste nous sélectionnons la variable **Patient** plutôt que **Weight**).

4.6 Recoder les variables

En réencodant les variables, nous pouvons regrouper un ensemble de valeurs dans certaines catégories prédéfinies en fonction du type de données. Ce processus permet d'économiser des efforts et du temps pour reconstruire les données collectées d'une meilleure manière.

Par exemple, supposons que nous voulions ré-encoder la variable **Weight**, où nous regrouperons chaque ensemble de valeurs comme suit :

1. Featherweight : [0-50]
2. Middleweight :]50-90]
3. Heavyweight :]90-120]
4. Super Heavyweight : >120

Voici comment le faire dans SPSS :

1. Choisissez **Transformer ➜ Création de variables** : la boîte de dialogue **Création de variables** apparaît.
2. Cliquez sur le bouton fléché pour déplacer la variable **Weight** vers la zone de travail à droite.
3. Nommez la nouvelle variable de sortie dans la case à droite comme **Weight-Cat ➜** Cliquez sur le bouton **Changer** pour enregistrer le nouveau nom de variable, comme le montre la Figure 4.7.

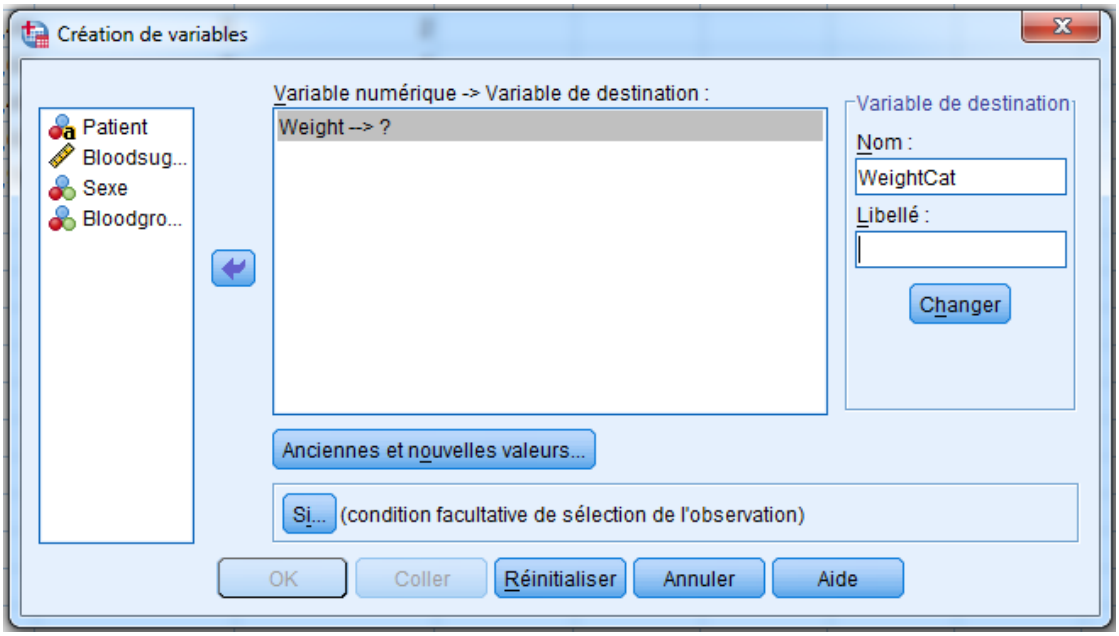


FIGURE 4.7 – Boîte de dialogue : Création de variables

4. Cliquez sur le bouton **Anciennes et nouvelles valeurs**.
5. Pour les catégories 1, 2 et 3 : on sélectionne le bouton radio **Plage ➜** Entrer les valeurs Min et Max **➜** A côté du bouton radio **Valeur** : on entre le numéro de la catégorie **➜** Cliquer sur le bouton **Ajouter** (Voir Figure 4.8).
6. Pour la catégorie 4 : on sélectionne le bouton radio **Plage, de la valeur au MAXIMUM ➜** Entrer la valeur Min (c.a.d. 120) **➜** A côté du bouton radio **Valeur** : on entre le numéro de la catégorie (c.a.d. 4) **➜** Cliquer sur le bouton **Ajouter** (Voir Figure 4.8).

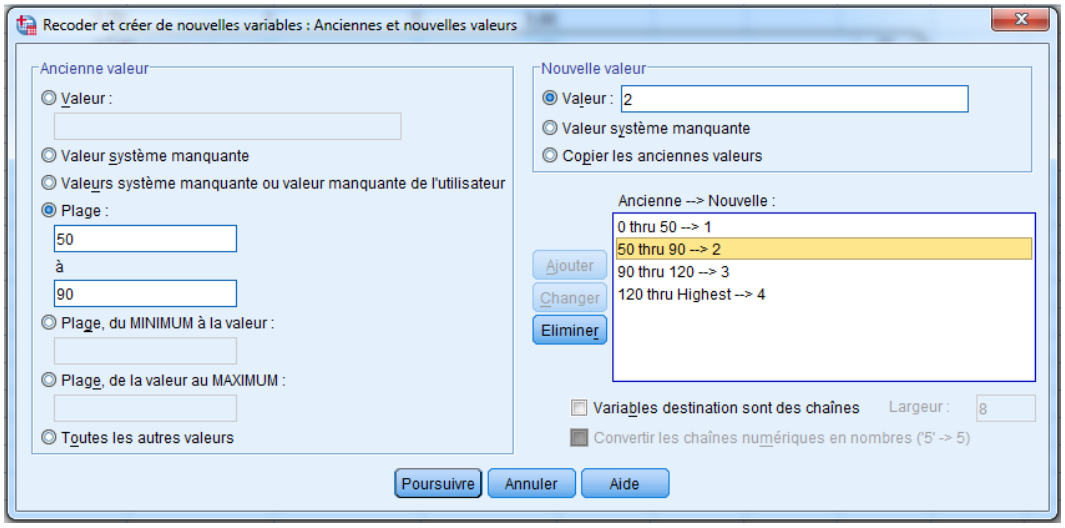


FIGURE 4.8 – Boîte de dialogue : Anciennes et nouvelles valeurs

7. Cliquez sur **Poursuivre**, puis sur **OK**. La figure 4.9 ci-dessous montre le résultat final.

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup	WeightCat
1	P1	53,00	1,20	1	1	2,00
2	P2	90,20	1,50	1	3	3,00
3	P3	88,00	1,40	2	2	2,00
4	P4	45,00	,60	1	4	1,00
5	P5	90,00	2,40	2	2	2,00
6	P6	175,00	3,60	2	1	4,00

FIGURE 4.9 – Résultat après le Recodage

Il sera utile de dire à quiconque regarde votre sortie ce que représentent ces valeurs recodées. Pour ce faire, cliquez sur l'onglet **Vue des variables** en bas de la feuille de calcul, puis cliquez dans la zone Values (sur la nouvelle variable **WeightCat**) et ajoutez des étiquettes de valeur comme indiqué précédemment (c.a.d. 1=Featherweight, 2=Middleweight, 3=Heavyweight, 4=Super Heavyweight).

4.7 Supprimer une Variable ou une Observation

➤ Supposons que nous voulions supprimer la nouvelle variable **WeightCat**. Pour supprimer une variable dans la vue des données, cliquez sur le nom de la variable et appuyez sur la touche **Suppr** du clavier, ou cliquez avec le bouton droit sur le nom de la variable et appuyez sur **Effacer**.

➤ Pour supprimer une observation (une ligne entière de données), suivez ces étapes : Cliquez sur le numéro d'observation à gauche (la ligne entière sera mise en surbrillance), Appuyez sur **Suppr** sur le clavier, ou cliquez avec le bouton droit sur le numéro d'observation et appuyez sur **Effacer**.

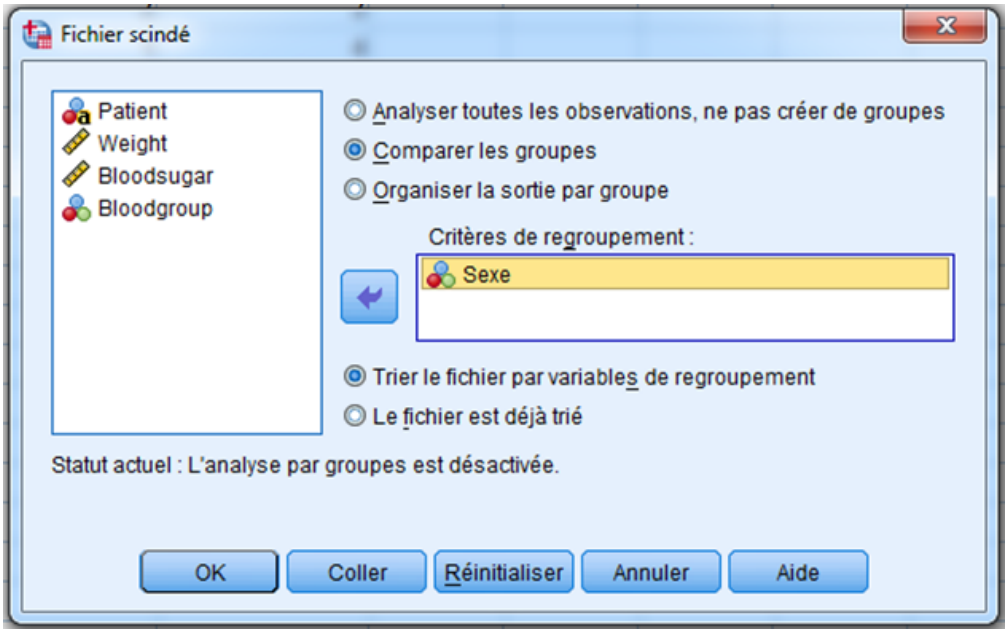


FIGURE 4.10 – Fichier scindé

4.8 Fractionnement des Données

Parfois, dans les études comparatives, nous divisons les données en plusieurs groupes en fonction de certains critères, puis effectuons nos analyses sur chaque sous-groupe séparément. Le fractionnement des données permet de réaliser ce processus.

1. Choisissez **Données ➜ Scinder un fichier** : la boîte de dialogue **Fichier scindé** apparaît.

2. Sélectionnez le bouton radio **Comparer les groupes**.

3. Choisissez **Sexe** comme variable de comparaison et cliquez sur **OK** (voir la Figure 4.10).

   Note : Après ce processus, nous remarquons qu'il n'y a pas de changement significatif, sauf que les données ont été à nouveau organisées en fonction du sexe des patients.

4. Choisissez **Analyse ➜ Statistiques descriptives ➜ Fréquences**.

5. Choisissez **Bloodgroup** et placez-la dans la zone Variable(s).

6. Cliquez sur **OK** : la sortie résultante, illustrée dans la Figure 4.11, est groupée par **Sexe**.

Fréquences

Statistiques

Bloodgroup

Homme	N	Valide	3
		Manquant	0
Femme	N	Valide	3
		Manquant	0

Bloodgroup

Sexe			Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Homme	Valide	AB	1	33,3	33,3	33,3
		B	1	33,3	33,3	66,7
		O	1	33,3	33,3	100,0
		Total	3	100,0	100,0	
Femme	Valide	AB	1	33,3	33,3	33,3
		A	2	66,7	66,7	100,0
		Total	3	100,0	100,0	

FIGURE 4.11 – Résultat des fréquences

Notes :
■ Le fractionnement peut être basé sur un ou plusieurs critères.
■ Pour annuler le fractionnement, vous devez :

1. Choisissez **Données ➜ Scinder un fichier**.
2. Sélectionnez le bouton radio **Analyser toutes les observations, ne pas créer de groupes**, puis cliquez sur le bouton **OK** (ou appuyez sur le bouton **Réinitialiser ➜ OK**).

4.9 Sélection des Données

4.9.1 Condition Logique Simple

1. Choisissez **Données ➜ Sélectionner des observations** : la boîte de dialogue Sélectionner les observations apparaît, comme illustré à la Figure 4.12.
2. Sélectionnez le bouton radio **Selon une condition logique**, puis cliquez sur le bouton **Si**... : Vous pouvez maintenant spécifier les critères de sélection (voir Figure 4.13).
3. Déplacez la variable **Sexe** de la liste de gauche vers la zone d'expression (en haut à gauche) : vous pouvez déplacer la variable soit en la faisant glisser, soit en la sélectionnant puis en cliquant sur le bouton fléché.
4. Utilisez votre clavier ou le clavier numérique à l'écran pour saisir **=2** dans la zone d'expression.

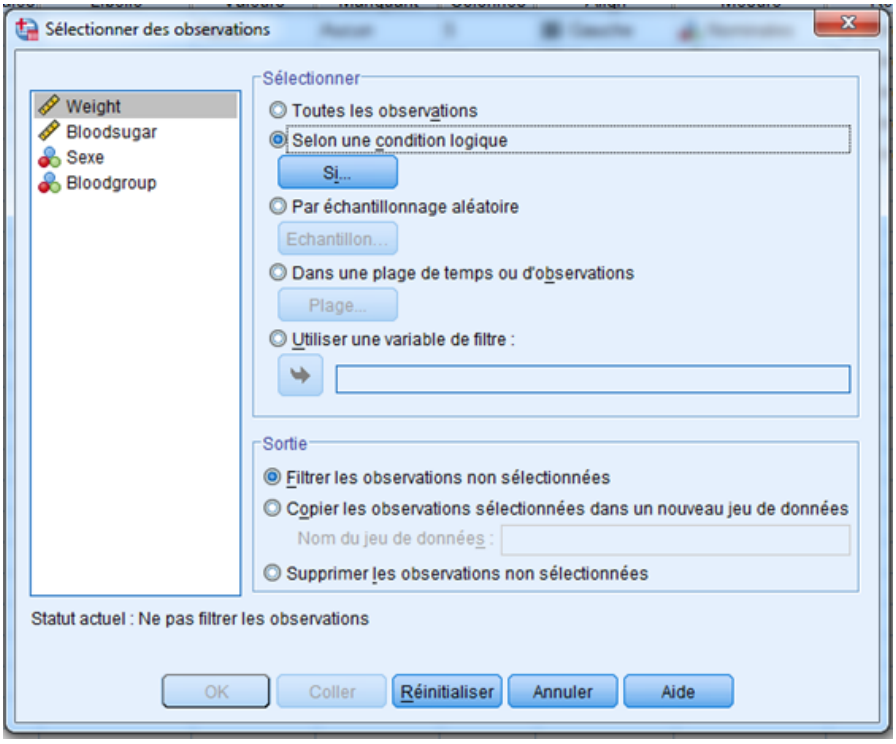


FIGURE 4.12 – Sélectionner des observations

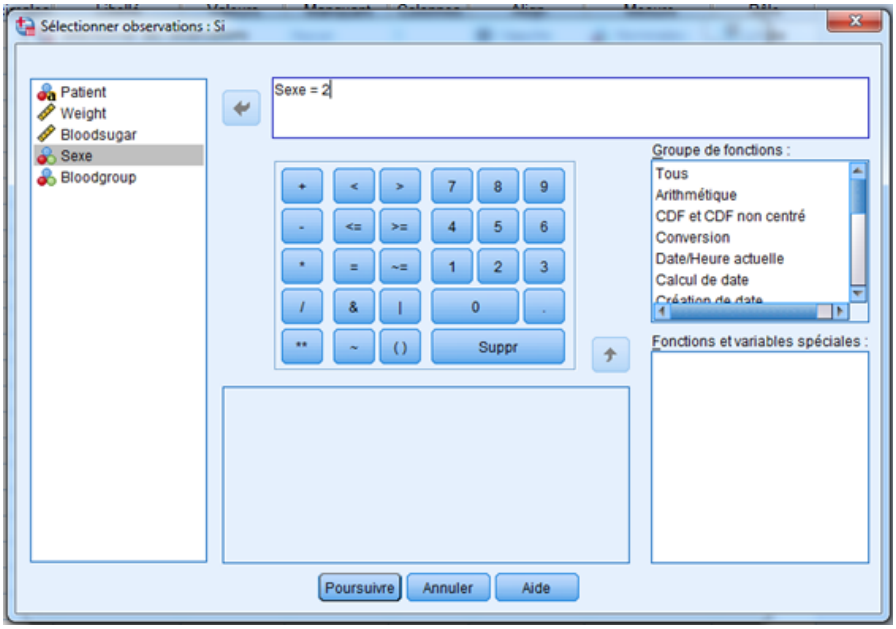


FIGURE 4.13 – Expression conditionnelle

5. Cliquez sur **Poursuivre**, puis sur **OK**. La figure ci-dessous montre le résultat final. Les barres obliques sur certains ID de ligne (dans la première colonne) indiquent que les patients (**hommes**) sont ignorés (pour le moment) et que seules les patientes (**femmes**) sont analysées. La variable $filter\_\$$ est également créée et comprend 0 et 1 pour les cas non sélectionnés et les cas sélectionnés, respectivement (voir Figure 4.14).

