الجمهورية الجزائرية الديمقراطية الشعبية,
REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
وزارة التعليم العالي والبحث العلمي
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Université 20 août 1955 – Skikda
Faculté des Sciences
Département de Mathématiques

جامعة 20 أوت 1955 - سكيكدة
كلية العلوم
قسم الرياضيات

Mémoire de Master

Domaine : Mathématiques et Informatique
Filière : Mathématiques
Spécialité : Commande optimal des systèmes dynamique

Thème

Analyses Statistiques avec R

Présenté par :
M^elle Bey Wided

Soutenu publiquement le : 01/07/2025

Devant le jury composé de :

Kimouche Karima	**M.C.A**	**Université de Skikda**	**Présidente**
Tilbi Djahida	**M.C.A,**	**Université de Skikda**	**Encadrante**
Benhadri Mimia	**M.C.A,**	**Université de Skikda**	**Examinatrice**

Année universitaire : 2024/2025

Table des matières

Résumé en English

Statistical Analyses Using R

Abstract

Statistical analysis is an essential element of data science, used to interpret data, identify trends, and make decisions based on data. R is one of the most popular programming languages for statistical computation due to its wide range of statistical packages, its flexibility, and its powerful data visualization capabilities. This work simply addresses the different aspects of descriptive statistics and statistical inference using the R software. It is also useful for anyone interested in understanding and using the main statistical methods with R.

Keywords : Descriptive statistics ; Statistical inference ; Statistical analysis ; R software.

Résumé en Français

Analyses Statistiques avec R

Résumé

L'analyse statistique est un élément essentiel de la science des données, utilisé pour interpréter les données, identifier les tendances et prendre des décisions basées sur les données. R est l'un des langages de programmation les plus populaires pour le calcul statistique en raison de sa vaste gamme de packages statistiques, de sa flexibilité et de ses puissantes capacités de visualisation des données. Ce travail aborde de manière simple les différents aspects de la statistique descriptive et de l'inférence statistique en utilisant le logiciel R. Il est également utile pour toute personne intéressée par la connaissance et l'utilisation des principales méthodes statistiques avec R.

Mots clés : Statistique descriptive ; Inférence statistique ; Analyse statistique ; Logiciel R.

Résumé en Arabe

التحليل الإحصائي باستخدام برنامج R

ملخص

يعد التحليل الإحصائي عنصراً أساسياً في علم البيانات ويستخدم لفهم البيانات، وتحديد الاتجاهات، واتخاذ القرارات بناء على معطى واقعي من أشهر لغات البرمجة المستخدمة في الحسابات الإحصائية، وذلك بفضل سهولة استخدامه وقدرته على تحليل البيانات يعد برنامج R مجموعة واسعة من الحزم الإحصائية وقدرته الرؤية على تحليل البيانات يتناول هذا العمل بشكل مبسط الجوانب المختلفة. كما أنه مفيد لأي شخص يهتم بفهم البيانات واستكشافها باستخدام الطرق الإحصائية تم التركيز في هذه الدراسة على الإحصاء الوصفي باستخدام برنامج R

الكلمات المفتاحية الإحصاء الوصفي، الاستدلال الإحصائي، برنامج R

Remerciements

Avant tout, le grand remercie à **Allah**, qui lui seul guide nos pas sur le bon sens et qui ma donné la volonté, la patience, et surtout la santé tout au long de mes années d'études.

je voudrer de remercier trés chaleureusement ma directrice de mémoire, **Madame Djahida Tilbi**, pour sa guidance, sa disponibilité,et sa qualité d'encadrement, et son soutien moral durant toute la préparation de ce travail.

Je remercie vivement les membres du jury, **le Dr. ben hadri mimiya** et **la Dr. Kimouche karima**, pour avoir accepté d'évaluer ce modeste travail.

Enfin, je souhaite exprimer ma gratitude à tous les enseignants du département de mathématiques de l'université du 20-Août-1955 de Skikda, qui ont contribué à ma formation académique.

Dédicace

Je dédie ce travail, avant tout, à Allah — Lui qui m'a donné la force d'avancer, la lumière dans les moments d'ombre, et la patience quand tout semblait s'arrêter. Que chaque ligne de ce travail soit une prière silencieuse de gratitude.

À ma mère, Khalef Yamina, douce lumière de ma vie.

À mon père, El-Aïdi, pilier de sagesse et de courage.

À mes sœurs : Siham, Radia et Leïla, pour leur affection constante et leur soutien discret.

À mes frères : Riad, Mourad, Redouane, Issam et Mohamed, pour leur force, leur présence, et leur amour fraternel.

À mon époux, Sofiane, compagnon fidèle, source de paix et d'encouragement.

À mon fils, Oussaid, trésor de mon cœur, raison de tant d'efforts et de rêves.

