L'UNIVERSITÉ D'ÉTAT DE MÉDECINE ET PHARMACIE
„NICOLAE TESTEMITANU" DE RÉPUBLIQUE DE MOLDOVA

1945
2025

Chair de médecine sociale et de management
«Nicolae Testemitanu»

Elena Raevschi, Olga Penina, Galina Obreja

Biostatistique de Base et Méthodologie de la Recherche

Chisinau
Centre Éditorial et de Publication *Medicina*
2025

CZU[311.2+001.891]:614.2(075.8)
R 19

Approuvé par le Comité de Gestion de la Qualité de l'Université d'État de Médecine et de Pharmacie « Nicolae Testemițanu », Procès-Verbal no. 2, 23.10.2024.

Auteurs :

Elena RAEVSCHI – Docteur Habilité en Sciences Médicales, Professeur, Chair de médecine sociale et de management « Nicolae Testemițanu ».

Olga PENINA – Docteur Habilité en Sciences Médicales, Maître de Conférences, Chair de médecine sociale et de management « Nicolae Testemițanu ».

Galina OBREJA – Docteur en Sciences Médicales, Maître de Conférences, Chair de médecine sociale et de management « Nicolae Testemițanu ».

Réviseurs :

Larisa SPINEI – Docteur Habilité en Sciences Médicales, Professeur, Chair de médecine sociale et de management « Nicolae Testemițanu ».

Alexandru CORLATEANU – Docteur Habilité en Sciences Médicales, Professeur, Département de Médecine Interne.

Redactor: În redacția autorilor

Cette publication est présentée dans sa forme originale.

Le présent compendium « Biostatistique de Base et Méthodologie de la Recherche » a été élaborée sur la base de l'expérience internationale et des normes didactiques mises à jour. Il correspond aux exigences du curriculum pour le cours de Biostatistique et Méthodologie de la Recherche Scientifique pour les étudiants en médecine de l'enseignement supérieur intégré et de la licence de l'Université d'État de Médecine et de Pharmacie « Nicolae Testemițanu ».

DESCRIEREA CIP A CAMEREI NAȚIONALE A CĂRȚII DIN REPUBLICA MOLDOVA

Raevschi, Elena.

Biostatistique de Base et Méthodologie de la Recherche / Elena Raevschi, Olga Penina, Galina Obreja; Universitatea de Stat de Medicină şi Farmacie „Nicolae Testemiţanu" din Republica Moldova, Chair de médecine sociale et de management. – Chişinău: CEP *Medicina*, 2025. – 180 p.: il., tab.

Bibliogr.: p. 176-177. – În red. aut. – 50 ex.

ISBN 978-9975-82-406-4.

[311.2+001.891]:614.2(075.8)

R 19

ISBN 978-9975-82-406-4

SOMMAIRE

Préface

La biostatistique et la méthodologie de la recherche, en tant que discipline, constitue un domaine interdisciplinaire qui intègre une vaste gamme de contributions en matière de connaissances et est essentielle pour la réalisation de recherches conformes aux normes internationales. Le compendium « Biostatistique de base et méthodologie de la recherche » est en adéquation avec le programme de la discipline et fournissent des informations détaillées et bien structurées sur chaque étape nécessaire à la conduite d'une étude scientifique rigoureuse. Ce guide se présente comme une introduction à la biostatistique et à la méthodologie de la recherche pour les étudiants en sciences de la santé, avec pour objectif de faciliter l'élaboration de leurs mémoires de licence.

Ce compendium est publié en quatre langues (roumain, russe, anglais et français) et offre une structure révisée, un contenu enrichi et des discussions approfondies sur divers sujets tout au long du livre. Il inclut des figures et des tableaux supplémentaires pour éclairer les concepts, notamment dans les domaines des statistiques descriptives et inférentielles. Le livre fournit des explications détaillées sur les tests d'hypothèses en utilisant des tests paramétriques et non paramétriques, ainsi que des analyses exhaustives de la corrélation et de la régression. Les tableaux de valeurs critiques pour la distribution t et la distribution chi-carré ont été ajoutés en Annexe A et Annexe B, respectivement.

Remerciements

Nous exprimons notre profonde gratitude aux professeurs ayant revu le manuscrit : Larisa Spinei, Docteur Habilité en Sciences Médicales, Professeur à la Chair de médecine sociale et de management « Nicolae Testemițanu », et Alexandru Corlateanu, Docteur Habilité en Sciences Médicales, Professeur au Département de Médecine Interne, tous deux affiliés à l'Université de Médecine et Pharmacie Nicolae Testemitanu. Nous tenons également à remercier tous les professeurs ayant contribué à l'enseignement du cours et ayant fourni des suggestions précieuses.

Les auteurs

CHAPITRE 1. INTRODUCTION AUX BIOSTATISTIQUES DE BASE ET À LA MÉTHODOLOGIE DE RECHERCHE MÉDICALE

Le compendium intitulé « Biostatistique de base et Méthodologie de la recherche » introduit les étudiants en médecine à l'étude des statistiques appliquées à la médecine et à d'autres disciplines du domaine de la santé. L'objectif principal est de développer des connaissances sur les méthodes contemporaines utilisées dans la recherche pratique. L'acquisition des connaissances nécessaires pour appliquer les méthodes modernes de documentation, l'assimilation des définitions théoriques pertinentes en recherche ainsi que le respect des normes de régulation sont essentiels pour mettre en évidence les résultats de recherche dans le cadre d'une thèse de premier cycle.

Le cours de Biostatistique et Méthodologie de la Recherche Scientifique englobe les aspects théoriques et pratiques relatifs à la réalisation de recherches scientifiques et à l'analyse des données statistiques. Ce cours, ayant un contenu comparable à celui des autres universités européennes avec des informations actualisées, fournit aux étudiants les connaissances nécessaires pour mener à bien des recherches scientifiques dans le domaine des sciences biomédicales. Il propose une approche principalement appliquée des méthodes statistiques nécessaires à la résolution de problèmes pratiques dans le domaine biomédical.

Objectif principal du cours :

Aider les étudiants à comprendre les concepts de base de la biostatistique et de la méthodologie de la recherche de manière à ce qu'ils puissent les utiliser pour planifier et analyser des données en recherche biomédicale.

Au niveau de la connaissance et de la compréhension :

- Connaître les concepts théoriques de la méthodologie de la recherche scientifique médicale.
- Développer une pensée claire et continue, capable de gérer et de traiter les données.
- Connaître les principes, technologies, méthodes et techniques utilisées en recherche médicale.
- Comprendre la corrélation entre les méthodes modernes utilisées en biostatistique et en méthodologie de la recherche médicale.
- Identifier les possibilités d'analyse et d'interprétation ainsi que les limites des méthodes modernes utilisées en recherche scientifique.

Au niveau de l'application:

- Analyser les définitions, les méthodes théoriques et pratiques de la méthodologie de la recherche scientifique.
- Utiliser des méthodes et des techniques statistiques dans le processus scientifique.
- Démontrer la capacité à analyser, interpréter et présenter les résultats de la recherche scientifique.
- Utiliser les connaissances de base en biostatistique nécessaires pour comprendre son application optimale dans la recherche scientifique.
- Maîtriser le langage et la terminologie spécifiques au style scientifique.
- Évaluer les informations contenues dans un article ou un rapport de spécialité et apprécier leur pertinence.
- Être capable de rechercher des informations scientifiques en utilisant des méthodes classiques ou informatiques pour la recherche et la sélection de données.

- Utiliser des méthodes modernes pour la rédaction et la présentation d'une proposition scientifique et d'un rapport des résultats finaux.

Au niveau de l'intégration :

- Apprécier la valeur théorique et applicable de la Méthodologie de la Recherche Médicale dans différentes disciplines du domaine de la santé.
- Évaluer la place et le rôle de la biostatistique et de la méthodologie de recherche dans la carrière professionnelle médicale.
- Intégrer les connaissances en biostatistique et méthodologie de recherche avec les disciplines cliniques.
- Être capable d'appliquer les connaissances acquises aux activités pratiques et de recherche.
- Être compétent dans l'utilisation critique des informations provenant des publications scientifiques dans ses propres recherches en utilisant les nouvelles technologies de l'information et de la communication.

Résultats de l'étude :

- Expliquer les concepts de base concernant l'organisation de la recherche scientifique et la publication des résultats.
- Développer un projet de recherche dans le domaine biomédical.
- Présenter la description des données expérimentales en fonction de leur nature et expliquer correctement les résultats de l'inférence statistique.
- Déterminer les méthodes statistiques pour l'analyse des données en tenant compte des caractéristiques du plan d'étude, de l'échelle de mesure et du nombre de variables impliquées.

- Caractériser les principales caractéristiques des conceptions d'étude épidémiologique (observationnelles et expérimentales), ainsi que leurs avantages et limites.
- Réaliser une étude épidémiologique (observationnelle ou expérimentale) et interpréter correctement ses résultats.
- Rédiger un article scientifique, y compris la thèse de licence, et tirer parti de ses résultats.
- Évaluer le rôle et l'importance de la biostatistique et de la méthodologie de recherche dans le contexte moderne de la médecine basée sur les preuves.
- Faire preuve d'ouverture à l'apprentissage tout au long de la vie.

La recherche en santé est un domaine interdisciplinaire qui repose sur un large éventail de connaissances. La biostatistique et la méthodologie de la recherche scientifique constituent une discipline qui intègre et analyse les connaissances acquises à partir des études fondamentales et appliquées. Cette discipline est essentielle pour l'évaluation des activités de recherche selon les normes contemporaines. En tant que domaine intégrateur, elle est en connexion avec d'autres disciplines utilisant les statistiques. Une compréhension approfondie des mathématiques de niveau secondaire ainsi que des bases de la biomédecine est indispensable pour maîtriser cette discipline.

CHAPITRE 2. STATISTIQUES DESCRIPTIVES : PRÉSENTATION DES DONNÉES

Concepts clés

- *Statistique :* Caractéristique ou valeur dérivée des données d'un échantillon.
- *Paramètre* : Caractéristique ou valeur dérivée des données d'une population.
- *Variable* : Caractéristique mesurée d'une unité d'observation.
- *Variable quantitative (numérique)* : Variable pour laquelle les valeurs attribuées sont ordonnées et significatives.
- *Variable qualitative (catégorielle)* : Variable pour laquelle les valeurs attribuées ont une signification nominale (étiquettes) mais pas de valeur numérique.
- *Variable alternative (dichotomique)* : Variable qualitative ayant seulement deux catégories.
- *Variable non-alternative* : Variable qualitative ayant plus de deux catégories.
- *Variable discrète* : Variable numérique ne prenant que des nombres entiers.
- *Variable continue* : Variable numérique pouvant prendre toutes les valeurs sur un continuum.
- *Données nominales* : Données classées en différentes catégories qualitatives pouvant être énumérées dans n'importe quel ordre.
- *Données ordinales* : Données classées en différentes catégories où l'ordre entre les catégories est significatif, mais sans information sur la distance quantitative entre les catégories.
- *Données d'intervalle* : Données avec des intervalles égaux entre les éléments mais sans zéro absolu.
- *Données de rapport (ou de ratio)* : Données avec des intervalles égaux et un zéro absolu.
- *Distribution de fréquence* : Ensemble des valeurs d'un échantillon sur une seule variable.

- *Histogramme* et *polygone de fréquence* : Graphiques utilisés pour afficher la distribution de fréquence des données numériques. Ils renseignent sur la forme de la distribution de fréquence.
- *Diagramme de dispersion* : Graphique utilisé pour illustrer la relation entre deux variables numériques.
- *Diagramme en barres* et *diagramme circulaire* : Graphiques utilisés pour afficher des données qualitatives.

Les statistiques descriptives facilitent la compréhension et le résumé de vastes ensembles de données en utilisant quelques indicateurs clés. Ces indicateurs permettent d'interpréter rapidement une grande quantité d'informations. Par exemple, au lieu d'examiner individuellement les résultats des tests de 100 étudiants, il est plus efficace de se référer à la moyenne pour obtenir une vue d'ensemble des performances de la classe. Ces chiffres clés permettent d'identifier des motifs et des tendances dans les données, ce qui simplifie l'interprétation et la communication des résultats importants de votre recherche. Cependant, il est crucial de noter que les statistiques descriptives ne fournissent pas d'explication sur les raisons d'existence de ces motifs ni sur leur signification. Elles offrent plutôt des indices qui peuvent mener à la formulation de questions ou d'hypothèses. Une hypothèse constitue une explication potentielle qui peut être vérifiée par des recherches complémentaires.

2.1 Notions de base

2.1.1 Variable

Définition de la variable : Une variable est une caractéristique d'intérêt qui présente des valeurs différentes pour divers sujets ou objets inclus dans une étude.

Exemples : Âge, date de naissance, nationalité, nombre d'enfants, pression artérielle.

Classification des variables : Pour présenter correctement les statistiques descriptives, il est nécessaire de comprendre les types de données couramment rencontrés dans les études de recherche (*Figure 2.1*).

Variables quantitatives (numériques) : Ces variables peuvent être quantifiées et se classifient comme suit :

1. *Variables Discrètes :* Elles représentent des quantités qui ne peuvent prendre que des valeurs spécifiques, distinctes, généralement des nombres entiers, sans valeurs intermédiaires possibles.

Exemples : nombre de patients, nombre de nouveaux cas de maladies cardiaques, nombre de nouveau-nés dans une année donnée, etc.

2. *Variables Continues :* Elles représentent des quantités qui peuvent prendre n'importe quelle valeur dans une plage, sans être limitées à des valeurs discrètes spécifiques.

Exemples : âge, niveau de glucose dans le sang, pression artérielle, etc.

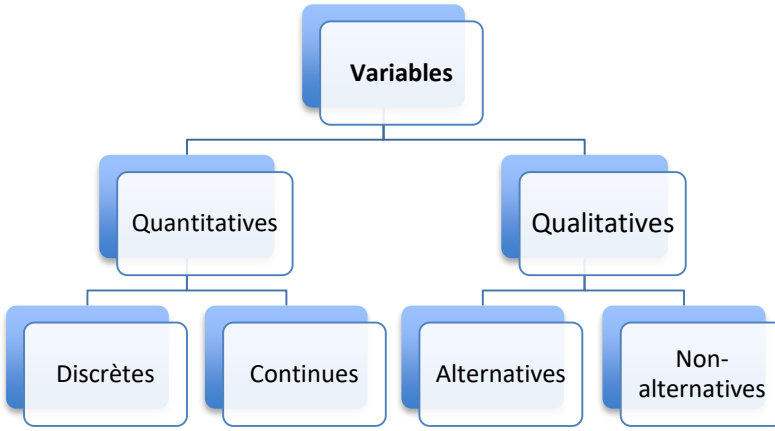

Figure 2.1 Types de variables

Variables Qualitatives (Catégorielles). Ces variables ne peuvent pas être quantifiées et sont classées comme suit :

1. *Variables alternatives (dichotomiques ou binaires) :* Représentent des catégories où les résultats ne peuvent prendre qu'une des deux valeurs possibles.

Exemples : « Oui » ou « Non » ; « Masculin » ou « Féminin », etc.

2. *Variables non-alternatives :* Représentent des catégories où les résultats peuvent prendre plusieurs valeurs.

Exemples : groupe sanguin, niveau de gravité d'une maladie, etc.

Dans les analyses, les variables qualitatives sont souvent codées par des valeurs numériques.

2.1.2 Population

Une population se réfère à l'ensemble des individus ou des éléments partageant une caractéristique commune et faisant l'objet d'une étude. Elle inclut chaque membre répondant aux critères définis par l'étude. Par exemple, une population pourrait être tous les élèves d'une école particulière, tous les patients d'un hôpital ou tous les arbres d'une forêt.

- *Population cible :* Il s'agit du groupe entier auquel les résultats de l'étude sont destinés à être généralisés. On l'appelle également la population théorique.
- *Population étudiée :* Il s'agit du segment spécifique de la population cible qui est accessible et dont les données sont effectivement collectées pour l'étude. On l'appelle également la population *accessible.*

2.1.3 Échantillon

L'échantillon représente un sous-ensemble de la population étudiée. *L'unité d'observation*, ou unité statistique, est l'entité à partir de laquelle les informations sont collectées, telle qu'un individu, un ménage, une communauté, une école, etc. Il est important d'identifier clairement l'unité d'observation pour un design logique du sondage, une collecte de données organisée et une analyse objective.

Population cible vs. population d'étude : Recherche sur les maladies cardiovasculaires.

- **Population cible :** Tous les adultes âgés de 40 ans et plus ayant été diagnostiqués avec une maladie cardiovasculaire dans le monde entier.
- **Population d'étude :** Adultes âgés de 40 ans et plus avec des maladies cardiovasculaires qui sont patients dans les principaux hôpitaux de la République de Moldavie**.**

La population d'étude sert de sous-ensemble représentatif de la population cible plus large.

2.1.4 Paramètres et statistiques

Un paramètre est une caractéristique ou une valeur dérivée des données de la population.

Une statistique est une caractéristique ou une valeur dérivée des données de l'échantillon.

Tableau 2.1 Symboles pour la population et l'échantillon

	Paramètres (données de la population)	**Statistiques (données de l'échantillon)**
	Symboles Grecs	*Symboles Romains*
Moyenne	μ	$\bar{x}$
Écart-type	σ	S
Variance	σ^2	S^2
Taille (nombre d'observations)	N	n

2.2 Échelles de mesure

Les échelles de mesure varient en fonction de la nature des variables. Elles définissent la manière dont les données sont affichées et résumées, ainsi que les méthodes statistiques employées pour leur analyse. En statistique, on distingue trois types d'échelles de mesure :

- Échelle nominale (de classification);
- Échelle ordinale (de classement);
- Échelle numérique (incluant les échelles d'intervalle et de rapport).

2.2.1 Échelle nominale

Les données sur l'échelle nominale se composent de catégories qui n'ont pas d'ordre intrinsèque. Une variable mesurée sur une échelle nominale peut comporter deux ou plusieurs sous-catégories, selon le degré de variation d'une variable qualitative.

Exemple 1. La variable « sexe » a généralement deux catégories : masculin et féminin. Les données nominales qui appartiennent à l'une des deux catégories distinctes, telles que masculin ou féminin, sont

considérées comme des variables catégoriques alternatives ou dichotomiques.

Exemple 2. Cependant, toutes les données nominales ne sont pas dichotomiques. De nombreuses variables nominales ont trois catégories ou plus. Par exemple, « anémie » peut être classée en plusieurs sous-catégories : anémie microcytaire, anémie macrocytaire ou mégaloblastique et anémie normocytaire. Cette variable est considérée comme une variable catégorique non-alternative. L'ordre dans lequel ces catégories sont listées n'a pas d'importance, car il n'y a pas de hiérarchie ou de relation entre elles.

Les données sur l'échelle nominale servent à étiqueter les variables sans fournir de valeur quantitative. Chaque catégorie est unique, et la séquence des catégories est sans importance.

2.2.2 Échelle ordinale

L'échelle ordinale, aussi connue sous le nom d'échelle de classement, permet de catégoriser les individus, objets, réponses ou propriétés en sous-groupes basés sur une caractéristique commune, puis de les classer dans un ordre précis.

Ces sous-groupes sont arrangés en ordre croissant ou décroissant selon l'ampleur de la variation de la variable en question. Lorsque l'ordre parmi les catégories devient significatif, les données sont alors considérées comme mesurées sur une échelle ordinale.

Exemple 1. Le « revenu » d'un patient peut être évalué à l'aide d'une variable qualitative avec les catégories « supérieur à la moyenne », « moyenne » et « inférieur à la moyenne ». La « distance » entre ces sous-catégories n'est pas égale, car aucune unité de mesure quantitative n'est utilisée.

Exemple 2. Le score Apgar, qui évalue la maturité des nouveau-nés, varie de 0 à 10. Des scores plus bas indiquent une dépression des

fonctions cardiorespiratoires et neurologiques, tandis que des scores plus élevés indiquent un bon fonctionnement. La différence entre les scores de 8 et 9 n'a pas les mêmes implications cliniques que celle entre les scores de 0 et 1. Cela montre qu'au sein des échelles ordinales, les intervalles entre les points de mesure ne sont pas égaux.

Les données de l'échelle ordinale permettent non seulement de catégoriser les variables, mais aussi de les classer, bien que les intervalles entre les rangs ne soient pas nécessairement uniformes ou mesurables.

2.2.3 Échelle numérique

L'échelle numérique est caractérisée par des intervalles égaux entre les points de mesure successifs. Il existe deux types d'échelles numériques :

⇒ *Échelle d'intervalle*

L'échelle d'intervalle englobe toutes les caractéristiques des échelles nominales et ordinales. De plus, elle permet de classer les données selon un ordre hiérarchique avec des distances égales entre les points. Les données d'une échelle d'intervalle ne possèdent pas de zéro absolu (c'est-à-dire un point qui signifierait l'absence complète de la variable), ce qui implique qu'elles peuvent prendre des valeurs significatives tant en dessous qu'au-dessus de zéro.

Example.

Échelle Celsius : 0°C - 100°C

Échelle Fahrenheit : 32°F - 212°F

Les échelles de température Celsius et Fahrenheit mesurent la température avec des intervalles égaux entre les degrés. Ces deux échelles comprennent des valeurs inférieures à zéro (pour l'échelle Celsius) ou en dessous du point de congélation de l'eau (32°F pour

l'échelle Fahrenheit), indiquant des températures inférieures au point de congélation de l'eau.

⇒ *Échelle de rapport*

L'échelle de rapport (ou de ratio) possède toutes les propriétés des échelles nominales, ordinales et d'intervalle, avec une caractéristique supplémentaire : un point zéro fixe. Cela signifie qu'elle a une valeur zéro absolue, indiquant l'absence totale de la variable (aucune valeur en dessous de zéro n'est possible).

Des exemples de variables mesurées sur une échelle de rapport incluent le taux de cholestérol sérique, le revenu, l'âge, la taille et le poids.

Le tableau ci-dessous souligne les différences essentielles entre les quatre échelles de mesure (*Tableau 2.2*).

Tableau 2.2 Différence entre les quatre échelles de mesure

Échelle	Indique la différence	Indique la direction de la différence	Indique la quantité de la différence	Zéro absolu
Nominal	+			
Ordinal	+	+		
D'intervalle	+	+	+	
De rapport	+	+	+	+

2.3 Tableaux

Un tableau constitue la méthode la plus simple pour synthétiser un ensemble de données et peut être employé pour tous les types de variables.

2.3.1 Distributions de fréquence

Une méthode fréquemment employée pour la synthèse des données est le tableau de distribution de fréquence. Ce tableau catégorise les

données et présente les décomptes numériques correspondants pour divers types de données. Pour les données nominales et ordinales, la distribution de fréquence organise les catégories avec leurs décomptes respectifs. Par exemple, le *Tableau 2.3* illustre la répartition des nouveau-nés selon le sexe en République de Moldavie pour l'année 2022, en utilisant des données nominales.

Tableau 2.3 Distribution de fréquence pour les données nominales : distribution des naissances par sexe en République de Moldavie en 2022

Sexe des nouveau-nés	Nombre de naissances
Masculin	13950
Féminin	13068

Pour représenter les données numériques, on utilise à la fois des distributions de fréquence non groupées et groupées. Par exemple, le Tableau 2.4 présente une distribution de fréquence non groupée, illustrant le nombre annuel de naissances en République de Moldavie de 2000 à 2020, en utilisant des données d'intervalle.

Tableau 2.4 Distribution de fréquence pour les données d'Intervalle : nombre annuel de naissances en République de Moldavie, de 2000 à 2020

Année	Nombre de naissances
2000	36939
2010	40474
2020	30834
2022	27018

Lorsque les données numériques sont très détaillées, les *distributions de fréquence groupées* sont particulièrement utiles. Cette méthode consiste à diviser la plage de valeurs en intervalles distincts et non chevauchants. Une fois ces intervalles établis, le nombre d'observations dans chaque intervalle est compté. Le Tableau 2.5 illustre cette méthode en détaillant la distribution de fréquence

groupée des naissances selon l'âge des mères en République de Moldavie pour l'année 2022, en utilisant des données de rapport.

Tableau 2.5 Distribution de fréquence groupée pour les données de rapport : nombre de naissances par âge de la mère en République de Moldavie en 2022

Groupe d'âge de la mère	Nombre de naissances
Moins de 24 ans	7117
25-34 ans	15032
35-44 ans	4843
45 ans et plus	26

Les tableaux sont plus efficaces lorsqu'ils sont clairs et bien organisés. Par conséquent, les tableaux et leurs colonnes doivent toujours être clairement étiquetés, et les unités de mesure doivent être précisées si nécessaire.

2.3.2 Fréquence relative

Il est parfois utile de connaître la proportion des valeurs qui se situent dans un intervalle donné d'une distribution de fréquences, plutôt que le nombre absolu. *La fréquence relative* pour un intervalle est la proportion du nombre total d'observations qui se trouvent dans cet intervalle. Cette fréquence relative est calculée en divisant le nombre de valeurs présentes dans l'intervalle par le nombre total de valeurs dans le tableau, exprimé en pourcentage. Les fréquences relatives sont utiles pour comparer des ensembles de données avec un nombre d'observations inégal.

La fréquence relative cumulative pour un intervalle représente le pourcentage du nombre total d'observations ayant une valeur inférieure ou égale à la limite supérieure de cet intervalle. Elle est calculée en faisant la somme des fréquences relatives pour l'intervalle spécifié et pour tous les intervalles précédents.

Le *tableau 2.6* présente les fréquences absolues, relatives et cumulatives de l'indice de choc pour 931 patients.

Tableau 2.6 Fréquences absolues, relatives et cumulatives de l'indice de choc pour 931 patients

Score de l'Indice de Choc	Fréquence (Nombre de Patients)	Fréquence Relative (%)	Fréquence Relative Cumulative (%)
0.30-0.39	38	4.1	4.1
0.40-0.49	104	11.2	15.3
0.50-0.59	198	21.3	36.6
0.60-0.69	199	21.4	58.0
0.70-0.79	155	16.6	74.6
0.80-0.89	102	11.0	85.6
0.90-0.99	60	6.4	92.0
1.00-1.09	37	4.0	96.0
1.10-1.19	19	2.0	98.0
1.20-1.29	19	2.0	100.0
Total	931	100.0	

2.4 Graphiques

Les graphiques (ou diagrammes) constituent une seconde méthode pour résumer et présenter des données. Ils sont souvent plus faciles à interpréter que les tableaux, bien qu'ils puissent offrir des informations moins détaillées. Les graphiques les plus informatifs sont généralement simples et auto-explicatifs. Comme pour les tableaux, les graphiques doivent être clairement étiquetés et les unités de mesure doivent être précisées.

2.4.1 Graphiques linéaires

Les graphiques linéaires sont couramment employés pour illustrer les tendances dans le temps pour des données numériques. Chaque valeur

sur l'axe des x correspond à une valeur spécifique sur l'axe des y, et les points adjacents sont reliés par des lignes droites (*Figure 2.2*).

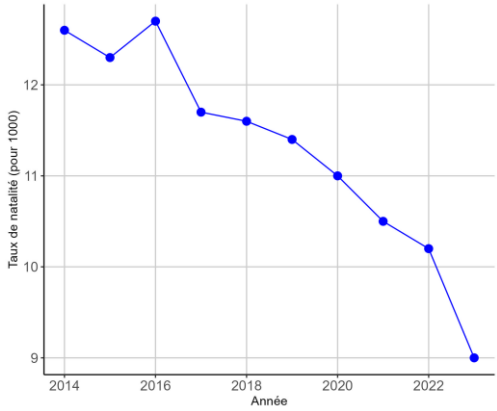

Figure 2.2 Taux de natalité dans la municipalité de Chisinau (2003-2023), pour 1000 habitants

2.4.2 Graphiques à barres

Les graphiques à barres, qu'ils soient verticaux ou horizontaux, constituent un outil largement utilisé pour représenter les distributions de fréquence des données nominales ou ordinales. Dans ces graphiques, les barres doivent être de largeur uniforme et espacées afin de prévenir une impression trompeuse de continuité. Les graphiques à barres permettent une comparaison efficace de plusieurs valeurs, les catégories étant affichées le long de l'axe vertical ou horizontal, comme le démontre la *Figure 2.3.*

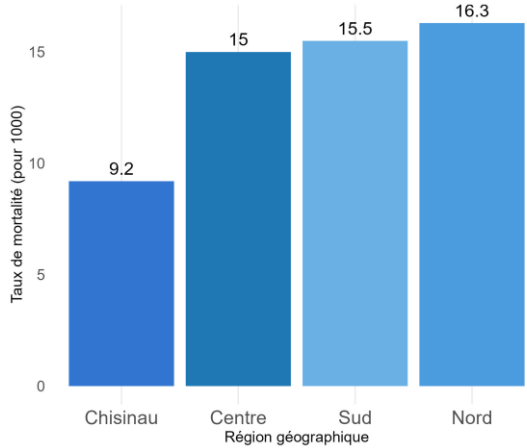

Figure 2.3 Taux de mortalité brut par régions géographiques en République de Moldavie en 2023, pour 1000 habitants

2.4.3 Graphiques circulaires

Les graphiques circulaires représentent la proportion de chaque valeur par rapport au total et sont utilisés pour illustrer les distributions de fréquence des données nominales.

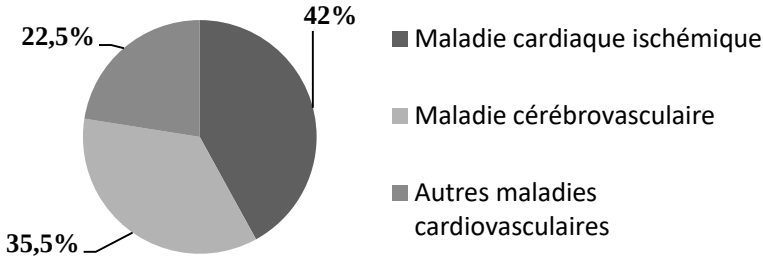

Figure 2.4 Répartition mondiale des décès cardiovasculaires dus aux infarctus du myocarde, aux accidents vasculaires cérébraux et à d'autres types de maladies cardiovasculaires, 2020

2.4.4 Histogrammes et Polygones de Fréquence

Les histogrammes constituent des méthodes particulièrement efficaces pour présenter les fréquences absolues et relatives des données numériques. Bien qu'ils ressemblent aux graphiques à barres, ils présentent des différences fondamentales :

- Les graphiques à barres illustrent les distributions de fréquence des données catégorielles (nominales ou ordinales), tandis que les histogrammes mettent en évidence les distributions de fréquence des données numériques (discrètes ou continues).
- Les histogrammes permettent de saisir la forme de la distribution des fréquences, alors que les graphiques à barres se contentent de comptabiliser les valeurs sans représenter la forme de la distribution.

Dans un histogramme, l'axe horizontal représente les limites des différents intervalles, tandis que l'axe vertical affiche la fréquence absolue ou relative.