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup	filter_$
1	P1	53,00	1,20	1	1	0
2	P2	90,20	1,50	1	3	0
3	P3	88,00	1,40	2	2	1
4	P4	45,00	,60	1	4	0
5	P5	90,00	2,40	2	2	1
6	P6	175,00	3,60	2	1	1

FIGURE 4.14 – Résultat après la sélection

Pour voir les effets de la sélection :

(a) Choisissez **Analyse ➜ Statistiques descriptives ➜ Fréquences**.

(b) Choisissez **Bloodgroup** et placez-la dans la zone Variable(s).

(c) Cliquez sur **OK** : la sortie résultante, illustrée dans la figure ci-dessous (Figure 4.15), SPSS affiche uniquement les résultats des groupes sanguins des patientes Femmes.

Fréquences

Statistiques

Bloodgroup

N	Valide	3
	Manquant	0

Bloodgroup

		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	AB	1	33,3	33,3	33,3
	A	2	66,7	66,7	100,0
	Total	3	100,0	100,0	

FIGURE 4.15 – Résultat des fréquences

4.9.2 Condition Logique Complexe

Conjonction

Table de vérité de la conjonction (**& : ET**) (T : True, F : False)

P	Q	P ∧ Q
T	T	T
T	F	F
F	T	F
F	F	F

Disjonction

Table de vérité de la disjonction (**| : OU**) (T : True, F : False)

P	Q	P ∨ Q
T	T	T
T	F	T
F	T	T
F	F	F

TABLE 4.1 – Conjonction et disjonction logique

➤ Essayons maintenant de sélectionner des hommes qui ont un poids supérieur ou égal à 50 ou des patients qui ont un groupe sanguin A.
➤ Taper la condition suivante : **(Sexe = 1 & Weight >= 50) | Bloodgroup = 2**, (voir Figure 4.16) (Les parenthèses entre les conditions sont très importantes).

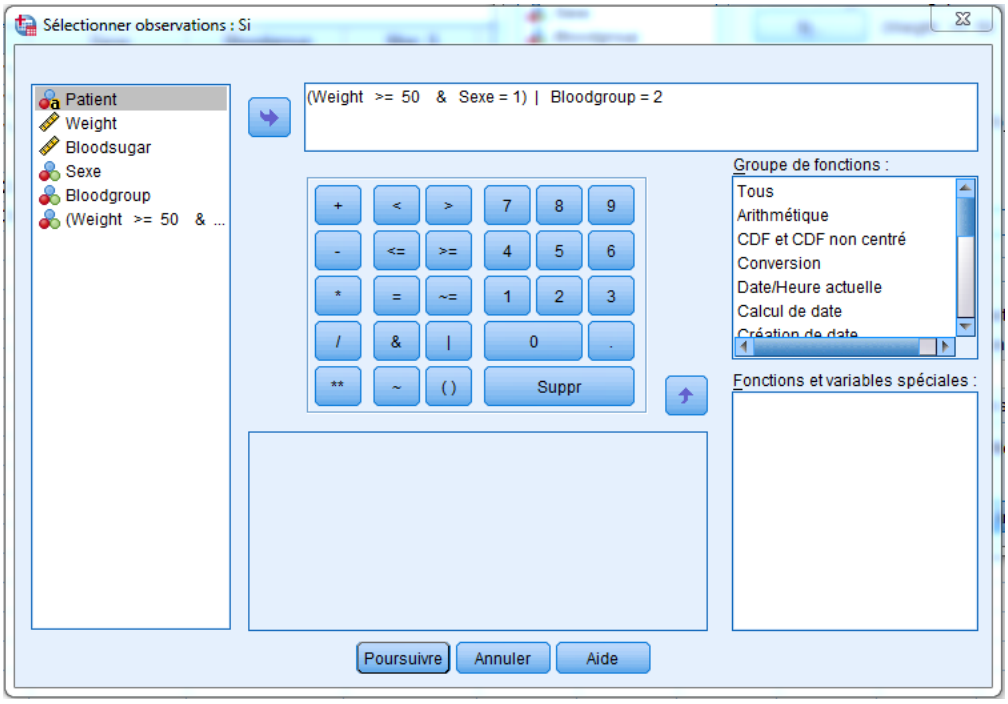


FIGURE 4.16 – Expression conditionnelle

Le résultat de la sélection sera :

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup	filter_$
1	P1	53,00	1,20	Homme	AB	Selected
2	P2	90,20	1,50	Homme	B	Selected
3	P3	88,00	1,40	Femme	A	Selected
~~4~~	P4	45,00	,60	Homme	O	Not Selected
5	P5	90,00	2,40	Femme	A	Selected
~~6~~	P6	175,00	3,60	Femme	AB	Not Selected

FIGURE 4.17 – Résultat après la sélection

Notes :

■ Pour annuler la sélection, vous devez :

1. Choisissez **Données ➜ Sélectionner des observations**.
2. Sélectionnez le bouton radio **Toutes les observations**, puis cliquez sur le bouton **OK** (ou appuyez sur le bouton **Réinitialiser ➜ OK**).

■ La sélection peut être basée sur un ou plusieurs critères.

4.10 Conclusion

Le prétraitement des données est une étape importante dans l'analyse des données, et SPSS fournit plusieurs outils et techniques pour un prétraitement efficace des données. SPSS peut aider à améliorer la qualité, la précision et la compatibilité des données avec les techniques d'analyse et les outils logiciels, et peut faciliter la création d'informations significatives à partir des données.

Chapitre 5

Analyse des Données

5.1 Introduction

Dans l'analyse des données biomédicales, les variables qualitatives et continues sont souvent analysées ensemble pour mieux comprendre des phénomènes complexes, tels que les facteurs de risque d'une maladie particulière ou l'efficacité d'un traitement médical. Par conséquent, il est important d'avoir une compréhension approfondie des méthodes statistiques pour les deux types de variables afin d'interpréter avec précision les données biomédicales et de prendre des décisions éclairées.

5.2 Collection de Données dans SPSS

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P1	63,00	1,20	1	1
2	P2	90,20	1,70	1	3
3	P3	88,00	1,40	2	2
4	P4	45,00	,60	1	4
5	P5	90,00	2,40	2	2
6	P6	175,00	3,60	2	1
7	P7	63,00	2,10	1	2

FIGURE 5.1 – Fichier de données SPSS

1. Télécharger le fichier de données SPSS appelé "**DataSPSS5.sav**" sur : `https://aboulesnane.net/wp-content/datafiles/DataSPSS5.sav`
2. Les données contiennent cinq variables nommées : Patient, Weight, Bloodsugar, Sexe, et Bloodgroup (voir la Figure 5.1).
   (a) La Variable "**Patient**" est une variable de type **Chaîne**.
   (b) La Variable "**Weight**" est une variable quantitative continue de type Numérique.

(c) La Variable "**Bloodsugar**" est une variable quantitative continue de type Numérique.

(d) Les valeurs possibles pour la variable qualitative "**Sexe**" : 1=Homme et 2=Femme.

(e) Les valeurs possibles pour la variable qualitative "**Bloodgroup**" : 1=AB , 2=A, 3=B et 4=O.

5.3 Prétraitement des Données dans SPSS

■ Les données des patients ont été triées en fonction de leurs valeurs de glycémie (**Bloodsugar**) et de poids (**Weight**). (voir section 4.5).

Le résultat du tri est :

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P4	45,00	,60	1	4
2	P7	60,00	1,20	1	2
3	P1	63,00	1,20	1	1
4	P3	88,00	1,40	2	2
5	P2	90,20	1,70	1	3
6	P5	90,00	2,40	2	2
7	P6	175,00	3,60	2	1

FIGURE 5.2 – Résultat après le tri

■ Les données des patients ont été fractionnées en fonction de leurs valeurs de groupe sanguin (**Bloodgroup**). (voir section 4.8).

Le résultat du fractionnement est :

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P1	63,00	1,20	1	1
2	P6	175,00	3,60	2	1
3	P7	60,00	1,20	1	2
4	P3	88,00	1,40	2	2
5	P5	90,00	2,40	2	2
6	P2	90,20	1,70	1	3
7	P4	45,00	,60	1	4

FIGURE 5.3 – Résultat après le fractionnement

5.4 Utilisation des Statistiques Descriptives

5.4.1 Fréquences pour les variables catégorielles (qualitatives)

■ La technique la plus courante pour décrire les données catégorielles - niveaux de mesure nominaux et ordinaux - consiste à demander **un tableau de fréquence**, qui fournit un résumé indiquant le nombre et le pourcentage de cas entrant dans chaque catégorie d'une variable. Les utilisateurs peuvent également demander des statistiques récapitulatives supplémentaires telles que le mode ou la médiane, entre autres.
■ Voici comment exécuter la procédure des fréquences afin de pouvoir créer un tableau des fréquences qui vous permettra d'obtenir des statistiques récapitulatives pour les variables qualitatives :

1. Choisissez **Analyse ➜ Statistiques descriptives ➜ Fréquences** : La boîte de dialogue Fréquences apparaît. Dans cet exemple, et basant sur le **fractionnement des données précédent**, vous souhaitez étudier la distribution de variable **Sexe** (Homme, Femme) pour chaque valeur de **Bloodgroup** (AB, A, B, O).

2. Sélectionnez la variables **Sexe**, et placez-la dans la zone Variable(s), comme illustrée à la Figure 5.4.

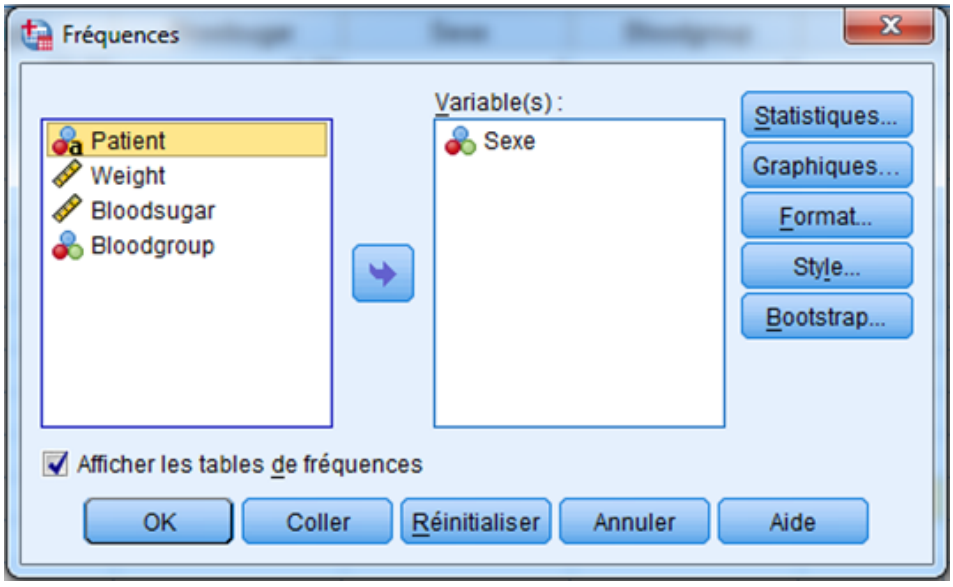


FIGURE 5.4 – Boîte de dialogue Fréquences

3. Cliquez sur le bouton **Statistiques** : la boîte de dialogue **Fréquences : Statistiques** s'affiche (voir Figure 5.5).
4. Dans la section **Tendance centrale**, cochez la case **Mode**, comme illustrée dans la figure ci-dessous. Cette boîte de dialogue fournit de nombreuses statistiques, mais il est essentiel que vous demandiez uniquement celles qui correspondent au niveau de mesure des variables que vous avez placées dans la zone Variable(s). Pour les variables nominales, la seule statistique appropriée est le **mode**.

5. Cliquez sur **Poursuivre**.

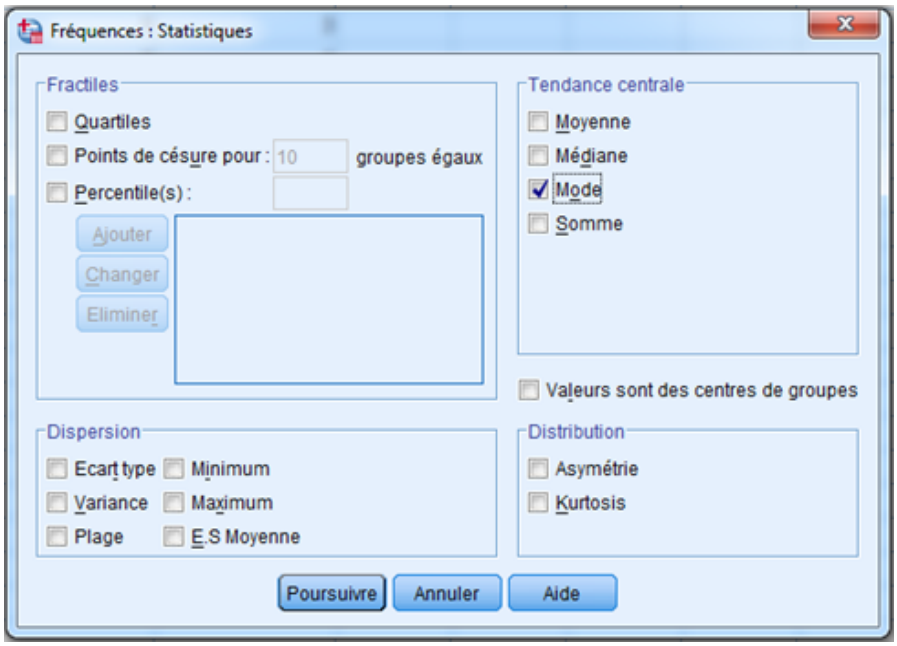


FIGURE 5.5 – Boîte de dialogue Fréquences :Statistiques

6. Cliquez sur le bouton **Graphiques** : la boîte de dialogue **Fréquences : Graphiques** s'affiche.
7. Dans la section **Type de graphique**, sélectionnez le bouton radio **Graphiques à barres** ; dans la section Valeurs du graphique, sélectionnez le bouton radio **Pourcentages** (voir Figure 5.6).

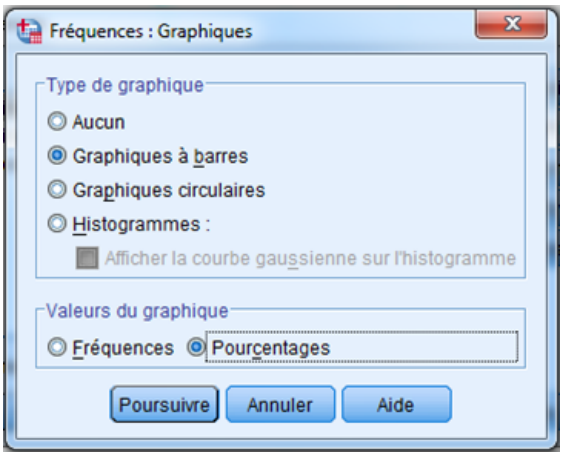


FIGURE 5.6 – Boîte de dialogue Fréquences :Graphiques

8. Cliquez sur **Poursuivre**, puis sur **OK** : SPSS exécute la procédure des fréquences et calcule les statistiques récapitulatives, le tableau des fréquences et le graphique à barres que vous avez demandés.
9. La sortie résultante, illustrée dans les figures ci-dessous, est groupée par **Bloodgroup** (voir Figure 5.7).

Fréquences

Statistiques

Sexe

AB	N	Valide	2
		Manquant	0
	Mode		1[a]
A	N	Valide	3
		Manquant	0
	Mode		2
B	N	Valide	1
		Manquant	0
	Mode		1
O	N	Valide	1
		Manquant	0
	Mode		1

a. Présence de plusieurs modes. La plus petite valeur est affichée.

Fréquences absolues

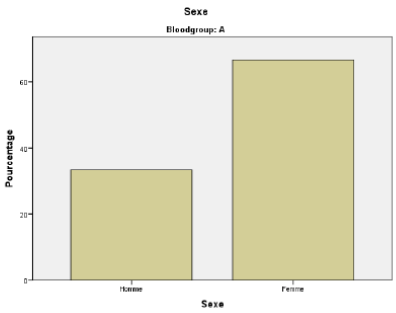
Fréquences relatives (%)

Fréquences cumulés (%)

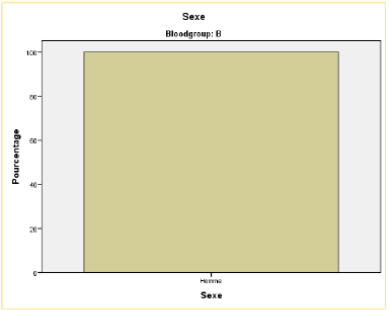
Sexe

Bloodgroup			Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
AB	Valide	Homme	1	50,0	50,0	50,0
		Femme	1	50,0	50,0	100,0
		Total	2	100,0	100,0	
A	Valide	Homme	1	33,3	33,3	33,3
		Femme	2	66,7	66,7	100,0
		Total	3	100,0	100,0	
B	Valide	Homme	1	100,0	100,0	100,0
O	Valide	Homme	1	100,0	100,0	100,0

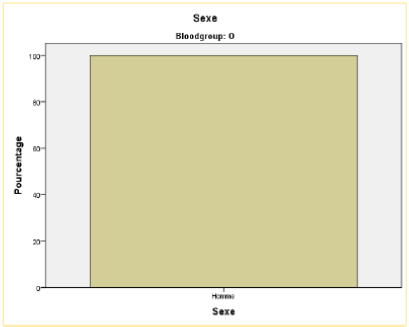
(a)



(b)



(c)



(d)

FIGURE 5.7 – Résultat de l'analyse

5.4.2 Fréquences pour les variables continues

■ Comme vous l'avez vu, les **tableaux de fréquence** affichent des nombres et des pourcentages, ce qui est extrêmement utile lorsque vous travaillez avec des variables qualitatives. Cependant, pour les variables continues, qui peuvent avoir de nombreuses valeurs, les tables de fréquences deviennent moins utiles.
■ Pour exécuter des fréquences pour des variables continues, procédez comme suit :

1. **Annulez le fractionnement des données**.
2. Choisissez **Analyse ➜ Statistiques descriptives ➜ Fréquences**.
3. Sélectionnez les variables **Weight**, et **Bloodsugar**, et placez-les dans la zone Variable(s).
4. Décochez la case **Afficher les tables de fréquences**, comme illustré dans la Figure 5.8 ci-dessous.

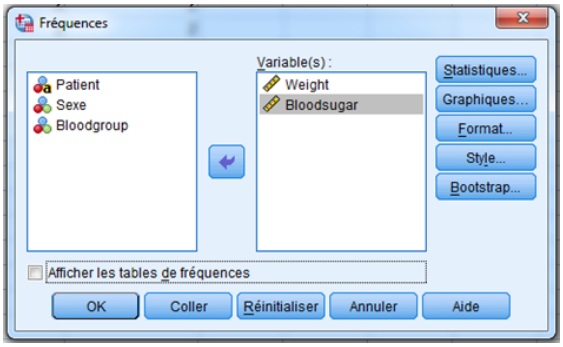


FIGURE 5.8 – Boîte de dialogue Fréquences

5. Cliquez sur le bouton **Statistiques**.
6. Dans la section **Tendance centrale**, cochez les cases **Moyenne, Médiane** et **Mode**. Dans la section **Dispersion**, sélectionnez **Ecart type, Variance, Minimum et Maximum**.
7. Cliquez sur **Poursuivre**.
8. Cliquez sur le bouton **Graphiques**.
9. Sélectionnez le bouton radio **Histogrammes** et cochez la case **Afficher la courbe gaussienne sur l'histogramme**, comme illustré dans la Figure 5.9.

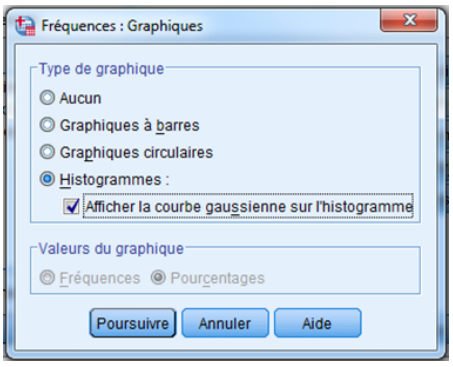


FIGURE 5.9 – Boîte de dialogue Fréquences :Graphiques

10. Cliquez sur **Poursuivre**, puis sur **OK** : SPSS exécute la procédure des fréquences et calcule les statistiques récapitulatives et l'histogramme que vous avez demandés.
11. La sortie résultante, illustrée dans la Figure 5.10.

Fréquences

Statistiques

		Weight	Bloodsugar
N	Valide	7	7
	Manquant	0	0
Moyenne		87,3143	1,7286
Médiane		88,0000	1,4000
Mode		45,00[a]	1,20
Ecart type		42,49029	,99115
Variance		1805,425	,982
Minimum		45,00	,60
Maximum		175,00	3,60

a. Présence de plusieurs modes. La plus petite valeur est affichée.

Histogramme

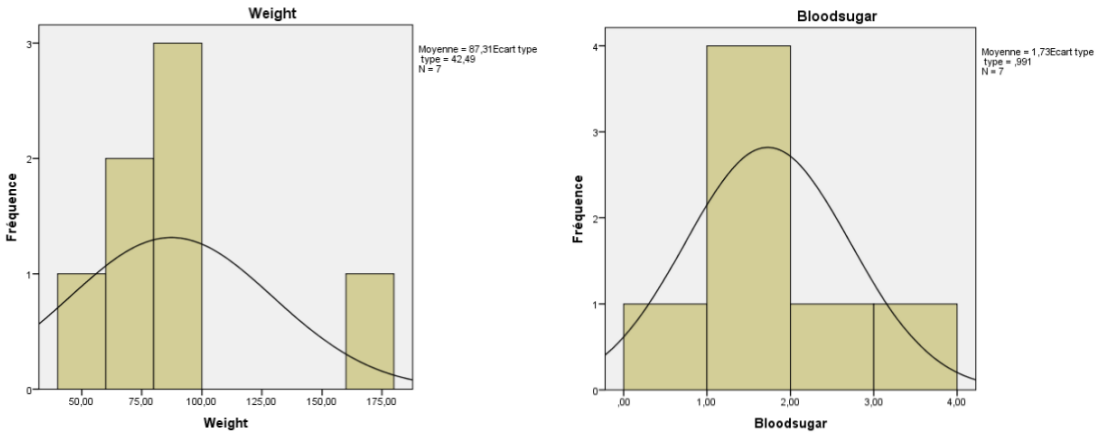


FIGURE 5.10 – Résultat de l'analyse

5.4.3 Résumer des variables continues avec la procédure descriptive

■ La procédure descriptive fournit un résumé succinct de diverses statistiques et le nombre de cas avec des valeurs valides pour chaque variable incluse dans le tableau.

■ Pour utiliser la procédure des descriptifs, procédez comme suit :

1. Choisissez **Analyse ➜ Statistiques descriptives ➜ Descriptives** : la boîte de dialogue Descriptives s'affiche.

2. Sélectionnez les variables **Weight**, et **Bloodsugar**, et placez-les dans la zone Variable(s).
3. Cliquez sur **OK** : SPSS exécute la procédure descriptive et calcule les statistiques récapitulatives comme indiqué ci-dessous (Figure 5.11).

Statistiques descriptives

	N	Minimum	Maximum	Moyenne	Ecart type
Weight	7	45,00	175,00	87,3143	42,49029
Bloodsugar	7	,60	3,60	1,7286	,99115
N valide (liste)	7				

FIGURE 5.11 – Résultat de statistiques descriptives

5.5 Conclusion

SPSS fournit une gamme de fonctions et d'outils qui permettent aux chercheurs d'analyser à la fois les variables qualitatives et continues dans les données biomédicales. Les chercheurs peuvent importer leurs données dans SPSS et utiliser la fonction "Descriptives" pour obtenir des statistiques descriptives pour les variables continues, tandis que la fonction "Fréquences" peut être utilisée pour analyser les deux types de variables (qualitatives et quantitatives).

Chapitre 6

Analyse des Relations entre les Variables Statistiques

6.1 Introduction

L'analyse des données biomédicales peut être utilisée pour étudier les relations entre les variables dans divers contextes, comme l'étude des facteurs de risque d'une maladie ou l'évaluation de l'efficacité d'un traitement médical. L'analyse des relations entre les variables consiste à examiner l'association ou la corrélation entre différentes variables, qui peut être catégorique ou continue. Des méthodes statistiques telles que les tableaux croisés et l'analyse de régression peuvent être utilisées pour identifier et quantifier les relations entre les variables qualitatives et continues, respectivement. De plus, l'analyse de corrélation peut être utilisée pour identifier la force et la direction de la relation entre deux variables continues. En comprenant les relations entre les variables, les chercheurs peuvent mieux comprendre les facteurs sous-jacents qui contribuent à un résultat ou à un phénomène particulier et prendre des décisions éclairées sur la base de leur analyse.

6.2 Collection de Données dans SPSS

1. Télécharger le fichier de données SPSS appelé "**DataSPSS6.sav**" sur : `https://aboulesnane.net/wp-content/datafiles/DataSPSS6.sav`
2. Les données contiennent cinq variables nommées : Patient, Weight, Bloodsugar, LungCancer, et Smoking (voir la Figure 6.1).
   (a) La Variable "**Patient**" est une variable de type **Chaîne**.
   (b) La Variable "**Weight**" est une variable quantitative continue de type Numérique.
   (c) La Variable "**Bloodsugar**" est une variable quantitative continue de type Numérique.
   (d) Les valeurs possibles pour la variable qualitative "**LungCancer**" : 0=Non et 1=Oui.
   (e) Les valeurs possibles pour la variable qualitative "**Smoking**" : 1=Non , 2=Parfois, 3=Beaucoup.

	Patient	Weight	Bloodsugar	LungCancer	Smoking
1	P1	63,00	1,20	1	2
2	P2	90,20	1,70	1	3
3	P3	88,00	1,40	0	2
4	P4	45,00	,60	1	3
5	P5	90,00	2,40	0	2
6	P6	175,00	3,60	0	1
7	P7	60,00	1,20	1	2
8	P8	120,00	1,92	0	1
9	P9	55,00	,70	0	1
10	P10	160,00	4,62	1	3

FIGURE 6.1 – Fichier de données SPSS

6.3 Prétraitement des données dans SPSS

■ Les données des patients ont été triées en fonction de leurs valeurs de poids (**Weight**). (voir section 4.5).

Le résultat du tri est :

	Patient	Weight	Bloodsugar	LungCancer	Smoking
1	P4	45,00	,60	1	3
2	P9	55,00	,70	0	1
3	P7	60,00	1,20	1	2
4	P1	63,00	1,20	1	2
5	P3	88,00	1,40	0	2
6	P5	90,00	2,40	0	2
7	P2	90,20	1,70	1	3
8	P8	120,00	1,92	0	1
9	P10	160,00	4,62	1	3
10	P6	175,00	3,60	0	1

FIGURE 6.2 – Résultat après le tri

6.4 Distributions Statistiques à Deux Caractères

6.4.1 Relations entre variables catégorielles (qualitatives)

L'un des moyens les plus courants d'analyser les données consiste à utiliser des tableaux croisés. Comme mentionné, vous utilisez un tableau croisé lorsque vous souhaitez étudier la relation entre deux ou plusieurs variables catégorielles. Par exemple, vous voudrez peut-être examiner l'impact (la relation) de la cigarette sur le cancer du poumon.

		Variables Independantes	
	Variables	**Qualitatives**	**Quantitative**
Variables Dependantes	**Qualitatives**	Tableaux Croises, Tests Non Parametriques	Regression Logistique, Analyse Discriminante
	Quantitative	T-Test, ANOVA	Corrélation, Régression Linéaire

TABLE 6.1 – Analyses des relations statistiques

Voici comment effectuer un tableau croisé à partir de nos données (entre la variable **LungCancer** et **Smoking**) :

1. Choisissez **Analyser ➜ Statistiques descriptives ➜ Tableaux croisés** : La boîte de dialogue Tableaux croisés s'affiche.
   Bien que vous puissiez placer les variables dans les zones Lignes ou Colonnes, il est habituel de placer la variable indépendante dans la colonne du tableau croisé. Dans les analyses bivariées, la variable indépendante est celle qui, en théorie, exerce une influence sur l'autre variable, appelée variable dépendante.

   La variable indépendante dans cette étude est l'acte de fumer, car on affirme que fumer a un impact direct sur le développement du cancer du poumon :
2. Sélectionnez la variable **LungCancer** et placez-la dans la zone Ligne(s).
3. Sélectionnez la variable **Smoking** et placez-la dans la zone Column(s), comme illustré à la Figure ci-dessous (Figure 6.3).

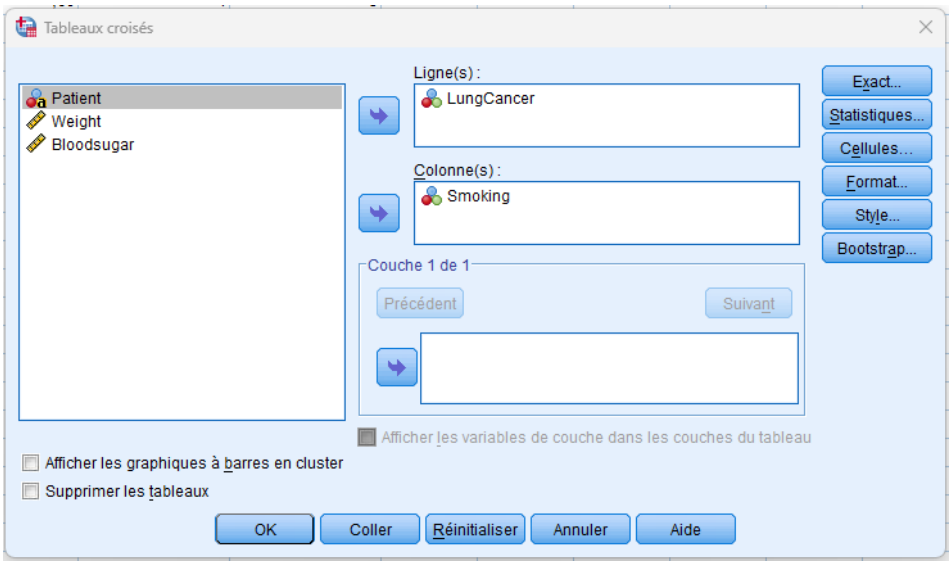
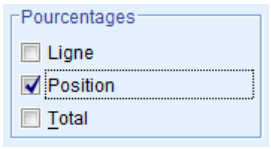


FIGURE 6.3 – Boîte de dialogue Tableaux croisés

4. Cliquez sur le bouton **Cellules** : sélectionnez les pourcentages de ligne, les pourcentages de colonne ou les deux.



5. Cliquez sur Poursuivre, puis sur **OK**.
6. La sortie résultante, illustrée dans les figures ci-dessous (Figure 6.4).

Tableaux croisés

Récapitulatif de traitement des observations

	Observations					
	Valide		Manquant		Total	
	N	Pourcentage	N	Pourcentage	N	Pourcentage
LungCancer * Smoking	10	100,0%	0	0,0%	10	100,0%

Tableau croisé LungCancer * Smoking

			Smoking			Total
			Non	Parfois	Beaucoup	
LungCancer	Non	Effectif	3	2	0	5
		% dans Smoking	100,0%	50,0%	0,0%	50,0%
	Oui	Effectif	0	2	3	5
		% dans Smoking	0,0%	50,0%	100,0%	50,0%
Total		Effectif	3	4	3	10
		% dans Smoking	100,0%	100,0%	100,0%	100,0%

FIGURE 6.4 – Résultat des tableaux croisés

6.4.2 Relations entre variables quantitatives

Les deux techniques statistiques les plus couramment utilisées pour analyser les relations entre les variables continues sont la corrélation de Pearson et la régression linéaire.
De nombreuses personnes utilisent le terme corrélation pour désigner l'idée d'une relation entre des variables dans un modèle. En d'autres termes, les variables sont corrélées entre elles parce que les changements d'une variable affectent l'autre.
Alors que la corrélation essaie simplement de déterminer si deux variables sont liées, la régression linéaire va encore plus loin et tente de prédire les valeurs d'une variable en fonction d'une autre variable.