Un polygone de fréquence, bien qu'il soit similaire à un histogramme, utilise des points reliés par des lignes droites plutôt que des barres, offrant ainsi une représentation graphique de la distribution du jeu de données. Pour élaborer un polygone de fréquence, placez les points médians de chaque intervalle sur l'axe horizontal et leurs fréquences correspondantes sur l'axe vertical, puis reliez ces points par des lignes droites. Cette méthode permet de visualiser la distribution des données sur les intervalles, de manière similaire à la représentation des fréquences des observations dans les histogrammes. Les polygones de fréquence et les histogrammes peuvent être superposés pour une analyse comparative, comme illustré à la *Figure 2.5* avec le jeu de données du *Tableau 2.6*.

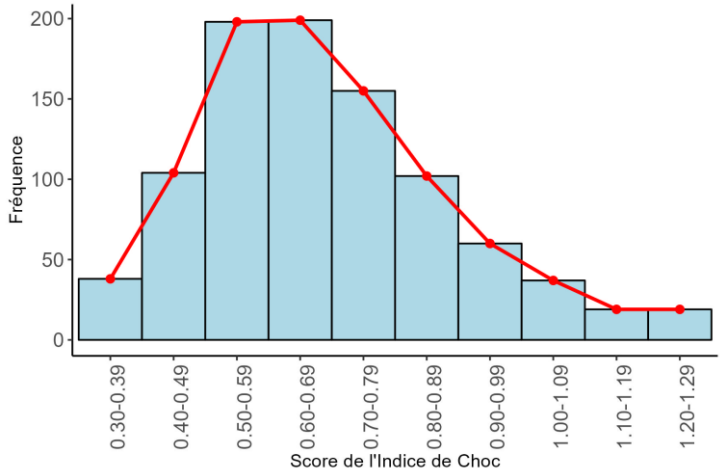

Figure 2.5 Histogramme et polygone de fréquence pour la distribution des scores de l'indice de choc chez 931 patients

2.4.5 Diagrammes en boîte

Les diagrammes en boîte, également appelés diagrammes en boîte et moustaches, sont des outils précieux pour résumer la distribution d'un ou plusieurs ensembles de données numériques. Contrairement à d'autres types de graphiques, les diagrammes en boîte offrent une vue succincte de la distribution des données sans afficher chaque point de donnée individuel.

La *Figure 2.6* illustre un diagramme en boîte ainsi que ses composants. La boîte centrale, qui peut être orientée verticalement ou horizontalement, s'étend du 25e au 75e percentile de l'ensemble de données. Une ligne à l'intérieur de la boîte représente la médiane (50e percentile), reflétant la tendance centrale des données. Lorsque la médiane est positionnée près du centre entre les quartiles, cela suggère une distribution symétrique des données.

Les lignes s'étendant depuis la boîte, appelées moustaches, représentent l'étendue des valeurs typiques dans l'ensemble de données. Ces moustaches s'étendent jusqu'à 1,5 fois l'intervalle interquartile (la différence entre le 75e et le 25e percentile) au-dessus et au-dessous de la boîte. Les points de données situés en dehors de ces moustaches sont marqués par des cercles, indiquant des valeurs aberrantes ou des valeurs qui dépassent l'étendue typique.

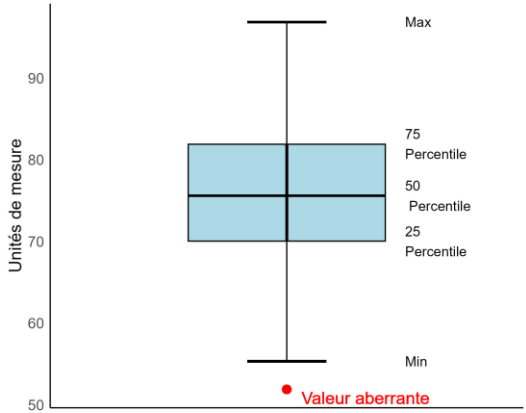

Figure 2.6 Diagramme en boîte et ses composants

2.4.6 Graphiques avec barres d'erreur

Les graphiques avec barres d'erreur sont couramment utilisés dans la recherche médicale pour comparer des groupes. Ils montrent la moyenne avec un cercle et illustrent la variabilité à l'aide de barres d'erreur ou d'écart-type. Ces graphiques offrent une vue d'ensemble des similitudes de distribution entre les groupes.

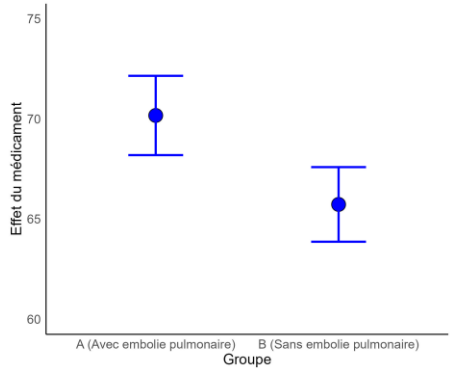

Figure 2.7 Graphiques avec barres d'erreur montrant l'effet du médicament chez les patients avec (A) et sans (B) embolie pulmonaire

2.4.7 Graphiques de dispersion

Un graphique de dispersion (ou nuage de points) est employé pour illustrer la relation entre deux variables numériques distinctes. Chaque point sur le graphique représente simultanément une paire de valeurs.

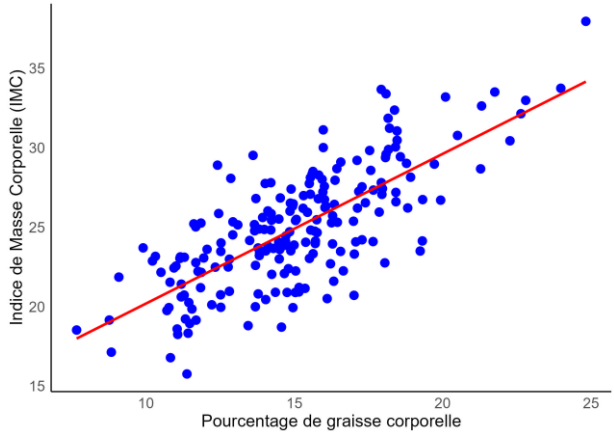

Figure 2.8 Graphique de dispersion illustrant la relation entre l'IMC et le pourcentage de graisse corporelle

Exercices de révision

1. Indiquez le type de variable et l'échelle de mesure appropriée pour les ensembles de données suivants:

a) Salaires de 125 médecins dans une clinique
b) Résultats des examens de tous les étudiants en médecine passant les examens d'hiver dans une année donnée
c) Niveau de cholestérol sérique des individus en bonne santé
d) Présence de diarrhée dans un groupe de nourrissons
e) Âge de début du cancer du sein chez les femmes
f) Température corporelle des patients
g) Issue des patients sortis de l'hôpital
h) Nombre de naissances dans une année donnée

2. Utilisez les données suivantes pour les représenter à l'aide de tous les graphiques appropriés. Justifiez votre choix.

a) Divergence des diagnostics cliniques et pathologiques à l'hôpital, 2017-2021

Années	2017	2018	2019	2020	2021
Divergence, %	11	9.8	8.0	9.2	8.2

b) Morbidité liée à l'hépatite virale aiguë en République de Moldova, 2021

Type	A	B	C	D	E
%	34.4	41.4	17.6	3.8	2.8

3. Proposez un ensemble de données qui peut être représenté par un graphique linéaire. Justifiez votre choix.
4. Proposez un ensemble de données qui peut être représenté par un graphique à barres. Justifiez votre choix.

Questions de révision

1. Quelles sont les statistiques descriptives ?
2. Quelles sont les classifications des variables ? Donnez des exemples.
3. En quoi une variable alternative diffère-t-elle d'une variable non alternative ? Donnez des exemples.
4. Définissez les échelles de mesure et classifiez-les. Donnez un exemple pour chaque type.
5. En quoi les données ordinales diffèrent-elles des données nominales ? Donnez des exemples.
6. Quels types de méthodes de présentation des données connaissez-vous ? Expliquez les différences entre elles.
7. Lors de la construction d'un tableau, dans quels cas il peut être avantageux d'utiliser des fréquences relatives plutôt que des fréquences absolues ?
8. Décrivez la présentation graphique des données : contenu et types. Donnez un exemple pour chaque type.
9. Quel est le graphique approprié pour la présentation des variables nominales ? Donnez des exemples.
10. Quel est le graphique approprié pour la présentation des variables ordinales ? Donnez des exemples.
11. Quel est le graphique approprié pour la présentation des variables numériques ? Donnez des exemples.

CHAPITRE 3. STATISTIQUES DESCRIPTIVES : RÉSUMÉ DES DONNÉES NUMÉRIQUES

Concepts clés

- *Les statistiques descriptives* servent uniquement à décrire, organiser ou résumer les données.
- *Le mode, la médiane et la moyenne* sont utilisés pour les données numériques (intervalles et rapports). La médiane est également appliquée pour les données ordinales.
- Le mode et la médiane sont *insensibles* aux valeurs aberrantes dans une distribution. La moyenne est *sensible* aux valeurs aberrantes.
- En l'absence des scores originaux d'une distribution, *la moyenne pondérée* peut être estimée à partir d'un tableau de fréquences.
- Dans une distribution *normale* (en forme de cloche), les trois mesures de tendance centrale sont identiques.
- Dans une distribution asymétrique *positive* (vers la droite), le mode est inférieur à la médiane.
- Dans une distribution asymétrique *négative* (vers la gauche), la médiane est inférieure au mode.
- *La moyenne* est utilisée pour les données numériques et les distributions normales.
- *La médiane* est utilisée pour les données ordinales ou les données numériques en cas de distribution asymétrique.
- Le mode est principalement utilisé pour les distributions bimodales.
- Une distribution *unimodale* présente un seul mode, une distribution *bimodale* en présente deux, tandis qu'une distribution *uniforme* ne possède pas de mode.
- *L'étendue* est sensible aux valeurs extrêmes dans une distribution.
- *L'intervalle interquartile* (IQR) permet de décrire les 50 % centraux d'une distribution, indépendamment de sa forme.
- L'IQR est défini comme la différence entre les percentiles 75 et 25.
- *La variance* (s^2) mesure la variabilité autour de la moyenne d'un ensemble de données.

- *L'écart type* (s) représente la distance moyenne entre les valeurs individuelles d'une distribution et la moyenne de celle-ci.
- *Le coefficient de variation* (CV) est une mesure de dispersion relative qui facilite la comparaison des observations mesurées sur différentes échelles.
- *L'écart type* est utilisé lorsqu'on se réfère à la moyenne (données numériques symétriques).
- L'IQR est utilisé en présence de la médiane (données ordinales ou données numériques asymétriques).
- *Une règle empirique* (règle 68-95-99,7) s'applique uniquement si les données suivent une distribution normale.
- *Le coefficient d'asymétrie de Pearson* permet de déterminer l'asymétrie dans un échantillon.

Les statistiques descriptives sont utilisées pour organiser et décrire les caractéristiques d'un ensemble de données. Contrairement aux statistiques inférentielles, les statistiques descriptives n'impliquent pas de test d'hypothèse ou d'analyse de données. Elles nous permettent de caractériser succinctement la distribution des valeurs dans leur ensemble.

Statistiques descriptives pour résumer les données numériques

⇒ *Mesures de tendance centrale* : Moyenne, Médiane et Mode

⇒ *Mesures de variabilité (dispersion)* : Étendue, Intervalle interquartile, Variance, Écart-type, Coefficient de variation

3.1 Mesures de tendance centrale

Les mesures de tendance centrale représentent des indicateurs statistiques qui caractérisent le point central d'un ensemble de données, là où les observations tendent à se concentrer. Les trois mesures couramment employées en médecine sont la moyenne, la médiane et le mode. Ces trois indicateurs servent à résumer des données numériques, tandis que la médiane est également utilisée pour les données ordinales.

3.1.1 Moyenne

La mesure de tendance centrale la plus fréquemment utilisée est la moyenne arithmétique. La moyenne est désignée par X-barre ($\bar{X}$), elle se calcule en divisant la somme (Σ) des valeurs individuelles (X_i) par le nombre total d'observations (n):

a) *Moyenne simple* – utilisée dans le cas d'une distribution de fréquence non groupée.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

b) *Moyenne pondérée* – utilisée pour une distribution de fréquence groupée :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\, f$$

Où f est la fréquence des valeurs individuelles.

La moyenne est généralement utilisée pour décrire des données numériques normalement distribuées. Elle est très sensible aux valeurs extrêmes, également appelées valeurs aberrantes. Par exemple, la moyenne de l'ensemble de données (1;2;2;3) est de 8/4, soit 2. Si le nombre 19 remplace le 3, l'ensemble de données devient (1;2;2;19), et

la moyenne est de 24/4, soit 6. Ainsi, une moyenne de 2 est plus appropriée pour cet ensemble de données qu'une moyenne de 6.

3.1.2 Médiane

La médiane (M_d) divise l'ensemble des données ordonnées en deux parties égales. La médiane est le point central dans l'ensemble des observations, avec une moitié des valeurs inférieures et l'autre moitié supérieures. Contrairement à la moyenne, la médiane est moins affectée par les valeurs extrêmes. Elle est couramment employée pour évaluer le centre d'une distribution de données ordinales ou numériques asymétriques. En cas de données asymétriques, la médiane constitue la mesure la plus appropriée de tendance centrale.

Pour déterminer la médiane, disposez les observations dans l'ordre croissant :

- ⇒ Si le nombre d'observations est *impair*, la médiane M_d correspondra à la valeur centrale de l'ensemble ordonné, soit la [(n+1)/2]ème observation. Par exemple, la médiane de l'ensemble de données avec n=5 (1; 2; 4; 5; 6) est $M_d = 4$.
- ⇒ Si le nombre d'observations est *pair*, la médiane M_d sera la valeur moyenne entre les deux observations centrales. Par exemple, la médiane de l'ensemble de données avec n=4 (1; 2; 4; 5) est $M_d = 3$.

3.1.3 Mode

Le mode (M_o) désigne la valeur qui apparaît le plus fréquemment dans un ensemble de données. Par exemple, dans l'ensemble de données (3; 4; 5; 6; 6; 6; 7; 8; 9), Mo=6. Si toutes les valeurs sont différentes (distribution uniforme), il n'y a pas de mode. Une distribution peut comporter plusieurs modes, auquel cas elle est dite bimodale ou multimodale. En pratique, le mode est rarement utilisé.

3.1.4 Relation empirique entre les mesures de tendance centrale

La mesure la plus appropriée de la tendance centrale pour un ensemble de données donné dépend de la forme de la distribution des données :

⇒ *Distribution normale :* Dans une distribution normale, les valeurs des données sont symétriques autour du centre et forment une courbe en cloche. La moyenne, la médiane et le mode devraient être approximativement les mêmes. Dans ce cas, la moyenne constitue la mesure la plus appropriée de la tendance centrale.

$$\bar{X} = M_d = M_o$$

⇒ *Distribution asymétrique :* Dans une distribution asymétrique, les valeurs extrêmes se produisent dans une seule direction. La médiane est la mesure la plus précise de la tendance centrale dans de telles distributions. La moyenne est sensible aux valeurs extrêmes et peut être considérablement surévaluée ou sous-évaluée.

Il existe deux types de distribution asymétrique *:*

- *Asymétrie négative* (vers la gauche) : Les valeurs extrêmes sont faibles.

$$\bar{X} < M_d < M_o$$

- *Asymétrie positive* (vers la droite) : Les valeurs extrêmes sont élevées.

$$M_o < M_d < \bar{X}$$

Lorsque les données sont asymétriques vers la droite, la moyenne se situe à droite de la médiane, et lorsqu'elles sont asymétriques vers la gauche, la moyenne se situe à gauche de la médiane *(Figure 3.1).*

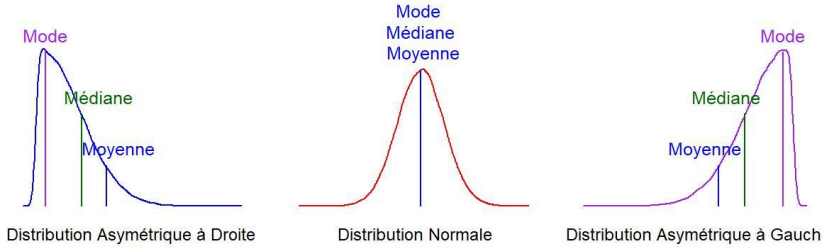

Figure 3.1 Distributions de données possibles

Recommandations pour le choix de la mesure appropriée de tendance centrale :

L'application correcte des mesures de tendance centrale dépend de l'échelle de mesure et de la forme de la distribution.

1. La moyenne est utilisée pour les données numériques avec une distribution symétrique (normale).
2. La médiane est utilisée pour les données ordinales ou les données numériques avec une distribution asymétrique.
3. Le mode est presque toujours utilisé pour les distributions bimodales.

3.2 Mesures de Variabilité

Les mesures de tendance centrale ne décrivent que le centre d'un ensemble de données. Pour évaluer la qualité d'une mesure de tendance centrale, il est nécessaire de comprendre la variabilité des données. Cela comprend la connaissance de la similitude des observations et leur proximité avec le centre, ou leur dispersion sur une large gamme de valeurs. Dans l'exemple suivant, nous observons deux distributions de valeurs de données très distinctes où la moyenne, la médiane et le mode sont identiques :

Jeu de données 1: -200; -20; -10; 7; 10; 20; 200 (n=7; $\bar{X}$=1; M_d=7)

Jeu de donnée 2: -20; -5; -2; 7; 2; 5; 20 (n=7; $\bar{X}$=1; M_d=7)

Malgré des différences significatives entre les ensembles de données, les mesures de tendance centrale restent les mêmes. Cet exemple souligne l'importance d'utiliser les mesures de tendance centrale en conjonction avec les mesures de variabilité pour décrire adéquatement un ensemble de données. Pour décrire de manière exhaustive la variabilité des données, nous devons utiliser des mesures telles que l'étendue (la plage), l'intervalle interquartile, la variance, l'écart-type et le coefficient de variation.

3.2.1 Étendue

L'étendue (ou la plage) représente la différence entre les valeurs maximales et minimales dans un ensemble de données.

Étendue (R) =Max (x_i) – Min (x_i)

Exemple :

Jeu de donnée 1 : (-200; -20; -10; 7; 10; 20; 200) ; n=7; $\bar{X}$= 1

Jeu de donnée 2 : (-20; -5; -2; 7; 2; 5; 20) ; n=7; $\bar{X}$=1

R_1 = 400

R_2 = 40

L'étendue est fortement influencée par les valeurs extrêmes et ne tient pas compte du reste de la distribution. Cet indicateur est extrêmement sensible aux valeurs exceptionnellement grandes ou petites et est utilisé pour les données numériques afin de mettre en évidence les valeurs extrêmes.

3.2.2 Intervalle interquartile

L'intervalle interquartile (IQR) réduit l'impact des valeurs extrêmes dans un ensemble de données. Les quartiles sont des valeurs qui partagent un ensemble de données ordonné en quatre segments égaux. L'IQR est

déterminé comme la différence entre le 75e percentile (Q3) et le 25e percentile (Q1), représentant les 50% d'observations centrales.

$$IQR = Q3 - Q1$$

Le 50e percentile (Q2) correspond à la médiane de l'ensemble de données, marquant ainsi son point central. *La figure 3.2* illustre les quartiles et l'intervalle interquartile via un diagramme en boîte.

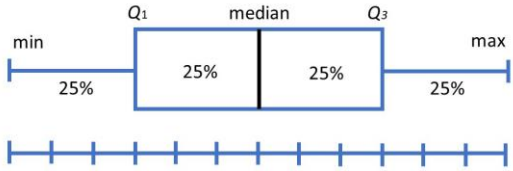

Figure 3.2 Présentation graphique des quartiles et de l'intervalle interquartile

Pour déterminer les quartiles, les données doivent être classées de la plus petite à la plus grande valeur :

$Q_1 = (n+1) / 4$

$Q_2 = (n+1) / 2$

$Q_3 = 3(n+1) / 4$

Le premier quartile correspond à la valeur médiane entre la plus petite valeur et la médiane d'une distribution. Il représente le 25e percentile. Le deuxième quartile, qui est la médiane de la distribution, représente le 50e percentile. Le troisième quartile est la valeur médiane entre la médiane et la plus grande valeur de la distribution, indiquant ainsi le 75e percentile.

L'intervalle interquartile est utilisé lorsque la médiane est utilisée (pour des données ordinales ou des données numériques asymétriques). Il est adapté pour décrire les 50 % centraux de la distribution, indépendamment de sa forme.

3.2.3 Variance et écart-type

La variance (s^2) quantifie la variabilité ou la dispersion autour de la moyenne. Elle est calculée comme la moyenne des carrés des écarts de chaque nombre par rapport à la moyenne Pour les données non groupées, la variance se calcule à l'aide de la formule suivante :

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

Pour les données groupées, la variance est déterminée en prenant en compte la fréquence :

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

Où :
X_i – valeurs individuelles de l'ensemble de données
$\bar{X}$ - moyenne
n – nombre d'observations
f_i – fréquence

L'écart-type *(S)* est la racine carrée de la variance et mesure la dispersion autour de la moyenne. Pour les données non groupées :

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Pour les données groupées :

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i}{n-1}}$$

Où :
S^2 – variance

L'écart-type est employé lorsque la moyenne est utilisée (données numériques symétriques) et, avec la moyenne, il résume les caractéristiques de l'ensemble de la distribution des valeurs. L'écart-type possède les mêmes unités de mesure que la moyenne.

3.2.4 Coefficient de variation

Le coefficient de variation (CV) est une mesure relative de la variabilité. Il est utilisé pour comparer les distributions mesurées sur différentes échelles. Le CV est défini comme l'écart-type divisé par la moyenne et multiplié par 100 %.

$$CV = \frac{S}{\overline{X}} \times 100\%$$

Où :

S – Écart-type

$\overline{X}$ – Moyenne

Le CV permet de comparer les niveaux de variabilité entre différents ensembles de données et est exprimé en pourcentage. Il peut être utilisé pour estimer le niveau de variabilité dans un ensemble de données selon l'échelle présentée dans le *Tableau 3.1*.

Tableau 3.1 Échelle du coefficient de variation

Coefficient de variation, %	Niveau de variabilité
<10	Faible
10-35	Moyen
>35	Élevé

Dans le cadre de la recherche scientifique, un ensemble de données doit présenter une variabilité faible ou moyenne afin d'assurer que la moyenne est représentative de l'ensemble.

Recommandations pour choisir la mesure de variabilité appropriée :

1. L'écart-type est utilisé lorsque la moyenne est utilisée (données numériques symétriques).

2. L'intervalle interquartile est utilisé :
   a) Lorsque la médiane est utilisée (données ordinales ou données numériques asymétriques).
   b) Pour comparer des observations individuelles avec un ensemble de normes.
   c) Pour décrire les 50 % centraux d'une distribution, indépendamment de sa forme.
3. L'étendue est utilisée avec des données numériques pour mettre en évidence les valeurs extrêmes.
4. Le coefficient de variation est employé pour comparer la variabilité de différents ensembles de données.

3.3 Distribution normale et ses propriétés

Les distributions normales ont des caractéristiques clés faciles à identifier sur un graphique, comme présenté dans la *Figure 3.3:*

1. La moyenne, la médiane et le mode sont égaux.
2. La distribution est symétrique par rapport à la moyenne : la moitié des valeurs se situe en dessous de la moyenne et l'autre moitié au-dessus.
3. La distribution peut être décrite par deux valeurs : la moyenne et l'écart-type.
4. La surface totale sous la courbe est égale à 1.

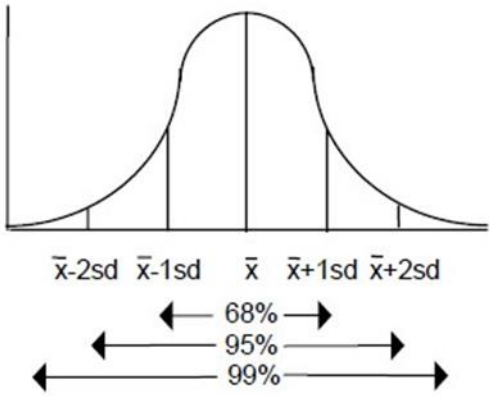

Figure 3.3 Propriétés de la distribution normale

Si les données suivent une distribution normale (formant une courbe en cloche), environ 68 % des données se situeront à l'intérieur d'un écart-type, environ 95 % à l'intérieur de deux écarts-types et environ 99,7 % à l'intérieur de trois écarts-types.

3.4 Asymétrie et kurtosis

L'asymétrie représente le degré de déviation d'une distribution dans un ensemble de données par rapport à la moyenne, indiquant la direction et l'ampleur avec lesquelles la distribution s'écarte d'une distribution normale symétrique (courbe en cloche). Il existe différentes formules pour mesurer l'asymétrie. Les deux coefficients d'asymétrie de Pearson suivants (S_k) sont les plus souvent utilisés pour les données numériques (intervalle ou rapport).

Premier coefficient d'asymétrie de Pearson (asymétrie par rapport au mode) :

$$S_k = \frac{\text{Moyenne} - \text{Mode}}{\text{Écart type}} = \frac{\overline{\text{X}} - \text{M}_\text{o}}{\text{S}}$$

Lorsque le premier coefficient d'asymétrie de Pearson est utilisé, les valeurs se situent entre -1 et +1. Une valeur de 0 indique une distribution parfaitement symétrique. Les valeurs proches de -1 ou +1 indiquent des distributions de plus en plus asymétriques négatives ou positives.

Deuxième coefficient d'asymétrie de Pearson (asymétrie par rapport à la médiane):

$$S_k = \frac{3 \times (\text{Moyenne} - \text{Mode})}{\text{Standard Deviation}} = \frac{3 \times (\overline{\text{X}} - \text{M}_\text{d})}{\text{S}}$$

Lorsque le deuxième coefficient d'asymétrie de Pearson est utilisé, les valeurs se situent entre -3 et +3. Une valeur de 0 indique une distribution parfaitement symétrique. Des valeurs proches de -1 ou +1 révèlent des distributions de plus en plus asymétriques, négatives ou positives.

Interprétation du second coefficient d'asymétrie de Pearson :

$S_k = 0$: La distribution est parfaitement symétrique.

S_k est entre -0.5 and 0.5: La distribution est presque symétrique.

S_k est entre -1 and -0.5: Asymétrie négative modérée.

S_k est entre 0.5 and 1: Asymétrie positive modérée.

S_k est inférieur à -1 ou supérieur à 1 : Les données sont fortement asymétriques.

La kurtosis constitue une autre mesure de la forme d'une distribution et se rapporte à la forme du pic de la courbe (*Figure 3.4*). Alors que l'asymétrie évalue le degré de dissymétrie, la kurtosis indique le degré d'acuité d'un pic de distribution de fréquence. La kurtosis est utilisée pour détecter la présence de valeurs aberrantes dans les données.

Il existe trois types de kurtosis :

⇒ *Platykurtique* – le pic de la courbe est aplati par rapport à la normale, et les queues sont longues (kurtosis négative).

⇒ *Mésokurtique* – le pic de la courbe est normal, et les queues de part et d'autre de la moyenne sont également normales.

⇒ *Leptokurtique* – le pic de la courbe est étroit, et les queues de part et d'autre de la moyenne sont courtes (kurtosis positive).

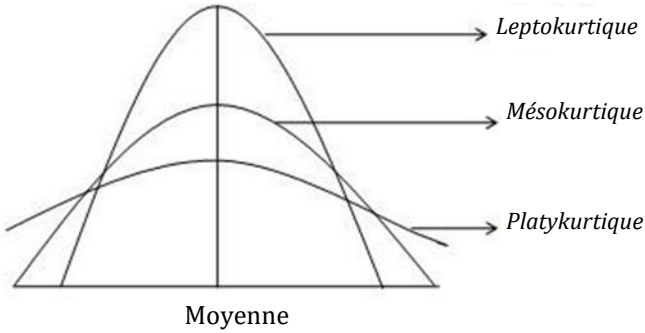

Figure 3.4 Types de Kurtosis

3.5. Exemple de Calcul

Supposons que les âges des 19 patients étudiés sont les suivants :
31; 24; 26; 30; 24; 35; 35; 31; 35; 33; 26; 33; 26; 31; 26; 30; 31; 31; 30.
Calculez les mesures de tendance centrale et de variabilité. Formulez votre conclusion.

1. Trier les données :

24; 24; 26; 26; 26; 26; 30; 30; 30; 31; 31; 31; 31; 31; 33; 33; 35; 35; 35.

2. Organiser les données dans un tableau de fréquences :

X_i	*Fréquence*, f	$x_i f_i$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 f_i$
24	2	48	-5.8	33.64	67.28
26	4	104	-3.8	14.44	57.76

X_i	*Fréquence, f*	$x_i f_i$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 f_i$
30	3	90	+0.2	0.04	0.12
31	5	153	+1.2	1.44	7.2
33	2	66	+3.2	10.24	20.48
35	3	105	+5.2	27.04	81.12
Total	n=19	566			233.96

3. Calculer la moyenne :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i f_i$$

$$\bar{X} = \frac{24 \times 2 + 26 \times 4 + 30 \times 3 + 31 \times 5 + 33 \times 2 + 35 \times 3}{19} = \frac{566}{19} = 29.8$$

4. Déterminer le mode :

$Mode\ (M_o) = 31$

5. Calculer la médiane :

Étant donné que n = 19 est impair, la position de la médiane est $\frac{n+1}{2} = \frac{19+1}{2} = 10$.

La 10ème observation dans les données triées est 31.

$Médiane\ (M_d) = 31$

6. Calculer la variance et l'écart-type :

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i = \frac{233.96}{18} = 13$$

$$s = \sqrt{s^2} = \sqrt{13} \approx 3.6$$

7. Calculer le coefficient de variation :

$$CV = \frac{S}{\bar{x}} \times 100\%$$

$$CV = \frac{3.6}{29.8} \times 100\% = 12.1\%$$

8. Conclusions :

- L'âge moyen des patients est de 29.8±3.6 ans.
- Le niveau de variabilité des données est moyen (CV=12.1%), ce qui est approprié pour la recherche scientifique.
- La distribution des données est presque symétrique, avec des valeurs du mode, de la médiane et de la moyenne proches les unes des autres (légèrement asymétrique vers la gauche).
- La moyenne est représentative de cet ensemble de données.

Exercices de révision

1. En utilisant l'ensemble de données IMC de 10 patients suivants :

   Données : 29, 22, 24, 37, 23, 35, 35, 27, 50, 37.

   - Trouvez les mesures de tendance centrale.
   - Analysez la relation entre les mesures de tendance centrale et identifiez la forme de la distribution. Formulez votre conclusion.

2. En utilisant l'ensemble de données des niveaux de glucose sanguin de 11 patients suivants :

   Données : 3.3, 12.0, 9.0, 6.0, 11.0, 11.8, 11.8, 11.0, 5.5, 3.3 (mmol/l).

   - Trouvez toutes les mesures de tendance centrale.
   - Analysez la relation entre les mesures de tendance centrale et identifiez la forme de la distribution. Formulez votre conclusion.

3. En utilisant l'ensemble de données IMC de l'exercice 1 :

   - Trouvez toutes les mesures de variabilité. Formulez votre conclusion.

- Évaluez la représentativité de la moyenne pour cet ensemble de données.
- En fonction de l'échelle de mesure et de la forme de la distribution, décidez quelles mesures de tendance centrale sont les plus appropriées.
- Construisez un diagramme en boîte.