❖ Exécution de la procédure bivariée

Le coefficient de corrélation de Pearson est une mesure de la mesure dans laquelle il existe une relation linéaire (ligne droite) entre deux variables. Il a des valeurs comprises entre -1 et $+1$, de sorte que plus la valeur absolue est grande, plus la corrélation est forte. Par exemple, une corrélation de $+1$ indique

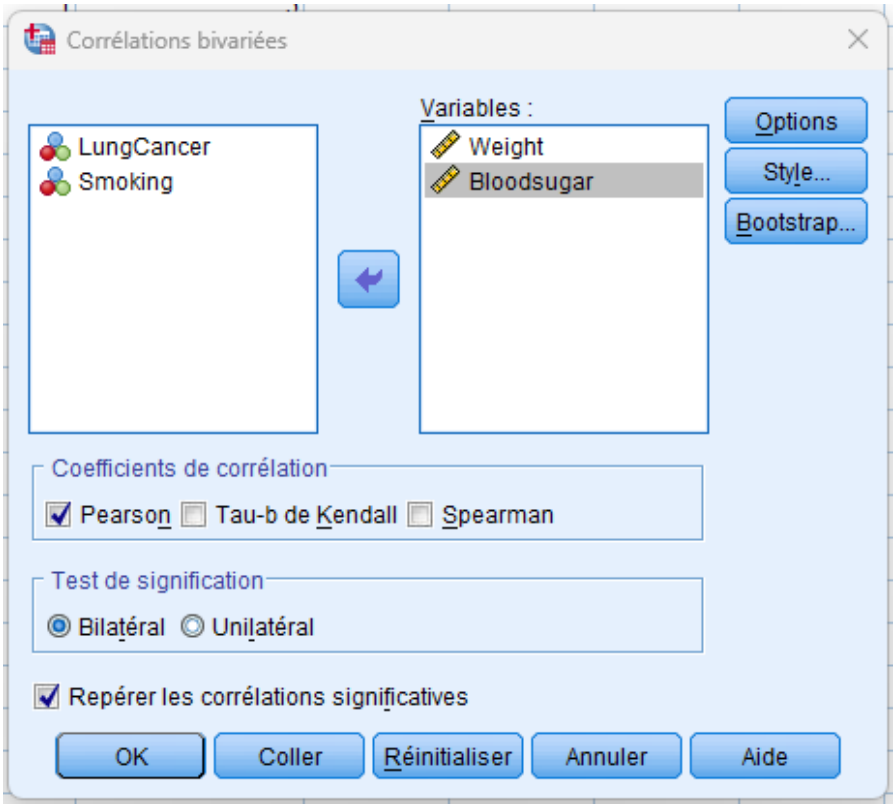
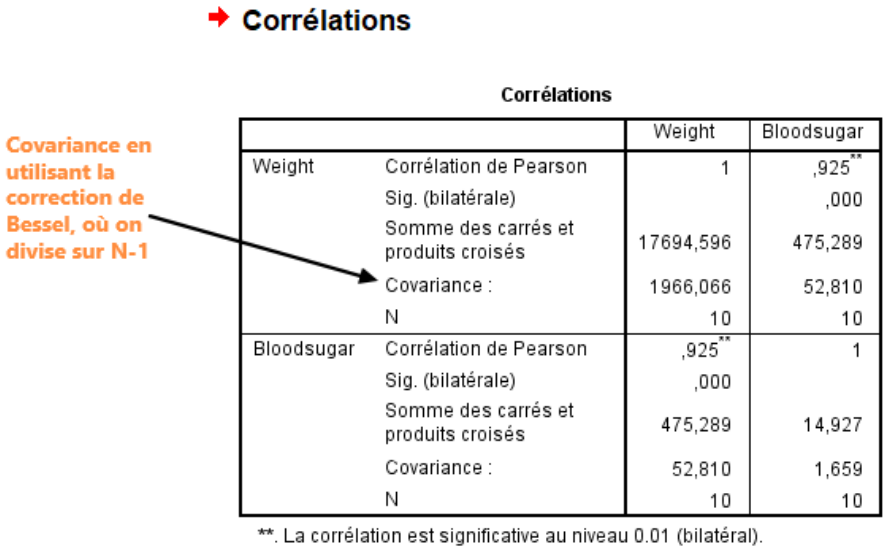


FIGURE 6.5 – Corrélations Bivariées

que les données tombent sur une ligne droite parfaite inclinée vers le haut (relation positive), et une corrélation de -1 représente les données formant une ligne droite parfaite inclinée vers le bas (relation négative). Une corrélation de 0 indique qu'il n'existe aucune relation linéaire.
Pour tester une corrélation, procédez comme suit :

1. Choisissez **Analyse ➜ Corrélation ➜ Bivarié**. La boîte de dialogue **Corrélations bivariées** s'affiche.
   Dans cet exemple, vous étudierez si la glycémie est liée au poids. Notez qu'il n'y a pas de désignation de variables dépendantes et indépendantes. Les corrélations seront calculées sur toutes les paires de variables.
2. Sélectionnez les variables **Weight** et **Bloodsugar** et placez-les dans la zone Variables, comme illustré à la Figure 6.5 ci-dessus.
3. Cliquez sur le bouton **Options** et cochez l'option « **Ecarts des produits croisés et covariances** »
4. Cliquez sur **Poursuivre**, puis sur **OK**.

5. La sortie résultante, illustrée dans la Figure 6.6 ci-dessous.



Corrélations

		Weight	Bloodsugar
Weight	Corrélation de Pearson	1	,925**
	Sig. (bilatérale)		,000
	Somme des carrés et produits croisés	17694,596	475,289
	Covariance :	1966,066	52,810
	N	10	10
Bloodsugar	Corrélation de Pearson	,925**	1
	Sig. (bilatérale)	,000	
	Somme des carrés et produits croisés	475,289	14,927
	Covariance :	52,810	1,659
	N	10	10

**. La corrélation est significative au niveau 0.01 (bilatéral).

FIGURE 6.6 – Tableau des corrélations

Dans cet exemple, vous avez une très forte corrélation positive (0,925) qui est statistiquement significative entre la glycémie et le poids.
D'une autre manière, du même tableau, nous pouvons calculer le coefficient de corrélation à partir de la matrice de covariance, comme suit :

$$r = \frac{XYcovariance}{\sqrt{Xvariance}\sqrt{Yvariance}}$$

$$r = \frac{\left(\frac{\sum(X-\bar{X})(Y-\bar{Y})}{N}\right)}{\sqrt{\frac{\sum(X-\bar{X})(X-\bar{X})}{N}} * \sqrt{\frac{\sum(Y-\bar{Y})(Y-\bar{Y})}{N}}}$$

$$r = \frac{\left(\frac{475{,}289}{10}\right)}{\sqrt{\frac{17694{,}596}{10}} * \sqrt{\frac{14{,}927}{10}}} = 0.925$$

❖ Exécution de la procédure de régression linéaire simple

Les corrélations vous permettent de déterminer si deux variables continues sont linéairement liées l'une à l'autre. L'analyse de régression consiste à prédire le futur (l'inconnu) sur la base de données recueillies dans le passé (le connu). La régression vous permet de quantifier davantage les relations en développant une **équation** afin que vous puissiez prédire, par exemple, la glycémie en fonction de poids corporel du patient.

La régression linéaire est une technique statistique utilisée pour prédire une variable **dépendante continue** à partir d'une ou plusieurs variables **indépendantes continues**.

Puisque nous avons une forte corrélation linéaire entre la glycémie et le poids, nous pouvons effectuer une régression linéaire, comme suit :

1. Sélectionnez **Analyse ➜ Régression ➜ Linéaire**.
   La boîte de dialogue Régression linéaire s'affiche. Dans cet exemple, vous souhaitez prédire la glycémie à partir de poids. Vous pouvez placer la variable dépendante, la glycémie (**Bloodsugar**), dans la case **Dépendant** ; il s'agit de la variable pour laquelle vous souhaitez définir une équation de prédiction. Vous pouvez placer la variable prédictive **Weight** dans la zone **Indépendantes** ; c'est la variable que vous utiliserez pour prédire la variable dépendante.
2. Sélectionnez la variable **Bloodsugar** et placez-la dans la zone Dépendant.
3. Sélectionnez la variable **Weight** et placez-la dans la zone indépendantes, comme illustré dans la Figure 6.7.

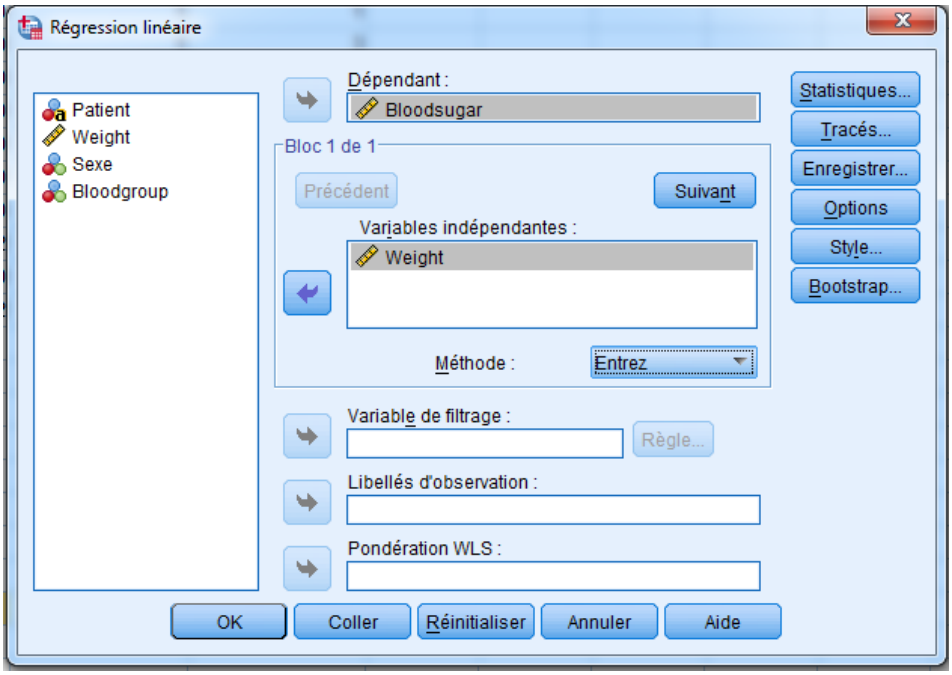


FIGURE 6.7 – Boîte de dialogue Régression linéaire

4. Cliquez sur **OK** : SPSS effectue la régression linéaire (voir Figure 6.8).

➔ **Régression**

Variables introduites/éliminées[a]

Modèle	Variables introduites	Variables éliminées	Méthode
1	Weight[b]	.	Introduire

a. Variable dépendante : Bloodsugar

b. Toutes les variables demandées ont été introduites.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,925[a]	,855	,837	,51969

a. Prédicteurs : (Constante), Weight

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	12,767	1	12,767	47,270	,000[b]
	Résidus	2,161	8	,270		
	Total	14,927	9			

a. Variable dépendante : Bloodsugar

b. Prédicteurs : (Constante), Weight

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	-,608	,405		-1,502	,172
	Weight	,027	,004	,925	6,875	,000

a. Variable dépendante : Bloodsugar

FIGURE 6.8 – Résultat de la régression linéaire

La colonne B contient les coefficients de régression (a : la pente, b : l'ordonnée à l'origine) que vous utiliseriez dans une équation de prédiction. Dans cet exemple, la glycémie peut être prédite avec l'équation suivante :

$$Bloodsugar = a * Weight + b$$

$$Bloodsugar = 0.027 * Weight - 0.608$$

6.5 Représentation graphique des données

Le menu **Graphiques** dans SPSS comporte trois options principales : **Générateur de graphiques**, **Sélecteur de modèles de représentations graphiques** et

Boîtes de dialogue ancienne version. Ces options sont différentes façons de faire le même travail. Les **Boîtes de dialogue ancienne version** sont les graphiques originaux de SPSS et sont principalement choisies par les personnes qui les utilisent depuis des années et qui trouvent trop difficile de passer à une autre option.

Sélecteur de modèles de représentations graphiques et **Générateur de graphiques** permettent de créer des graphiques de différentes manières. Dans « **Sélecteur de modèles de représentations graphiques** », vous sélectionnez d'abord les variables que vous souhaitez afficher. Sur la base de ces informations, différentes options de graphique sont proposées. **Générateur de graphiques** commence par présenter différents types de graphiques. Après avoir sélectionné un graphique, vous spécifiez les variables que vous utiliserez.

6.5.1 Construire des graphiques à la manière du Générateur de graphiques

SPSS contient **Générateur de graphiques**, qui utilise un affichage graphique pour vous guider à travers les étapes de construction de graphiques. Le programme vérifie continuellement vos actions et bloque l'utilisation de fonctionnalités qui ne sont pas compatibles ou qui ne fonctionneront pas correctement.

A travers un exemple, nous allons voir comment générer un graphe. Supposons que nous voulions tracer la **boîte à moustaches** de la variable « **Bloodsugar** ».

1. Choisissez **Graphiques ➜ Générateur de graphiques**. Un avertissement s'affiche, vous informant qu'avant d'utiliser cette boîte de dialogue, le niveau de mesure doit être défini correctement pour chaque variable de votre graphique. (Nous avons défini le niveau de mesure correct, vous pouvez donc continuer.)
2. Cliquez sur **OK** : la boîte de dialogue **Générateur de graphiques** s'affiche.
3. Assurez-vous que l'onglet **Galerie** est sélectionné : dans la liste "**Choisir parmi**", sélectionnez "**Boites à moustaches**" comme type de graphique.
4. Différents types de boîtes à moustaches apparaissent dans la galerie à droite de la liste, comme illustré dans la Figure 6.9 ci-dessous.

FIGURE 6.9 – Générateur de graphiques

5. Sélectionnez et faites glisser « **Boîte à Moustaches : 1D** » vers le panneau d'affichage. L'onglet **Propriétés des éléments** apparaît maintenant

à droite du panneau d'aperçu en haut. Cet onglet vous permet de savoir quelles fonctionnalités de l'élément vous pouvez modifier. Par exemple, vous pouvez changer la statistique à afficher ou le style des graphes. Dans cet exemple, vous n'utilisez pas l'onglet **Propriétés des éléments**, alors fermez-le simplement.

6. Dans la liste des variables, sélectionnez la variable **Bloodsugar** et faites-la glisser vers l'étiquette de l'axe Y dans le diagramme. L'affichage graphique à l'intérieur de la fenêtre d'aperçu du graphique ne représente jamais vos données réelles, même après avoir inséré des noms de variables.
7. Cliquez sur le bouton **OK** pour produire le graphique : la sortie résultante, illustrée dans la Figure 6.10 ci-dessous.

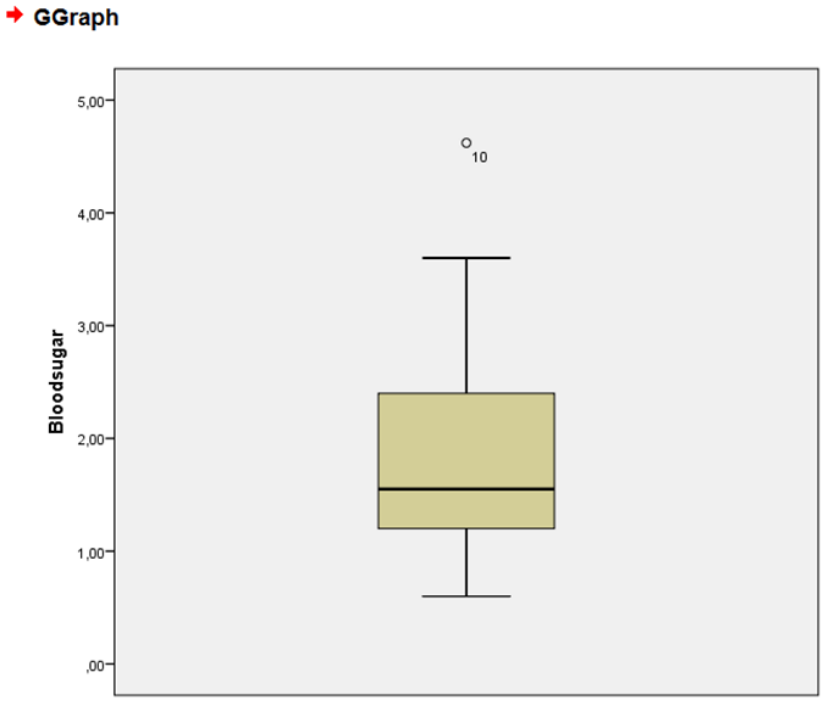


FIGURE 6.10 – Boîte à Moustaches :1D pour la variable Bloodsugar

6.5.2 Affichage d'une relation linéaire

Les étapes suivantes vous montrent comment construire un nuage de points simple :

1. Choisissez **Graphiques ➜ Générateur de graphiques**.
2. Cliquez sur le bouton **OK** puis le bouton **Réinitialiser**.
3. Dans la liste « **Choisir parmi** », sélectionnez **Dispersion/Points**.
4. Sélectionnez le premier diagramme de nuage de points (Diagramme de dispersion : simple) et faites-le glisser vers le panneau en haut.
5. Dans la liste Variables, sélectionnez **Weight** et faites-le glisser vers le rectangle intitulé X-Axis dans le diagramme.
6. Dans la liste Variables, sélectionnez **Bloodsugar** et faites-le glisser vers le rectangle intitulé Y-Axis dans le diagramme.
7. Cliquez sur **OK** : le graphique dans la Figure 6.11 s'affiche.

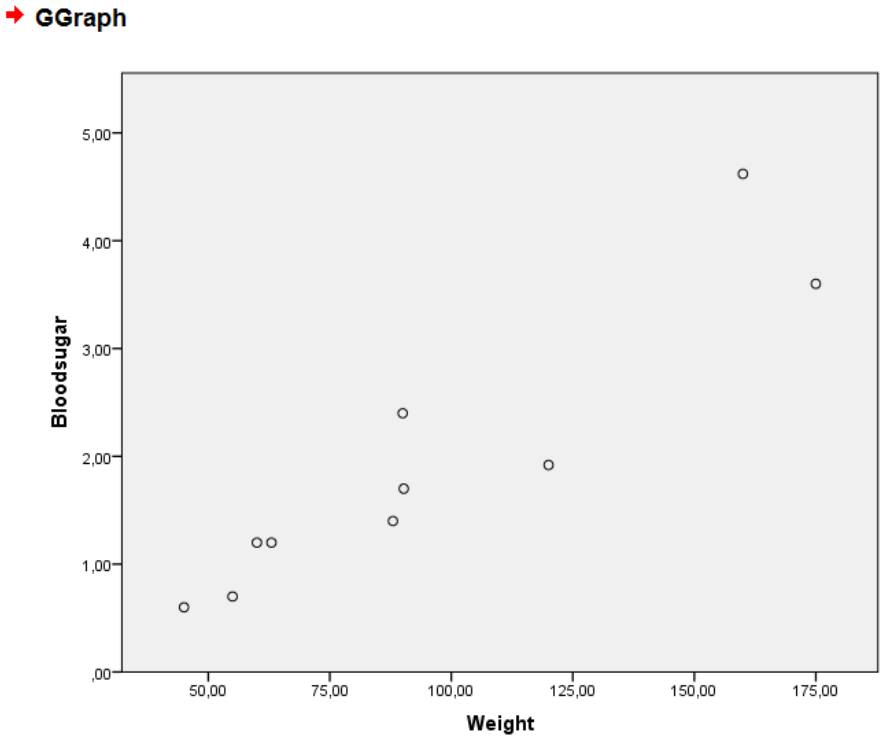


FIGURE 6.11 – Nuage de points de deux variables quantitatives

8. Double-cliquez sur le graphique produit : la boîte de dialogue « **Editeur de graphiques** » s'affiche.
9. Appuyez sur l'icône « **Ajouter une courbe d'ajustement au total** », puis sur le bouton Fermer, et enfin fermez la fenêtre « **Editeur de graphiques**».

10. La sortie résultante, illustrée dans la Figure 6.12 ci-dessous.

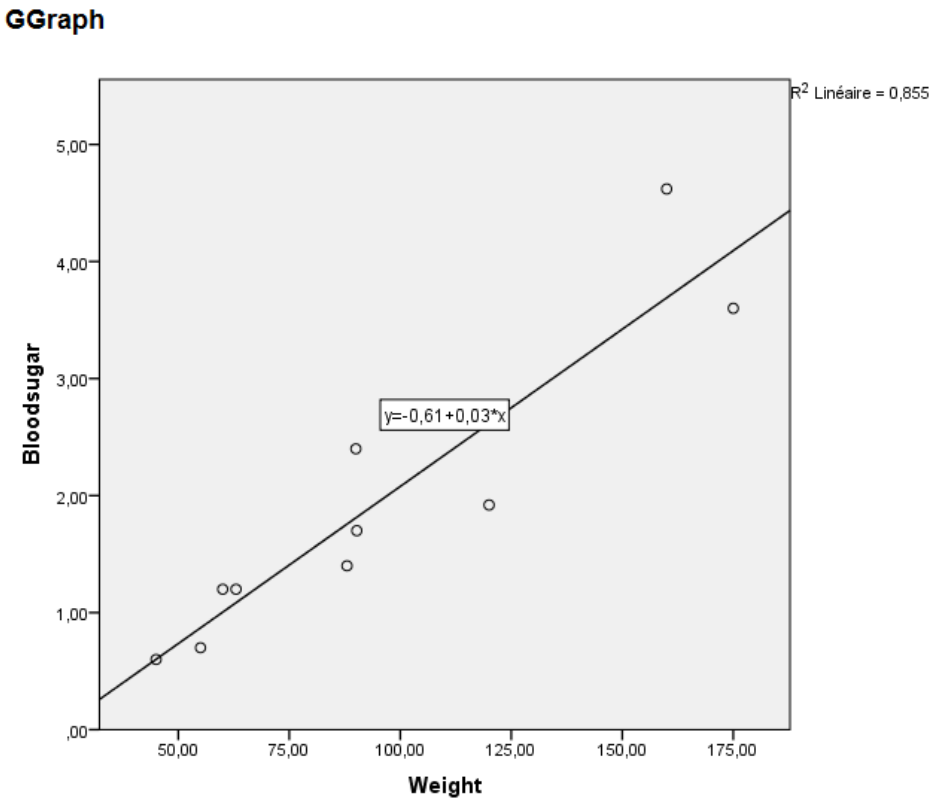


FIGURE 6.12 – Droite de régression linéaire

La droite superposée au nuage de points est la meilleure droite qui décrit la relation linéaire de l'équation : **y = 0.03*x- 0.61**, où y représente **Bloodsugar** et x est le **Weight**.

6.6 Conclusion

Les chercheurs peuvent utiliser SPSS pour étudier les relations entre les variables en analysant à la fois les variables qualitatives et continues. SPSS fournit divers outils et fonctions, tels que la tabulation croisée et l'analyse de corrélation, pour examiner l'association ou la corrélation entre les variables. L'analyse de tableaux croisés peut être utilisée pour identifier les relations entre deux ou plusieurs variables qualitatives, tandis que l'analyse de corrélation peut être utilisée pour déterminer la force et la direction de la relation entre deux variables continues. De plus, l'analyse de régression peut être utilisée pour examiner la relation entre une variable dépendante et une ou plusieurs variables indépendantes, et pour créer un modèle qui prédit les valeurs de la variable dépendante en fonction des variables indépendantes. En utilisant ces fonctions et outils, les chercheurs peuvent étudier les relations entre les variables dans leurs données biomédicales et prendre des décisions éclairées en fonction de leur analyse.

Chapitre 7

Distributions de Probabilité avec SPSS

7.1 Introduction

SPSS comprend de nombreuses distributions standard intégrées couramment utilisées dans l'analyse statistique et la construction de modèles, qui peuvent remplacer les tableaux statistiques traditionnels. Parmi la multitude de distributions disponibles, les distributions **binomiale** et **normale** se dressent comme des piliers, offrant des insights inestimables dans les données **discrètes** et **continues**, respectivement.

Les trois éléments fondamentaux qui peuvent être calculés pour une distribution statistique sont :

1. **Fonction de densité de probabilité (PDF)** : La fonction PDF pour les distributions binomiales et normales est utilisée pour décrire la probabilité ou la vraisemblance d'observer des résultats ou des valeurs spécifiques au sein des distributions respectives.
2. **Fonction de distribution cumulative (CDF)** : elle représente la probabilité qu'une variable aléatoire prenne une valeur inférieure ou égale à une valeur donnée. En d'autres termes, le CDF est la somme cumulée des probabilités jusqu'à un certain point.
3. **Fonction de distribution inverse (IDF)** : elle est également appelée fonction quantile et fournit la valeur d'une variable aléatoire qui correspond à une probabilité donnée. En d'autres termes, l'IDF donne la valeur de x telle que la probabilité d'obtenir une valeur inférieure ou égale à x soit égale à une probabilité donnée.

Pour toutes les distributions implémentées dans SPSS, il existe une fonction pour chacun des trois éléments énumérés ci-dessus. La convention de dénomination de ces fonctions consiste à préfixer le nom par "Pdf" pour la fonction de densité de probabilité (PDF), "Cdf" pour la fonction de distribution cumulative (CDF) et "Idf" pour la fonction de distribution cumulative inverse (IDF).

7.2 Distribution Binomiales (distribution discrète finie)

Pour calculer les probabilités binomiales de la forme $P(x \mid n,p)$, vous devez créer une colonne de données contenant les valeurs de x que vous souhaitez utiliser. Une fois que vous avez cette colonne, vous pouvez l'utiliser comme entrée dans la formule de probabilité binomiale pour calculer les probabilités pour chaque valeur donnée.
Pour effectuer les probabilités binomiales, nous utilisons les fonctions SPSS suivantes :
➤ *La fonction de densité de probabilité* : calcule la probabilité que la variable aléatoire binomiale prenne une valeur particulière x (**=*x***), compte tenu du nombre d'essais n et de la probabilité de succès p.

$$PDF.BINOM(x, n, p)$$

➤ *La fonction de distribution cumulative binomiale* : calcule la probabilité que la variable aléatoire binomiale prenne une valeur **inférieure ou égale à *x***, compte tenu du nombre d'essais n et de la probabilité de succès p.

$$CDF.BINOM(x, n, p)$$

Où x est le nombre de succès, n est la taille de l'échantillon, et p représente la probabilité de succès.

Exemple : Quelle est la probabilité d'obtenir k faces avec une pièce lancée n fois, ou avec n pièces lancées simultanément ?

- Calculer P(X=k) avec k=0, 1, 2, 3, 4, 5 et n=5.
- Représenter graphiquement les probabilités obtenues.

Procédure SPSS pour les probabilités binomiales :

1. Dans l'onglet Vue des variables, créez une nouvelle variable d'échelle (discrète) nommée **k**.
2. Tapez le nombre de succès de 0 à 5 dans la variable **k**.
3. Choisissez **Transformer ➜ Calculer la variable** : la boîte de dialogue **Calculer la variable**...apparaît.
4. Dans le groupe de fonctions, sélectionnez **PDF et PDF non centré ➜** Dans la zone fonctions et variables spéciales, double-cliquez sur la fonction **Pdf.Binom**, et la formule apparaîtra dans la zone d'expression numérique, voir Figure 7.1.

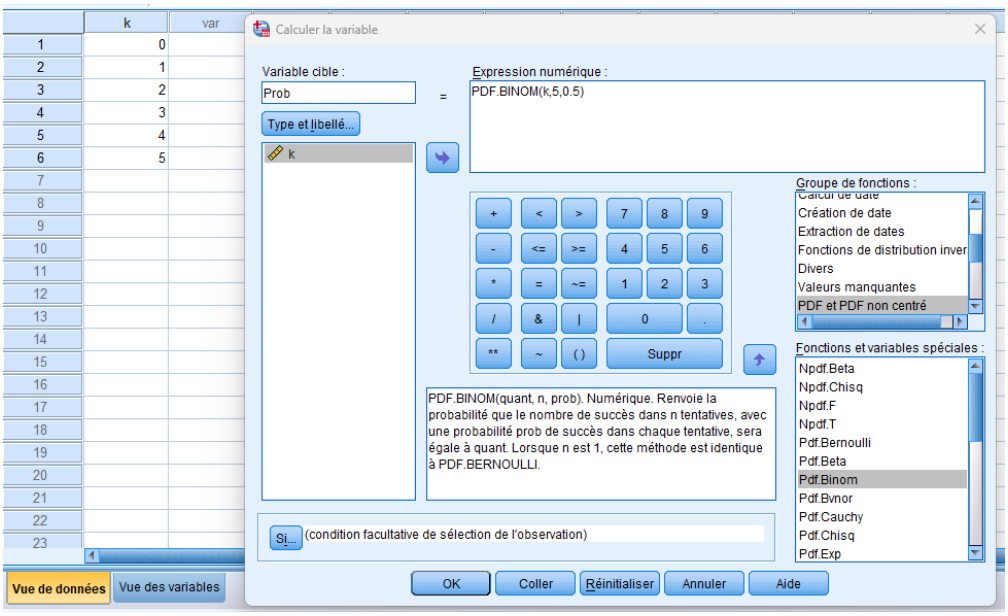


FIGURE 7.1 – Boîte de dialogue : Calculer la variable

5. Donnez la variable cible le nom : "**Prob**" et tapez l'expression **PDF.BINOM (k,5,0.5)** ➜ Cliquez sur **OK**.
6. En conséquence, une nouvelle variable apparaîtra avec des probabilités binomiales, voir la Figure 7.2 (définir le nombre décimal sur 4 pour la nouvelle variable **Prob**).

	k	Prob
1	0	,0313
2	1	,1563
3	2	,3125
4	3	,3125
5	4	,1563
6	5	,0313

FIGURE 7.2 – Résultat après le calcul des probabilités binomiales

7. Choisissez **Graphiques ➜ Générateur de graphiques ➜** Cliquez sur le bouton **OK**.
8. Dans la liste « **Choisir parmi** », sélectionnez **Courbe**.
9. Sélectionnez le premier diagramme de nuage de points (**Diagramme en Lignes : simple**) et faites-le glisser vers le panneau en haut.
10. Dans la liste Variables, sélectionnez **k** et faites-le glisser vers le rectangle intitulé X-Axis dans le diagramme.
11. Dans la liste Variables, sélectionnez **Prob** et faites-le glisser vers le rectangle intitulé Y-Axis dans le diagramme.

12. Cliquez sur **OK** : le graphique dans la Figure 7.3 s'affiche.

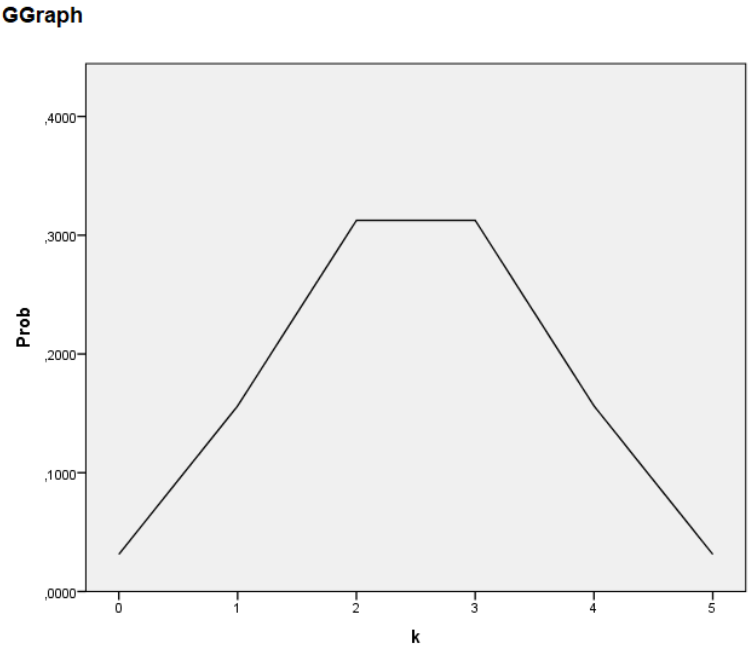


FIGURE 7.3 – Représentation graphique des probabilités

7.3 Distribution Normale (distribution continue)

La distribution normale est largement considérée comme la distribution de probabilité la plus importante dans les statistiques, car elle décrit avec précision de nombreux phénomènes naturels, tels que le poids, l'âge, les tempêtes, les inondations et autres.