4. En utilisant l'ensemble de données des niveaux de glucose sanguin de l'exercice 2 :
    - Trouvez toutes les mesures de variabilité. Formulez votre conclusion.
    - Évaluez la représentativité de la moyenne pour cet ensemble de données.
    - En fonction de l'échelle de mesure et de la forme de la distribution, décidez quelles mesures de tendance centrale sont les plus appropriées.

Questions de révision

1. La moyenne : définition, types et règles de calcul. Donnez un exemple.
2. La médiane : définition et règles de calcul. Donnez un exemple.
3. Le mode : définition et règles de calcul. Donnez un exemple.
4. Comparez la moyenne, la médiane et le mode en tant que mesures de tendance centrale.
5. Dans quelles conditions l'utilisation de la moyenne est-elle préférée ?
6. Dans quelles conditions l'utilisation de la médiane est-elle préférée ?
7. Dans quelles conditions l'utilisation du mode est-elle préférée ?
8. Mesures de variabilité : raisons d'application.
9. Étendue : définition, caractéristiques et conditions d'utilisation préférées. Règle de calcul.
10. Quartiles : définition, caractéristiques et règle de calcul.
11. Intervalle interquartile : définition, caractéristiques et conditions d'utilisation préférées. Règle de calcul.
12. Écart-type : signification, caractéristiques et conditions d'utilisation préférées. Règle de calcul.
13. Coefficient de variation : définition, caractéristiques et conditions d'utilisation préférées. Règle de calcul.
14. Définissez la distribution normale et ses propriétés.

CHAPITRE 4. STATISTIQUES DESCRIPTIVES : RÉSUMER DES DONNÉES NOMINALES ET ORDINALES

Concepts clés

- *Les données qualitatives* peuvent être mesurées à l'aide de plusieurs méthodes : ratios, proportions et taux.
- *Un ratio* est le nombre d'observations présentant une caractéristique donnée « a » divisé par le nombre d'observations sans cette caractéristique « b ».
- *Une proportion* est le nombre d'observations présentant une caractéristique donnée « a » divisé par le nombre total d'observations « a+b ».
- *Un taux* est comparable à une proportion mais utilise un multiplicateur (par exemple, 1,000, 10,000 ou 100,000) et est calculé sur une période déterminée.
- *Les graphiques en barres/colonnes et les graphiques linéaires* sont utilisés pour représenter graphiquement les ratios et les taux.
- *Un graphique en secteurs* est employé pour représenter graphiquement les proportions.
- *Une proportion* reflète la structure d'un phénomène et constitue un indicateur statique.
- *Un taux* indique la fréquence (niveau ou intensité) d'un phénomène au fil du temps et est un indicateur dynamique.
- *Les taux* sont essentiels en épidémiologie et en médecine basée sur les preuves ; ils servent de base pour le calcul des statistiques vitales, qui décrivent l'état de santé des populations.
- *Les taux* sont généralement calculés sur une base annuelle (taux annuels).
- *Les taux* peuvent être bruts, spécifiques ou standardisés.

- ❖ *Les taux* doivent être standardisés lorsqu'on compare des populations avec des différences significatives dans leur structure (par exemple, en fonction de l'âge et du sexe).
- ❖ *La prévalence et l'incidence* sont deux mesures importantes de la morbidité.

4.1 Méthodes de description des données catégorielles

Dans la recherche, les informations statistiques sont généralement présentées sous forme de valeurs absolues. Ces valeurs peuvent être difficiles à interpréter car elles ne permettent pas toujours des comparaisons significatives, une synthèse ou une corrélation entre différentes caractéristiques. Pour rendre les comparaisons entre groupes plus pertinentes, il est souvent préférable d'utiliser des valeurs relatives plutôt que des chiffres absolus. Les données catégorielles (qualitatives) peuvent être mesurées par trois méthodes :

- Proportions et pourcentages
- Rapports
- Taux

4.1.1 Proportions et pourcentages

Les proportions sont des indicateurs statistiques qui reflètent la structure d'un phénomène. Elles sont définies comme la fraction d'un phénomène par rapport au total. La proportion se calcule en divisant le nombre d'observations possédant une caractéristique d'intérêt (a) par le nombre total d'observations (a+b). Un pourcentage est simplement la proportion multipliée par 100 %.

$$Proportion\ (Pourcentage) = \left(\frac{a}{a+b}\right) \times 100\%$$

Où :

- a est le nombre d'observations avec la caractéristique d'intérêt

- b est le nombre d'observations sans la caractéristique.

Exemple de proportion :

$$Mortalité\ proportionnelle = \frac{Nombre\ de\ décès\ dus\ à\ une\ cause\ X\ dans\ une\ année}{Nombre\ total\ de\ décès\ dans\ une\ année} \times 100$$

Nombre de décès dus au cancer en République de Moldavie en 2023 = 5,959
Nombre total de décès en République de Moldavie en 2023 = 33,733

$$Mortalité\ proportionnelle\ due\ au\ cancer\ en\ 2023 = \frac{5{,}959}{33{,}733} \times 100 = 18\%$$

Les proportions sont considérées comme des indicateurs extensifs car elles illustrent la structure d'un phénomène. Ce sont des indicateurs statiques qui offrent une vue d'ensemble à un moment donné sans refléter les variations dynamiques. Elles sont particulièrement utiles pour les données ordinales et numériques ainsi que pour les données nominales, surtout lorsque les observations sont groupées dans un tableau de fréquence.
Présentation graphique : Diagramme circulaire.

4.1.2 Rapports

Un rapport est défini comme une partie divisée par une autre partie représentant deux phénomènes indépendants. Le rapport se calcule en divisant le nombre d'observations dans un groupe possédant une caractéristique donnée (a) par le nombre d'observations dépourvues de cette caractéristique (b) :

$$Rapport = \frac{a}{b}$$

Exemples de rapports :

$$Rapport\ de\ sexe = \frac{Nombre\ d'hommes}{Nombre\ de\ femmes}$$

Nombre de nouveau-nés de sexe masculin en République de Moldavie en 2023 = 12,239

Nombre de nouveau-nés de sexe féminin en République de Moldavie en 2023 = 11,794

$$Rapport\ de\ sexe\ \text{à}\ la\ naissance\ en\ 2023 = \frac{12{,}239}{11{,}794} = 1.04$$

Dans cet exemple, le rapport est sans unité. Toutefois, il peut être redimensionné en le multipliant par un multiplicateur, tel que 100 ou 1,000. Le rapport de sexe à la naissance en République de Moldavie en 2023 était de 1.04 ou 104 % (1.04 x 100 %), ce qui signifie 104 nouveau-nés de sexe masculin pour chaque 100 nouveau-nés de sexe féminin.

$$Approvisionnement\ médical = \frac{Nombre\ de\ médecins}{Population\ totale} \times multiplicateur$$

Nombre de médecins en République de Moldavie en 2022 = 12,600

Population totale en République de Moldavie en 2022 = 2,565,030

$$Approvisionnement\ médical\ en\ 2023 = \frac{12{,}600}{2{,}565{,}030} \times 1{,}000$$
$$= 5\ médecins\ pour\ 1{,}000\ habitants$$

Présentation graphique : Graphique linéaire, diagramme à barres/colonnes.

4.1.3 Taux

Un taux est semblable à une proportion mais comprend un multiplicateur et une dimension temporelle. Le taux est toujours calculé pour une période déterminée, généralement une année (taux annuel). Les taux sont des indicateurs statistiques intensifs qui expriment la

fréquence ou le niveau d'un phénomène sur une période donnée. Ils sont calculés selon la formule suivante :

$$Taux = \left(\frac{a}{a+b}\right) \times multiplicateur$$

Où :

- *a* représente le nombre d'observations avec une caractéristique spécifique (par exemple, le nombre de décès ou de naissances dans une année et un lieu déterminés);
- *a+b* est le nombre total d'observations (par exemple, la population totale pour une année donnée);
- le multiplicateur est un facteur de mise à l'échelle (par exemple, 100; 1,000; 10,000; 100,000).

Exemple de taux :

Par exemple, si une étude a duré un an et que la proportion de patients ayant développé une maladie était de 0.02, le taux pour 1,000 patients serait (0.02) x (1,000), soit 20 cas de maladie pour 1,000 patients sur un an.

Présentation graphique : Graphique linéaire, diagramme à barres/colonnes.

4.2 Indicateurs de l'état de santé en statistiques descriptives

Les indicateurs de l'état de santé évaluent la condition sanitaire d'une population en utilisant des statistiques vitales. On distingue principalement trois types d'indicateurs :

⇒ Indicateurs de mortalité ;

⇒ Indicateurs de morbidité ;

⇒ Indicateurs de handicap.

La majorité des indicateurs de santé sont exprimés sous forme de taux et, parfois, de proportions et de ratios. Les taux de mortalité et de morbidité sont les plus fréquemment utilisés.

4.2.1 Taux de mortalités

Le taux de mortalité ou taux de décès est défini comme le nombre de décès survenus au cours d'une période déterminée, divisé par le nombre total de personnes exposées au risque de décès durant cette même période.

- ⇒ *Taux brut :* Calculé pour l'ensemble des individus d'une population donnée, sans tenir compte des variations dues à l'âge, au sexe, à la race, etc.
- ⇒ *Taux spécifique* : Calculé au sein de sous-groupes de population relativement restreints et bien définis. Par exemple :
  - Taux de décès spécifiques à l'âge ;
  - Taux de décès spécifiques au sexe ;
  - Taux de décès spécifiques à la cause.

4.2.2 Taux de morbidité

Le taux de morbidité est défini comme le nombre d'individus ayant contracté une maladie au cours d'une période donnée, divisé par le nombre total de personnes à risque pendant cette même période.

L'incidence et la prévalence sont les principales mesures de morbidité et sont couramment utilisées pour évaluer l'état de santé de la population dans de nombreuses études médicales et épidémiologiques.

Incidence : Le nombre de nouveaux cas apparus au cours d'une période spécifiée, divisé par le nombre total de personnes à risque pendant cette période.

Prévalence : Le nombre d'individus présentant une maladie particulière à un moment donné, divisé par la population totale à risque pour cette maladie à ce moment précis.

Les taux de morbidité fournissent une méthode normalisée pour évaluer tant les taux globaux que spécifiques.

Exemple de calcul :

Dans une année et une localité spécifique, la population est de 75 000 personnes. Cette année-là, 897 personnes sont décédées. Dans cette localité, il y avait 40 médecins : 20 généralistes, 10 chirurgiens et 10 autres. Calculez les indicateurs statistiques résumant les données nominales.

1. *Proportion et pourcentage*

$$\text{Proportion (pourcentage) de médecins} = \frac{20}{40} \times 100\% = 50\%$$

2. *Ratio*

$$\text{Approvisionnement médical} = \frac{40}{75{,}000} \times 10{,}000$$
$$= 5.3 \text{ médecins pour } 10\,000 \text{ habitants}$$

3. *Taux*

$$\text{Taux brut de mortalité} = \frac{897}{75{,}000} \times 1{,}000$$
$$= 11.9 \text{ décès pour } 1{,}000 \text{ habitants}$$

4.3 Taux standardisés : méthode directe de standardisation

Les taux bruts peuvent être utilisés pour comparer deux populations différentes uniquement si ces populations sont homogènes dans toutes leurs caractéristiques. Cependant, si les populations diffèrent par des caractéristiques telles que le sexe ou l'âge, l'utilisation des taux bruts peut conduire à des résultats incorrects. Cela est dû à l'influence

importante de la structure par âge sur les taux bruts. Par exemple, une population peut avoir une proportion plus élevée de personnes âgées que l'autre. Dans de tels cas, les taux bruts doivent être ajustés ou standardisés pour permettre des comparaisons valides.

La standardisation ajuste les taux bruts pour éliminer les effets des différences dans la composition de la population, principalement selon l'âge et le sexe. Deux méthodes principales de standardisation existent : la standardisation directe et la standardisation indirecte. La méthode directe de standardisation calcule les taux qui résulteraient si tous les groupes comparés avaient la même composition standard. Les taux standardisés sont donc des valeurs conventionnelles conçues uniquement à des fins de comparaison et ne peuvent pas être utilisés de manière autonome.

La méthode directe de standardisation des taux se compose de quatre étapes :

1. Calcul des taux spécifiques pour chaque groupe ;
2. Sélection de la population standard ;
3. Calcul du nombre attendu d'événements pour chaque groupe ;
4. Calcul des taux standardisés.

Un exemple détaillé du calcul des taux de mortalité standardisés pour les deux régions est présenté ci-dessous.

Calcul exemple : méthode directe de standardisation des taux de mortalité.

Étape 1. Calcul des taux spécifiques pour chaque groupe

Sexe	**Région A**		**Région B**		**Taux de mortalité spécifique par sexe (pour 1,000)**	
	Personnes	Décès	Personnes	Décès	Région A	Région B
Hommes	50	1	170	4	**20**	**24**
Femmes	200	10	30	3	**50**	**100**
Total	250	11	200	7	**44**	**35**

Par exemple, le taux de mortalité spécifique pour les hommes dans la Région A = $\frac{1}{50} \times 1{,}000 = 20$ décès pour 1,000 hommes.

Étape 2. Sélection de la population standard

Sexe	**Région A**		**Région B**		**Population standard**
	Personnes	Décès	Personnes	Décès	Personnes de A + Personnes de B
Hommes	50	1	170	4	**220**
Femmes	200	10	30	3	**230**
Total	250	11	200	7	**450**

Dans cet exemple, la population standard est la somme des populations des deux régions. En d'autres termes, nous supposons que la structure de la population par sexe est identique dans les deux régions.

Étape 3. Calcul du nombre attendu d'événements pour chaque groupe

Sexe	**Taux de mortalité spécifique par sexe (pour 1,000)**		**Population standard**	**Événements attendus (décès)**	
	Région A	Région B	Personnes de A + Personnes de B	Région A	Région B
Hommes	20	24	220	4	5
Femmes	50	100	230	12	23
Total				**16**	**28**

$$\text{Nombre d'événements attendus} = \frac{Taux\ spécifique \times\ population\ standard}{1{,}000}$$

Par exemple, le nombre de décès attendus pour la région A chez les hommes = $\frac{20\times220}{1{,}000} = 4$ décès. Autrement dit, si la population masculine dans la région A représentait la population standard (220 personnes), nous attendrions 4 décès ici.

Étape 4. Calcul du taux standardisé (ajusté) pour chaque groupe (région)

Sexe	**Standard population**	**Événements attendus (décès)**		**Taux standardisé (pour 1,000)**	
		Région A	Région B	Région A	Région B
Hommes	220	4	5		
Femmes	230	12	23		
Total	450	16	28	**36**	**62**

$$\text{Taux standardisé} = \frac{(\text{Nombre total d'événements attendus}}{\text{Population standard totale}} \times 1{,}000$$

Par exemple, le taux de mortalité standardisé pour la région A = $\frac{16}{450} \times 1{,}000 = 36$ décès pour 1,000 de population standard.

Conclusion: Comparaison des taux bruts et standardisés

Taux brut (pour 1,000)		**Taux standardisé (pour 1,000)**	
Région A	Région B	Région A	Région B
44	**35**	**35**	**63**

L'intensité (niveau) de la mortalité est plus élevée dans la région B. Le niveau de mortalité dans la région B est 1.8 fois plus élevé que dans la région A (63 ÷ 35 = 1.8).

Exercices de révision

1. Dans la localité A, au cours d'une année donnée, 2,500 cas de maladies ont été enregistrés : 800 étaient des maladies cardiovasculaires ; 500 des maladies pulmonaires ; 450 des traumatismes ; et 750 d'autres maladies. La population totale est de 900,000 habitants.
   - Calculez toutes les statistiques vitales possibles.
   - Réalisez une présentation graphique appropriée pour ces données.

2. Dans la localité B, au cours d'une année donnée, la population est de 78 000 habitants. Cette année-là, 110 personnes sont décédées et 400 personnes ont développé pour la première fois une maladie cardiovasculaire.
   - Calculez toutes les statistiques vitales possibles.
   - Réalisez une présentation graphique appropriée pour ces données.

3. Analysez les données suivantes comparant la létalité de l'abdomen aigu dans les hôpitaux « A » et « B » :

Durée d'hospitalisation, heures	Hôpital « A »		Hôpital « B »	
	Nombre de patients	Nombre de cas létaux	Nombre de patients	Nombre de cas létaux
< 6	650	72	490	34
6-12	450	83	380	66
>24	131	23	736	206
Total	1,231	178	1,606	306

- Calculez les taux bruts et comparez-les.
- Comment les taux ajustés diffèrent-ils des taux bruts dans chacun de ces hôpitaux ? Expliquez ces différences (interprétation et conclusions).

4. Analysez les données suivantes comparant la mortalité hospitalière dans les hôpitaux « A » et « B » :

Maladie	Hôpital « A »		Hôpital « B »	
	Nombre de patients	Nombre de décès	Nombre de patients	Nombre de décès
Gastro-intestinal	1,200	24	1,700	40
Tumeur maligne	190	55	100	30
Cardiovasculaire	160	100	1100	72
Total	1,650	179	2,900	142

- Calculez les taux bruts et comparez-les.
- Comment les taux ajustés diffèrent-ils des taux bruts dans chacun de ces hôpitaux ? Expliquez ces différences (interprétation et conclusions).

Questions de révision

1. Valeurs absolues et relatives : leur signification et leur application en biostatistique. Dans quelles conditions l'utilisation des valeurs relatives est-elle préférable ? Donnez un exemple.
2. Types de valeurs relatives : quelles sont les différences et les similitudes entre eux ? Donnez un exemple pour chaque type.
3. Proportions, ratios et taux : spécificités, règles de calcul, conditions d'application et présentation graphique appropriée. Donnez un exemple.
4. Statistiques vitales : définition et principaux taux utilisés.
5. Quelle est la différence entre les taux bruts et les taux spécifiques ?
6. Quelle est la différence entre les taux de mortalité et les taux de morbidité ?
7. Quelles sont les différences et les similitudes entre la prévalence et l'incidence ?
8. Dans quelles conditions l'utilisation du taux standardisé est-elle recommandée ?
9. Méthode directe de standardisation : définition et étapes du processus.
10. Dans quelles circonstances doit-on utiliser les taux bruts, spécifiques et standardisés ?

CHAPITRE 5. CORRÉLATION ET RÉGRESSION

Concepts clés

- *Coefficient de corrélation (r)* : évalue la force et la direction de la relation linéaire entre deux variables.
- *Variable X* : souvent appelée variable indépendante ou explicative.
- *Variable Y* : désignée comme variable dépendante ou de résultat.
- La valeur du coefficient de corrélation varie de -1 à +1.
- *Direction de la corrélation* : déterminée par le signe (+ ou -) du coefficient.
- *Force de la corrélation* : indiquée par la magnitude du coefficient.
- *Corrélation positive* (+) : les valeurs élevées d'une variable sont corrélées avec des valeurs élevées de l'autre variable.
- *Corrélation négative* (-) : les valeurs élevées d'une variable sont corrélées avec des valeurs faibles de l'autre variable.
- *Coefficient de corrélation de Pearson (r)* : mesure la relation linéaire entre deux variables numériques.
- *Coefficient de corrélation de Spearman (r_s)* : évalue la relation entre deux variables selon leur ordre de rang.
- *Régression* : technique de prévision de la valeur d'une variable dépendante (Y) basée sur une ou plusieurs variables indépendantes (X).
- *La régression linéaire* peut être simple ou multiple.

La recherche biomédicale explore souvent la relation entre deux ou plusieurs variables. Par exemple, existe-t-il une relation entre la consommation de sel et la pression artérielle ou entre le tabagisme et l'espérance de vie ? Comprendre ces relations est crucial pour découvrir des associations et faire des prédictions. Pour examiner ces relations,

deux techniques statistiques fondamentales sont couramment utilisées : la corrélation et la régression.

1. *Corrélation :* Cette technique est utilisée pour établir et quantifier la force et la direction de la relation entre deux variables.
2. *Régression* : Cette méthode est employée pour exprimer la relation fonctionnelle entre deux ou plusieurs variables. L'analyse de régression permet au chercheur de prédire la valeur d'une variable en fonction de la valeur d'une autre variable.

5.1 Corrélation

5.1.1 Types de coefficient de corrélation

Coefficient de corrélation de Pearson

Le coefficient de corrélation de Pearson constitue une mesure paramétrique de la relation entre deux variables numériques ayant une distribution normale. La variable indépendante ou explicative est désignée par « X », tandis que la variable dépendante ou de résultat est « Y ».

Le coefficient de corrélation est noté par « r » et est calculé à l'aide de la formule suivante :

$$r = \frac{\Sigma\ (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma}(Y_i - \bar{Y})^2}$$

Où:

X_i: le score individuel de la variable indépendante

Y_i: le score individuel de la variable dépendante

$\bar{X}$: la moyenne de la variable indépendante

$\bar{Y}$: la moyenne de la variable dépendante

Le coefficient de corrélation est un nombre sans dimension, ce qui signifie qu'il n'a pas d'unité de mesure. Sa valeur varie de -1 à 1. Pour tout ensemble de données, le coefficient de corrélation (r) satisfait l'inégalité suivante :

$-1 \leq r \leq 1$

La direction de la relation est déterminée par le signe du coefficient, tandis que la force de la relation est déterminée par l'amplitude du coefficient.

Direction de la corrélation :

- ⇒ *Une corrélation positive* $(0 < r \leq 1)$ indique qu'à mesure qu'une variable augmente, l'autre tend également à augmenter. Par exemple, une consommation élevée de sel est associée à une pression artérielle plus élevée.
- ⇒ *Une corrélation négative* $(-1 \leq r < 0)$ indique qu'à mesure qu'une variable augmente, l'autre tend à diminuer. Par exemple, une consommation élevée de cigarettes est associée à une espérance de vie plus courte.

Valeurs spécifiques :

$r = 1$: Indique une relation linéaire positive parfaite.

$r = -1$: Indique une relation linéaire négative parfaite.

$r = 0$: Indique l'absence de relation linéaire (absence de corrélation) entre les variables.

Interprétation de la force de la corrélation :

- ⇒ 0 à 0.25 (±) : Corrélation nulle ou faible ;
- ⇒ 0.25 à 0.50 (±) : Corrélation modérée *;*
- ⇒ 0.50 à 0.75 (±) : Corrélation forte ;
- ⇒ 0.75 à 1 (±) : Corrélation très forte ;
- ⇒ ± 1 : Corrélation parfaite.

Cette échelle aide à comprendre à quel point les deux variables sont liées. Des valeurs absolues de r plus élevées indiquent une relation plus forte, tandis que des valeurs proches de zéro indiquent une relation plus faible.

Diagrammes de dispersion

Les diagrammes de dispersion (nuages de points) sont un outil visuel utile pour illustrer la relation entre deux variables numériques. Ils permettent d'évaluer la linéarité de cette relation et de repérer les valeurs aberrantes.

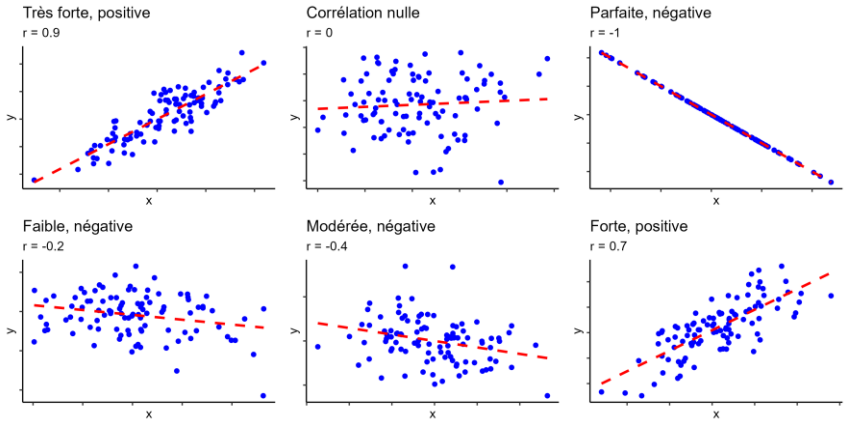

Figure 5.1 Diagrammes de dispersion illustrant différentes forces de corrélation

Exemple de calcul du coefficient de corrélation de Pearson

Considérons la relation entre la longueur (variable indépendante) et le poids (variable dépendante) de neuf nouveau-nés (*Tableau 5.1*). Pour calculer le coefficient de corrélation de Pearson :

1. Calculer la moyenne des variables X et Y.

$$\bar{X} = \frac{\sum X_i}{n} = \frac{437}{9} = 48.6 \qquad \bar{Y} = \frac{\sum Y_i}{n} = \frac{25.7}{9} = 2.9$$

2. Calculer les écarts pour chaque variable X et Y.

$$(X_i - \bar{X}) \qquad \text{et} \qquad (Y_i - \bar{Y})$$

3. Élever au carré les écarts des variables X et Y:

$$(X_i - \bar{X})^2 \qquad \text{et} \qquad (Y_i - \bar{Y})^2$$

4. Calculer le coefficient en utilisant la formule :

$$r = \frac{\Sigma\,(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma (Y_i - \bar{Y})^2}} = \frac{17.5}{\sqrt{106.2 \times 3.5}} = \frac{17.5}{19.2} = 0.91$$

Tableau 5.1. Données sur la longueur (cm) (X_i) et le poids (kg) (Y_i) de neuf nouveau-nés

Numéro de l'enfant	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	43.0	2.2	30.9	0.4	3.6	30.9	0.4
2	45.0	2.1	12.6	0.6	2.7	12.6	0.6
3	47.0	2.4	2.4	0.2	0.7	2.4	0.2
4	48.0	2.4	0.3	0.2	0.3	0.3	0.2
5	49.0	3.0	0.2	0.0	0.1	0.2	0.0
6	49.0	2.7	0.2	0.0	-0.1	0.2	0.0
7	50.0	3.5	2.1	0.4	0.9	2.1	0.4
8	50.0	3.4	2.1	0.3	0.8	2.1	0.3
9	56.0	4.0	55.4	1.3	8.5	55.4	1.3
Total	**437**	**25.7**	**106.2**	**3.5**	**17.5**	**106.2**	**3.5**

Conclusion : Le coefficient de corrélation r est de 0,91. La relation entre la longueur et le poids des nouveau-nés est positive et très forte.

Le coefficient de corrélation des rangs de Spearman

La corrélation des rangs de Spearman est une mesure non paramétrique utilisée pour les types de données suivants :

- Deux variables ordinales ;
- Une variable ordinale et une variable numérique ;
- Deux variables numériques si l'une d'elles n'est pas distribuée normalement.

Ce coefficient est symbolisé par r_s et se calcule en classant les valeurs de chaque variable, puis en appliquant la formule de Pearson

aux rangs. La formule de corrélation des rangs de Spearman est la suivante :

$$r_s = \frac{\Sigma\,(R_x - \bar{R}_x)\left(R_y - \bar{R}_y\right)}{\sqrt{\Sigma(R_x - \bar{R}_x)^2 \Sigma\left(R_y - \bar{R}_y\right)^2}}$$

Où :

R_x et R_y sont les rangs des variables X et Y

$\bar{R}_x$ et $\bar{R}_y$ sont les rangs moyens pour les variables X et Y

Comme le coefficient de corrélation de Pearson, le coefficient de corrélation des rangs de Spearman varie entre -1 et 1. Des valeurs de r_s proches des extrêmes indiquent un haut degré de corrélation entre X et Y ; des valeurs proches de 0 impliquent une relation plus faible.

Autres types de corrélations

- Le coefficient de corrélation point-bisériale : utilisé lorsque l'une des variables est numérique et l'autre est dichotomique.
- Le coefficient Phi : utilisé pour deux variables dichotomiques.

Considérations importantes

⇒ *La corrélation ne signifie pas causalité.* Une corrélation significative entre deux variables n'implique pas que les variations dans une variable entraînent des variations dans l'autre. Le coefficient de corrélation n'est qu'une mesure de la relation entre deux variables. Déduire une relation causale à partir d'une corrélation est une erreur courante et fondamentale.

⇒ La présence d'une corrélation entre deux variables dans un échantillon ne garantit pas que cette corrélation se retrouve dans l'ensemble de la population. Des tests statistiques, tels que les tests t, sont nécessaires pour évaluer la signification de la corrélation (voir Chapitre 6).

5.1.2 Coefficient de détermination

Le coefficient de détermination, noté R^2, correspond au carré du coefficient de corrélation. Il indique la proportion de la variance d'une variable qui est expliquée par la variance de l'autre variable. Lorsque les deux variables sont corrélées, il existe une certaine quantité de variance partagée entre elles. Plus la corrélation est forte, plus la variance partagée est élevée, ce qui accroît le coefficient de détermination. La valeur de R^2 varie de 0 à 100 % (*Figure 5.2*).

Exemple de calcul du coefficient de détermination

Considérons une étude qui analyse la relation entre le nombre d'heures d'étude par semaine (variable X) et les résultats aux examens (variable Y) chez les étudiants en médecine. Supposons que l'étude trouve une corrélation r de 0,70. Pour calculer le coefficient de détermination :
$R^2 = (0.70)^2 = 0.49$ ou 49%

Ce résultat montre que 49 % de la variance des résultats aux examens est expliquée par la variance du nombre d'heures d'étude. Ainsi, près de la moitié des différences dans les résultats des examens entre les étudiants peut être attribuée aux variations dans leurs habitudes d'étude.

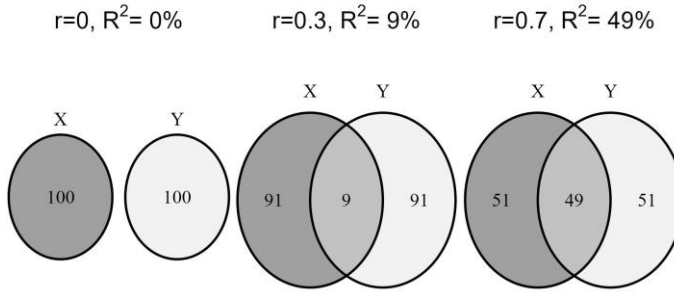

Figure 5.2 Explication graphique du coefficient de détermination (R^2)

5.2 Régression : approches générales

5.2.1 Régression linéaire simple

Si deux variables sont fortement corrélées, il est possible de prédire la valeur de la variable dépendante à partir de la valeur de la variable indépendante en utilisant des méthodes de régression. Les analyses de corrélation ne font pas cette distinction, car les deux variables sont traitées symétriquement. Comme les analyses de corrélation, la régression linéaire simple est une technique utilisée pour explorer la nature de la relation entre deux variables continues. La principale différence entre ces deux méthodes statistiques est que la régression permet d'étudier le changement dans une variable Y (variable dépendante ou de résultat) en fonction d'un changement donné dans l'autre variable X, connue comme variable indépendante ou explicative, par une *équation de régression* qui quantifie la relation linéaire entre les deux variables. Cette ligne droite, ou ligne de régression, est la même « ligne de meilleur ajustement » pour le nuage de points que celle utilisée pour calculer le coefficient de corrélation.

Dans le cas de l'équation linéaire simple, la valeur d'une variable (X) est utilisée pour prédire la valeur de l'autre variable (Y) à l'aide de l'équation de régression. L'équation pour la régression linéaire simple est :

$$Valeur\ prédite\ Y = \alpha + \beta X$$

Où :

Y – la valeur attendue de Y (variable dépendante)
α – une constante appelée « constante d'interception » (intercept)
β – la « constante de pente » de la ligne de régression (slope)
X – la valeur de la variable X (variable indépendante)

L'équation de la régression linéaire simple est connue sous le nom de forme pente-interception. La pente (β) st le nombre multiplié par la valeur de X et l'interception (α) est le nombre ajouté ou soustrait. La constante de pente indique le changement dans Y lorsque X augmente de 1 unité. La constante d'interception est la valeur de Y lorsque X est 0 (c'est le point où la ligne de régression intersecte l'axe des Y).

Les interprétations du coefficient de régression (β) sont les suivantes :

Si $\beta = 0$, la variable Y ne dépend pas de la variable X

Si $\beta \neq 0$, la variable Y dépend de la variable X comme suit :

$\beta > 0$, indique une relation positive entre Y et X

$\beta < 0$, indique une relation négative entre Y et X

Une fois les valeurs de α et β déterminées, la valeur attendue de Y peut être prédite pour toute valeur donnée de X. Par exemple, il a été montré que la vitesse de clairance hépatique de la lidocaïne (variable Y, mL/min/kg) peut être prédite en fonction de la vitesse de clairance hépatique du colorant vert d'indocyanine (variable X, mL/min/kg), en utilisant l'équation $Y = 0.30 + 1.07X$. T Cette approche permet aux anesthésistes de réduire le risque de surdosage en lidocaïne en évaluant la clairance du colorant (Pagano M., Gauvreau K., 2000).

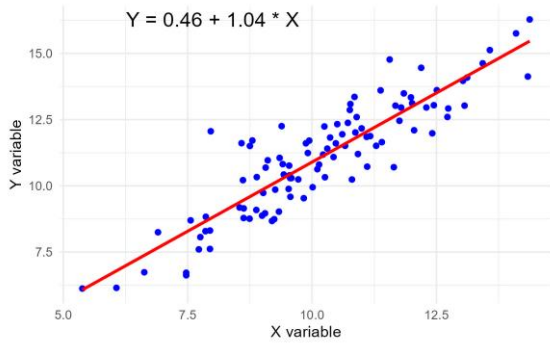

Figure 5.3 Diagramme de dispersion avec la ligne de régression pour un ensemble de données hypothétiques

5.2.2 Régression linéaire multiple

Les techniques de régression linéaire multiple sont appliquées lorsqu'une ou plusieurs variables continues X sont utilisées pour prédire la valeur attendue de Y. L'équation de régression multiple est formulée comme suit :

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

5.2.3 Régression logistique

Dans le cas de la régression linéaire, la variable de réponse Y est numérique et supposée suivre une distribution normale. *La régression logistique* traite des situations où les variables prédicteurs (indépendantes X) sont numériques et les variables prédites (dépendantes Y) sont nominales (par exemple, survie vs décès, guéri vs non guéri). Au lieu de prédire une valeur moyenne, la régression logistique estime la probabilité associée à la réponse dichotomique pour différentes valeurs d'une variable explicative.

Exercices de révision

1. Un groupe de 10 étudiants a été observé pour le nombre d'heures d'étude et leurs notes correspondantes

Ensemble de données 1. Heures d'étude vs. Notes aux examens

Nr. d'observations	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Heures d'étude	2	3	4	5	6	7	8	9	10	11
Notes aux examens	56	61	68	70	72	75	78	80	85	88

- Calculez le coefficient de corrélation approprié et expliquez votre choix.
- Interprétez le coefficient de corrélation calculé.
- Réalisez un diagramme de dispersion à deux variables pour ces données.

2. Un groupe de 10 individus a enregistré leur temps quotidien d'exercice (en minutes) et leur perte de poids correspondante (en kg) sur une période d'un mois.

Ensemble de données 2. Temps d'exercice quotidien vs. Perte de poids

Nr. d'observations	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Temps d'exercice quotidien (minutes)	15	20	25	30	35	40	45	50	55	60
Perte de poids (kg)	0.5	0.8	1.0	1.3	1.5	1.8	2.0	2.3	2.5	2.8

- Calculez le coefficient de corrélation approprié et expliquez votre choix.
- Interprétez le coefficient de corrélation calculé.
- Réalisez un diagramme de dispersion à deux variables pour ces données.

3. Si la relation entre deux mesures est linéaire et que le coefficient de corrélation est proche de 1, le diagramme de dispersion des observations :

a) Est une ligne droite horizontale
b) Est une ligne droite verticale
c) Est une ligne droite qui n'est ni horizontale ni verticale
d) A une pente négative
e) A une pente positive

4. Si la relation entre deux mesures est linéaire et que le coefficient de corrélation est proche de -1, le diagramme de dispersion des observations :

a) Est une ligne droite horizontale
b) Est une ligne droite verticale
c) Est une ligne droite qui n'est ni horizontale ni verticale
d) A une pente négative
e) A une pente positive

Questions de révision

1. Dans quelles circonstances l'utilisation de la corrélation est-elle préférable ?
2. Quels sont les avantages et les limites du coefficient de corrélation de Pearson ?
3. En quoi la corrélation des rangs de Spearman se distingue-t-elle de la corrélation de Pearson ?
4. Pourquoi est-il crucial de créer un graphique de dispersion lors de l'examen de la relation entre deux variables continues ?
5. Si un test d'hypothèse révèle que la corrélation entre deux variables n'est pas significativement différente de zéro, cela signifie-t-il nécessairement que les variables sont indépendantes ? Justifiez votre réponse.
6. Quelle est la principale différence entre les analyses de corrélation et les analyses de régression ?
7. Dans quelles conditions l'emploi de la régression linéaire simple est-il approprié ?
8. Dans quelles situations l'application de la régression linéaire multiple est-elle préférable ?
9. Dans quel contexte l'utilisation de la régression logistique est-elle recommandée ?

CHAPITRE 6. STATISTIQUES INFÉRENTIELLES : THÉORIE DES PROBABILITÉS ET TESTS D'HYPOTHÈSES

Concepts clés

- *Les statistiques inférentielles* sont obtenues à partir des données d'échantillon et permettent de formuler des inférences (conclusions) concernant les paramètres de la population.
- Les statistiques inférentielles ne sont applicables qu'aux échantillons *probabilistes*. Pour les échantillons *non probabilistes*, seules les statistiques descriptives peuvent être utilisées.
- *La probabilité* joue un rôle fondamental dans les statistiques inférentielles. Pour évaluer si un résultat d'étude est statistiquement significatif, il est nécessaire de s'appuyer sur la probabilité pour effectuer cette détermination.
- *La distribution des moyennes d'échantillons* tend à être normale, quelle que soit la forme de la distribution de la population d'origine des échantillons (*Théorème de la limite centrale*).
- *L'erreur standard* représente la différence moyenne entre la moyenne de la population et la moyenne d'un échantillon individuel.
- *L'erreur standard* est inversement proportionnelle à la racine carrée de la taille de l'échantillon (n). Des échantillons réduits génèrent des erreurs standards élevées.
- *L'hypothèse nulle* (H_0) postule qu'il n'y a pas d'effet dans la population (par exemple, les moyennes de deux populations ne diffèrent pas).
- *L'hypothèse alternative* (H_1) prévoit l'existence de différences entre les groupes (par exemple, les moyennes de deux populations diffèrent).
- Une hypothèse alternative *bilatérale* ou *non directionnelle* ne spécule pas sur la valeur qui sera la plus grande.
- Une hypothèse alternative *unilatérale* ou *directionnelle* précise quelle valeur sera la plus grande.

- La règle « AAA » pour une erreur de type I : l'erreur Alpha accepte l'hypothèse alternative incorrecte (*Alpha error Accepts the Alternative*).
- La règle « BEAN » pour une erreur de type II : l'erreur Beta accepte l'hypothèse nulle incorrecte (*Beta Error Accepts the Null*).
- *Le niveau de confiance* est la probabilité d'accepter l'hypothèse nulle lorsqu'elle est effectivement vraie.
- *La puissance de l'étude* est la probabilité de rejeter l'hypothèse nulle lorsqu'elle est effectivement fausse.
- *Le niveau de signification (niveau α)* représente la probabilité de commettre une erreur de type I, fixée avant le calcul du test statistique.
- Le niveau α est généralement fixé à 0,05 ou moins.
- *La valeur p* correspond à la probabilité de commettre une erreur de type I, déterminée après le calcul du test statistique.
- Si la valeur p < niveau α (valeur p < 0,05), il est possible de rejeter l'hypothèse nulle. Vous pouvez conclure que la différence entre les deux moyennes est statistiquement significative.
- *Un intervalle de confiance* est un intervalle calculé à partir des statistiques d'échantillon qui contient le paramètre de la population (par exemple, la moyenne) avec un certain degré de confiance (par exemple, 95 % ou 99 % de confiance).

Les statistiques inférentielles utilisent les données d'échantillon pour tirer des conclusions sur les paramètres de la population. Elles permettent de déterminer si un phénomène observé dans un échantillon existe réellement dans la population plus large dont l'échantillon est issu.

6.1 Théorie des probabilités

6.1.1 Concepts généraux

La probabilité est essentielle en statistiques inférentielles, notamment pour déterminer si un résultat d'étude est statistiquement significatif, c'est-à-dire si notre résultat n'est pas dû au hasard. La théorie des probabilités est fondamentale pour analyser de grands ensembles de données et des expériences répétées, appelées essais, qui produisent divers résultats.

La définition classique de la probabilité (p) stipule que la probabilité d'un événement est le nombre de résultats favorables (m) divisé par le nombre total de résultats possibles (n) :

$$p = \frac{m}{n}$$

La probabilité qu'un événement ne se produise pas (q) est :

$$q = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - p$$

Ainsi, la somme des probabilités qu'un événement se produise ou non est égale à 1 (ou 100 %). La valeur de p varie de 0 à 1 (0 % à 100 %), une valeur de p plus élevée indiquant une probabilité plus grande de réalisation de l'événement.

Deux concepts mathématiques importants en théorie des probabilités sont la Loi des grands nombres et le Théorème central limite.

6.1.2 Loi des grands nombres

La Loi des grands nombres stipule qu'à mesure que le nombre d'essais dans une expérience augmente, la moyenne des résultats se rapproche de la valeur attendue. En augmentant le nombre d'essais, la moyenne des résultats expérimentaux devient aussi proche que possible de la valeur attendue.

6.1.3 Théorème central limite

Le Théorème central limite décrit les propriétés de la distribution d'échantillonnage des moyennes d'échantillons :

1. La moyenne de la distribution d'échantillonnage est égale à la moyenne de la population (μ).
2. L'écart-type de la distribution d'échantillonnage des moyennes est appelé erreur standard de la moyenne ($SE_{\bar{x}}$;). L'erreur standard de la moyenne est calculée en divisant l'écart-type de la population (σ) par la racine carrée de la taille de l'échantillon (*n*):

$$SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

En d'autres termes, l'erreur standard représente la différence moyenne entre la moyenne de la population et la moyenne d'un échantillon individuel.

Où :

$SE_{\bar{x}}$: erreur standard de la moyenne

σ: écart-type de la population

n – taille de l'échantillon

3. À condition que *n* soit suffisamment grand, la forme de la distribution d'échantillonnage est approximativement normale, quelle que soit la forme de la distribution de la population dont sont extraits les échantillons.

6.1.4 Utilisation de l'erreur standard

L'erreur standard dépend de la taille de l'échantillon : plus l'échantillon est grand, plus la moyenne de l'échantillon ($\bar{X}$) représente fidèlement la moyenne de la population (μ). L'erreur standard indique l'incertitude autour de l'estimation de la moyenne, reflétant combien la moyenne de l'échantillon pourrait varier avec des échantillonnages répétés. Des erreurs standards importantes proviennent d'échantillons petits avec de

grandes écarts-types. L'erreur standard est utilisée pour calculer les intervalles de confiance, ce qui aide à inférer les paramètres de la population.

6.2 Échantillonnage

6.2.1 Définition de l'échantillonnage

La théorie des probabilités nous permet de tirer des conclusions sur les caractéristiques de la population en utilisant des données d'échantillon. L'échantillonnage consiste à sélectionner un groupe au sein de la population afin de collecter des données pour la recherche. Les raisons de l'échantillonnage incluent la réduction des coûts et du temps, l'obtention de résultats plus précis, la diminution de l'hétérogénéité et l'estimation des erreurs.

6.2.2 Méthodes d'échantillonnage

Plusieurs méthodes d'échantillonnage sont utilisées dans la recherche médicale, chacune nécessitant un choix aléatoire pour utiliser efficacement les statistiques inférentielles. Les échantillons probabilistes garantissent que la probabilité d'inclusion de chaque sujet est connue, ce qui permet des inférences valides. Les échantillons non probabilistes ne permettent que des statistiques descriptives.

A. Échantillonnage probabiliste :

⇒ *Échantillonnage aléatoire simple* : Chaque membre de la population a une chance égale d'être sélectionné. Les unités sont sélectionnées au hasard jusqu'à atteindre la taille de l'échantillon.

⇒ *Échantillonnage systématique* : Chaque k-ième élément est sélectionné dans une liste de population, avec k déterminé en divisant la taille de la population par la taille d'échantillon souhaitée.

⇒ *Échantillonnage stratifié* : La population est divisée en groupes mutuellement exclusifs (strates), et des échantillons aléatoires sont tirés de chaque strate.

⇒ *Échantillonnage en grappes* : Au départ, un ensemble de groupes ou « grappes » est sélectionné au hasard dans une population, puis des cas sont sélectionnés au hasard parmi les grappes. Les grappes sont généralement basées sur des zones géographiques.

B. Échantillonnage non probabiliste :

Les échantillons non probabilistes sont ceux pour lesquels la probabilité qu'une unité soit sélectionnée est inconnue.

⇒ Échantillonnage de commodité : Le chercheur sélectionne les membres de la population les plus faciles à obtenir des informations.

⇒ Échantillonnage par quotas : Le chercheur interroge un nombre prescrit de personnes dans chacune de plusieurs catégories.

6.3 Estimation et test des hypothèses

L'estimation et le test des hypothèses sont des éléments clés de la statistique inférentielle, permettant aux chercheurs de tirer des conclusions sur les données et les relations entre les variables.

Estimation :

- *Estimation Ponctuelle :* Utilise les données de l'échantillon pour calculer un seul nombre afin d'estimer le paramètre d'intérêt, comme la moyenne de la population, sans fournir d'informations sur la variabilité de l'estimation. On ne sait pas à quel point la moyenne de l'échantillon ($\bar{X}$) est proche de la moyenne de la population (μ).
- *Estimation par Intervalle :* Utilise une plage de valeurs pour estimer le paramètre, fournissant un intervalle de confiance (IC) qui contient la moyenne de la population (μ) avec un certain degré de confiance.

Test des Hypothèses : Consiste à formuler une hypothèse nulle (H_0) et une hypothèse alternative (H_1), puis à réaliser un test statistique pour déterminer quelle hypothèse accepter. L'objectif est de rejeter l'hypothèse nulle et d'accepter l'hypothèse alternative.

6.4 Intervalles de confiance

Un intervalle de confiance est une estimation par intervalle d'un paramètre de population, représenté par la moyenne, la proportion, le coefficient de corrélation ou les différences entre deux moyennes ou proportions. Les extrémités de l'intervalle sont appelées limites de confiance. Les intervalles de confiance (IC) sont calculés en utilisant la moyenne de l'échantillon et l'erreur standard :

$$CI = \bar{X} \pm z \times SE_{\bar{x}}$$

Où :

$\overline{X}$: Moyenne de l'échantillon

$SE_{\bar{x}}$: Erreur standard

z : La valeur z égale à 1.96 pour CI_{95} et 2.56 pour CI_{99}.

Exemple de calcul de l'intervalle de confiance

Cet exemple illustre la méthode pour déterminer l'Intervalle de Confiance avec un degré de confiance de 95% (IC_{95}) pour le jeu de données suivant.

Tableau 6.1 Données sur le niveau de cholestérol collectées pour 10 patients

Unité d'observation	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Niveau de cholestérol sanguin (mg/dl), x_i	168	258	228	230	247	156	172	165	210	264

1. Calculez la moyenne de l'échantillon.

$\bar{X} = \frac{\sum x_i}{n}$ $\bar{X} = 210\ mg/dl$

2. Déterminez l'écart-type.

$s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$ $s = 41.1\ mg/dl$

3. Calculez l'erreur standard (SE).

$SE = \frac{s}{\sqrt{n}}$ $SE = \frac{41.1}{3{,}16} = 13.1\ mg/dl$

4. Déterminez les limites inférieure et supérieure de l'intervalle de confiance IC_{95}

$IC_{95} = \overline{X} \pm z \times SE$

$IC_{95} = 210 \pm 1.96 \times 13.1$

IC_{95}: de 183.5 (limite inférieure) à 236 (limite supérieure)

Interprétation des résultats : Si vous répétez l'expérience 100 fois, la moyenne de la population (μ) se situera en dehors des limites de l'intervalle de confiance seulement 5 fois sur 100. Dans 95 cas sur 100, la moyenne de la population sera comprise entre les limites inférieure et supérieure de l'intervalle de confiance. Pour IC_{95}, la probabilité de trouver la moyenne de la population à l'intérieur de ces limites est de 0,95. La probabilité de la trouver en dehors est de 0,05.

Comparer deux moyennes à l'aide des intervalles de confiance

Dans la recherche médicale, comparer les moyennes de différents groupes est une tâche courante pour déterminer si un traitement ou une condition a un effet significatif. En comparant les intervalles de confiance (IC) de différents groupes, nous pouvons conclure si la différence entre les deux moyennes est significative, c'est-à-dire statistiquement significative, ou non.

La figure 6.1 compare les fréquences cardiaques de trois groupes de patients : le groupe B et le groupe C sont comparés au groupe de référence A. Les fréquences cardiaques sont représentées par leurs moyennes et les intervalles de confiance à 95 % (barres d'erreur). La fréquence cardiaque du groupe B est significativement plus élevée que celle du groupe A, car l'IC à 95 % pour le groupe B ne chevauche pas l'IC à 95 % pour le groupe A (p-valeur pour le test t < 0,05). La différence de fréquences cardiaques entre le groupe A et le groupe C n'est pas statistiquement significative ; le chevauchement des IC suggère que la différence observée pourrait être due à une variation aléatoire (p-valeur pour le test t > 0,05).

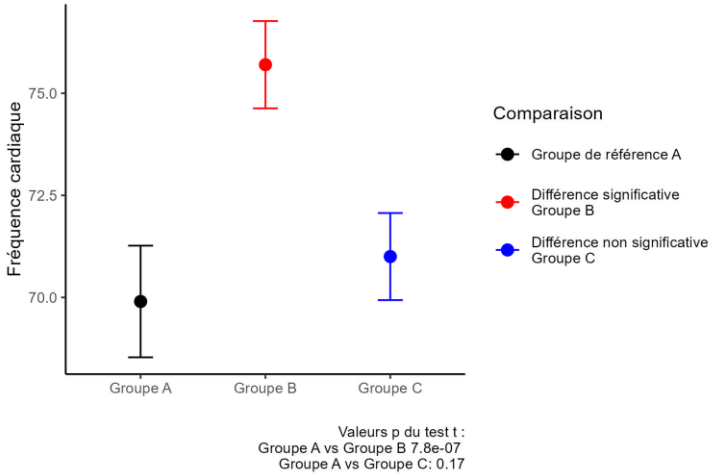

Figure 6.1 Comparaison des fréquences cardiaques entre trois groupes utilisant des intervalles de confiance à 95 %

Note : Comparé au groupe A (groupe de référence), la différence de la fréquence cardiaque moyenne est statistiquement significative seulement pour le groupe B (absence de chevauchement des IC), mais pas pour le groupe C (chevauchement des IC).

Le tableau 6.2 présente un cadre pour interpréter le chevauchement des intervalles de confiance et ses implications pour la signification statistique des différences entre les groupes.

Tableau 6.2 Comparaison de deux moyennes à l'aide des intervalles de confiance (IC)

Chevauchement des IC	Signification statistique entre les groupes comparés
Aucun	Différence hautement significative
Léger	Différence potentiellement significative mais non hautement significative
Large	Certainement non significative

6.5 Test des hypothèses : Concepts théoriques de base

6.5.1 Définition de l'hypothèse

Une hypothèse est une supposition fondée ou une conjecture concernant un phénomène. Elle se concentre sur la recherche, apportant clarté, précision et objectivité.

6.5.2 Types d'hypothèses

⇒ *Hypothèse nulle (H_0) :* Affirme qu'il n'existe pas de différence significative ou de relation entre les variables (par exemple, les moyennes de deux populations sont identiques).

$$H_0: \mu_1 = \mu_2$$

⇒ *Hypothèse alternative (H_1) :* Contredit l'hypothèse nulle en affirmant qu'il existe des différences entre les groupes (par exemple, les moyennes de deux populations sont différentes).

- *Hypothèse alternative non directionnelle* ou *bilatéral (à deux queues) :* Indique qu'une différence existe sans préciser laquelle des valeurs est plus élevée, testée par un test bilatéral.

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

- *Hypothèse alternative directionnelle* ou *unilatéral (à une seule queue) :* Précise la direction attendue de la différence, en indiquant laquelle des valeurs est plus élevée, testée par un test unilatéral.

$$H_1 : \mu_1 > \mu_2 \qquad \text{or} \qquad H_1 : \mu_1 > \mu_2$$

Ensemble, l'hypothèse nulle et l'hypothèse alternative couvrent toutes les valeurs possibles de la moyenne de la population (μ) ; par conséquent, l'une des deux affirmations doit nécessairement être vraie.

6.5.3 Erreur de type I et erreur de type II

Deux types d'erreurs peuvent se produire lors du test des hypothèses :

⇒ *Erreur de Type I :* Rejet de l'hypothèse nulle lorsqu'elle est en réalité vraie (faux positif). La probabilité de commettre une erreur de Type I (α) est également connue sous le nom *d'erreur de rejet.*

⇒ *Erreur de Type II :* Acceptation de l'hypothèse nulle lorsqu'elle est en réalité fausse (faux négatif). La probabilité de commettre une erreur de Type II (β) est également connue sous le nom d'erreur d'acceptation.

6.5.4 Puissance de l'étude

La puissance d'une étude se réfère à sa capacité à détecter une différence réelle lorsqu'elle existe. Plus précisément, c'est la probabilité de rejeter l'hypothèse nulle lorsqu'elle est fausse, ce qui permet de

conclure que l'hypothèse alternative est vraie lorsqu'elle l'est vraiment. Ainsi, la puissance d'une étude représente la probabilité d'éviter une erreur de type II (β). Mathématiquement, cela s'exprime ainsi :

$$\text{Puissance de l'étude} = 1 - \beta$$

Où : β est la probabilité de commettre une erreur de type II.

Pour qu'une étude soit considérée comme acceptable, il est généralement nécessaire qu'elle présente une puissance d'au moins 0.8 (ou un β de 0.2). Autrement dit, une étude doit avoir au moins 80 % de chances de détecter une différence réelle si elle existe. Une étude avec une puissance inférieure à 80 % est typiquement inacceptable.

L'un des moyens les plus pratiques et importants pour augmenter la puissance d'une étude est d'augmenter la taille de l'échantillon. Des échantillons plus grands réduisent l'erreur standard, rendant plus facile la détection de différences ou d'effets réels.

6.5.5 Niveau de confiance

Le niveau de confiance exprime la capacité d'une étude à ne pas détecter une différence inexistante. Il s'agit de la probabilité d'accepter l'hypothèse nulle lorsqu'elle est effectivement vraie. Le niveau de confiance se définit comme :

$$\text{Niveau de confiance} = 1 - \alpha$$

Où : α est la probabilité de commettre une erreur de type I.

Le tableau 6.3 présente les quatre résultats possibles lors d'un test d'hypothèse. Il est crucial de rappeler que nous testons toujours l'hypothèse nulle, qui peut être soit vraie, soit fausse dans une réalité inconnue.

Tableau 6.3 Quatre résultats possibles du test de l'hypothèse nulle (H_0)

	H_0 est VRAIE	**H_0 est FAUSSE**
ACCEPTER H_0	Décision correcte *Niveau de confiance* 1-α	Erreur de type II β
REJETER H_0	Erreur de type I α	Décision correcte *Puissance de l'étude* 1-β

6.5.6 Niveau de signification

Avant de pouvoir rejeter l'hypothèse nulle, il est nécessaire d'avoir une certitude raisonnable que toute différence observée entre la statistique de l'échantillon ($\bar{X}$) et le paramètre de la population (μ) n'est pas due au hasard. À quel point une moyenne d'échantillon doit-elle différer d'une moyenne de population pour que cette différence soit considérée comme significative ou statistiquement significative ? Ce critère est appelé *niveau de signification* ou *niveau α*, qui indique la probabilité de commettre une erreur de type I, c'est-à-dire de rejeter l'hypothèse nulle alors qu'elle est en réalité vraie.

Le niveau de signification (niveau α) est la probabilité de commettre une erreur de type I, déterminée avant le calcul du test statistique. Afin de minimiser le risque de rejeter incorrectement l'hypothèse nulle, le niveau de signification doit être suffisamment bas. Il est couramment fixé à 0,05 ou moins (0,01 ou 0,001). Plus le niveau α est bas, moins il est probable de commettre une erreur de type I.

6.5.7 Valeur p

La valeur p est une mesure étroitement liée au niveau de signification (niveau α). Elle exprime la probabilité d'obtenir les résultats observés en

supposant que l'hypothèse nulle est vraie. Il s'agit de la probabilité de commettre une erreur de type I après le calcul du test statistique.

Après avoir réalisé un test statistique, la valeur p est comparée au niveau α. Si la valeur p est inférieure au niveau α (par exemple, valeur p < 0.05), il est possible de rejeter l'hypothèse nulle et d'accepter l'hypothèse alternative. Cela signifie que la différence entre les deux moyennes est statistiquement significative et qu'il est peu probable qu'elle soit due au hasard.

À l'inverse, si la valeur p est supérieure au niveau α (par exemple, valeur p > 0.05), il faut accepter l'hypothèse nulle et conclure que la différence entre les deux moyennes n'est pas statistiquement significative et qu'elle est probablement attribuable au hasard.

6.6 Processus de test d'hypothèse : approches générales

Lors de la réalisation d'un test d'hypothèse, notre objectif est de tirer des conclusions sur un paramètre de population à partir de données d'échantillon. Une des méthodes consiste à construire un intervalle de confiance pour la moyenne d'une population (μ); une autre consiste à effectuer un test statistique. Les principales étapes du processus de test d'hypothèse incluent :

1. Formuler les hypothèses : hypothèse nulle (H_0) et hypothèse alternative (H_1);
2. Sélectionner le test statistique approprié ;
3. Déterminer le niveau de signification (niveau α) ;
4. Identifier la valeur critique nécessaire pour que le test statistique soit considéré comme significatif ;
5. Effectuer les calculs nécessaires ;
6. Formuler les conclusions.

Les étapes détaillées du test d'hypothèse, accompagnées d'exemples, seront présentées en détail au chapitre 7.

Exercices de révision

1. Quel est l'objectif d'un test d'hypothèse ?
2. Expliquez brièvement la relation entre les intervalles de confiance et les tests d'hypothèse.
3. Dans quelles circonstances utiliseriez-vous un test d'hypothèse unilatéral plutôt qu'un test bilatéral ?
4. Décrivez les deux types d'erreurs possibles lors de la réalisation d'un test d'hypothèse.
5. Le niveau de cholestérol sérique dans un échantillon de 400 hommes adultes présente une distribution asymétrique à gauche. La distribution d'échantillonnage des moyennes de cholestérol sérique est-elle :
    a) Asymétrique à gauche
    b) Asymétrique à droite
    c) Normale
    d) Impossible à déterminer
6. L'erreur standard d'une statistique est :
    a) La moyenne de la distribution de l'échantillon
    b) L'écart-type de la distribution de l'échantillon
    c) La moyenne divisée par la racine carrée de la taille de l'échantillon (n)
7. La pression artérielle systolique moyenne dans un groupe de 500 individus sélectionnés parmi un total de 100 000 individus de la localité A est de 130 mmHg avec un écart-type de 15 mmHg. Calculez : l'erreur standard et l'intervalle de confiance à 95 % pour la moyenne de la population. Interprétez vos résultats.
8. Une étude a été réalisée concernant la pression artérielle systolique chez des hommes de 60 ans atteints de diabète sucré. Dans cette étude, un échantillon aléatoire de 300 hommes de 60

ans atteints de diabète a été sélectionné, avec une moyenne de pression artérielle systolique de 160 mmHg et un écart-type de l'échantillon de 25 mmHg.

a) Calculez un intervalle de confiance à 95 % pour la pression artérielle systolique moyenne parmi la population des hommes de 60 ans atteints de diabète.
b) Supposons que la taille de l'échantillon ait été de 150 au lieu de 300, tout en maintenant la même moyenne et les mêmes écarts-types d'échantillon. L'intervalle de confiance s'élargit-il ou se rétrécit-il ? Pourquoi ?

Questions de révision

1. Quel est le concept de la statistique inférentielle ?
2. Quel est le rôle de la théorie des probabilités en statistique ?
3. Décrivez la loi des grands nombres et sa pertinence pour les applications de recherche.
4. Expliquez le théorème de la limite centrale.
5. Qu'est-ce que l'échantillonnage et pourquoi est-il essentiel dans l'analyse statistique ?
6. Pourquoi l'échantillonnage aléatoire est-il crucial pour garantir la validité des résultats de recherche ?
7. Dans quels cas l'échantillonnage systématique peut-il être préféré à l'échantillonnage aléatoire simple ?
8. Quand la méthode de l'échantillonnage stratifié est-elle la plus appropriée ?
9. Dans quelles conditions l'échantillonnage par grappes est-il la méthode la plus adaptée ?
10. Quel est le but principal de la réalisation de tests d'hypothèses ?

11. Quelle est la différence entre l'estimation ponctuelle et l'estimation par intervalle ?
12. Définissez une hypothèse et expliquez les différents types d'hypothèses.
13. Quels sont les intervalles de confiance et les limites de confiance, et comment sont-ils utilisés dans l'analyse statistique ?
14. Quels facteurs influencent la largeur d'un intervalle de confiance pour une moyenne ?
15. Décrivez les deux principaux types d'erreurs qui peuvent survenir lors des tests d'hypothèses.
16. Clarifiez la distinction entre les erreurs de type I et les erreurs de type II dans les tests d'hypothèses.
17. Que signifie la puissance d'une étude et comment peut-elle être décrite ?
18. Quel est le niveau de confiance et comment doit-il être interprété ?
19. Que représente le niveau de signification et comment peut-il être expliqué ?
20. Qu'est-ce que la valeur p et que signifie-t-elle dans le contexte des tests statistiques ?