7.3.1 Probabilités Normales

Pour effectuer les probabilités normales, nous utilisons la fonction SPSS suivante :

$$CDF.NORMAL(x, m, d)$$

Où x représente la quantité donnée, m représente la moyenne et d est l'écart type.

Exemple : Une variable aléatoire continue X suit la distribution normale avec une moyenne m = 50 et une variance σ^2 = 100. Trouvez la probabilité que :

1. P(X < 44) ?
2. P(29 < X < 53) ?
3. P(35 < X < 43) ?
4. P(X < 63) ?
5. P(X > 37,2) ?
6. P(58,1 < X < 69,4) ?

Procédure SPSS pour calculer la probabilité à partir de la distribution normale :

1. Dans l'onglet Vue des variables, créez une nouvelle variable d'échelle (continue) nommée *Prob* avec quatre chiffres derrière la virgule (Décimal=4).
2. Tapez le chiffre 0 dans la cellule **1 :Prob**.
3. Choisissez **Transformer ➜ Calculer la variable** : la boîte de dialogue **Calculer la variable**. . .apparaît.
4. Dans le groupe de fonctions, sélectionnez **CDF et CDF non centré ➜** Dans la zone fonctions et variables spéciales, double-cliquez sur la fonction **Cdf.Normal**, et la formule apparaîtra dans la zone d'expression numérique, voir Figure 7.4.

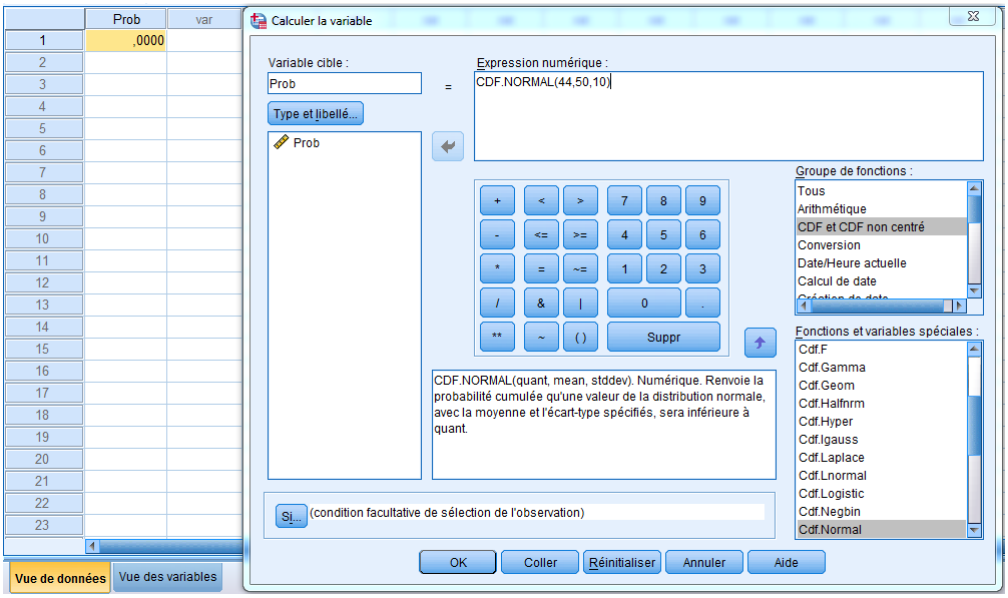


FIGURE 7.4 – Boîte de dialogue : Calculer la variable

(a) **Pour P(X < 44) :** Donnez la variable cible le nom : "**Prob**" et tapez l'expression **CDF.NORMAL (44,50,10) ➜** Cliquez sur **OK ➜** La cellule **1 :Prob** affichera une probabilité de 0,2743.

(b) **Pour P(29 < X < 53) :** Donnez la variable cible le nom : "**Prob**" et tapez l'expression **CDF.NORMAL(53,50,10)-CDF.NORMAL(29,50,10) ➜** Cliquez sur **OK ➜** La cellule **1 :Prob** affichera une probabilité de 0,6000.

(c) **Pour P(X > 37,2) :** Donnez la variable cible le nom : "**Prob**" et tapez l'expression **1-CDF.NORMAL (37.2,50,10) ➜** Cliquez sur **OK ➜** La cellule **1 :Prob** affichera une probabilité de 0,8997.

7.3.2 Centiles Normaux

Étant donné une probabilité, pour déterminer quelle valeur correspond à une probabilité unilatérale à gauche, nous utilisons la fonction SPSS suivante :

$$IDF.NORMAL(p, m, d)$$

Où p représente la probabilité donnée, m représente la moyenne et d est l'écart type.
Dans P(X <= x) = p, où X est une variable aléatoire avec une distribution normale et p est une probabilité donnée, la fonction IDF.NORMAL aide à trouver la valeur de x qui correspond à une probabilité ou à un centile spécifique dans une distribution normale.

Exemple : Les scores au test IELTS ces dernières années suivent approximativement la distribution N(504,111). Quelle note un étudiant doit-il obtenir pour se classer parmi les 10% de tous les étudiants qui passent l'IELTS ?
Procédure SPSS pour les centiles Normaux :

1. Dans l'onglet vue des variables, créez trois nouvelles variables d'échelles (continues) nommées **Prob**, **Moyenne** et **Sd**. Changez le nombre de décimales pour **Moyenne** et **Sd** par 0.
2. Allez dans l'onglet vue de données ➜ Tapez 0,90 sous la variable **Prob** dans la première cellule. Dans une distribution normale, qui est symétrique autour de sa moyenne, la zone sous la courbe à gauche d'un certain point représente la probabilité cumulative jusqu'à ce point. Par conséquent, si nous voulons trouver le score qui correspond aux 10 % supérieurs, nous cherchons essentiellement un score qui laisse 90 % de la distribution en dessous de lui (Voir la Figure 7.5).
3. Saisissez 504 sous la variable **Moyenne**. Tapez 111 sous la variable **Sd**.
4. Choisissez **Transformer ➜ Calculer la variable** : la boîte de dialogue **Calculer la variable**...apparaît.
5. Dans le groupe de fonctions, sélectionnez **Functions de distribucion inverse ➜** Dans la zone fonctions et variables spéciales, double-cliquez sur la fonction **Idf.NORMAL**, et la formule apparaîtra dans la zone d'expression numérique.

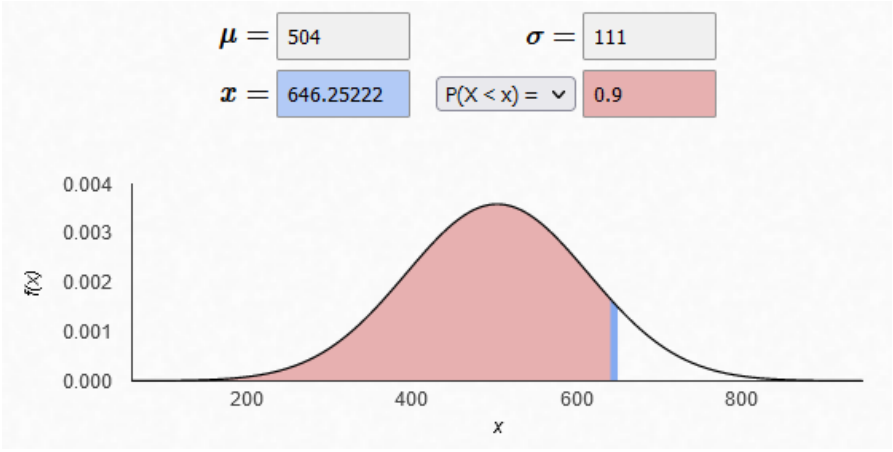


FIGURE 7.5 – La zone sous la courbe à gauche de x ($p(X < x) = 0,9$) [uiowa, 2024].

6. Donnez la variable cible le nom : "**RES**" et tapez l'expression **IDF.NORMAL (Prob,Moyenne,Sd)**, voir Figure 7.6.

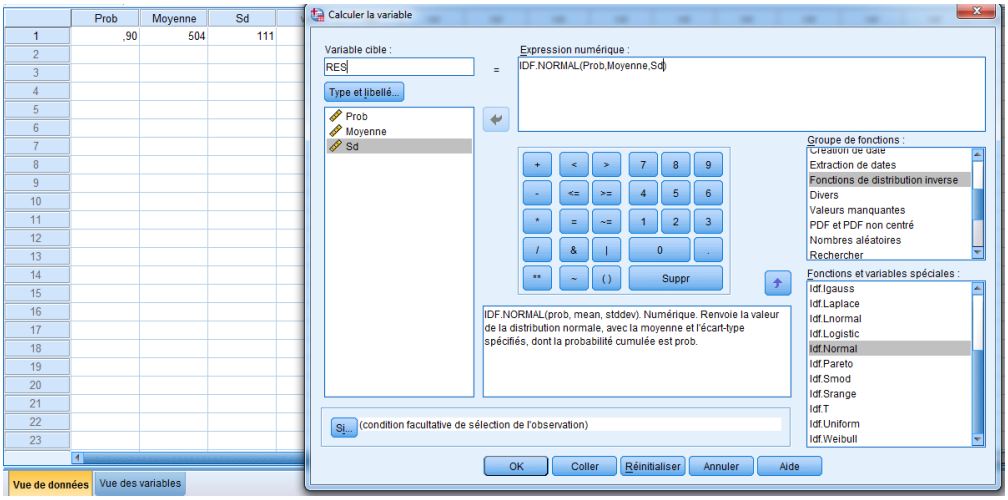


FIGURE 7.6 – Boîte de dialogue : Calculer la variable

7. Cliquez sur **OK** ➜ La cellule **RES** affichera une valeur de 646,25, voir la Figure 7.7.

	Prob	Moyenne	Sd	RES
1	,90	504	111	646,25

FIGURE 7.7 – Résultat après l'application de la fonction IDF.NORMAL

Donc, un étudiant doit obtenir un score d'environ 646,25 ou plus pour se classer parmi les 10 % des meilleurs étudiants passant l'IELTS.

7.3.3 Estimation par Intervalles de confiance

Pour estimer la moyenne d'une population avec un intervalle de confiance, vous devez suivre ces étapes :

1. Déterminez la taille de l'échantillon (n) et la moyenne de l'échantillon ($\bar{x}$) des données.
2. Choisissez un niveau de confiance (généralement 95% ou 99%) et déterminez la valeur z correspondante pour ce niveau de confiance. Par exemple, un niveau de confiance de 95% correspondrait à une valeur z de 1,96.
3. Calculez l'erreur type de la moyenne (*SEM*), qui est l'écart type de l'échantillon (*s*) divisé par la racine carrée de la taille de l'échantillon (*n*) : $SEM = s/\sqrt{n}$.
4. Calculez la marge d'erreur (*ME*), qui est le produit de la *SEM* et de la valeur z :$ME = \text{z} * SEM$.
5. Calculez les bornes inférieure et supérieure de l'intervalle de confiance en soustrayant et en ajoutant la marge d'erreur à la moyenne de l'échantillon : Borne inférieure = $\bar{x} - ME$, Borne supérieure = $\bar{x} + ME$.

Par exemple, supposons que nous voulions estimer la taille moyenne d'une population sur la base d'un échantillon de 100 personnes, avec une moyenne d'échantillon de 68 pouces et un écart type d'échantillon de 3 pouces. Nous choisissons un niveau de confiance de 95%.

n = 100, $\bar{x}$ = 68 pouces
Pour un niveau de confiance de 95%, z = 1,96.
$SEM = s/\sqrt{n}$ = $3/\sqrt{100} = 0,3$ pouces
$ME =$ z $* SEM = 1,96 * 0,3 = 0,588$ pouces
Limite inférieure $= 68 - 0,588 = 67,412$ pouces, Limite supérieure $= 68 + 0,588 = 68,588$ pouces

Par conséquent, nous pouvons dire avec une confiance de 95% que la taille moyenne de la population se situe entre 67,412 et 68,588 pouces sur la base de nos données d'échantillon.

Procédure SPSS pour l'estimation de μ avec un intervalle de confiance de 95% :

1. Dans l'onglet vue des variables, créez trois nouvelles variables d'échelles (continues) nommées **xbar**, **sigma** et **n** (voir Figure 7.8 ci-dessous).

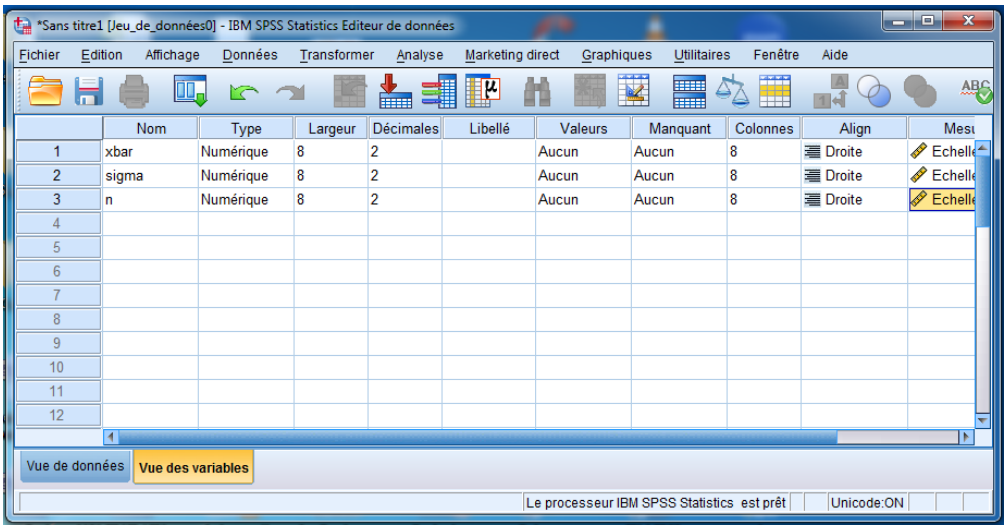


FIGURE 7.8 – Vue des variables

2. Dans l'onglet vue des données, sous la variable **xbar**, tapez « 68 » pour la moyenne de l'échantillon, sous la variable **sigma**, tapez « 3 » pour l'écart type de la population, et sous la variable **n**, tapez « 100 » pour la taille de l'échantillon.
3. Choisissez **Transformer ➜ Calculer la variable** : la boîte de dialogue **Calculer la variable**...apparaît.
4. Donnez la variable cible le nom : **lowerbound** et tapez l'expression **xbar – 1.96 * sigma/SQRT(n)**, voir Figure 7.9.

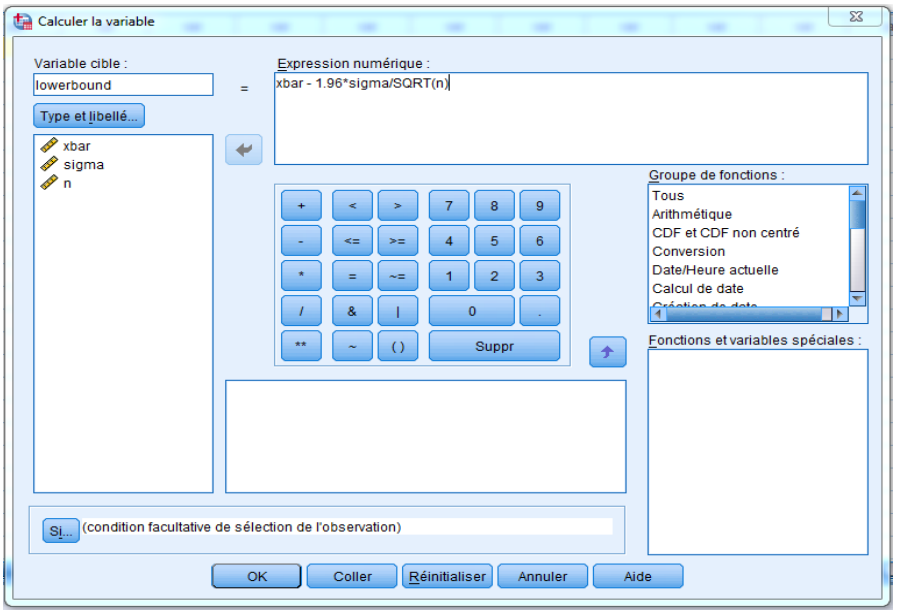


FIGURE 7.9 – Boîte de dialogue : Calculer la variable

5. Cliquez sur **OK**.
6. Répétez les étapes 3 à 5, en utilisant **upperbound** et l'expression : **xbar + 1,96 * sigma/SQRT(n)**.
7. Cliquez sur **OK ➜** Les cellules **lowerbound** et **upperbound** afficheront les valeur de **67,41** et **68,59**, respectivement, voir la Figure 7.10.