CHAPITRE 7. TESTS D'HYPOTHÈSES : MÉTHODES PARAMÉTRIQUES ET NON-PARAMÉTRIQUES

Concepts clés

- Les tests statistiques *paramétriques* sont utilisés pour les données numériques (échelle d'intervalle ou de ratio), en supposant que les données de la population suivent une distribution normale.
- *Les tests t* sont des exemples de tests statistiques paramétriques.
- Les tests statistiques *non-paramétriques* sont utilisés pour les données catégoriques (nominales ou ordinales).
- *Le test du chi carré* (χ^2) est un exemple de test statistique non paramétrique. Ce test est utilisé pour vérifier les différences de proportions et les associations entre deux variables.
- Les tests statistiques *non-paramétriques* sont moins puissants que les tests paramétriques.
- *Le type de test d'hypothèse*—bilatéral, unilatéral à gauche ou unilatéral à droite—dépend de l'hypothèse alternative.
- Si la valeur absolue de la statistique de test est supérieure à la valeur critique, l'hypothèse nulle est rejetée. Cela signifie que la p-valeur est inférieure au niveau α.
- Si la valeur absolue de la statistique de test est inférieure à la valeur critique, l'hypothèse nulle n'est pas rejetée. Cela signifie que la p-valeur est supérieure au niveau α.

7.1 Tests paramétriques et non-paramétriques

Les méthodes de test d'hypothèse se divisent en deux catégories : paramétriques et non paramétriques, selon le type de données analysées. Les méthodes paramétriques sont adaptées aux données d'échelle d'intervalle et de rapport, en supposant que ces données suivent une distribution normale. À l'inverse, les méthodes non-

paramétriques conviennent aux données d'échelle ordinale ou nominale et n'exigent pas l'hypothèse d'une distribution normale des données dans une population.

Le choix des méthodes statistiques appropriées dépend principalement de plusieurs facteurs clés :

⇒ Type de données (numériques, nominales ou ordinales) ;
⇒ Nature des échantillons (indépendants ou appariés) ;
⇒ Taille de l'échantillon (n>30 ou n<30) ;
⇒ Nombre de groupes (un, deux ou plus) ;
⇒ Type d'hypothèse alternative (directionnelle ou non-directionnelle) ;
⇒ Type de distribution des données (normale ou asymétrique) ;
⇒ Homogénéité des variances.

La *Figure 7.1* illustre des exemples de tests statistiques paramétriques et non paramétriques.

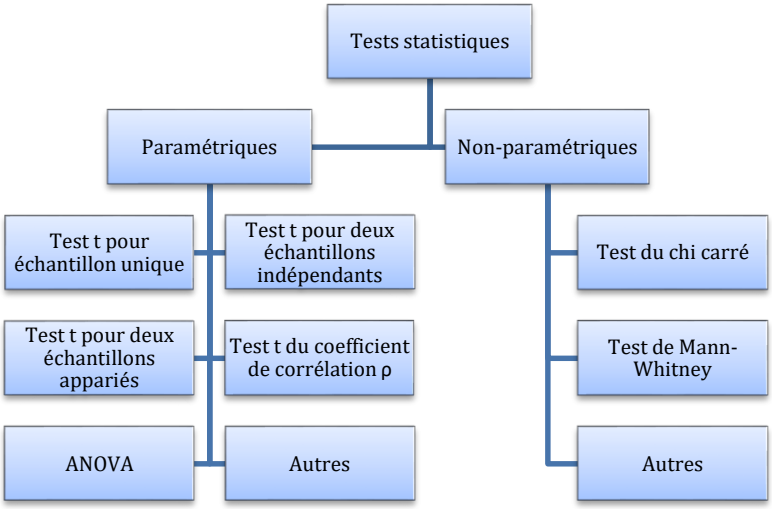

Figure 7.1 Tests Statistiques Paramétriques vs. Non-Paramétriques

7.2 Approche générale de la vérification des hypothèses

7.2.1 Étapes de la vérification des hypothèses

Bien qu'il existe de nombreux tests statistiques, les étapes de la vérification des hypothèses restent les mêmes.

Étape 1. Formulation des hypothèses :

- Hypothèse nulle (H_0): C'est l'hypothèse selon laquelle il n'y a aucun effet ou aucune différence. C'est l'hypothèse que le chercheur cherche à rejeter.
- Hypothèse alternative (H_1): C'est l'hypothèse selon laquelle il y a un effet ou une différence. C'est l'hypothèse que le chercheur cherche à prouver.

Étape 2. Sélection du test statistique approprié :

- Le choix du test statistique dépend du type de données et de la conception de l'étude. Les tests fréquemment utilisés comprennent les tests t, le test du chi carré, l'ANOVA, etc.
  - Test t pour échantillon unique : Utilisé pour comparer la moyenne d'un échantillon unique à une moyenne de référence.
  - Test t pour deux échantillons indépendants : Utilisé pour comparer les moyennes entre deux groupes indépendants.
  - Test t pour deux échantillons appariés : Utilisé pour comparer les moyennes au sein du même groupe à différents moments.
  - Test t pour le coefficient de corrélation ρ (rho): Utilisé pour évaluer la force et la direction de la relation entre deux variables continues.

- Test du chi carré : Utilisé pour les données catégorielles afin d'évaluer la probabilité que la distribution observée soit due au hasard.
- ANOVA : Utilisé pour comparer les moyennes entre trois groupes ou plus.

Étape 3. Détermination du niveau de signification (niveau α) :

- Le niveau α représente la probabilité de rejeter l'hypothèse nulle lorsqu'elle est vraie (erreur de type I).
- Les niveaux α couramment utilisés sont 0.05, 0.01 et 0.001.

Étape 4. Détermination des degrés de liberté et de la valeur critique :

- Degrés de liberté (df): Ils dépendent du test statistique et de la taille de l'échantillon (n).
  - Pour le test t pour échantillon unique : $\mathrm{df} = \mathrm{n} - 1$.
  - Pour le test t pour deux échantillons indépendants : $\mathrm{df} = \mathrm{n}_1 + \mathrm{n}_2 - 2$.
  - Pour le test t du coefficient de corrélation ρ : $\mathrm{df} = \mathrm{n} - 2$.
  - Pour le test du chi carré, les df sont généralement calculés en fonction du nombre de catégories.
- *Valeur critique* (t-critique) : Cette valeur est obtenue à partir des tables statistiques (distribution t, distribution chi carré, etc.), en fonction du niveau α et des degrés de liberté. Les tables statistiques sont disponibles en Annexe A (tableau de la distribution t) et en Annexe B (tableau du chi carré).

Étape 5. Réalisation des calculs :

- Calculez la statistique de test (t-calculée) en utilisant la formule adéquate pour le test sélectionné.

Étape 6. Formulation des conclusions :

- Comparez la valeur absolue de la statistique de test (|t-calculée|) avec la valeur absolue de la valeur critique (|t-critique|).
  - Si la valeur absolue de la statistique de test (|t-calculée|) dépasse la valeur critique (|t-critique|), *rejetez l'hypothèse nulle*. Cela indique que la valeur p est inférieure au niveau α, ce qui traduit une signification statistique (valeur p < niveau α).
  - Si la valeur absolue de la statistique de test (|t-calculée|) est inférieure à la valeur critique, *ne rejetez l'hypothèse nulle*. Cela signifie que la valeur p est supérieure au niveau α, indiquant l'absence de signification statistique (valeur p > niveau α).

7.2.2 Tests d'hypothèses : tests bilatéraux, unilatéraux gauche et unilatéraux droite

La vérification des hypothèses permet de déterminer si une affirmation concernant un paramètre de la population est vraie. Ce processus implique de formuler une hypothèse nulle (H_0) et une hypothèse alternative (H_1). Selon la nature de l'hypothèse alternative, on distingue trois types de tests d'hypothèses :

⇒ Test bilatéral : L'hypothèse alternative utilise le symbole « ≠ ».

⇒ Test unilatéral à gauche : L'hypothèse alternative utilise le symbole « < ».

⇒ Test unilatéral à droite : L'hypothèse alternative utilise le symbole « > ».

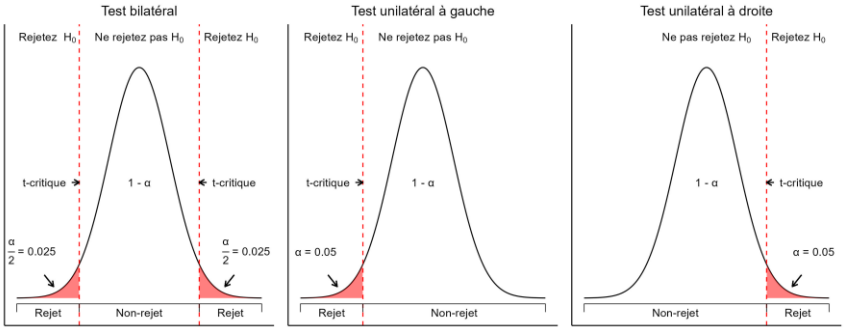

Figure 7.2 Illustration des valeurs critiques, des régions de rejet et de non-rejet dans les tests d'hypothèses

La *Figure 7.2* illustre les tests bilatéraux, unilatéraux à gauche et unilatéraux à droite à l'aide d'une représentation graphique d'une courbe normale en forme de cloche. L'aire sous cette courbe est égale à 1 ou 100 %. Une valeur critique t (représentée par une ligne pointillée dans la figure 7.2) divise cette aire en régions de rejet et de non-rejet de l'hypothèse nulle.

Pour *un test bilatéral*, l'hypothèse nulle est rejetée si la statistique de test (t-calculé) est soit trop faible, soit trop élevée, créant ainsi deux régions de rejet : une à gauche et une à droite. L'hypothèse nulle est rejetée si la statistique de test est inférieure ou supérieure aux valeurs critiques.

Pour *un test unilatéral à gauche*, l'hypothèse nulle est rejetée si la statistique de test est trop faible, plaçant ainsi la région de rejet à gauche de la distribution. L'hypothèse nulle est rejetée si la statistique de test est inférieure à la valeur critique à gauche.

Pour *un test unilatéral à droite*, l'hypothèse nulle est rejetée si la statistique de test est trop élevée, plaçant ainsi la région de rejet à droite de la distribution. L'hypothèse nulle est rejetée si la statistique de test dépasse la valeur critique à droite.

En utilisant l'approche des valeurs critiques, nous déterminons si la statistique de test calculée est plus extrême que la valeur critique. La statistique de test calculée est comparée à la valeur critique, qui sert de seuil de décision. Si la statistique de test dépasse cette valeur critique, l'hypothèse nulle est rejetée. Si la statistique de test n'est pas aussi extrême que la valeur critique, l'hypothèse nulle n'est pas rejetée.

7.3 Tests paramétriques

7.3.1. Test t pour échantillon unique

Définition : Le test t pour un échantillon est utilisé pour déterminer si la moyenne d'un échantillon unique est significativement différente d'une moyenne de population connue ou hypothétique. Ce test paramétrique est utile pour comparer la moyenne de l'échantillon à une valeur standard ou attendue.

Conditions d'utilisation du test t pour échantillon unique

- Les données doivent être continues (échelle d'intervalle/rapport).
- L'échantillon doit être sélectionné de manière aléatoire.
- Les données de l'échantillon doivent être approximativement distribuées normalement.
- L'écart type de la population est inconnu.

Example

Considérons un échantillon de 15 étudiants en médecine qui rapportent leurs heures de sommeil durant la semaine des examens finaux. Les données collectées sont les suivantes :

7, 6, 5, 6, 7, 8, 6, 7, 5, 6, 7, 8, 5, 6, 7

Vous souhaitez tester si le nombre moyen d'heures de sommeil pendant la semaine des examens finaux ($\bar{X}$) est significativement

différent des 8 heures recommandées (moyenne hypothétique de la population, μ).

Solution

1. **Formulation des hypothèses :**

- Hypothèse nulle (H_0) : $\mu = 8$ ((La durée moyenne de sommeil est de 8 heures)
- Hypothèse alternative (H_1) : $\mu \neq 8$ (La durée moyenne de sommeil n'est pas de 8 heures). Dans ce cas, l'hypothèse alternative est bilatérale, ce qui signifie que le test statistique est bilatéral.

2. **Sélection du test statistique approprié :**

- Utiliser un test t pour un échantillon puisque nous comparons la moyenne de l'échantillon à une moyenne connue de la population.

3. **Détermination du niveau de signification (niveau α) :**

- Choisir $\alpha = 0.05$.

4. **Détermination des degrés de liberté et de la valeur critique :**

- $df = 15 - 1 = 14$
- La valeur critique t pour un test bilatéral avec df=14 et α=0,05 est obtenue à l'aide du tableau de distribution (voir Annexe A). t- critique = ±2.145.

5. **Réalisation des calculs :**

- Calculer la moyenne de l'échantillon ($\bar{X}$):

$$\bar{X} = \frac{\sum X_i}{n} = \frac{96}{15} = 6.4$$

- Calculer l'écart-type de l'échantillon (s):

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{\sum (7-6.4)^2 + (6-6.4)^2 \ldots + (7-6.4)^2}{15-1}} \approx 1.12$$

- Calculer la statistique du test t en appliquant la formule du test t à un échantillon :

$$t = \frac{\bar{X} - \mu}{SE_{\bar{X}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{6.4 - 8}{1.12/\sqrt{15}} = \frac{-1.6}{0.289} \approx -5.54$$

6. Formulation des conclusions :

- Comparer la valeur t calculée à la valeur critique t :
  - La valeur t calculée est de -5.54
  - La valeur critique t est ±2.145.
- Conclusion : Selon nos calculs, la statistique du test est de -5.54. La valeur absolue de la statistique du test (5.54) dépasse celle de la valeur critique (2.145). Ainsi, nous *rejetons l'hypothèse nulle* et concluons que le nombre moyen d'heures de sommeil des étudiants en médecine durant leur semaine d'examens finaux est significativement différent des 8 heures recommandées (valeur $p < 0.05$).

Illustration des valeurs critiques et des régions de rejet

Dans notre exemple, le test t pour un échantillon est bilatéral, comme l'indiquent les deux valeurs critiques de ±2.145, correspondant à un niveau de signification (α) de 0.05 et *df* de 14 (*Figure 7.3*). Les zones ombragées représentent les régions de rejet, où l'hypothèse nulle serait rejetée. La région centrale non ombragée indique où l'hypothèse nulle ne serait pas rejetée. La valeur t calculée (-5.54) se trouve dans la région de rejet (au-delà de la valeur critique de -2.145), ce qui conduit au rejet de l'hypothèse nulle.

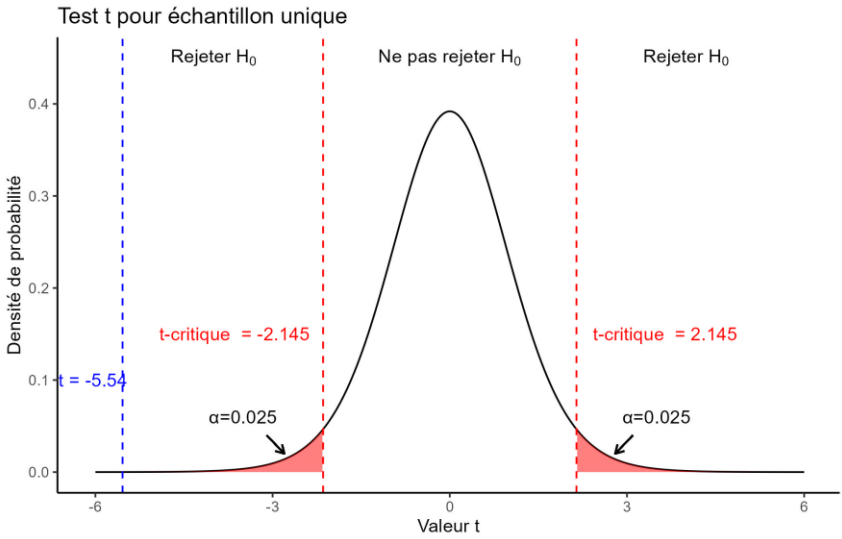

Figure 7.3 Test t pour échantillon unique : visualisation des valeurs critiques et des régions de rejet pour un test bilatéral

7.3.4 Test t pour deux échantillons indépendants

Définition : Le test t pour deux échantillons indépendants compare les moyennes de deux groupes indépendants pour déterminer s'il existe des preuves statistiques que les moyennes des populations associées sont significativement différentes.

Conditions d'application du test t pour deux échantillons indépendants

- Les deux échantillons doivent être sélectionnés de manière aléatoire
- Les échantillons doivent être indépendants l'un de l'autre.
- La variable dépendante doit être mesurée sur une échelle d'intervalle ou de rapport.
- Les données dans chaque groupe doivent être approximativement normalement distribuées.

- Le test peut être effectué avec des variances égales ou inégales entre les deux groupes. Différentes formules sont utilisées en fonction de l'égalité des variances.

Example

Un chercheur souhaite déterminer s'il existe une différence significative dans le temps moyen de récupération (en jours) entre les patients traités avec le Médicament A et ceux traités avec le Médicament B pour une maladie spécifique. Le chercheur collecte des données sur le temps de récupération de deux groupes indépendants de patients : un groupe traité avec le Médicament A et l'autre groupe traité avec le Médicament B.

Données

- Groupe A (Médicament A) :
  - Moyenne de l'échantillon ($\bar{X}_A$): 8 jours
  - Écart-type de l'échantillon(s_A): 2 jours
  - Taille de l'échantillon (n_A): 31
- Groupe B (Drug B):
  - Moyenne de l'échantillon ($\bar{X}_B$): 6 jours
  - Écart-type de l'échantillon (s_B): 1.5 jours
  - Taille de l'échantillon (n_B): 31

Solution

Nous appliquerons un test unilatéral, en supposant des variances inégales.

1. **Formulation des hypothèses :**

- Hypothèse nulle (H_0) : $\mu_A = \mu_B$. Il n'existe pas de différence significative dans le temps moyen de récupération entre les patients traités avec le Médicament A et ceux traités avec le

Médicament B. H_0 peut également être formulée comme $\mu_A - \mu_B = 0$.

- Hypothèse alternative (H_1): $\mu_A > \mu_B$. Le temps moyen de récupération des patients traités avec le Médicament A est plus élevé que celui des patients traités avec le Médicament B. On peut également écrire H_1 comme $\mu_A - \mu_B > 0$.

2. Sélection du test statistique approprié :

- Nous utilisons un test t pour deux échantillons indépendants avec des variances inégales (test de Welch).

3. Détermination du niveau de signification (niveau α) :

- Choisissez $\propto = 0.05$.

4. Détermination des degrés de liberté (df) et de la valeur critique :

- $df = 31 + 31 - 2 = 60$
- La valeur critique t pour un test unilatéral avec df=60 et α=0.05 peut être trouvée en utilisant le tableau de distribution t (voir Annexe A). t critique = 1.671.

5. Réalisation des calculs :

- Calculez la statistique du test t en utilisant la formule du test t pour échantillons indépendants avec variances inégales :

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{{s_A}^2}{n_A} + \frac{{s_B}^2}{n_B}}} = \frac{8 - 6}{\sqrt{\frac{2^2}{31} + \frac{1.5^5}{31}}} = \frac{2}{\sqrt{0.129 + 0.072}} \approx 4.46$$

6. Formulation des conclusions :

- Comparez la valeur t calculée à la valeur t critique :
  - La valeur t calculée est de 4.46
  - La valeur t critique est de 1.671.
- Conclusion : Selon nos calculs, la statistique de test est de 4.46. La valeur absolue de la statistique de test est supérieure à la valeur critique de 1.671. Par conséquent, nous *rejetons*

l'hypothèse nulle et concluons qu'il existe des preuves significatives suggérant que le temps de récupération moyen pour les patients traités avec le Médicament A est supérieur à celui des patients traités avec le Médicament B (valeur $p < 0.05$).

Illustration des valeurs critiques et des régions de rejet

Dans notre exemple, la valeur t calculée (4.46) se situe dans la région de rejet (elle est supérieure à la valeur critique de 1.671). Cela conduit au rejet de l'hypothèse nulle. Ce test t indépendant à deux échantillons est un test unilatéral droit. La zone ombragée à droite représente la région de rejet, où l'hypothèse nulle serait rejetée. La région non ombragée restante indique où l'hypothèse nulle ne serait pas rejetée (*Figure 7.4*).

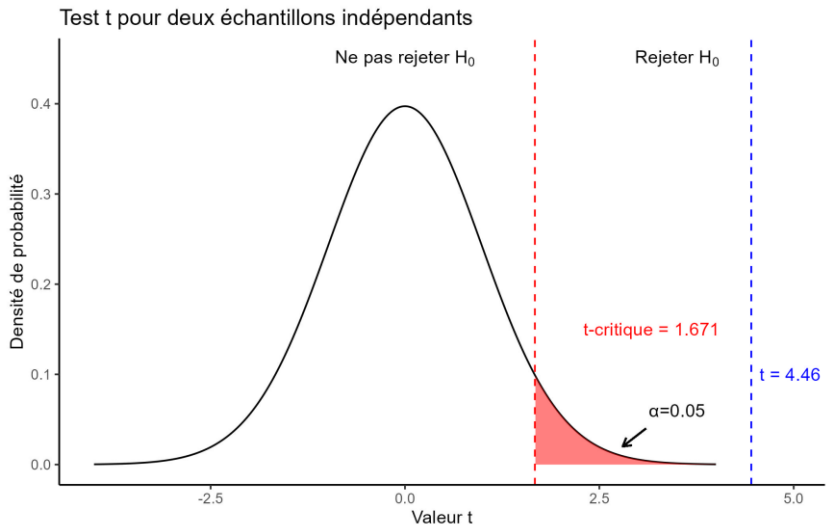

Figure 7.4 Test t pour deux échantillons indépendants : visualisation des valeurs critiques et des régions de rejet pour un test unilatéral à droit

7.3.5 Test t pour le coefficient de corrélation

Définition : Le test t du coefficient de corrélation ρ (rho) évalue l'existence d'une relation linéaire statistiquement significative entre deux variables continues.

Conditions d'application du test du coefficient de corrélation

- Les deux variables doivent être de nature continue.
- Les données doivent suivre une distribution bivariée.
- L'échantillon doit être sélectionné de manière aléatoire.
- La relation entre les variables doit être linéaire.

Example

Un chercheur étudie la relation entre le nombre d'heures d'activité physique par semaine des patients et leurs niveaux de cholestérol sanguin. Des données ont été collectées auprès de 30 patients. Pour chaque patient, le nombre d'heures d'exercice par semaine et le taux de cholestérol (mg/dl) ont été enregistrés. Le coefficient de corrélation calculé (r) est de -0.15. L'objectif est de déterminer s'il existe une corrélation négative statistiquement significative entre la durée de l'exercice physique et les niveaux de cholestérol

Solution

1. **Formulation des hypothèses :**

- Hypothèse nulle (H_0) : $\rho = 0$. Il n'y a pas de corrélation entre les heures d'exercice et les niveaux de cholestérol.
- Hypothèse alternative (H_1) : $\rho < 0$. Il y a une corrélation négative entre les heures d'exercice et les niveaux de cholestérol.

2. **Sélection du test statistique approprié :**

- On utilise le test de corrélation de Pearson.

3. **Détermination du niveau de signification (niveau α) :**

- Choisissez $\propto = 0.05$.

4. Détermination des degrés de liberté (df) et de la valeur critique :

- $df = 30 - 2 = 28$
- La valeur critique t pour un test unilatéral avec df=28 et α=0.05 se trouve dans le tableau de distribution t (voir Annexe A). t critique = 1.701.

5. Réalisation des calculs :

- Étant donné r = -0.15, calculez la statistique du test t en utilisant la formule suivante :

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$t = \frac{-0.15\sqrt{30-2}}{\sqrt{1-(-0.15)^2}} = \frac{-0.15\sqrt{28}}{\sqrt{1-0.0225}} = \frac{-0.15 \times 5.2915}{0.9887} = \frac{-0.7937}{0.9887} = -0.803$$

6. Formulation des conclusions :

- Comparez la valeur t calculée à la valeur t critique :
  - La valeur t calculée est de -0.803
  - La valeur t critique est de 1.701.
- Conclusion : Selon nos calculs, la statistique du test est de -0.803. La valeur absolue de cette statistique (0.803) est inférieure à la valeur critique de 1.701. Ainsi, nous ne rejetons pas l'hypothèse nulle et concluons qu'il n'existe pas de corrélation négative statistiquement significative entre les heures d'exercice et les niveaux de cholestérol.

Illustration des valeurs critiques et des régions de rejet

La *Figure 7.5* illustre le test unilatéral à gauche pour le coefficient de corrélation. La valeur critique t est indiquée par une ligne en pointillés à

-1.701. Si la valeur t calculée tombe dans la zone ombragée, l'hypothèse nulle est rejetée. Dans cet exemple, la valeur t calculée de -0.803 se situe dans la région de non-rejet, ce qui implique qu'il n'existe pas de corrélation négative statistiquement significative entre les heures d'exercice et les niveaux de cholestérol.

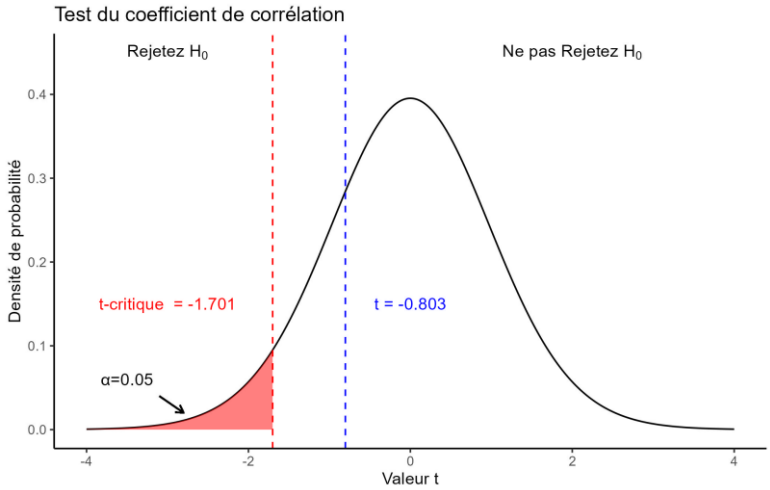

Figure 7.5 Test du coefficient de corrélation : visualisation des valeurs critiques et des régions de rejet pour un test unilatéral à gauche

7.4 Tests non paramétriques

Les tests non paramétriques ne supposent pas que les données d'une population sont distribuées normalement, c'est pourquoi ils sont appelés tests sans distribution. Ces tests sont utilisés pour examiner des données nominales, ordinales ou des données numériques asymétriques.

7.4.1 Test du Chi-Carré

Définition : Le test du Chi-Carré est une méthode statistique utilisée pour déterminer s'il existe une association significative entre deux variables catégorielles.

Conditions d'utilisation du test du Chi-Carré

- Les données doivent être présentées sous forme de dénombrements ou de fréquences dans un tableau de contingence 2x2.
- Les catégories doivent être mutuellement exclusives.
- La taille de l'échantillon doit être suffisamment grande (les fréquences attendues dans chaque cellule doivent être d'au moins 5).

Example

Un chercheur examine la relation entre le statut de fumeur (fumeur contre non-fumeur) et l'incidence du cancer du poumon (présent contre absent). Il a collecté des données auprès de 100 patients, comme le montre le *Tableau 7.1*. Il est nécessaire de déterminer s'il existe une association statistiquement significative entre le statut de fumeur et le cancer du poumon.

Tableau 7.1 Fréquences observées du statut de fumeur et du cancer du poumon dans un tableau 2x2

	Cancer du poumon présent	Cancer du poumon absent	Total
Fumeur	30	20	50
Non-fumeur	10	40	50
Total	40	60	100

Solution

1. **Formulation des hypothèses :**

- Hypothèse nulle (H_0) : Il n'existe pas d'association entre le statut de fumeur et le cancer du poumon.
- Hypothèse alternative (H_1) : Il existe une association entre le statut de fumeur et le cancer du poumon.

2. **Sélection du test statistique approprié :**

- Utilisation du test du Chi-Carré.

3. Détermination du niveau de signification (niveau α) :

- Choisissez $\propto = 0.05$.

4. Détermination des degrés de liberté (df) et de la valeur critique :

- $df = (nombre\ de\ lignes - 1) \times (nombre\ de\ colonnes - 1)$
- $df = (2-1) \times (2-1) = (2-1) \times (2-1) = 1$
- La valeur critique pour *df*=1 et α=0.05 est déterminée à l'aide du tableau de distribution du Chi-Carré (voir Annexe B). Elle est de 3.841.

5. Réalisation des calculs :

- Calculez des fréquences attendues pour chaque cellule :

$$E_{ij} = \frac{(Total\ de\ la\ ligne) \times (Total\ de\ la\ colonne)}{Total\ général}$$

Par exemple, pour la cellule « Fumeur & Cancer du poumon présent » :

$$E_{1,1} = \frac{50 \times 40}{100} = 20$$

De la même manière, calculez les valeurs attendues pour les autres cellules comme indiqué dans le Tableau 7.2.

Tableau 7.2 Fréquences attendues du statut de fumeur et du cancer du poumon dans un tableau 2x2

	Cancer du poumon présent	Cancer du poumon absent	Total
Fumeur	20	30	50
Non-fumeur	20	30	50
Total	40	60	100

- Calculez du test du Chi-Carré (χ^2) en utilisant les fréquences observées (O_{ij}) et attendues (E_{ij}) :

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\chi^2 = \frac{(30-20)^2}{20} + \frac{(20-30)^2}{30} + \frac{(10-20)^2}{20} + \frac{(40-30)^2}{30} = 16.66$$

6. Formulation des conclusions :

- Comparez la valeur calculée du Chi-Carré à la valeur critique :
  - La valeur calculée du Chi-Carré est de 16.66.
  - La valeur critique est de 3.841.
- Conclusion : Le test du Chi-Carré calculé (16.66) est supérieur à la valeur critique de 3.841. Par conséquent, nous *rejetons l'hypothèse nulle* et concluons qu'il existe une association statistiquement significative entre le statut de fumeur et le cancer du poumon ($p < 0.05$).

Exercices de révision

En utilisant les ensembles de données suivants :

1. Calculez la moyenne pour les deux groupes.
2. Calculez les mesures de variation et déterminez si les moyennes sont représentatives.
3. Calculez l'intervalle de confiance pour $\alpha = 0.05$.
4. Comparez les moyennes en utilisant un test t et formulez des conclusions concernant la signification statistique de la différence entre les moyennes.

Hypothèses :

H_0 : La différence entre les moyennes des échantillons n'est pas statistiquement significative.

H_1 : La différence entre les moyennes des échantillons est statistiquement significative.

$p > 0.05 \Rightarrow H_0$ acceptée

$p < 0.05 \Rightarrow H_0$ rejetée

Ensemble de données 1: Résultats des niveaux de cholestérol sanguin pour deux échantillons indépendants (n=15 chacun):

No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Group 1	168	258	228	247	156	172	165	210	264	220	258	200	195	245	189
Group 2	136	148	125	121	157	148	116	140	161	122	128	122	137	139	128

Ensemble de données 2 : Résultats de la pression artérielle systolique (PAS) pour deux échantillons indépendants (n=12 chacun):

No	1	2	3	5	6	7	8	9	10	11	12
Group 1	130	130	120	110	90	120	125	115	135	140	120
Group 2	170	175	160	170	170	190	185	185	170	160	190

Questions de révision

1. Quelles sont les méthodes paramétriques de test d'hypothèses ?
2. Quelles sont les méthodes non paramétriques de test d'hypothèses ?
3. Dans quelles situations serait-il préférable d'appliquer les méthodes paramétriques de test d'hypothèses ?
4. Dans quelles situations serait-il préférable d'appliquer les méthodes non paramétriques de test d'hypothèses ?
5. Précisez les principales considérations concernant les méthodes paramétriques par rapport aux méthodes non paramétriques.
6. Quand devez-vous utiliser le test t pour un échantillon ?
7. Quand devez-vous utiliser le test t pour deux échantillons indépendants ?
8. Quand devez-vous utiliser le test t pour le coefficient de corrélation ?
9. Quand devez-vous utiliser le test du Chi-Carré ?

CHAPITRE 8. INTRODUCTION À LA MÉTHODOLOGIE DE RECHERCHE

Concepts clés

- Du point de vue de *l'application de la recherche,* elle peut être classée en recherche fondamentale et recherche appliquée.
- Du point de vue de *la méthodologie de recherche* utilisée pour répondre à la question de recherche, les recherches peuvent être quantitatives et qualitatives.
- Les conceptions d'études en médecine se divisent en deux catégories : les études *observationnelles*, où les sujets sont simplement observés, et les études *expérimentales*, où l'effet d'une intervention est mesuré.
- Les conceptions d'études *descriptives* ne permettent que de formuler une hypothèse sur une relation potentielle entre un facteur de risque et un résultat.
- Les conceptions d'études *analytiques* permettent d'évaluer ou d'analyser la relation entre un facteur de risque et une maladie, en testant des hypothèses de causalité.
- La *Revue Systématique* constitue une étude secondaire qui offre un résumé exhaustif de la littérature clinique existante.
- Le *biais* se manifeste lorsque la conception ou la réalisation d'une étude engendre des erreurs dans les résultats et les conclusions. Ce biais peut résulter de la sélection inappropriée des sujets ou des méthodes inadéquates de collecte et d'analyse des données.
- Le *biais de rappel* est une erreur systématique qui survient lorsque les participants à une étude sont systématiquement plus ou moins susceptibles de se souvenir des informations concernant l'exposition, en fonction de leur état de santé.

8.1 Définition, caractéristiques et types de recherche

8.1.1 Définition et caractéristiques de la recherche

> *La recherche* est une activité systématique qui utilise des méthodologies scientifiques appropriées pour aborder des problèmes et générer des connaissances nouvelles, applicables à grande échelle.

La recherche implique la collecte, l'analyse et l'interprétation d'informations pour répondre à des questions spécifiques. Elle se distingue par les *caractéristiques* suivantes :

1. *Objectivité :* La recherche doit être dépourvue de biais personnels et autres. Les biais peuvent survenir si l'étude est conçue ou conduite de manière à induire des erreurs dans les résultats et les conclusions. Ces biais peuvent découler d'une sélection inappropriée des sujets ou de méthodes de collecte et d'analyse des données non adaptées.
2. *Validité :* La validité est synonyme de précision dans la recherche (précision des procédures, des instruments de recherche, des tests, etc.). Deux types de validité sont reconnus :
    - *Validité interne :* Les résultats sont valides pour l'échantillon de sujets étudiés.
    - *Validité externe :* Les résultats sont valides pour la population dont l'échantillon a été tiré. Si les résultats peuvent être étendus à d'autres populations, l'étude est considérée comme *généralisable.*
3. *Fiabilité* : La fiabilité, synonyme de reproductibilité ou répétabilité, implique que les résultats peuvent être reproduits et vérifiés par le chercheur initial et d'autres chercheurs.
4. *Comparabilité :* Le processus d'investigation doit être rigoureux et exempt de défauts, permettant aux conclusions et résultats de la

recherche de résister à un examen critique et à la comparaison avec d'autres études.

5. *Approche systématique :* Toutes les procédures d'investigation doivent suivre une séquence logique et ordonnée, garantissant la cohérence et la constance tout au long du processus de recherche.
6. *Pertinence* : Il existe deux types principaux de pertinence :

⇒ *Pertinence scientifique* : L'étude fait progresser notre compréhension d'un concept scientifique spécifique, d'un processus ou d'une maladie, contribuant à l'ensemble des connaissances scientifiques ;

⇒ *Pertinence sociétale :* L'étude apporte des bénéfices directs à la société, tels que l'amélioration de la santé publique, l'information des décisions politiques ou l'amélioration de la qualité de vie grâce à des applications pratiques des résultats de la recherche.

8.1.2 Erreurs aléatoires et biais systématiques dans la recherche

Dans la recherche, les biais peuvent considérablement affecter la validité et la fiabilité des résultats. Comprendre la différence entre les erreurs aléatoires et les biais systématiques est essentiel pour concevoir des études et interpréter les résultats de manière précise.

⇒ **Erreurs aléatoires**

Les erreurs aléatoires sont des variations imprévisibles qui surviennent lors du processus de mesure. Ces erreurs résultent de fluctuations des instruments de mesure ou d'autres facteurs imprévisibles. Bien que les erreurs aléatoires puissent affecter la précision des résultats, elles ne conduisent généralement pas à des biais constants dans une direction spécifique. Les erreurs aléatoires n'affectent pas lourdement les résultats d'une étude.

Exemple : La variabilité des mesures de la pression artérielle peut être due à de légères différences dans la position du brassard ou dans la

posture du sujet pendant les mesures. Chaque mesure peut différer légèrement, mais en moyenne, elles devraient se concentrer autour de la valeur réelle.

⇒ **Biais systématiques**

Les biais systématiques, également appelés erreurs systématiques, sont des inexactitudes constantes et répétables qui se produisent en raison de défauts dans la conception de l'étude, la collecte des données ou les méthodes d'analyse. Ces biais peuvent conduire à des conclusions incorrectes en faussant systématiquement les résultats dans une direction particulière. Les biais systématiques affectent fortement les résultats d'une étude.

En général, on distingue trois types de biais systématiques :

1. **Confusion** : Le biais survient lorsque le principal facteur de risque étudié est mélangé avec une autre variable, ce qui rend difficile l'isolement de l'effet réel du facteur de risque principal.

Exemple : Dans une étude qui examine la relation entre la consommation de café et les maladies cardiaques, si les buveurs de café ont également tendance à fumer davantage, il devient difficile d'isoler l'impact de la consommation de café sur le risque de maladies cardiaques.

2. **Biais de sélection** : C'est une erreur systématique qui survient lorsque les participants ou sujets inclus dans une étude ne sont pas représentatifs de la population cible, ce qui conduit à des résultats biaisés.

Exemples :

- ✓ *Biais de non-réponse* : Si les individus qui ne répondent pas à un sondage diffèrent significativement de ceux qui répondent, les résultats peuvent ne pas être représentatifs de l'ensemble de la population.

- ✓ *Biais d'exclusion* : Si certains groupes sont systématiquement exclus de l'étude, les résultats peuvent ne pas être généralisables. Par exemple, l'exclusion des patients âgés d'un essai clinique pourrait conduire à des résultats inapplicables à la population plus âgée.
- ✓ *Biais lié à la taille de l'échantillon* : Si la taille de l'échantillon est trop réduite ou insuffisamment randomisée, les résultats risquent de ne pas refléter fidèlement les caractéristiques de la population plus large.

3. **Biais d'information :** Cette distorsion se produit lors de la collecte des données, souvent en raison d'erreurs de mesure systématiques ou de classifications incorrectes des sujets.

Exemples :

- ✓ *Biais d'interview* : Si les enquêteurs ne sont pas suffisamment formés ou s'ils influencent involontairement les réponses, les données collectées peuvent être biaisées.
- ✓ *Biais de rappel* : Dans les études rétrospectives, les participants peuvent ne pas se souvenir précisément des événements ou expositions passés, ce qui conduit à des données biaisées. Par exemple, les patients atteints d'une maladie peuvent se souvenir différemment de leur exposition aux facteurs de risque par rapport aux individus en bonne santé.
- ✓ *Biais de déclaration* : Si les participants révèlent sélectivement des informations, comme sous-déclarer des comportements socialement indésirables (par exemple, le tabagisme ou la consommation d'alcool), les résultats de l'étude peuvent être faussés.

8.1.3 Types de recherche

La recherche peut être classifiée selon trois perspectives principales, comme l'indique le *tableau 8.1* :

- Application de la recherche
- Méthodologie de la recherche
- Objectifs de la recherche

Tableau 8.1 Types de recherche par perspectives

Application de la recherche	*Méthodologie de la recherche*	*Objectifs de la recherche*
1. pure	1. quantitative	1. historique
2. appliquée	2. qualitative	2. descriptive
		3. corrélationnelle
		4. expérimentale
		5. exploratoire

Types de recherches selon leur domaine d'application

⇒ *Recherche pure* : Ce type de recherche est axé sur le développement et la validation de théories et d'hypothèses intellectuellement stimulantes. Bien qu'elle ne présente pas d'applications pratiques immédiates, elle contribue à l'avancement des connaissances théoriques.

⇒ *Recherche appliquée* : Cette recherche vise à répondre à des questions pratiques spécifiques. Elle a pour objectif de résoudre des problèmes concrets.

Types de recherches en fonction des méthodologies employées

⇒ *Recherche quantitative* : Cette approche quantifie l'ampleur d'un problème, d'une question ou d'un phénomène par la mesure de variables et l'analyse statistique. La question centrale est : « Combien ? »

⇒ *Recherche qualitative* : Cette approche examine la nature d'un problème, d'une question ou d'un phénomène sans procéder à une quantification. Elle se concentre sur la compréhension du «

comment » et du « pourquoi » d'une situation à travers des données descriptives.

Les approches quantitative et qualitative présentent des forces distinctes et sont souvent complémentaires.

Types de recherches selon les objectifs poursuivis

⇒ *Recherche historique :* Cherche à tirer des conclusions sur des événements passés, des tendances, des causes ou des effets. Elle implique souvent l'analyse de sources primaires ou l'interview de témoins oculaires. Ce type de recherche aide à comprendre les événements actuels et à prédire les tendances futures.

⇒ *Recherche descriptive :* Décrit systématiquement une situation, un problème ou un phénomène. Elle se concentre sur la collecte de données pour répondre à des questions sur l'état actuel du sujet. Les méthodes incluent des questionnaires, des interviews ou des observations pour recueillir des informations sur les conditions présentes.

⇒ *Recherche corrélationnelle :* Cherche à identifier et mesurer les relations entre deux ou plusieurs variables. Elle va au-delà de la simple description pour explorer comment les variables sont liées et peut être utilisée pour faire des prédictions ou tester des hypothèses. Ce type de recherche est souvent utilisé pour valider des outils et instruments prédictifs.

⇒ *Recherche expérimentale :* Établit la causalité en manipulant activement les variables et en observant les effets. Contrairement à la recherche corrélationnelle, la recherche expérimentale implique l'intervention du chercheur pour déterminer les relations de cause à effet.

⇒ *Recherche exploratoire :* Réalisée lorsque peu d'informations sont disponibles sur un sujet, elle vise à évaluer la faisabilité d'une étude

ou à identifier des domaines pour des recherches ultérieures. Souvent utilisée comme étude préliminaire ou étude pilote pour explorer de nouvelles idées et recueillir des données préliminaires.

8.2 Les étapes du processus de recherche

Dans le cadre de la réalisation de la recherche, deux décisions fondamentales doivent être prises :

1. Ce que vous souhaitez découvrir.
2. La manière dont vous allez procéder pour effectuer cette découverte.

Pour obtenir des réponses à vos questions de recherche, il est nécessaire de suivre une série d'étapes pratiques. L'approche adoptée pour répondre à ces questions est appelée méthodologie de recherche. À chaque étape du processus de recherche, vous sélectionnerez parmi une variété de méthodes et de techniques conçues pour optimiser l'atteinte de vos objectifs de recherche. Pour garantir un processus de recherche systématique et efficace, suivez les étapes suivantes :

1. **Définir le problème de recherche** : Formulez clairement le problème ou la question que vous souhaitez explorer. Cela implique d'identifier les lacunes dans la recherche existante et de préciser ce que vous visez à étudier.
2. **Réaliser une revue de littérature** : Analysez les recherches et les publications existantes en rapport avec votre sujet. Cela permet de comprendre l'état actuel des connaissances, de raffiner le problème de recherche et d'identifier les lacunes que votre étude pourrait combler.
3. **Formuler le but principal et les objectifs** : Définissez clairement le but principal de votre recherche et décomposez-le en objectifs spécifiques et mesurables. Ces objectifs guideront le

développement de votre conception de recherche et de votre méthodologie.

4. **Élaborer le plan de recherche** : Concevez un plan détaillé qui décrit la méthodologie de recherche, y compris la conception de l'étude, les méthodes d'échantillonnage, les procédures de collecte des données et les techniques d'analyse.
5. **Collecter les données** : Rassemblez les données nécessaires en utilisant les méthodes choisies. Cela peut inclure des enquêtes, des entretiens, des expériences ou l'analyse de données secondaires, en fonction de la conception de votre recherche.
6. **Analyser les données** : Traitez et analysez les données collectées. Employez des techniques statistiques appropriées en fonction des objectifs de recherche et des types de données.
7. **Généraliser et interpréter les résultats** : Formulez des conclusions sur la base de l'analyse des données. Interprétez les résultats par rapport aux questions de recherche et proposez des recommandations fondées sur ces résultats.
8. **Présenter les résultats** : Communiquez les résultats de votre recherche à travers un rapport bien structuré et, le cas échéant, une présentation orale.

8.3 Formulation du problème de recherche

La formulation du problème de recherche constitue la première étape fondamentale du processus de recherche. Un problème de recherche bien défini offre une direction et une concentration pour l'ensemble de l'étude. Lors de la sélection d'un problème de recherche, il est essentiel de considérer les facteurs suivants :

⇒ *Intérêt* : Optez pour un sujet qui suscite un véritable intérêt. La passion pour le sujet aidera à maintenir la motivation tout au long du processus de recherche.

⇒ *Magnitude* : Veillez à ce que le sujet soit abordable dans le cadre des ressources et du temps disponibles.

⇒ *Pertinence* : La recherche doit apporter une contribution significative au corpus existant de connaissances.

⇒ *Disponibilité des données* : Vérifiez que les données nécessaires sont accessibles ou peuvent être collectées.

⇒ *Considérations éthiques* : Assurez-vous que la recherche respecte les normes éthiques établies.

Le processus de formulation d'un problème de recherche comporte plusieurs étapes clés. Une compréhension approfondie du domaine général est cruciale pour définir clairement et efficacement le problème de recherche. Suivez ces étapes pour formuler un problème de recherche robuste :

1. *Identifiez un domaine d'intérêt général* : Débutez en sélectionnant une zone générale correspondant à vos intérêts et à votre expertise.
2. *Décomposez le domaine général en sous-domaines* : Divisez le domaine général en sous-domaines plus spécifiques pour affiner l'axe de recherche.

3. *Sélectionnez les sous-domaines d'intérêt* : Choisissez les sous-domaines les plus pertinents et captivants pour vos objectifs de recherche.
4. *Formulez des questions de recherche* : Élaborez des questions précises auxquelles votre recherche doit répondre. Ces questions orienteront la direction de votre étude.
5. *Établissez des hypothèses* : En fonction des questions de recherche, développez des hypothèses proposant des réponses ou des explications possibles.
6. *Vérifiez* : Assurez-vous que le problème de recherche est clairement défini, réalisable et en adéquation avec vos objectifs de recherche.

8.4 Revue de la littérature

La revue de la littérature constitue une étape initiale fondamentale dans le processus de recherche, essentielle pour appréhender le corpus de connaissances existantes dans votre domaine d'intérêt. Cette étape fournit non seulement le contexte nécessaire à votre recherche, mais contribue également de manière significative à l'affinement et à l'orientation de chaque phase de votre étude.

Fonctions d'une revue de littérature :

1. Clarifiez et précisez le problème de recherche.
2. Améliorez la méthodologie de l'étude.
3. Élargissez le champ des connaissances.
4. Contextualisez les résultats obtenus.

Procédures pour réaliser une revue de littérature :

⇒ *Recherchez la littérature existante* : Identifiez et rassemblez la littérature pertinente en lien avec votre domaine de recherche. Utilisez les bases de données académiques et les bibliothèques pour localiser les sources.

⇒ *Examinez la littérature sélectionnée* : Analysez et synthétisez de manière critique la littérature afin de comprendre l'état actuel des recherches sur votre sujet.

⇒ *Développez un cadre théorique* : Élaborer un cadre théorique qui définit les théories et concepts devant guider votre recherche.

⇒ *Développez un cadre pratique* : Concevez un cadre pratique intégrant les méthodologies, techniques et approches pertinentes pour votre étude.

⇒ *Organisez et documentez les sources*: Cataloguez systématiquement la littérature examinée.

Outils et ressources :

Pour effectuer une recherche bibliographique efficace, commencez par une compréhension claire de votre sujet de recherche et définissez les paramètres de votre recherche. Utilisez les bases de données et ressources en ligne suivantes :

- MedLine https://www.nlm.nih.gov/medline/medline_overview.html
- Research4Life https://portal.research4life.org/
- HINARI https://www.emro.who.int/information-resources/hinari/hinari.html
- PubMed https://pubmed.ncbi.nlm.nih.gov/
- Et bien d'autres…

Pour un article académique, il est recommandé d'utiliser des livres, des articles ainsi que des sites web spécialisés qui recueillent des informations pertinentes sur votre sujet. Dans la rédaction académique, il est essentiel de recourir à une variété de sources, incluant des ouvrages, des articles scientifiques et des sites web fiables. Au cours de

votre revue, prenez des notes détaillées sur les méthodologies et les instruments utilisés dans les recherches antérieures. Cela facilitera le choix des approches appropriées pour votre propre étude, telles que les techniques d'échantillonnage, les méthodes de collecte de données et les procédures d'analyse. Veillez à citer vos sources selon un format standardisé. La bibliographie de votre revue de littérature doit constituer une liste complète et claire de toutes les sources citées, organisée par ordre alphabétique des noms des auteurs.

Les trois styles de citation les plus fréquemment employés sont :

⇒ ***Système Harvard :*** Un style de citation « auteur-date ».
- Citation dans le texte : (Smith, 2020)
- Liste de références : Smith, J. (2020). *Understanding Modern Research*. Oxford University Press.

⇒ ***Système Vancouver :*** Un style de citation numérique introduit au Canada en 1978.
- Citation dans le texte : (1)
- Liste de références : Smith, J. *Understanding Modern Research*. Oxford University Press; 2020.

⇒ ***Systèmes lettre-numéro :*** Un système de citation hybride alliant éléments des systèmes numériques et alphabétiques.
- Citation dans le texte : (Smith 2020a)
- Liste de références : Smith J. *Understanding Modern Research*. Oxford University Press; 2020a.

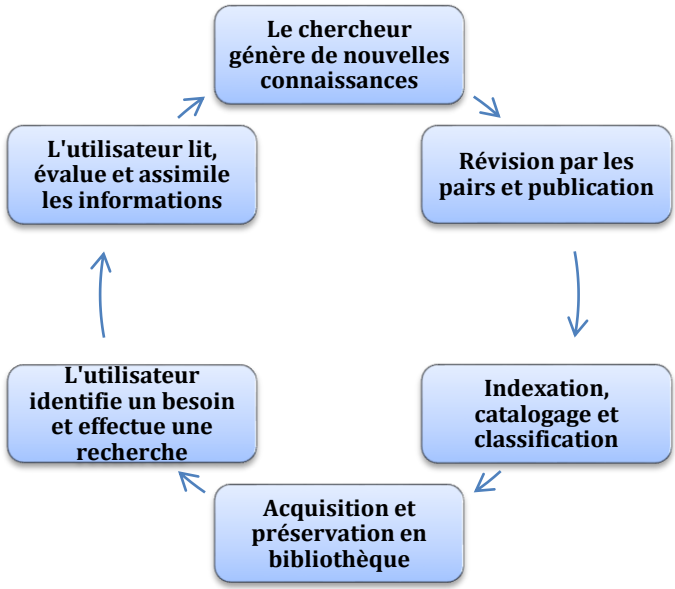

Figure 8.1 Cycle de l'information

Le diagramme 8.1 illustre comment l'information évolue à travers différentes phases pour devenir une partie intégrante des connaissances publiées. Ces connaissances deviennent alors accessibles aux chercheurs pour être approfondies, inspirer de nouveaux domaines d'investigation et générer de nouvelles connaissances.

8.5 Formulation du but et des objectifs de l'étude

Le but représente l'intention générale que l'étude se fixe, agissant comme une déclaration globale de son intention. Les objectifs, en revanche, sont les tâches spécifiques nécessaires pour atteindre ce but.

Il est essentiel de formuler ces objectifs de manière claire et précise. Les objectifs doivent être énoncés de manière ordonnée, chaque objectif abordant un seul aspect de l'étude. Lors de la formulation des

objectifs, il est recommandé d'utiliser des verbes d'action. Ainsi, les objectifs devraient commencer par des termes tels que : déterminer, découvrir, établir, mesurer, explorer, etc.

8.6 Préparation de la conception de recherche et collecte des données

8.6.1 Définition et étapes de la conception de recherche

> *La conception (ou le design) de recherche* constitue le cadre conceptuel dans lequel la recherche est menée.

Elle a pour but d'assurer la collecte d'informations pertinentes avec un minimum d'effort, de temps et de dépenses. La préparation d'une conception de recherche adéquat pour un problème spécifique implique les étapes suivantes :

- ⇒ Déterminer *le design de l'échantillon*
- ⇒ Élaborer les outils pour la collecte des données
- ⇒ Adopter *une conception d'étude*

8.6.3 Détermination du design de l'échantillon

Les chercheurs tirent généralement des conclusions sur une population en étudiant un échantillon. La conception de l'échantillon repose sur trois décisions clés :

1. Qui sera inclus dans l'enquête ? (*L'échantillon*)
2. Combien de personnes seront incluses dans l'enquête ? (*Taille de l'échantillon*)
3. Comment l'échantillon doit-il être sélectionné ? (*Type d'échantillonnage*)

8.6.4 Outil de collecte de données

La construction d'un outil de recherche pour la collecte de données constitue un aspect fondamental d'un protocole de recherche. Les

résultats et les conclusions s'appuient sur les données collectées, lesquelles dépendent directement des questions formulées dans le questionnaire.

Directives pour l'élaboration d'un questionnaire :

Étape I : Définir avec précision et énumérer tous les objectifs spécifiques ou les questions de recherche pour l'étude.

Étape II : Pour chaque objectif ou question de recherche, répertorier toutes les questions associées que vous souhaitez examiner dans le cadre de l'étude.

Étape III : Pour chaque question de recherche listée à l'étape I et chaque objectif énoncé à l'étape II, identifier les informations nécessaires pour y répondre.

Étape IV : Formuler des questions visant à obtenir ces informations.

Un questionnaire se compose d'un ensemble de questions adressées aux répondants. Bien qu'il existe diverses méthodes pour poser des questions, il est essentiel d'assurer la clarté et d'obtenir les informations requises. Le questionnaire doit être soigneusement élaboré et testé avant d'être déployé à grande échelle.

Types de structure des questions :

- ⇒ Fermées
- ⇒ Ouvertes
- ⇒ Combinaison des deux

Questions fermées : Ces questions englobent toutes les réponses possibles au sein de catégories préétablies, parmi lesquelles les répondants sélectionnent (par exemple, questions à choix multiples, questions à échelle). Les questions fermées sont employées pour générer des statistiques dans les recherches quantitatives. Leur principal avantage

réside dans la facilité avec laquelle les réponses peuvent être analysées et rapportées.

Questions ouvertes : Ces questions permettent aux répondants de s'exprimer avec leurs propres mots. L'avantage majeur est la possibilité de capturer les pensées des répondants dans leur propre langage. Cependant, l'analyse des données peut s'avérer plus complexe.

Questions combinées: Cette structure débute par une série de questions fermées et se conclut par des questions ouvertes pour obtenir des réponses plus détaillées.

La majorité des enquêtes utilisent des questionnaires auto-administrés – en personne, par courrier, e-mail ou entretiens – à nouveau en personne ou par téléphone. Chaque méthode présente ses avantages et ses inconvénients, dont certains sont illustrés dans le *Tableau 8.3.*

Tableau 8.2 Questions ouvertes versus questions fermées

Critères	Questions ouvertes	Questions fermées
Objectif	Capturer des mots ou des citations réels	Réponses les plus courantes
Répondants	Fournissent des réponses approfondies et détaillées	Préfèrent des réponses rapides et faciles
Contexte des questions	Les choix sont inconnus	Les choix peuvent être anticipés
Analyse	Analyse de contenu ; chronophage	Comptage ou notation
Rapport	Réponses individuelles ou groupées	Données statistiques

Source D'après Dawson and Trapp, 2004

Tableau 8.3 Avantages et inconvénients des différentes méthodes d'enquête

	Auto-administré par courrier/email	Auto-administré en personne	Entretien par téléphone	Entretien en personne
Coût	++	+	-	-
Temps	++	+	-	-
Standardisation	+	+	+/-	+/-
Profondeur/détail	-	-	+	++
Taux de réponse	-	++	+	++
Réponses manquantes	-	+	++	++
+ Avantages ; - Inconvénients ; +/- Neutre				

Source : Dawson and Trapp, 2004.

Structure générale du questionnaire:

⇒ Titre
⇒ Instructions
⇒ Informations générales sur le répondant
⇒ Questions
⇒ Note de remerciement

8.6.5 Classification des types de conception d'étude

Il existe plusieurs schémas visant à classifier les méthodes de conception des études. L'un de ces schémas est illustré dans le *Tableau 8.4*. Cette classification divise les conceptions d'étude en trois grandes catégories : les études observationnelles (où les participants sont observés sans aucune intervention), les études expérimentales (où une intervention spécifique est mise en œuvre) et les études secondaires (où les études primaires sont préalablement évaluées ou filtrées).

Tableau 8.4 Classification des types de conception d'étude

CONCEPTION	MÉTHODE	TYPES
I. Études primaires :		
– Études observationnelles	Descriptives	– Séries de cas/rapport – Études transversales – Études écologiques (au niveau de la population)
	Analytiques	– Étude cas-témoins – Étude de cohorte
– Études expérimentales	Analytiques	– Essai clinique – Essai clinique au niveau de la population
II. Études secondaires :		
– Études préalablement évaluées ou filtrées	Quantitatives Qualitatives	– Revues systématiques – Revues narratives – Méta-analyses

Aperçu des conceptions d'étude

Les conceptions d'étude sont des méthodes utilisées pour recueillir des informations et évaluer la relation entre les facteurs de risque (expositions ou causes) et les maladies (résultats ou effets). L'objectif principal est de déterminer la causalité, c'est-à-dire la relation entre l'exposition et le résultat ou entre la cause et l'effet. Le choix de la conception d'étude est influencé par les questions de recherche, les

préoccupations relatives à la validité et les considérations pratiques ou éthiques.

⇒ ***Études observationnelles*** : Ces études fournissent des informations sur les expositions dans des contextes naturels, en évitant les problèmes éthiques associés aux études expérimentales. Elles comprennent:

- **Études de série de cas** : Documentent une série de cas présentant une condition similaire, sans inclure de groupe de comparaison.
- **Études transversales** : Évaluent les données à un instant précis pour offrir un aperçu de la relation entre les facteurs de risque et les résultats.
- **Études cas-témoins** : Comparent les individus atteints d'une maladie (cas) avec ceux non atteints (témoins) afin d'identifier les différences d'exposition.
- **Études de cohorte** : Suivent un groupe exposé à un facteur de risque au fil du temps pour observer les résultats sur la santé.

Les études de séries de cas et les études transversales sont *descriptives* ; elles émettent des hypothèses sur les relations entre les facteurs de risque et les résultats sans établir de lien causal. À l'inverse, les études cas-témoins et les études de cohorte sont *analytiques* ; elles testent des hypothèses pour établir la causalité. Les études de cohorte, qui sont *longitudinales*, collectent des données sur les facteurs de risque et les résultats à plusieurs moments dans le temps, tandis que les études *transversales* recueillent des données à un moment unique. Les études de cohorte sont *prospectives*, suivant l'évolution des facteurs de risque vers les résultats, tandis que les études cas-témoins sont *rétrospectives*, traçant le chemin des résultats vers les facteurs de risque. Les études

transversales n'ont pas de direction temporelle, car les données sont collectées en un seul point (*Figure 8.2)*.

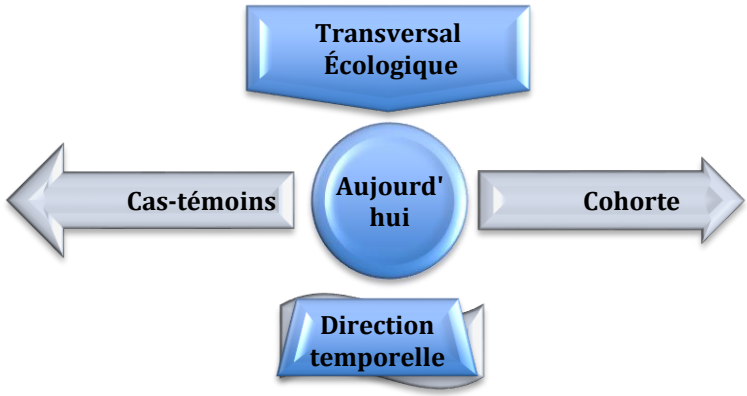

Figure 8.2 Relation temporelle entre les différents types de conception d'études observationnelles

⇒ ***Études expérimentales (essais cliniques)*** : Ces études impliquent des interventions telles que des médicaments ou des traitements, et sont conçues pour évaluer leurs effets.

⇒ ***Revue systématique*** : Ce type d'étude secondaire évalue et synthétise de manière critique la littérature clinique sur un sujet précis en suivant une méthodologie structurée. Les chercheurs localisent, rassemblent et évaluent systématiquement les études primaires pertinentes selon des critères prédéfinis. L'objectif est de répondre à une question de recherche clairement définie en utilisant à la fois des méthodes qualitatives et quantitatives.

⇒ ***Revues narratives*** : Ces revues ont un champ d'application plus vaste et peuvent ne pas respecter des critères rigoureux pour la sélection ou l'évaluation des articles. Elles manquent souvent de critères

explicites d'inclusion et peuvent négliger l'évaluation de la validité des études examinées, ce qui peut introduire un biais.

⇒ ***Méta-analyse*** : Il s'agit d'un type d'étude secondaire qui combine quantitativement les résultats de plusieurs études primaires pour offrir un résumé exhaustif et des conclusions. Les méta-analyses sont souvent orientées vers l'évaluation de l'efficacité thérapeutique ou la planification de recherches complémentaires.

8.7 Force des preuves dans la conception des études

La pyramide des preuves hiérarchise les conceptions d'études en fonction de leur solidité, avec les études de la plus haute qualité en haut et celles fournissant des preuves plus faibles en bas. Comme le montre la *Figure 8.3*, les études fournissant des preuves robustes sont relativement rares.