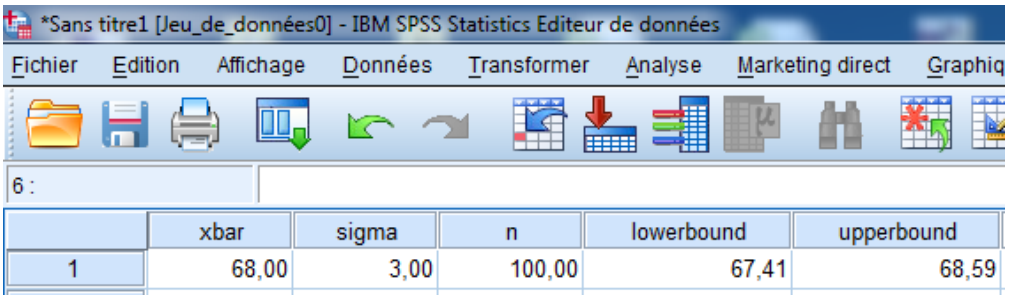


FIGURE 7.10 – Résultat de l'estimation

7.4 Conclusion

En conclusion, l'intégration des fonctions de densité de probabilité (PDF), de distribution cumulative (CDF) et de distribution inverse (IDF) de SPSS pour les distributions binomiales et normales a grandement facilité l'analyse et l'interprétation des données. Ces puissants outils statistiques ont permis une exploration approfondie de la probabilité de résultats ou de valeurs spécifiques au sein des distributions, facilitant ainsi la formulation de conclusions pertinentes. De plus, maîtriser l'estimation de la moyenne en utilisant des intervalles de confiance à l'aide de SPSS renforce notre capacité à exploiter efficacement les fonctionnalités du logiciel, améliorant ainsi la précision et la fiabilité de nos analyses statistiques.

Travaux Pratiques

TP 1

But

Dans ce TP, nous apprendrons à créer un fichier Excel contenant plusieurs variables quantitatives et qualitatives. Nous visons également à apprendre à valider les données saisies.

Exercice

	A	B	C	D	E
1	Patient	Age	Sexe	Blood	Temperature
2	P1	77	Femme	B	37,4
3	P2	64	Homme	A	38
4	P3	87	Homme	AB	38,5
5	P4	70	Femme	A	38,2
6	P5	85	Femme	O	38,1

1. Créer un fichier de données EXCEL avec différents types de variables (qualitatives et quantitatives) :
   (a) Créer un nouveau fichier de données EXCEL.
   (b) Les données doivent contenir cinq variables nommées : Patient, Age, Sexe, Blood, et Temperature (voir la Figure ci-dessus).
   (c) Saisir les données de la variable "Patient" de P1 à P5 par glissement ?
2. Avant de saisir les données des variables Age, Sexe, Blood, et Temperature, vous devez vous assurer que chaque variable est valide.
   (a) Vérifier la validité des données en variables quantitatives (Age, Temperature) ?
   (b) Vérifier la validité des données en variables qualitatives (Sexe, Blood) ?
3. Enregistrez le fichier Excel obtenu sous le nom « Tp1.xlsx ».
4. Supprimez la variable « Température » et enregistrez les données dans un autre fichier appelé « Tp1_2.xlsx ».

Solution

1. (a) Ouvrez le logiciel EXCEL.
   (b) Dans la première ligne, entrez les noms de cinq variables dans chaque colonne. (A1 : Patient, B1 : Age, C1 : Sexe, D1 : Blood, et E1 : Temperature).
   (c) Sélectionner la cellule A2 ➜ Faire glisser la poignée de recopie vers le bas (jusqu'à la cellule A6).
2. (a) Pour la variable quantitative discrète Age : Sélectionner les cellules où vous allez saisir les valeurs de cette variable B2 :B6 ➜ Données ➜ Validation des données ➜ Validation des données ➜ Choisir Autoriser : Nombre entier ➜ Données : comprise entre ➜ Minimum :1 ➜ Maximum :120 ➜ OK.
   (b) Pour la variable quantitative continue Temperature : Sélectionner les cellules où vous allez saisir les valeurs de cette variable E2 :E6 ➜ Données ➜ Validation des données ➜ Validation des données ➜ Choisir Autoriser : Décimal ➜ Données : comprise entre ➜ Minimum :29.5 ➜ Maximum :42,3 ➜ OK.
   (c) Pour la variable qualitative Sexe : Sélectionner les cellules où vous allez saisir les valeurs de cette variable C2 :C6 ➜ Données ➜ Validation des données ➜ Validation des données ➜ Choisir Autoriser : Liste ➜ Source : Homme ;Femme ➜ OK.
   (d) Pour la variable qualitative Blood : Sélectionner les cellules où vous allez saisir les valeurs de cette variable D2 :D6 ➜ Données ➜ Validation des données ➜ Validation des données ➜ Choisir Autoriser : Liste ➜ Source : AB ;A ;B ;O ➜ OK.
   (e) Nous remplissons les données comme montré dans la Figure ci-dessus.
3. Bouton Office ➜ Enregistrer ➜ Nom de fichier : « Tp1.xlsx » ➜ Enregistrer.
4. Sélectionnez la colonne E ➜ appuyez sur le bouton "Suppr" ➜ Bouton Office ➜ Enregistrer sous ➜ Nom de fichier : « Tp1_2.xlsx » ➜ Enregistrer.

Pour plus de détails, reportez-vous au chapitre 2

TP 2

But

Dans ce TP, nous apprendrons à saisir des données et à définir correctement des variables à l'aide du programme SPSS. De plus, il est important de savoir comment gérer les fichiers SPSS et comment transférer des données d'Excel vers SPSS.

Exercice

	Patient	Sexe	Temperature	Bloodsugar	Hypertension
1	P1	1	38,10	1,20	1
2	P2	1	37,10	1,70	2
3	P3	2	38,00	1,40	1
4	P4	1	38,10	,60	1

1. Créer un fichier de données SPSS avec différents types de variables (qualitatives et quantitatives) :
   (a) Créer un nouveau fichier de données SPSS appelé « Tp2.sav ».
   (b) Les données doivent contenir cinq variables nommées : Patient, Sexe, Temperature, Bloodsugar, et Hypertension (voir la Figure ci-dessus).
   (c) La Variable « Patient » est une variable de type Chaîne.
   (d) Entrez les valeurs possibles pour la variable qualitative « Sexe » : 1=Homme et 2=Femme.
   (e) La Variable « Temperature » est une variable quantitative continue de type Numérique.
   (f) La Variable « Bloodsugar » est une variable quantitative continue de type Numérique.
   (g) Entrez les valeurs possibles pour la variable qualitative « Hypertension » : 1=Yes , 2=No.
2. Enregistrez le fichier SPSS obtenu.
3. Créez un autre fichier Excel et essayez de l'ouvrir à partir de SPSS.

Solution

1. Pour saisir des données dans SPSS, vous pouvez suivre ces étapes :
   (a) Ouvrez le logiciel SPSS et cliquez sur l'onglet "Vue des variables".
   (b) Saisissez les noms des variables dans la première colonne du tableau. Chaque nom de variable doit être unique et descriptif.
   (c) Entrez les données pour chaque variable dans la colonne correspondante (Nom, Type, Largeur, Décimales, Libellé, Valeurs, Manquant, Colonnes, Align, Mesure).

2. Une fois que vous avez entré toutes les données, enregistrez le fichier en cliquant sur "Enregistrer" dans le menu "Fichier".
3. Vous pouvez également importer des données dans SPSS à partir d'autres logiciels ou formats de fichiers, tels qu'Excel. Pour cela, cliquez sur Fichier ➜ Ouvrir ➜ Données ➜ Type de fichier Excel (*.xls *.xlsx, *.xlsm) ➜ sélectionnez le fichier Excel ➜ Ouvrir.

   Après avoir importé les données, vous devrez peut-être modifier les types de variables ou recoder les données pour répondre à vos besoins d'analyse.

Pour plus de détails, reportez-vous au chapitre 3

TP 3

But

1. Aider l'étudiant à comprendre les techniques appropriées pour collecter et saisir des données avec précision à l'aide du logiciel SPSS.
2. Le TP vise à mettre en évidence les meilleures pratiques pour le traitement des données dans SPSS, y compris les stratégies de traitement des données manquantes, de tri des données, de recodage des variables, de fractionnement des données et de sélection des observations.

Exercice 1

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P1	53,00	1,20	1	1
2	P2	44,00	-99,00	1	3
3	P3	88,00	1,40	2	2
4	P4	45,00	,60	1	4
5	P5	90,00	2,40	2	2
6	P6	175,00	3,60	2	1

1. Créer un nouveau fichier de données SPSS appelé « Tp3.sav ».
2. Les données doivent contenir cinq variables nommées : Patient, Weight, Bloodsugar, Sexe, et Bloodgroup (voir la Figure ci-dessus).
3. La Variable « Patient » est une variable de type Chaîne.
4. La Variable « Weight » est une variable quantitative continue de type Numérique.
5. La Variable « Bloodsugar » est une variable quantitative continue de type Numérique. (Pour la variable Bloodsugar, on représente les valeurs manquantes par le nombre -99).
6. Entrez les valeurs possibles pour la variable qualitative « Sexe » : 1=Homme et 2=Femme.
7. Entrez les valeurs possibles pour la variable qualitative « Bloodgroup » : 1=AB , 2=A, 3=B et 4=O.

Exercice 2

1. Dans la variable Bloodsugar, remplacez les valeurs manquantes par une valeur moyenne de la série ?
2. Triez les données des patients en fonction de leurs valeurs de glycémie (Bloodsugar) ?
3. Dans une nouvelle variable nommée "BloodsugarCat" : essayez de recoder les valeurs de la variable "Bloodsugar" comme suit : Classe 1 : Normal $\leq$0,99 ; Classe 2 : Prédiabète]0,99 - 1,25] ; Classe 3 : Diabète > 1,25.

4. Supprimer la nouvelle variable "BloodsugarCat"?
5. Comparez séparément les données des patients en fonction de leurs valeurs de Bloodgroup?
6. Annuler le fractionnement des données?
7. Sélectionnez des patients avec un poids $\geq$70 et une glycémie <1 ou des patients hommes avec un groupe sanguin égal à AB?

Solution

1. Choisissez Transformer ➜ Remplacer les valeurs manquantes ➜ Passez la variable Bloodsugar vers la zone «Nouvelles variables» ➜ OK ➜ Remplacer manuellement la valeur -99 dans la colonne de Bloodsugar par 1,84 ➜ Supprimer la colonne Bloodsugar_1.
2. Choisissez Données ➜ Trier les observations ➜ Trier par : Bloodsugar ➜ OK.
3. Choisissez Transformer ➜ Création de variables ➜ Déplacer la variable Bloodsugar vers la zone de travail à droite ➜ Dans Nom : Nommez la nouvelle variable comme BloodsugarCat ➜ Changer ➜ Anciennes et nouvelles valeurs ➜ Pour la catégorie 1 : on sélectionne le bouton radio Plage, du MINIMUM à la valeur ➜ Entrer la valeur Max (c.a.d. 0,99) ➜ A côté du bouton radio Valeur : on entre le numéro de la catégorie (c.a.d. 1) ➜ Cliquer sur le bouton Ajouter ➜ Pour la catégorie 2 : on sélectionne le bouton radio Plage ➜ Entrer les valeurs Min(c.a.d. 0,99) et Max(c.a.d. 1,25) ➜ A côté du bouton radio Valeur : on entre le numéro de la catégorie (c.a.d. 2) ➜ Cliquer sur le bouton Ajouter ➜ Pour la catégorie 3 : on sélectionne le bouton radio Plage, de la valeur au MAXIMUM ➜ Entrer la valeur Min (c.a.d. 1,25) ➜ A côté du bouton radio Valeur : on entre le numéro de la catégorie (c.a.d. 3) ➜ Cliquer sur le bouton Ajouter ➜ Poursuivre ➜ OK.
4. Dans l'onglet vue des données, cliquez sur le nom de la variable "BloodsugarCat" ➜ Appuyez sur la touche Suppr du clavier.
5. Choisissez Données ➜ Scinder un fichier ➜ Sélectionnez le bouton radio "Comparer les groupes" ➜ Choisissez Bloodgroup comme variable de comparaison ➜ OK ➜ Choisissez Analyse ➜ Statistiques descriptives ➜ Fréquences ➜ Choisissez Sexe et placez-la dans la zone Variable(s) ➜ OK.
6. Choisissez Données ➜ Scinder un fichier ➜ Réinitialiser ➜ OK.
7. Choisissez Données ➜ Sélectionner des observations ➜ Sélectionnez le bouton radio "Selon une condition logique" ➜ Cliquez sur le bouton Si... ➜ Dans la zone d'expression, Écrivez "(Weight >= 70 & Bloodsugar<1) | (Sexe = 1 & Bloodgroup = 1)" ➜ Poursuivre ➜ OK.

Pour plus de détails, reportez-vous au chapitre 4

TP 4

But

1. Aider l'étudiant à comprendre les techniques appropriées pour collecter et saisir des données avec précision à l'aide du logiciel SPSS.
2. Réviser certaines des méthodes utilisées dans le traitement des données SPSS.
3. Présenter les techniques utilisées pour analyser les données qualitatives et quantitatives à l'aide de procédures de fréquence et descriptives spécifiquement dans SPSS.

Exercice 1

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P1	63,00	1,20	1	1
2	P2	90,20	1,70	1	3
3	P3	88,00	1,40	2	2
4	P4	45,00	,60	1	4
5	P5	90,00	2,40	2	2
6	P6	175,00	3,60	2	1
7	P7	63,00	2,10	1	2

1. Créer un nouveau fichier de données SPSS appelé « Tp4.sav ».
2. Les données doivent contenir cinq variables nommées : Patient, Weight, Bloodsugar, Sexe, et Bloodgroup (voir la Figure ci-dessus).
3. La Variable « Patient » est une variable de type Chaîne.
4. La Variable « Weight » est une variable quantitative continue de type Numérique.
5. La Variable « Bloodsugar » est une variable quantitative continue de type Numérique.
6. Entrez les valeurs possibles pour la variable qualitative « Sexe » : 1=Homme et 2=Femme.
7. Entrez les valeurs possibles pour la variable qualitative « Bloodgroup » : 1=AB , 2=A, 3=B et 4=O.

Exercice 2

1. Triez les données des patients en fonction de leurs valeurs de poids (Weight) et de glycémie (Bloodsugar). Que remarquez-vous ? (Focus sur les patients P1 et P7).
2. Comparez séparément les données des patients en fonction de leurs valeurs de sexe.

3. Basant sur le fractionnement des données précédent, nous souhaitons étudier la distribution de variable Bloodgroup (AB, A, B, O) pour chaque valeur de Sexe (Homme, Femme). Analyser statistiquement et graphiquement la variable qualitative Bloodgroup avec la procédure de fréquence ?
4. Annuler le fractionnement des données ?
5. Analyser statistiquement et graphiquement la variable quantitative Weight avec la procédure de fréquence ?
6. Résumer statistiquement les variables continues avec la procédure descriptive ?