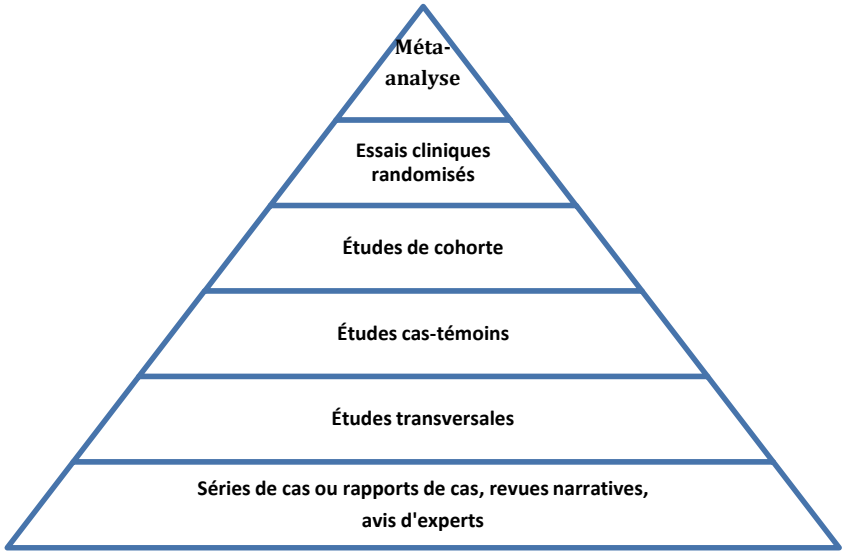

Figure 8.3 Pyramide de la force des preuves selon les designs d'étude

Les études secondaires, telles que les *revues systémiques* et les *méta-analyses*, offrent la validité interne la plus élevée. Ces études synthétisent les résultats de multiples études primaires pour fournir des preuves solides de causalité.

Les études expérimentales, telles que les essais contrôlés randomisés (ECR), sont particulièrement efficaces pour établir des relations causales à travers des interventions directes. Elles sont plus contrôlées que les études observationnelles, mais leur fréquence est réduite en raison des défis pratiques et éthiques associés.

Les études de cohorte suivent des groupes exposés à des facteurs de risque au fil du temps pour observer les résultats de santé, fournissant ainsi des preuves substantielles de causalité.

Les études cas-témoins comparent des individus présentant une maladie à ceux qui n'en présentent pas, afin d'identifier les différences d'exposition. Bien que précieuses, les études cas-témoins sont considérées comme moins solides que les études de cohorte en raison de leur nature rétrospective.

Les études transversales examinent les données d'un groupe de sujets à un moment donné, offrant une vue d'ensemble des associations sans établir de causalité.

Les études de série de cas, *les revues narratives* et *les opinions d'experts* sont positionnées en bas de la pyramide. Ces études, étant moins contrôlées, fournissent des preuves moins robustes de causalité.

Exercices de révision

1. Un gestionnaire de clinique souhaite mener une enquête auprès d'un échantillon aléatoire de patients afin de recueillir leur avis sur les changements récents apportés au fonctionnement de la clinique. Le gestionnaire a élaboré un questionnaire et demande votre révision. L'une des questions posées est : « Êtes-vous d'accord pour affirmer que les nouvelles heures de fonctionnement de la clinique représentent une amélioration par rapport aux anciennes ? »
    - Quel conseil donneriez-vous au gestionnaire concernant la formulation de cette question ? Justifiez votre choix.
2. Supposons que vous souhaitiez déterminer jusqu'à quelle distance les médecins sont prêts à se déplacer pour assister à des cours de formation continue, en considérant qu'un certain nombre d'heures est requis chaque année. Par ailleurs, vous souhaitez connaître les thèmes qu'ils aimeraient voir inclus dans les programmes futurs. Comment sélectionneriez-vous l'échantillon de médecins à inclure dans votre enquête ? Justifiez votre choix.
    a) Tous les médecins ayant assisté aux programmes de l'année précédente
    b) Tous les médecins participant aux deux programmes à venir
    c) Un échantillon aléatoire de médecins ayant assisté aux programmes de l'année précédente
    d) Un échantillon aléatoire de médecins obtenu à partir d'une liste maintenue par la société médicale de l'État
    e) Un échantillon aléatoire de médecins dans chaque district, obtenu à partir d'une liste maintenue par les sociétés médicales des districts

Questions de révision

1. La définition et les caractéristiques de la recherche.
2. Le rôle de la validité dans le processus de recherche.
3. Les caractéristiques de la recherche : leur signification. Fournissez un exemple pour chacune d'elles.
4. La classification des types de recherche.
5. La classification des types de recherche selon leur application : types et signification.
6. La classification des types de recherche en fonction des objectifs : types et signification.
7. Les étapes du processus de recherche : leur contenu et spécificités.
8. La formulation du problème de recherche : fonction principale et critères de sélection.
9. Les étapes de la formulation d'un problème de recherche : contenu et spécificités.
10. La revue de la littérature : fonctions, procédures et systèmes de citation des références.
11. Les objectifs et buts : définition et règles de formulation.
12. La définition et les étapes du design de recherche.
13. Les étapes de la construction d'un questionnaire.
14. Les types fondamentaux de structure des questions. Leur contenu.
15. Les types de méthodes d'enquête. Leur contenu.
16. La classification de la conception d'étude.
17. La conception d'étude observationnelle versus expérimentale : signification et spécificités. Avantages et inconvénients.
18. Quel type de conception d'étude est le plus pertinent en fonction des questions de recherche ?

a) Question de thérapie
b) Diagnostic/dépistage
c) Pronostic
d) Occurrence
e) Causalité

19. Énoncez la différence principale entre les conceptions d'étude suivantes : observationnelle descriptive et observationnelle analytique.
20. Énoncez la différence principale entre les conceptions d'étude suivantes : étude cas-témoins et étude de cohorte.
21. Énoncez la différence principale entre les conceptions d'étude suivantes : séries de cas et étude cas-témoins.
22. Classez les conceptions d'étude en fonction de la robustesse des preuves.

CHAPITRE 9. ÉTUDES OBSERVATIONNELLES DESCRIPTIVES

Concepts Clés

❖ *Les études de séries* de cas et les études transversales appartiennent à la catégorie des études observationnelles et descriptives.

❖ Une étude transversale est aussi désignée sous les termes *d'étude de prévalence* ou *enquête communautaire*.

❖ *Une étude transversale* collecte des informations sur l'exposition et les résultats à un seul moment, sans suivi temporel.

❖ Chaque type de conception d'étude présente des avantages et des inconvénients spécifiques.

9.1 Étude de séries de cas / rapport de cas

Une étude de séries de cas ou un rapport de cas est le design d'étude le plus simple, où l'auteur décrit des observations intéressantes ou inhabituelles survenues chez un petit nombre de patients (étude de séries de cas) ou même chez un seul patient présentant une condition rare (rapport de cas). Lorsque certaines caractéristiques d'un groupe (ou série) de patients (ou cas) sont décrites dans un rapport publié, on parle d'étude de séries de cas. Ce type de conception d'étude génère souvent des hypothèses qui peuvent ensuite être étudiées dans des études cas-témoins ou des études de cohorte.

Utilisations :

⇒ Identification de nouvelles pathologies ou résultats.

⇒ Élaboration d'hypothèses.

Avantages :

1. Facile à rédiger.
2. Les observations peuvent être extrêmement utiles pour d'autres chercheurs.

Inconvénients :

1. Sensible à de nombreux biais.
2. Non approprié pour des décisions définitives.

9.2 Étude transversale

Une étude transversale, également appelée étude transversale, analyse les données collectées auprès d'un groupe de sujets à un moment donné plutôt que sur une période. Elle est conçue pour répondre à la question : « Que se passe-t-il actuellement ? » Les sujets sont sélectionnés, et les informations concernant l'exposition aux facteurs de risque et les résultats (maladies) sont obtenues sur une période courte, comme illustré dans la *Figure 9.1*.

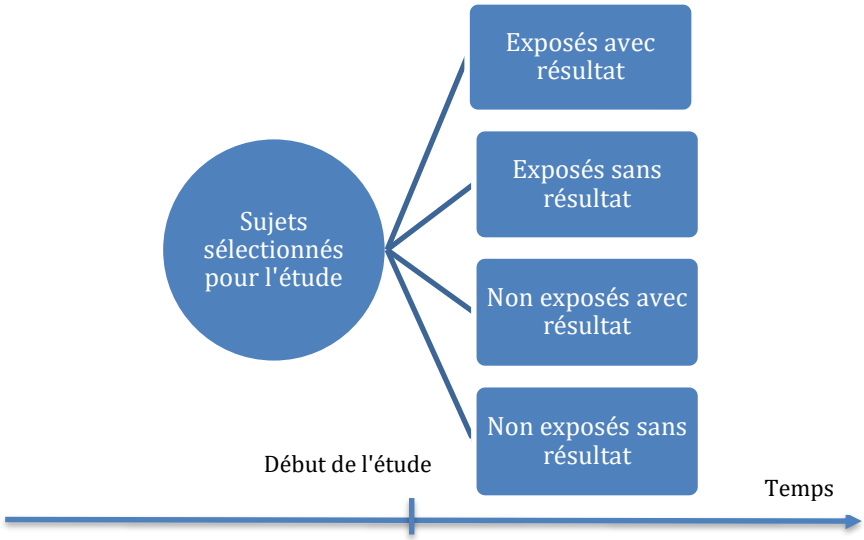

Figure 9.1 Diagramme de flux de la conception de l'étude transversale

Une étude transversale est utilisée pour mesurer la prévalence d'une maladie et examiner les facteurs de risque ou les causes potentiels. Les études transversales, qui analysent la relation entre l'exposition et la prévalence de la maladie dans une population définie à un moment donné, sont également connues sous le nom d'études de *prévalence*. Les *enquêtes communautaires* sont généralement des études transversales, bien que celles-ci puissent aussi faire partie d'études de cohorte ou cas-témoins.

La conception d'une étude transversale est optimal pour le diagnostic/dépistage, l'occurrence d'une maladie (prévalence), les enquêtes ou pour la formulation de questions de recherche.

Procédures statistiques pour l'analyse des données d'une étude transversale :

- Calcul des proportions, des taux (y compris ajustés) et des rapports.
- Calcul des intervalles de confiance pour les proportions ou les moyennes.
- Analyses de corrélation et de régression, y compris la régression logistique.
- Application de tests paramétriques tels que le test t, l'analyse de variance, le test du chi carré et autres tests non paramétriques.

Avantages :

1. Utile pour évaluer l'impact d'une maladie dans un groupe – permet de déterminer le taux de prévalence.
2. Économique et rapide.
3. Précieux pour l'évaluation des procédures diagnostiques.
4. Aide à l'étude des facteurs de risque communs.
5. Utile pour l'analyse des maladies fréquentes.

Inconvénients :

1. Les participants peuvent être moins disposés à collaborer.
2. Ne montre pas la séquence des événements.
3. Montre une association entre les facteurs de risque et la maladie étudiée, sans indiquer de causalité.
4. Pas utile pour identifier les causes des résultats (maladies).
5. Principalement utile pour l'étude des maladies chroniques.
6. Les facteurs de confusion peuvent être distribués de manière inégale.
7. Les tailles des groupes peuvent être inégales.
8. Biais de rappel.

Questions de révision

1. Définissez une étude de série de cas ainsi que son contenu.
2. Quels sont les avantages d'une étude de série de cas ?
3. Quels sont les inconvénients d'une étude de série de cas ?
4. Quelle analyse statistique est appropriée pour une étude de série de cas ?
5. Définissez une étude transversale et donnez ses synonymes.
6. Décrivez le contenu du diagramme pour une étude transversale.
7. Quels sont les avantages d'une étude transversale ?
8. Quels sont les inconvénients d'une étude transversale ?
9. Quelle analyse statistique est adéquate pour une étude transversale ?

CHAPITRE 10. ÉTUDES OBSERVATIONNELLES ANALYTIQUES

Concepts clés

- Une étude cas-témoins inclut un groupe de cas (ceux avec un résultat) et un groupe de témoins (ceux sans résultat).
- Dans une étude cas-témoins, les informations sur l'exposition sont recueillies de manière rétrospective.
- Bien qu'une étude cas-témoins fournisse des informations sur la causalité, le biais de rappel est un problème fréquent.
- Le jumelage dans une étude cas-témoins réduit l'influence des facteurs de confusion.
- Le rapport de cotes montre combien de fois un cas est plus susceptible d'avoir été exposé à un facteur de risque par rapport à un témoin.
- Une cohorte est un groupe de personnes qui n'ont pas la maladie d'intérêt, sélectionnées puis observées pendant une période prolongée.
- Dans une étude de cohorte, le chercheur recueille des informations sur les nouveaux cas de maladie (incidence) lors d'examens réguliers futurs (prospectifs).
- Dans une étude longitudinale, deux ou plusieurs séries d'observations sont recueillies pour chaque sujet au fil du temps.
- Une étude de cohorte est également appelée étude prospective, longitudinale, de suivi ou d'incidence.
- Le risque relatif montre combien de fois une personne exposée est plus susceptible de contracter la maladie par rapport à une personne non exposée.
- Le risque attribuable indique la proportion de l'incidence de la maladie qui peut être attribuée à une exposition spécifique (parmi les personnes exposées).
- Dans une étude cas-témoins, nous connaissons le résultat et recherchons l'exposition dans le passé, rétrospectivement.

- ❖ Dans une étude de cohorte, nous connaissons l'exposition et suivons les sujets sur une période donnée pour observer le résultat à venir (dans le futur, prospectivement).

10.1 Étude cas-témoins

L'étude cas-témoins constitue un design de recherche rétrospectif dans lequel les informations sur les facteurs de risque sont recueillies en examinant l'historique des participants. Bien qu'une étude cas-témoins puisse offrir des aperçus précieux sur la causalité, le biais de rappel demeure un problème récurrent.

Conception de l'étude

Cas : Individus présentant une maladie ou un résultat spécifique.
Témoins : Individus ne présentant pas la maladie ou le résultat en question.

Les chercheurs doivent recourir à l'appariement pour associer les témoins aux cas en fonction de caractéristiques telles que l'âge et le sexe. Les cas et les témoins doivent être comparables à l'exception de leur exposition au facteur de risque étudié, ce qui permet de réduire l'influence des variables confondantes potentielles.

Les participants sont sélectionnés sur la base de leur statut de maladie (variable dépendante), et les deux groupes sont interrogés sur leur exposition aux facteurs de risque potentiels (variable indépendante). Les études cas-témoins sont principalement utilisées pour explorer les causes potentielles des maladies (question de recherche causale).

Méthodes de collecte des données

- ⇒ Documents disponibles des hôpitaux, statistiques vitales et autres registres.
- ⇒ Interviews.

⇒ Questionnaires auto-administrés.

⇒ Mesures directes.

Les historiques d'exposition des cas et des témoins sont ensuite comparés. Les études cas-témoins visent à répondre à la question : « Que s'est-il passé ? ». En connaissant le résultat, les chercheurs investiguent les expositions passées de manière rétrospective, comme le montre la *Figure 10.1*

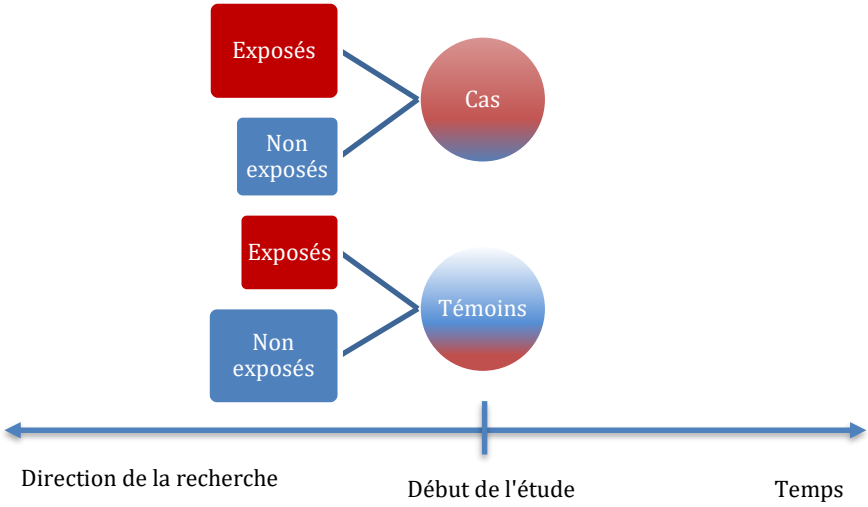

Figure 10.1 Diagramme de flux de la conception d'étude cas-témoins

Analyse des données

L'analyse d'une étude cas-témoin implique le calcul d'une mesure d'association appelée le Rapport de Cotes (*Odds Ratio*). Les cotes sont définies comme la probabilité qu'un événement se produise divisée par la probabilité que le même événement ne se produise pas : p / (1-p). Le Rapport de Cotes se calcule ainsi :

$$Rapport\ de\ Cotes = \frac{\text{Cotes qu'un cas ait été exposé au facteur de risque}}{\text{Cotes qu'un témoin ait été exposé au facteur de risque}}$$

Le Rapport de Cotes permet d'évaluer le risque dans les études cas-témoins, montrant combien de fois un cas est plus susceptible d'avoir été exposé à un facteur de risque comparé à un témoin. Il est facile à calculer lorsque les données sont présentées dans un tableau de contingence 2x2. Un tableau de contingence est un type spécial de tableau de distribution de fréquences où deux variables sont montrées simultanément (*Tableau 10.1*).

Tableau 10.1 2x2 Tableau de contingence 2x2 pour une étude cas-témoin

Facteur d'exposition	**Résultat (maladie)**		**Total**
	Cas	Témoins	
Exposés	a	b	a + b
Non exposés	c	d	c + d
Total	a + c	b + d	a + b + c + d

$$\text{Rapport de Cote (OR)} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Le Rapport de Cotes est également connu sous le nom de rapport des produits croisés, car il peut être défini comme le rapport du produit des diagonales dans un tableau 2x2.

Interprétation du rapport de cotes

⇒ OR = 1 : Aucune association (aucune différence d'exposition entre les cas et les témoins).

⇒ OR > 1 : Exposition dangereuse.

⇒ OR < 1 : Exposition bénéfique.

Test de signification

Le Rapport de Cotes peut être testé pour sa signification en utilisant le calcul d'un intervalle de confiance, qui est une plage de valeurs susceptibles de contenir un OR de population avec un certain niveau de confiance. Si l'intervalle de confiance à 95 % pour l'OR inclut 1, les résultats ne sont pas statistiquement significatifs.

Exemple d'interprétation de l'OR

L'interprétation d'un Rapport de Cote (OR) statistiquement significatif est la suivante :

> Si OR = 3,23, les personnes atteintes de la maladie sont 3,23 fois plus susceptibles d'avoir été exposées par rapport à celles qui ne sont pas atteintes de la maladie.

Avantages :

1. Permet d'examiner plusieurs facteurs de risque.
2. Permet d'étudier les effets à long terme de l'exposition sur une courte période.
3. Nécessite un nombre réduit de sujets.
4. Relativement rapide et peu coûteux.
5. Adapté aux maladies rares.

Inconvénients :

1. Risque plus élevé de biais et de confusions en raison de la collecte de données rétrospectives.
2. Difficulté à sélectionner un groupe témoin approprié.
3. Biais de mémoire possible en raison de la nature rétrospective.
4. Incapacité à déterminer l'incidence ou la prévalence.
5. Difficulté à établir une relation temporelle entre l'exposition et le résultat.

10.2 Étude de cohorte

Une étude de cohorte, également connue sous les noms d'étude d'incidence, étude longitudinale ou étude prospective, suit un groupe d'individus sur une période prolongée pour observer le développement de nouveaux cas d'une maladie et les facteurs de risque associés.

Conception de l'étude

Une cohorte est un groupe de personnes partageant une caractéristique commune mais ne présentant pas la maladie d'intérêt au début de l'étude. Elles sont observées au fil du temps pour surveiller l'incidence de nouveaux cas. Les chercheurs collectent des données sur les expositions et suivent la cohorte dans le temps pour observer de nouveaux cas de la maladie (incidence) lors d'examens réguliers futurs (nature prospective). Dans une étude de cohorte, la question posée est : « Que se passera-t-il ? ». Les chercheurs connaissent le statut d'exposition au début et suivent les sujets dans le temps pour observer les résultats, comme illustré dans la *Figure 10.2.*

Les études de cohorte typiques sont généralement prospectives car les résultats de santé sont observés après le début de l'étude. Les études de cohorte utilisent des groupes qui sont similaires à tous égards, sauf en ce qui concerne l'exposition.

Les études de cohorte sont utilisées pour :

⇒ Mesurer l'incidence des maladies.
⇒ Enquêter sur les causes des maladies.
⇒ Déterminer le pronostic.
⇒ Établir le moment et la direction des événements.

Méthodes de collecte des données

Les données dans une étude de cohorte peuvent être obtenues par des interviews personnelles, des examens médicaux et des enquêtes environnementales.

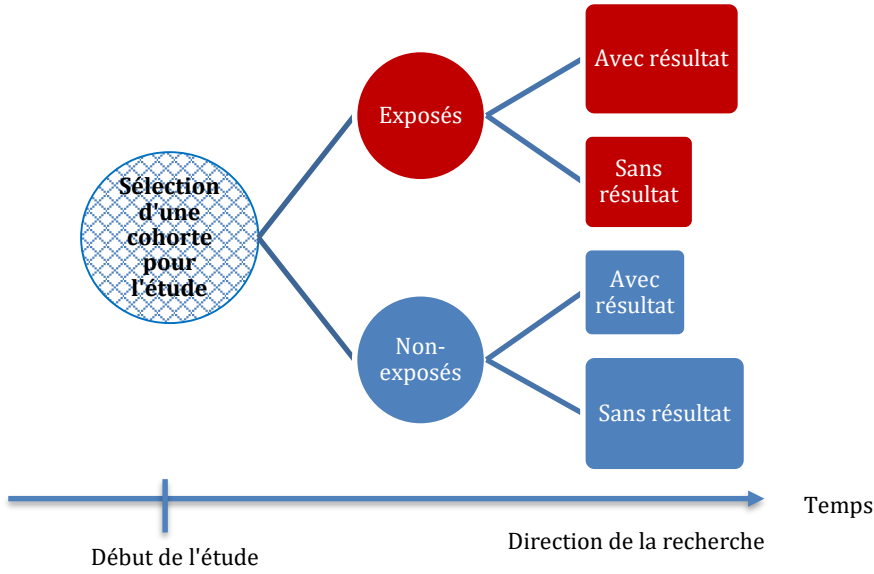

Figure 10.2 Diagramme de flux de la conception d'une étude de cohorte

Analyse des données

L'objectif principal de l'analyse des données des études de cohorte est de comparer l'occurrence des résultats dans les groupes exposés et non exposés.

Les mesures d'association suivantes sont utilisées pour estimer la relation entre un facteur de risque et l'apparition d'un résultat donné :

⇒ Risque relatif (RR)

⇒ Risque attribuable (AR)

Le Risque Relatif (RR) est un rapport entre le taux d'incidence dans le groupe exposé et le taux d'incidence dans le groupe non exposé. Il est facile de calculer le RR pour une étude de cohorte lorsque les données sont présentées dans un tableau 2x2 :

Tableau 10.2 Tableau de contingence 2x2 pour l'étude de cohorte

Facteur d'exposition	Résultat (maladie)		Total
	Oui	Non	
Exposés	a	b	a + b
Non exposés	c	d	c + d
Total	a + c	b + d	a + b + c + d

$$Risque\ relatif\ (RR) = \frac{\text{Incidence des exposés}}{\text{Incidence des non} - \text{exposés}} = \frac{a\ /\ (a+b)}{c\ /\ (c+d)}$$

Le risque relatif indique combien de fois une personne exposée est plus susceptible d'avoir un résultat par rapport à une personne non exposée.

Interprétation du Risque Relatif

- $\Rightarrow$ RR = 1: Pas d'association (pas de différence dans l'incidence de la maladie entre les groupes exposés et non exposés).
- $\Rightarrow$ RR > 1: Exposition dangereuse.
- $\Rightarrow$ RR<1: Exposition bénéfique.

Test de signification

Le risque relatif (RR) peut être testé pour sa signification en utilisant des intervalles de confiance. Si l'intervalle de confiance à 95 % pour RR inclut 1, les résultats ne sont pas statistiquement significatifs.

Exemple d'interprétation du RR

> Si RR = 3.23, les individus exposés ont 3.23 fois plus de chances de développer la maladie par rapport à ceux qui ne sont pas exposés.

Le Risque Attribuable (AR) mesure la proportion d'incidence de la maladie parmi les exposés qui est due à l'exposition. C'est le rapport entre la différence d'incidence entre les exposés et les non-exposés et l'incidence des exposés, exprimé en pourcentage.

$$Risque\ Attribuable\ (AR) = \frac{\text{Incidence des Exposés} - \text{ Incidence des Nonexposés}}{\text{Incidence des Exposés}}\ x\ 100\%$$
$$= \frac{(a\ /\ (a+b)) - \ (c\ /\ (c+d))}{a\ /\ (a+b)} \times\ 100\%$$

Exemple d'interprétation de l'AR

> Si AR = 80%, alors 80% de l'incidence de la maladie parmi les exposés peut être attribuée à l'exposition.

Avantages

1. Permet de mesurer les facteurs de risque avant l'apparition de la maladie, fournissant des preuves de causalité.
2. Permet d'étudier plusieurs résultats de la maladie.
3. Fournit des taux d'incidence et des estimations du risque relatif.
4. Adapté à l'étude des expositions rares.
5. Minimise le biais de sélection et le biais d'information.

Inconvénients

1. Coût élevé.
2. Inefficace pour l'étude des résultats rares.
3. Nécessite un suivi prolongé et/ou une grande population.
4. Les pertes au suivi peuvent compromettre la validité des résultats.
5. Inefficace pour les maladies rares.

6. Problèmes éthiques.

Exercices de révision

Scénario 1 : Étude de cohorte sur la radiation solaire et le cancer de la peau, Tableau 2x2

Exposition à la radiation solaire brutale	Résultat (cancer de la peau)		Total
	Oui	Non	
Oui	39	12	51
Non	10	64	74
Total	49	76	125

1. Selon ces résultats, essayez de décrire les scénarios de l'étude en mots.
2. Calculez toutes les mesures d'association possibles.
3. Interprétez les résultats obtenus.

Scénario 2 : Étude de cohorte sur la pratique du sport et la maladie ischémique du cœur, Tableau 2x2

Exposition à la pratique du sport	Résultat (maladie ischémique du cœur)		Total
	Oui	Non	
Oui	1024	2376	3400
Non	1205	604	1809
Total	2229	2980	5209

1. Selon ces résultats, essayez de décrire les scénarios de l'étude en mots.
2. Calculez toutes les mesures d'association possibles.
3. Interprétez les résultats obtenus.

Scénario 3 : Étude cas-témoins sur le diabète et l'infarctus du myocarde, Tableau 2x2

Exposition au diabète	Résultat (infarctus du myocarde)		Total
	Oui	Non	
Oui	60	40	100
Non	340	360	700
Total	400	400	800

1. Selon ces résultats, essayez de décrire les scénarios de l'étude en mots.
2. Calculez toutes les mesures d'association possibles.
3. Interprétez les résultats obtenus.

Questions de révision

1. Définissez la conception d'une étude cas-témoins et proposez ses synonymes.
2. Quels sont les principaux indicateurs d'association utilisés dans une étude cas-témoins ? Définissez-les et expliquez leur interprétation.
3. Quels sont les avantages d'une étude cas-témoins ?
4. Quels sont les inconvénients d'une étude cas-témoins ?
5. Définissez la conception d'une étude de cohorte et proposez ses synonymes.
6. Quels sont les principaux indicateurs d'association utilisés dans une étude de cohorte ? Définissez-les et expliquez leur interprétation.
7. Quels sont les avantages d'une étude de cohorte ?
8. Quels sont les inconvénients d'une étude de cohorte ?
9. Quelle est la principale différence entre une étude cas-témoins et une étude de cohorte ? Donnez un exemple.

CHAPITRE 11. ÉTUDES EXPÉRIMENTALES

Concepts clés

- *Les essais cliniques* ou études expérimentales se répartissent en deux catégories principales : les essais contrôlés et les essais non contrôlés.
- *Les essais contrôlés* comparent un traitement expérimental à un groupe témoin (traitement standard, traitement antérieur ou placebo). Ces essais sont jugés plus valides en raison de leur capacité à minimiser les biais et à évaluer avec précision l'efficacité de l'intervention.
- *Les essais non contrôlés* se caractérisent par l'absence de groupe de contrôle, ce qui diminue leur robustesse méthodologique et leur validité, en raison de l'absence de référence comparative.
- *Les essais cliniques contrôlés en groupes parallèles* impliquent une évaluation simultanée des groupes expérimentaux et de contrôle au sein d'une même étude, garantissant que les écarts observés sont directement imputables à l'intervention. Les essais en aveugle (simple ou double) sont utilisés pour réduire les biais.
- *Les essais contrôlés randomisés* (ECR) sont considérés comme le standard d'excellence en matière d'essais cliniques, avec une répartition aléatoire des participants dans les différents groupes, ce qui réduit les biais de sélection et améliore la fiabilité des résultats.
- *Les essais contrôlés non randomisés* ne bénéficient pas de l'assignation aléatoire, rendant plus difficile la garantie que les différences observées entre les groupes sont exclusivement dues à l'intervention, ce qui affaiblit les preuves obtenues.

- Dans *les essais croisés*, les patients reçoivent successivement les deux traitements (expérimental et témoin), jouant ainsi le rôle de leur propre contrôle, ce qui réduit la variabilité interindividuelle.
- *Les essais avec contrôles externes* comparent les résultats obtenus à des données historiques provenant d'autres études. Bien que ces essais puissent offrir des perspectives intéressantes, leur fiabilité est moindre en raison des différences potentielles dans les conditions expérimentales et les populations étudiées.
- *Les mesures statistiques* utilisées dans les essais cliniques comprennent le taux d'événements expérimentaux (EER), le taux d'événements dans le groupe témoin (CER), le risque relatif (RR), la réduction absolue du risque (ARR), la réduction relative du risque (RRR) et le nombre de sujets à traiter (NNT).

11.1 Classification des essais cliniques

Les essais cliniques sont des recherches expérimentales impliquant des sujets humains, dont l'objectif est d'évaluer l'efficacité des interventions médicales. Ils jouent un rôle crucial dans la résolution des questions liées à la thérapie et se divisent généralement en deux grandes catégories : les essais contrôlés et les essais non contrôlés.

Classification des essais cliniques

I. Essais contrôlés
   - 1.1 Contrôles parallèles
     - *a) Randomisés*
     - *b) Non randomisés*
   - 1.2 Contrôles séquentiels
     - *a) Auto-contrôle*
     - *b) Croisé*
   - 1.3 Contrôles externes

II. Essais non contrôlés

Les essais contrôlés comparent une intervention ou un médicament expérimental à un autre traitement, qu'il s'agisse d'un traitement standard, d'une thérapie déjà acceptée, ou d'un placebo.

Les essais non contrôlés, quant à eux, décrivent les résultats d'une intervention ou d'un traitement expérimental sans comparaison avec un autre traitement. Ces études, en l'absence de groupe témoin, sont généralement considérées comme moins fiables en termes de validité.

Les essais contrôlés jouissent d'une plus grande validité en recherche médicale, car ils sont conçus pour isoler les effets de l'intervention, réduisant ainsi les biais, et permettant une évaluation plus précise de l'efficacité réelle de l'intervention. Les études non

contrôlées, bien qu'utiles dans certaines situations, ne fournissent pas le même niveau de preuve, en raison de l'absence d'un point de comparaison rigoureux.