Solution

1. Choisissez Données ➜ Trier les observations ➜ Trier par : Weight et Bloodsugar, successivement➜ OK. (On voit que P1 est positionné avant P7)
2. Choisissez Données ➜ Scinder un fichier ➜ Sélectionnez le bouton radio "Comparer les groupes" ➜ Choisissez Sexe comme variable de comparaison ➜ OK.
3. Choisissez Analyse ➜ Statistiques descriptives ➜ Fréquences ➜ placez Bloodgroup dans la zone Variable(s) ➜ Statistiques ➜ Cochez la case Mode ➜ Poursuivre ➜ Graphiques ➜ Sélectionnez le bouton radio Graphiques à barres ➜ Sélectionnez le bouton radio Pourcentages ➜ Poursuivre ➜ OK.
4. Choisissez Données ➜ Scinder un fichier ➜ Réinitialiser ➜ OK.
5. Choisissez Analyse ➜ Statistiques descriptives ➜ Fréquences ➜ Placez la variable Weight dans la zone Variable(s) ➜ Statistiques ➜ Cochez les cases : Moyenne, Médiane, Mode, Ecart type, Variance, Minimum et Maximum ➜ Poursuivre ➜ Graphiques ➜ Sélectionnez le bouton radio Histogrammes ➜ Cochez la case "Afficher la courbe gaussienne sur l'histogramme" ➜ Poursuivre ➜ Décochez la case "Afficher les tables de fréquences" ➜ OK.
6. Choisissez Analyse ➜ Statistiques descriptives ➜ Descriptives ➜ Placez les variables Weight et Bloodsugar dans la zone Variable(s) ➜ OK.

Pour plus de détails, reportez-vous au chapitre 5

TP 5

But

1. Aider l'étudiant à comprendre les techniques appropriées pour collecter et saisir des données avec précision à l'aide du logiciel SPSS.
2. Pour montrer comment analyser les relations entre les variables qualitatives à l'aide de tableaux croisés : Les tableaux croisés sont utilisés pour créer des tableaux de contingence qui résument la relation entre deux ou plusieurs variables catégorielles. Le TP vise à montrer comment créer et interpréter ces tableaux dans SPSS.
3. Démontrer comment analyser les relations entre des variables quantitatives à l'aide de la corrélation et de la régression linéaire : La corrélation et la régression linéaire sont des méthodes statistiques utilisées pour modéliser la relation entre deux ou plusieurs variables continues. Le TP vise à montrer comment utiliser ces méthodes pour créer et interpréter des modèles de régression dans SPSS.
4. Démontrer comment créer des tracés et des graphiques dans SPSS : Le TP vise à montrer comment utiliser SPSS pour créer des visualisations de données afin de mieux comprendre les relations entre les variables ou de communiquer les résultats aux autres.

Exercice 1

	Patient	Weight	Bloodsugar	Sexe	Bloodgroup
1	P1	63,00	1,20	1	1
2	P2	90,20	1,70	1	3
3	P3	88,00	1,40	2	2
4	P4	45,00	,60	1	4
5	P5	90,00	2,40	2	2
6	P6	175,00	3,60	2	1
7	P7	60,00	1,20	1	2
8	P8	120,00	1,92	1	1
9	P9	55,00	,70	2	4
10	P10	160,00	4,62	2	3

1. Créer un nouveau fichier de données SPSS appelé "Tp5.sav".
2. Les données doivent contenir cinq variables nommées : Patient, Weight, Bloodsugar, Sexe, et Bloodgroup (voir la Figure ci-dessus).
3. La Variable "Patient" est une variable de type Chaîne.
4. La Variable "Weight" est une variable quantitative continue de type Numérique.
5. La Variable "Bloodsugar" est une variable quantitative continue de type Numérique.

6. Entrez les valeurs possibles pour la variable qualitative "Sexe" : 1=Homme et 2=Femme.
7. Entrez les valeurs possibles pour la variable qualitative "Bloodgroup" : 1=AB , 2=A, 3=B et 4=O.

Exercice 2

1. Triez les données des patients en fonction de leurs valeurs de poids (Weight) ?
2. En utilisant les tableaux croisés, y a-t-il une relation entre les variables qualitatives Sexe et Bloodgroup ?
3. Prouver qu'il existe une corrélation linéaire entre la variable Bloodsugar et Weight ? ; Si oui, essayez d'extraire cette équation linéaire en utilisant la technique de régression linéaire.
4. représenter graphiquement la relation linéaire entre la variable Bloodsugar et Weight ?

Solution

1. Choisissez Données ➜ Trier les observations ➜ Trier par : Weight ➜ OK.
2. Choisissez Analyser ➜ Statistiques descriptives ➜ Tableaux croisés ➜ Placez la variable Sexe dans la zone Ligne(s) et la variable Bloodgroup dans la zone Column(s) ➜ Cellules ➜ Sélectionnez le pourcentage de Position ➜ Poursuivre ➜ OK.
3. (a) Choisissez Analyse ➜ Corrélation ➜ Bivarié ➜ Placez les variables Weight et Bloodsugar dans la zone Variables ➜ Options ➜ cochez l'option « Ecarts des produits croisés et covariances » ➜ Poursuivre ➜ OK.
   (b) Sélectionnez Analyse ➜ Régression ➜ Linéaire ➜ Placez la variable Bloodsugar dans la zone Dépendant et la variable Weight dans la zone Indépendantes ➜ OK.
4. Choisissez Graphiques ➜ Générateur de graphiques ➜ OK ➜ Réinitialiser ➜ Dispersion/Points ➜ Diagramme de dispersion : simple ➜ Sélectionnez Weight et faites-le glisser vers le rectangle intitulé X-Axis dans le diagramme ➜ Sélectionnez Bloodsugar et faites-le glisser vers le rectangle intitulé Y-Axis dans le diagramme ➜ OK ➜ Double-cliquez sur le graphique produit ➜ Appuyez sur l'icône « Ajouter une courbe d'ajustement au total » ➜ Fermer ➜ Fermez la fenêtre « Editeur de graphiques ».

Pour plus de détails, reportez-vous au chapitre 6

TP 6

But

1. Comprendre les concepts de distributions binomiales et normales.
2. Savoir comment utiliser SPSS pour effectuer des distributions binomiales et normales.
3. Comprendre les fonctions de densité de probabilité (PDF), de fonction de distribution cumulative (CDF) et de fonction de distribution inverse (IDF).
4. Être capable d'interpréter les résultats des distributions binomiales et normales à partir des sorties de SPSS.
5. Apprendre à estimer la moyenne d'une population avec un intervalle de confiance.

Exercice 1

On sait que 60 % des souris inoculées avec un sérum sont protégées d'une certaine maladie. Si 5 souris sont inoculées, quelle est la probabilité que :

1. exactement 3
2. au moins 3
3. au plus 3

des souris contractent la maladie ?

Exercice 2

Si X est une variable aléatoire de loi normale standard (càd centrée réduite : $\mu = 0$ et $\sigma = 1$) :

1. Trouver P(X<-2) ?
2. Trouver P(-1<X<0,5) ?
3. Trouver P(4X≥-3) ?
4. Trouver la valeur de u_0 telle que P(|X|< u_0)= 0,82 ?
5. Trouver la valeur de v_0 telle que P(X< $-v_0$)= 0,61 ?
6. Trouver la valeur de x_0 telle que P(X≥x_0)= 0,05 ?

Exercice 3

Pour un échantillon de taille 144 on a trouvé une moyenne de 127,1 et un écart-type s = 6,17.
Déterminer l'intervalle de confiance au risque de 5%.

Solution

Ex1

Soit X le nombre de souris qui contractent la maladie, alors P(une souris attrape la maladie) = p = 0,4 $\Rightarrow$ q = 1 - p = 0,6.
Nombre de toutes les souris = n = 5

1. Pour P(X=3) :
   (a) Dans l'onglet Vue des variables, créez une nouvelle variable d'échelle (continue) nommée *Prob1* avec quatre chiffres derrière la virgule (Décimal=4).
   (b) Tapez le chiffre 0 dans la cellule 1 :*Prob1*.
   (c) Choisissez Transformer ➜ Calculer la variable ➜ PDF et PDF non centré ➜ Double-cliquez sur la fonction Pdf.Binom ➜ Donnez la variable cible le nom : *Prob1* ➜ Tapez l'expression PDF.BINOM (3,5,0.4) ➜ Cliquez sur OK.
   (d) On retrouve dans la vue de données le résultat : *Prob1*=0,2304.
2. Pour P(X$\geq$3) = 1 - P(X$\leq$2) :
   (a) Dans l'onglet Vue des variables, créez une nouvelle variable d'échelle (continue) nommée *Prob2* avec quatre chiffres derrière la virgule (Décimal=4).
   (b) Tapez le chiffre 0 dans la cellule 1 :*Prob2*.
   (c) Choisissez Transformer ➜ Calculer la variable ➜ CDF et CDF non centré ➜ Double-cliquez sur la fonction Cdf.Binom ➜ Donnez la variable cible le nom : *Prob2* ➜ Tapez l'expression 1 - CDF.BINOM (2,5,0.4) ➜ Cliquez sur OK.
   (d) On retrouve dans la vue de données le résultat : *Prob2*=0,3174.
3. Pour P(X$\leq$3) :
   (a) Dans l'onglet Vue des variables, créez une nouvelle variable d'échelle (continue) nommée *Prob3* avec quatre chiffres derrière la virgule (Décimal=4).
   (b) Tapez le chiffre 0 dans la cellule 1 :*Prob3*.
   (c) Choisissez Transformer ➜ Calculer la variable ➜ CDF et CDF non centré ➜ Double-cliquez sur la fonction Cdf.Binom ➜ Donnez la variable cible le nom : *Prob3* ➜ Tapez l'expression CDF.BINOM (3,5,0.4) ➜ Cliquez sur OK.
   (d) On retrouve dans la vue de données le résultat : *Prob3*=0,9130.

Ex2

(a) Dans l'onglet Vue des variables, créez une nouvelle variable d'échelle (continue) nommée *Prob* avec quatre chiffres derrière la virgule (Décimal=4).
(b) Tapez le chiffre 0 dans la cellule 1 :*Prob*.

1. Pour P(X<-2) :
   (c) Choisissez Transformer ➜ Calculer la variable ➜ CDF et CDF non centré ➜ Double-cliquez sur la fonction Cdf.NORMAL ➜ Donnez la variable cible le nom : *Prob* ➜ Tapez l'expression CDF.NORMAL (-2,0,1) ➜ Cliquez sur OK.

(d) On retrouve dans la vue de données le résultat : *Prob*=0,0228.

2. P(-1<X<0,5) :

   (c) Choisissez Transformer ➜ Calculer la variable ➜ CDF et CDF non centré ➜ Double-cliquez sur la fonction Cdf.NORMAL ➜ Donnez la variable cible le nom : *Prob* ➜ Tapez l'expression CDF.NORMAL (0.5,0,1) - CDF.NORMAL (-1,0,1) ➜ Cliquez sur OK.

   (d) On retrouve dans la vue de données le résultat : *Prob*=0,5328.

3. Pour P(4X≥-3) :

   (c) Choisissez Transformer ➜ Calculer la variable ➜ CDF et CDF non centré ➜ Double-cliquez sur la fonction Cdf.NORMAL ➜ Donnez la variable cible le nom : *Prob* ➜ Tapez l'expression 1 - CDF.NORMAL (-3/4,0,1) ➜ Cliquez sur OK.

   (d) On retrouve dans la vue de données le résultat : *Prob*=0,7734.

4. Pour la valeur de u_0 telle que P(|X|$< u_0$)= 0,82 :

   (a) Dans l'onglet Vue des variables, créez une nouvelle variable d'échelle (continue) nommée *U0* avec quatre chiffres derrière la virgule (Décimal=4).

   (b) Tapez le chiffre 0 dans la cellule 1 :*U0*.

   (c) Choisissez Transformer ➜ Calculer la variable ➜ Functions de distribucion inverse ➜ Double-cliquez sur la fonction Idf.NORMAL ➜ Donnez la variable cible le nom : *U0* ➜ Tapez l'expression IDF.NORMAL (0.91,0,1) ➜ Cliquez sur OK.
   P(|X|$< u_0$)= 1 - P(|X|$\geq u_0$)
   P(|X|$\geq u_0$)= P(X$< -u_0 \cup$ X$> u_0$)
   P(|X|$< u_0$)= 1 - (P(X$< -u_0 \cup$ X$> u_0$))
   P(|X|$< u_0$)= 1 - (P(X$< -u_0$) + P(X$> u_0$))
   P(|X|$< u_0$)= 1 - (P(X$< -u_0$) + P(X$< -u_0$))
   P(|X|$< u_0$)= 1 - 2*P(X$< -u_0$)
   0,82 = 1 - 2*P(X$< -u_0$)
   P(X$< -u_0$) = 0,09
   P(X$< u_0$) = 1 - 0,09 = 0,91

   (d) On retrouve dans la vue de données le résultat : *U0*=1,3408.

5. Pour la valeur de v_0 telle que P(X$< -v_0$)= 0,61 :

   (a) Dans l'onglet Vue des variables, créez une nouvelle variable d'échelle (continue) nommée *V0* avec quatre chiffres derrière la virgule (Décimal=4).

   (b) Tapez le chiffre 0 dans la cellule 1 :*V0*.

   (c) Choisissez Transformer ➜ Calculer la variable ➜ Functions de distribucion inverse ➜ Double-cliquez sur la fonction Idf.NORMAL ➜ Donnez la variable cible le nom : *V0* ➜ Tapez l'expression -IDF.NORMAL (0.61,0,1) ➜ Cliquez sur OK.

   (d) On retrouve dans la vue de données le résultat : *V0*=-0,2793.

6. Pour la valeur de x_0 telle que P(X$\geq x_0$)= 0,05 :

(a) Dans l'onglet Vue des variables, créez une nouvelle variable d'échelle (continue) nommée *X0* avec quatre chiffres derrière la virgule (Décimal=4).

(b) Tapez le chiffre 0 dans la cellule 1 :*X0*.

(c) Choisissez Transformer ➜ Calculer la variable ➜ Functions de distribucion inverse ➜ Double-cliquez sur la fonction Idf.NORMAL ➜ Donnez la variable cible le nom : *X0* ➜ Tapez l'expression IDF.NORMAL (1-0.05,0,1) ➜ Cliquez sur OK.

(d) On retrouve dans la vue de données le résultat : *X0*=1,6449.

Ex3

1. Dans l'onglet vue des variables, créez trois nouvelles variables d'échelles (continues) nommées xbar, sigma et n.
2. Dans l'onglet vue des données, sous la variable xbar, tapez « 127,1 » pour la moyenne de l'échantillon, sous la variable sigma, tapez « 6,17 » pour l'écart type de la population, et sous la variable n, tapez « 144 » pour la taille de l'échantillon.
3. Choisissez Transformer ➜ Calculer la variable.
4. Donnez la variable cible le nom : lowerbound et tapez l'expression xbar – 1.96 * sigma/SQRT(n) ➜ Cliquez sur OK.
5. Répétez les étapes 3 à 4, en utilisant upperbound et l'expression : xbar + 1,96 * sigma/SQRT(n) ➜ Cliquez sur OK.
6. Les cellules lowerbound et upperbound afficheront les valeur de 126,09 et 128,11, respectivement.

Pour plus de détails, reportez-vous au chapitre 7

Références bibliographiques

[ChatGPT, 2023] CHATGPT (2023). Chat.openai.com. https://chat.openai.com/. Accessed : 2020-01-01.

[Cronk, 2019] CRONK, B. C. (2019). *How to use SPSS® : A step-by-step guide to analysis and interpretation*. Routledge.

[Denis, 2018] DENIS, D. J. (2018). *SPSS data analysis for univariate, bivariate, and multivariate statistics*. John Wiley & Sons.

[Jean-Pierre, 2016] JEAN-PIERRE, L. (2016). Statistiques et probabilités : Cours et exercices corrigés.

[Microsoft Corporation,] MICROSOFT CORPORATION. Microsoft excel. https://support.microsoft.com/en-us/excel. Accessed : 2019-09-01.

[Nasir *et al.*, 2022] NASIR, M. A., BAKOUCH, H. S. et JAMAL, F. (2022). *Introductory Statistical Procedures with SPSS*. Bentham Science Publishers.

[Salcedo et McCormick, 2020] SALCEDO, J. et MCCORMICK, K. (2020). *SPSS Statistics for Dummies*. John Wiley & Sons.

[Twigg, 2010] TWIGG, M. (2010). Discovering statistics using spss.

[uiowa, 2024] UIOWA (2024). Normal distribution applet/calculator. https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html. Accessed : 2024-02-01.