11.2 Essais contrôlés en groupes parallèles

Une méthode courante pour réaliser un essai clinique contrôlé consiste à établir deux groupes de sujets : le groupe expérimental, recevant la procédure expérimentale, et le groupe de contrôle, recevant la procédure standard ou un placebo, comme illustré à la *Figure 11.1.*

Pour garantir la validité des résultats, il est essentiel que les groupes expérimental et de contrôle soient aussi homogènes que possible, afin que toute différence observée puisse être exclusivement attribuée à l'intervention. Le contrôle concomitant signifie que les interventions pour les deux groupes sont menées simultanément dans le cadre de la même étude.

Afin de minimiser les biais, les chercheurs peuvent concevoir des essais en aveugle :

- *Essais en simple aveugle :* Les sujets ne savent pas quelle intervention leur est administrée.
- *Essais en double aveugle :* Ni les sujets ni les chercheurs ne savent quel groupe reçoit l'intervention expérimentale ou le traitement de contrôle.

Pour des raisons éthiques, les essais cliniques ne sont autorisés que lorsque les interventions sont censées apporter des bénéfices.

⇒ **Essais contrôlés randomisés (ECR)**

Les ECR représentent le standard de référence pour établir la causalité, car ils offrent la garantie la plus forte que les résultats observés sont attribuables exclusivement à l'intervention.

Dans un essai contrôlé randomisé (ECR), les participants sont assignés aléatoirement soit au groupe expérimental, soit au groupe de contrôle, assurant ainsi que chaque individu a une probabilité égale de recevoir l'une des interventions possibles. Ce processus de randomisation contribue à éliminer le biais de sélection, équilibrant les facteurs de confusion connus et inconnus entre les groupes. De plus, dans les ECR en aveugle, ni les participants ni les chercheurs ne savent quelle intervention est administrée, ce qui réduit davantage le biais et améliore la fiabilité des résultats.

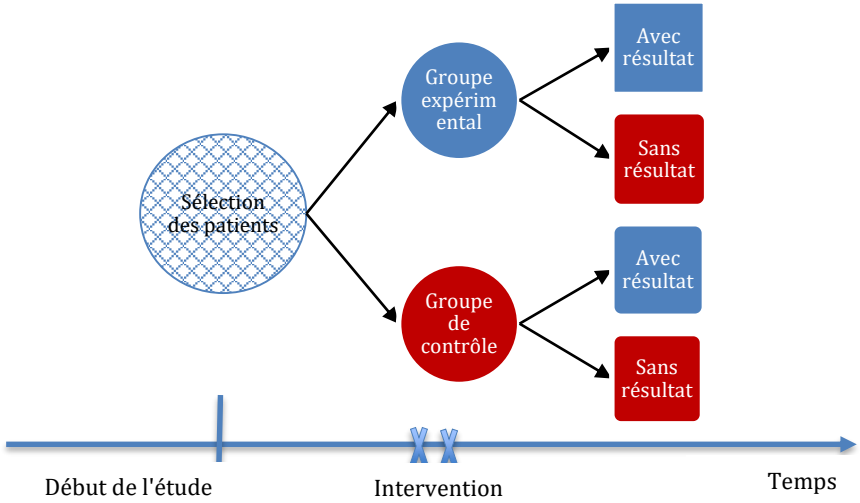

Figure 11.1 Diagramme de flux d'un essai clinique contrôlé en groupes parallèles

⇒ **Essais contrôlés non randomisés**

Les essais contrôlés non randomisés sont des études où les sujets ne sont pas attribués aléatoirement aux groupes expérimentaux ou de

contrôle. Ces essais, également désignés sous le terme d'essais cliniques ou d'études comparatives sans randomisation, sont considérés comme moins fiables car ils ne permettent pas de réduire le biais dans l'attribution des patients, compliquant ainsi la vérification que les différences observées entre les groupes sont exclusivement dues à l'intervention.

11.3 Essais contrôlés séquentiels

⇒ ***Essais en auto-contrôle***

Les essais en auto-contrôle sont des études dans lesquelles le même groupe de sujets est à la fois groupe expérimental et groupe témoin. Ce design permet un contrôle modéré en comparant les résultats au sein du même groupe d'individus dans différentes conditions.

⇒ ***Essais croisés***

Les essais croisés impliquent deux groupes de patients : un groupe reçoit d'abord le traitement expérimental, tandis que l'autre groupe reçoit le placebo ou le traitement de contrôle. Après une période définie, les deux traitements sont suspendus pendant une période de « *lavage* » (*washout*), durant laquelle aucun traitement n'est administré afin de garantir que les effets des traitements initiaux se sont dissipés. Après la période de lavage, les groupes échangent leurs traitements : le premier groupe reçoit maintenant le placebo, et le second groupe reçoit le traitement expérimental (*Figure* 11.2).

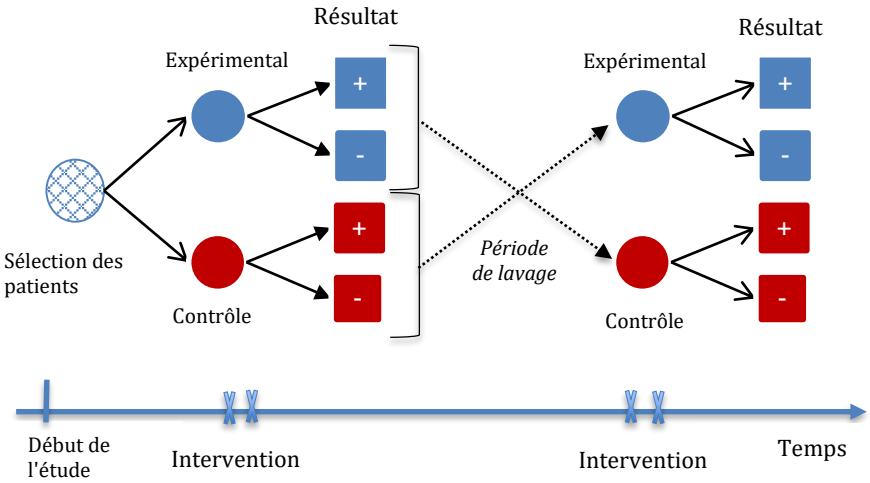

Figure 11.2 Diagramme de flux d'un essai croisé ou crossover

L'essai croisé, également connu sous le nom de design « *intra-individuel* », est particulièrement puissant lorsqu'il est correctement appliqué. Dans ce design d'étude, chaque patient reçoit à différents moments le traitement actif et le placebo, permettant des comparaisons directes au sein du même individu. Par conséquent, chaque patient sert de témoin à lui-même, ce qui renforce la fiabilité et la validité des résultats en réduisant la variabilité entre les différents sujets.

11.4 Essais à contrôle externe

Les essais à *contrôle externe* consistent à comparer les résultats d'un traitement expérimental avec des données provenant d'autres études ou de données collectées antérieurement, comme le présente la *Figure* 11.3.

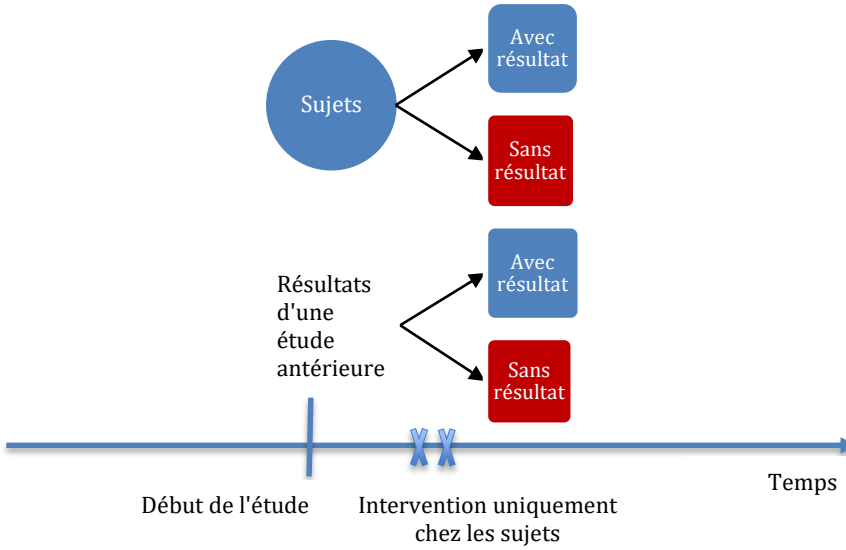

Figure 11.3 Diagramme de flux d'un essai clinique à contrôle externe

Ces comparaisons, également connues sous le nom de contrôles historiques, permettent aux chercheurs d'évaluer l'efficacité d'une intervention en utilisant les données existantes comme référence. Bien que cette méthode puisse offrir des informations précieuses, elle est généralement jugée moins fiable que les essais avec contrôles simultanés en raison des différences potentielles dans les conditions d'étude et les populations.

11.5 Analyse statistique des essais cliniques

L'analyse statistique dans les essais cliniques implique le calcul de différentes mesures pour évaluer l'efficacité des interventions. Ces mesures incluent :

- Taux d'événements expérimentaux (EER, *Experimental Event Rate*)
- Taux d'événements de contrôle (CER, *Control Event Rate*)
- Risque relatif (*Relative Risk*, RR)
- Réduction absolue du risque (ARR, *Absolute Risk Reduction*)
- Réduction relative du risque (RRR, *Relative Risk Reduction*)
- Nombre nécessaire de sujets à traiter (NNT, *Number Needed to Treat*)

Pour calculer ces mesures d'association, les résultats des essais cliniques sont généralement organisés dans un tableau 2x2 :

Tableau 11.1 Tableau 2x2 pour l'étude clinique

Facteur d'exposition (Intervention)	**Résultat *(résultat négatif)***		**Total**
	Oui	Non	
Exposés : Traitement expérimental	a	b	a + b
Non exposés : Placebo	c	d	c + d
Total	a + c	b + d	a + b + c + d

⇒ ***Taux d'événements expérimentaux (EER)*** : Le taux d'événements (risque d'un résultat indésirable, tel qu'une maladie ou un décès) dans le groupe recevant le traitement expérimental :

$$EER = \frac{a}{a+b}$$

⇒ ***Taux d'événements de contrôle (CER)***: Le taux d'événements dans le groupe recevant le placebo ou le traitement de contrôle :

$$CER = \frac{c}{c+d}$$

⇒ ***Relative Risk (RR)***: Le rapport entre le risque de maladie dans le groupe expérimental et le risque de maladie dans le groupe de contrôle. Il indique la probabilité pour le groupe exposé de présenter le résultat comparativement au groupe non exposé.

$$RR = \frac{EER}{CER}$$

Interprétation du RR :

> RR = 1 : L'intervention n'a aucun effet.
> RR > 1 : L'intervention est un facteur de risque.
> RR < 1 : L'intervention est un facteur bénéfique/protecteur.

⇒ ***Réduction absolue du risque (ARR)*** : Mesure la diminution du risque due à l'intervention par rapport au risque initial et indique combien de sujets évitent l'événement pour chaque 100 sujets traités. L'ARR est calculée comme la différence entre le taux d'événements de contrôle (CER) et le taux d'événements expérimentaux (EER):

$$ARR = CER - EER$$

Exemple : Dans une étude, si l'EER pour la maladie cardiovasculaire (CVD) dans le groupe aspirine est de 0,14 et le CER dans le groupe de contrôle est de 0,30, alors :

$$ARR = 0.30 - 0.14 = 0.16$$

Interprétation de l'ARR :

> Le risque de CVD est de 14 sujets sur 100 dans le groupe prenant de l'aspirine et de 30 sur 100 dans le groupe prenant le placebo.

Ainsi, 16 personnes sur 100 évitent la CVD en prenant de l'aspirine.

⇒ ***Réduction relative du risque (RRR)***: Indique la proportion de réduction du risque par rapport au risque initial :

$$RRR = \frac{CER - EER}{CER} = \frac{ARR}{CER}$$

Exemple : En utilisant l'étude précédente avec l'aspirine,

$$RRR = \frac{0.16}{0.30} = 0.53 \text{ or } 53\%$$

Interprétation de la RRR :

La réduction relative du risque est de 53 %. Cela signifie que, par rapport au risque initial de 30 cas de CVD pour 100 personnes, l'aspirine réduit le risque de 53 %.

⇒ ***Nombre nécessaire de sujets à traiter (NNT)*** : Indique le nombre de patients à traiter pour éviter un événement indésirable.

$$NNT = \frac{1}{ARR}$$

Exemple : Pour l'étude avec l'aspirine, le NNT se calcule comme suit :

$$NNT = \frac{1}{0.16} = 6.25$$

Interprétation du NNT :

Ainsi, environ 6,25 patients doivent être traités avec de l'aspirine pour éviter un cas de CVD. Ces informations aident les cliniciens à évaluer les bénéfices et les risques d'un traitement.

Avantages :

1. Fournit des preuves solides de causalité.
2. Réduit les biais.
3. Permet d'utiliser des contrôles historiques pour les études préliminaires.

Inconvénients :

1. Coûteux.
2. Problèmes éthiques.
3. Nécessite du temps.
4. Dépend de la conformité des participants.

Exercices de révision

1. Les résultats d'un essai clinique randomisé contrôlé sont présentés dans le tableau 2x2 suivant :

Facteur d'exposition (Intervention)	**Résultat**		**Total**
	Cholestérol stable	Cholestérol diminuant de 20 mg/dl	
Exposés : Traitement expérimental : Un médicament inhibiteur de l'enzyme 3-hydroxy-3-méthylglutarate coenzyme A réductase	3	43	46
Non exposés : Placebo	8	39	47
Total	11	82	93

1.1. Décrivez les scénarios de l'étude en utilisant ces résultats.

1.2. Calculez toutes les mesures d'association possibles.

1.3. Interprétez les résultats obtenus.

2. Sélectionnez une étude récente publiée dans une revue scientifique évaluée par des pairs qui vous paraît intéressante.
   2.1. Identifiez la question principale de recherche qu'elle aborde.
   2.2. Déterminez quel design d'étude (par exemple, essai clinique randomisé, étude de cohorte) serait le plus approprié pour répondre à la question de recherche.
   2.3. Vérifiez si l'étude a adopté le design que vous considérez comme optimal. Sinon, examinez les raisons pour lesquelles les auteurs ont choisi leur design.

Questions de révision

1. Quelle est la définition et la classification des essais cliniques ?
2. En quoi un essai clinique contrôlé se distingue-t-il d'un essai clinique non contrôlé ?
3. Qu'est-ce qui définit un essai clinique randomisé ?
4. Quels sont les essais cliniques contrôlés en groupes parallèles ? Fournissez des définitions, des types et un diagramme du design adéquat.
5. Quels sont les essais cliniques contrôlés séquentiels ? Fournissez des définitions, des types et un diagramme du design adéquat.
6. Quels sont les essais cliniques à contrôle historique ? Fournissez une définition et un diagramme du design adéquat.
7. Quelles sont les principales mesures d'association dans l'analyse des essais cliniques ? Définissez chaque mesure et expliquez son interprétation.

CHAPITRE 12. PRÉSENTATION DES RÉSULTATS DE RECHERCHE : APPROCHES GÉNÉRALES

12.1 La Rédaction d'un Rapport de Recherche

La rédaction du rapport de recherche constitue la phase ultime du processus de recherche. Ce rapport a pour objectif de communiquer les finalités de votre étude, les méthodes employées, les résultats obtenus, ainsi que les conclusions qui en découlent. Il doit être rédigé dans un style académique rigoureux, en évitant tout recours à un langage familier ou journalistique.

Un rapport de recherche traditionnel se structure généralement selon le format suivant :

A. Page de titre

- Titre du projet de recherche
- Nom du chercheur
- Nom de l'institution
- Date de publication

B. Contenu du projet

- **Introduction** : Présente le contexte de l'étude et la question de recherche ou l'hypothèse.
- **Revue de la littérature** : Fait le point sur les recherches existantes pertinentes à l'étude.
- **Matériel et méthodes** : Décrit la conception de la recherche, les procédures et les outils employés.
- **Résultats** : Expose l'analyse des données et l'interprétation des résultats.
- **Discussions** : Synthétise les résultats, en explicite les implications et les met en relation avec la littérature existante.

- **Conclusions** : Propose des conclusions définitives fondées sur les résultats et la discussion.
- **Recommandations** : Émet des suggestions pour des applications pratiques ou pour des recherches futures en s'appuyant sur les résultats.
- **Références / Bibliographie** : Énumère toutes les sources citées dans le rapport.
- **Annexes** : Comprend des documents supplémentaires tels que les données brutes, les questionnaires ou des calculs détaillés.

12.2 Présentation publique des résultats de recherche médicale

L'objectif d'une présentation orale est de communiquer efficacement les résultats scientifiques de votre recherche médicale à un public.

Conception générale d'une présentation orale :

- **Diapositive de titre** : Titre de la présentation et noms des auteurs.
- **Introduction** : 1-2 diapositives fournissant le contexte et les informations de base.
- **Objectifs** : 1-2 diapositives décrivant les buts de la recherche.
- **Matériel et Méthodes** : 1-2 diapositives décrivant le design et la méthodologie de l'étude.
- **Résultats** : 2-3 diapositives mettant en avant les découvertes les plus importantes.
- **Discussion** : 1 diapositive résumant les principales interprétations et implications.
- **Conclusions** : 1 diapositive avec les principales conclusions tirées de l'étude.
- **Clôture** : 1 diapositive pour les pensées finales et les remerciements.

Conseils pour la présentation :

- **Durée** : Visez une présentation de 10 minutes avec 8-10 diapositives. Généralement, allouez environ une minute par diapositive.
- **Présentation graphique** : Utilisez des graphiques et des tableaux pour illustrer vos données, car ils sont souvent plus efficaces que le texte. Assurez-vous que tous les visuels sont clairs et pertinents pour votre message.

Recommandations pratiques :

- **Diapositive de titre** : Utilisez une ligne unique avec des couleurs en gras ou contrastantes pour la visibilité.
- **Texte** : Assurez-vous que le texte est lisible depuis l'arrière de la salle. Limitez le texte à 5-7 lignes par diapositive, en suivant la règle des 7x7 : pas plus de 7 lignes et 7 mots par ligne.
- **Graphiques et tableaux** : Maintenez la clarté et la simplicité. Les tableaux ne doivent pas contenir plus de 3-4 colonnes et 5-7 lignes pour éviter l'encombrement.

12.3 Structure et principes de développement du mémoire de fin d'études à l'USMF Nicolae Testemitanu

Le mémoire de fin d'études doit démontrer la capacité de l'étudiant à travailler efficacement avec la littérature pertinente à son sujet. Il doit être méthodologiquement solide, avec une analyse et une interprétation rigoureuse des données, et suivre une structure logique. De plus, le mémoire doit être rédigé dans un langage scientifique, en respectant les normes académiques et les directives de rédaction scientifique établies par l'Université d'État de Médecine et de Pharmacie Nicolae Testemitanu. Le respect de ces normes garantit que le mémoire satisfait aux exigences universitaires pour l'élaboration et la soutenance des mémoires de fin d'études. Pour des directives détaillées, consultez la documentation officielle de l'université sur www.usmf.md.

CHAPITRE 13. INTRODUCTION A L'ETHIQUE DE LA RECHERCHE

13.1 Définition et objectifs de l'éthique de la recherche

L'éthique est l'ensemble des règles qui régulent nos attentes concernant notre propre comportement et celui des autres.

> *L'éthique de la recherche* est le cadre éthique qui guide la manière dont la recherche scientifique doit être réalisée et diffusée.

Objectifs de l'éthique de la recherche :

1. Protéger les participants en garantissant leur dignité, leurs droits et leur bien-être.
2. Veiller à ce que la recherche soit conduite de manière à favoriser le bien-être des individus, des groupes et de la société dans son ensemble.
3. Évaluer les activités de recherche spécifiques en termes d'intégrité éthique.

13.2 Principes de l'éthique de la recherche

L'éthique de la recherche repose sur trois approches fondamentales :

⇒ **Respect des personnes** : Reconnaissance de l'autonomie et des droits des individus.

⇒ **Bienfaisance** : Maximisation des bénéfices et minimisation des risques pour les participants.

⇒ **Justice** : Garantie d'une distribution **équitable des bénéfices et des charges de la recherche**.

Les cinq principes principaux de l'éthique de la recherche, fondés sur ces approches fondamentales, sont :

1. **Minimisation des risques de préjudice** : Mise en œuvre de mesures pour éviter les dommages aux participants.
2. **Obtention du consentement éclairé** : Garantie que les participants sont pleinement informés des objectifs de la recherche et consentent volontairement à y participer.
3. **Protection de l'anonymat et de la confidentialité** : Assurer la sauvegarde des informations privées des participants.
4. **Évitement des pratiques trompeuses** : Maintien de l'honnêteté et de la transparence vis-à-vis des participants.
5. **Droit de retrait** : Permettre aux participants de se retirer de l'étude à tout moment sans subir de pénalités.

Conseils essentiels pour garantir une recherche éthique :

- Recueillir les faits et discuter ouvertement des questions de propriété intellectuelle.
- Identifier et résoudre les problèmes éthiques.
- Reconnaître les parties prenantes et considérer leurs intérêts.
- Évaluer les sacrifices et les risques potentiels.
- Reconnaître les responsabilités (principes, droits, justice).
- Réfléchir à l'intégrité personnelle et à l'honnêteté.
- Adopter une approche créative pour les actions possibles.
- Respecter la vie privée et la confidentialité.
- Prendre des décisions éthiques appropriées et être prêt à gérer les opinions divergentes.

CHAPITRE 14. PRÉVENIR LE PLAGIAT : PRINCIPES CLÉS

14.1 Signification et Types de Plagiat

Le plagiat se définit comme l'utilisation des travaux ou des idées d'autrui sans leur accorder le crédit académique requis, que ce soit par des citations, des mentions dans les listes de références, ou dans les remerciements.

Types de plagiat :

⇒ **Plagiat direct :** Utiliser une transcription mot pour mot du travail de quelqu'un d'autre sans citation ni guillemets.
Il existe plusieurs types de plagiat direct :

- **Plagiat global** : Appropriation intégrale d'un texte comme s'il s'agissait du sien.
- **Plagiat par paraphrase** : Réécriture des idées d'un autre, présentées comme étant les siennes.
- **Plagiat mosaïque** : Combinaison de fragments de divers travaux pour constituer le sien.

⇒ **Auto-plagiat :** Réutilisation d'un travail déjà publié ou soumis dans un contexte académique. Si vous devez réintégrer un texte, une idée ou des données précédemment soumis, veillez à vous auto-citer.

⇒ **Plagiat accidentel :** Appropriation non intentionnelle des idées et contenus d'autrui, résultant d'une méconnaissance des normes de citation et de documentation. Même involontaire, cela constitue du plagiat et reste inacceptable.

14.2 Stratégies pour Éviter le Plagiat

Il existe plusieurs méthodes simples pour éviter le plagiat dans vos écrits académiques :

⇒ Soyez sûr de bien comprendre ce qu'est le plagiat ;

⇒ Apportez vos propres idées en trouvant quelque chose de nouveau à dire ;
⇒ Utilisez des citations pour souligner que vous utilisez les idées des autres ;
⇒ Donnez un crédit académique complet à toutes les sources que vous utilisez par des citations correctes et en les incluant dans la liste de références ;
⇒ Faites attention à la paraphrase : lorsque vous paraphrasez, vous devez le faire avec vos propres mots et ne pouvez pas simplement retirer un mot et le remplacer ; même ainsi, vous devez toujours donner un crédit académique approprié ;
⇒ Citez-vous également ;
⇒ Utilisez un logiciel de détection de plagiat.

BIBLIOGRAPHIE

BERRY G., MATTHEWS JNS, ARMITAGE P. Statistical Methods in Medical Research, 4th Edition, Blackwell Scientific, 2001.

COLTON T. *Statistics in Medicine*, Little, Brown, 1974.

COMSTOCK G. *Research Ethics: A Philosophical Guide to the Responsible Conduct of Research,* 1st Edition. Cambridge University Press, 2013.

DANIEL W.W. Biostatistics: *A Foundation for Analysis in the Heal.th Sciences*, 7th ed. Wiley, 1998

DAWSON B., TRAPP G. R. *Basic and Clinical Biostatistics,* Fourth Edition, McGraw-Hill Companies, Inc., USA, 2004.

FEINSTEIN A.R. *Clinical Epidemiology: The Architecture of Research*, WB Saunders, 1985.

FISHER LD, VAN BELLE G. *Biostatistics: A Methodology for Health Sciences,* Wiley,1996.

FLEISS JL. *Design and Analysis of Clinical Experiments*, Wiley, 1999.

FLEISS JL. *Statistical Methods for Rates and Proportion*, 2nd Edition, Wiley, 1981.

GLANTZ, STANTON A. *Primer of Biostatistics,* University of California. 4th Edition, McGraw-Hill, Inc, 1994: *перевод на русский язык, Издательский дом «Практика», 1999.*

GLASER, ANTONY N. *High-Yield Biostatistics,* Medical University of South Carolina. 4th Edition, Lippincott Williams & Wilkins, a Wolters Kluwer, Philadelphia, 2014.

GREENBERG RS. *Prospective studies.* In Kotz S, Johnson NL (editors): *Encyclopedia of Statistics Sciences*, Vol. 7, pp.315-319. Wiley, 1986.

GREENBERG RS. *Retrospective studies.* In Kotz S, Johnson NL (editors): *Encyclopedia of Statistics Sciences*, Vol. 8, pp.120-124. Wiley, 1988.

HENNESEY DESENA L. *Preventing plagiarism. Tips and Techniques,* National Council of Teachers of English, 2007

HULLEY SB (ED), CUMMINGS SR, BROWNER WS ET AL. *Designing Clinical Research*, 2nd Edition Lippincott Williams and Wilkins, 2001.

INGELFINGER JA, WARE JH, THIBODEAU LA. *Biostatistics in Clinical Medicine*, 3rd Edition, Macmillian, 1994.

KANE RL. *Understanding Health Care Outcomes Research*, Aspen Publishers, 1997.

KRUGER RA, CASEY MA. *Focus Groups: A Practical Guide for Applied Research*. Sage, 2000.

LANDRIVON G., DELAHAYE F. *La Recherche Clinique. De l'idee a la publication*. RECIF. Masson, Paris, 1995: traducere limba română Edit DAN, 2002.

NAGESVARO RAO G. *Biostatistics and Research Methodology,* PharmaMed Press, 2018.

PAGANO M., GAUVREAU K., *Principles of Biostatistics*, Second Edition, Belmont, CA, USA, 2000.

RAEVSCHI E., TINTIUC D., *Biostatistics & Research Methodology*, Nicolae Testemitanu SUMPh, CEP Medicina, Chisinau, 2012.

REA LM, PARKER RA: *Designing and Conducting Survey Research: A comprehensive Guide*, 2nd Edition Jossey-Bass, 1997

SCHLESSELMAN JJ: *Case-Control Studies: Design, Conduct, Analysis*. Oxford, 1982.

TAYLOR B. R. Medical Writing: A Guide for Clinicians, Educators, and Researchers, 3rd Edition, Springer, 2018.

TINTIUC D., BADAN V., RAEVSCHI E., GROSSU IU., GREJDEANU T., ET AL. *Biostatistica si Metodologia Cercetarii Stiintifice*, USMF „Nicolae Testemitanu", CEP Medicina, Chisinau, 2011.

WEINSTEIN MC, FINEBERG HV: *Clinical Decision Analysis*, WB Saunders, 1998.

ANNEXE A: **Valeurs critiques pour la distribution « t »**

Degrés de liberté (df)	Test unilatéral				
	0.05	0.025	0.01	0.005	0.0005
	Test bilatéral				
	0.10	0.05	0.02	0.01	0.001
1	6.314	12.706	31.821	63.657	636.62
2	2.920	4.303	6.965	9.925	31.598
3	2.353	3.182	4.541	5.841	12.924
4	2.132	2.776	3.747	4.604	8.610
5	2.015	2.571	3.365	4.032	6.869
6	1.943	2.447	3.143	3.707	5.959
7	1.895	2.365	2.998	3.499	5.408
8	1.860	2.306	2.896	3.355	5.041
9	1.833	2.262	2.821	3.250	4.781
10	1.812	2.228	2.764	3.169	4.587
11	1.796	2.201	2.718	3.106	4.437
12	1.782	2.179	2.681	3.055	4.318
13	1.771	2.160	2.650	3.012	4.221
14	1.761	2.145	2.624	2.977	4.140
15	1.753	2.131	2.602	2.947	4.073
16	1.746	2.120	2.583	2.921	4.015
17	1.740	2.110	2.567	2.898	3.965
18	1.734	2.101	2.552	2.878	3.922
19	1.729	2.903	2.539	2.861	3.883
20	1.725	2.086	2.528	2.865	3.850
21	1.721	2.080	2.518	2.831	3.819
22	1.717	2.074	2.508	2.819	3.792
23	1.714	2.069	2.500	2.807	3.767
24	1.711	2.064	2.492	2.797	3.745
25	1.708	2.060	2.485	2.787	3.725
26	1.706	2.056	2.479	2.779	3.707
27	1.703	2.052	2.473	2.771	3.690
28	1.701	2.048	2.467	2.763	3.674
29	1.699	2.045	2.462	2.756	3.659
30	1.697	2.042	2.457	2.750	3.646
40	1.684	2.021	2.423	2.704	3.551
60	1.671	2.000	2.390	2.660	3.460
120	1.658	1.980	2.358	2.617	3.373
⋈	1.645	1.960	2.326	2.576	3.291

ANNEXE B: **Valeurs critiques pour la distribution chi carré**

Degrés de liberté (df)	Niveau de signification (α)								
	0.995	0.00	0.975	0.95	0.9	0.1	0.05	0.025	0.01
1	0	0	0	0	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.1	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.3	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.2	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.7	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.6	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.27
12	3.07	3.57	4.4	5.23	6.3	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.6	5.23	6.26	7.26	8.55	22.31	25	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.3	28.85	32
17	5.7	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.2	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.4	13.85	15.66	33.2	36.42	39.36	42.98
26	11.16	12.2	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	16.49	20.6	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.5	17.79	19.81	21.66	23.95	44.9	48.6	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.9	61.16
42	22.14	23.65	26	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.2
50	27.99	29.71	32.36	34.76	37.69	63.17	67.5	71.42	76.15
55	31.73	33.57	36.4	38.96	42.06	68.8	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.4	79.08	83.3	88.38
65	39.38	41.44	44.6	47.45	50.88	79.97	84.82	89.18	94.42
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
85	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24
90	59.2	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
95	63.25	65.9	69.92	73.52	77.82	113.04	118.75	123.86	129.97
100	67.33	70.09	74.22	77.93	82.36	118.5	124.34	129.56	135.81

USMF „Nicolae Testemițanu”
Centrul Editorial-Poligrafic *Medicina*
Formatul hârtiei 60x84 $^{1}/_{16}$ Tiraj: 50 ex.
Coli de autor 5,8 Comanda nr. 33
Chișinău, bd. Ștefan cel Mare și Sfânt, 165