À mes neveux et nièces, qui embellissent mes jours, et tout particulièrement à Rahma, Tahar et kassouara, éclats de bonheur et de tendresse.

À mes beaux-parents, Houssein et Ratiba, pour leur bienveillance et leur présence réconfortante.

À mes amies : Sara, Imen, Chaima, Bouchra, Saïda, Meriam, Basma, Hadjar, Rania et Faten — merci d'avoir été là, chacune à votre manière.

Et tout spécialement à Chaima, qui a marché à mes côtés du début à la fin, avec fidélité et cœur. Tu n'as pas seulement accompagné ce parcours, tu en as été l'une des causes les plus sincères.

Enfin, à tous ceux et celles qui ont contribué d'une manière ou d'une autre à ce parcours par un mot, un geste ou une pensée, merci du fond du cœur.

Introduction générale

L'analyse statistique est devenue un outil incontournable dans de nombreux domaines tels que la santé, l'économie, les sciences sociales ou encore l'ingénierie. Avec l'augmentation massive des volumes de données disponibles, les logiciels de calcul sont aujourd'hui essentiels pour mener à bien des analyses statistiques précises, rapides et fiables. En effet, ces outils permettent non seulement d'automatiser des calculs souvent complexes, mais aussi de gagner un temps considérable dans le traitement de grands ensembles de données, tout en réduisant les risques d'erreurs humaines (Moore & McCabe, 2017).

Les logiciels statistiques offrent également des fonctionnalités avancées pour la visualisation graphique (histogrammes, courbes, nuages de points, etc.), l'application de méthodes sophistiquées telles que la régression multiple, les tests d'hypothèse, la classification ou encore l'analyse en composantes principales (ACP), ainsi que pour la manipulation et le nettoyage des données. Ces avantages rendent leur utilisation indispensable tant dans la recherche scientifique que dans les milieux professionnels.

Parmi les logiciels les plus utilisés, on retrouve **Excel**, pratique pour les statistiques de base ; **SPSS** et **Stata**, largement employés dans les sciences sociales ; **SAS**, très répandu dans les secteurs de la santé et de la finance ; **Python** et **R**, puissants langages de programmation open-source adaptés à l'analyse de données, à la visualisation et au machine learning. En particulier, **R** est reconnu pour sa richesse en bibliothèques statistiques, sa flexibilité et sa forte adoption dans le monde académique et la recherche (Wickham & Grolemund, 2017).

Dans le cadre de ce travail, nous nous intéressons à l'application pratique des principales fonctions statistiques offertes par le logiciel **R**. L'objectif est de simplifier leur usage pour l'analyse, l'estimation, la vérification d'hypothèses, et la visualisation des données.

Le manuscrit est structuré en **quatre chapitres**.

— Le **premier chapitre** est consacré à la **statistique descriptive**. Il présente les outils fondamentaux permettant de résumer et visualiser un ensemble de données, notamment à l'aide des mesures de tendance centrale (moyenne, médiane) et de dispersion (variance, écart-type), ainsi que par l'intermédiaire de représentations graphiques.

— Le **deuxième chapitre** aborde **l'estimation des paramètres**, une composante essentielle de l'inférence statistique. Il s'agit d'utiliser les données d'un échantillon pour tirer des conclusions sur une population cible, à travers des estimateurs ponctuels et des intervalles de confiance.

— Le **troisième chapitre** est dédié aux **tests d'hypothèse**, qui permettent de valider ou de rejeter des conjectures statistiques concernant les paramètres d'une population, en s'appuyant sur des règles de décision rigoureuses et le contrôle des erreurs de type I et II.

— Enfin, le **quatrième chapitre** se concentre sur **l'application concrète de l'analyse statistique**. À travers l'utilisation du logiciel R et de bibliothèques spécialisées, nous illustrons des méthodes telles que l'estimation par maximum de vraisemblance, l'exploration visuelle de données, et des éléments d'intelligence artificielle appliqués à l'analyse automatisée de jeux de données réels.

Ce travail vise ainsi à fournir un aperçu structuré et pratique de l'analyse statistique avec R, en mettant l'accent sur l'accessibilité des outils, la rigueur des méthodes, et la pertinence des résultats obtenus dans des contextes variés.

Chapitre 1

Statistique Descriptive

1.1 Introduction

L'analyse statistique commence généralement par une étape essentielle : la statistique descriptive. Celle-ci permet de résumer, d'organiser et de présenter les données d'une manière claire et significative. Avant toute interprétation ou modélisation, il est fondamental de comprendre la nature des données disponibles, leur structure, ainsi que leurs principales caractéristiques.

Dans ce chapitre, nous allons présenter les notions de base de la statistique descriptive, à travers une définition générale, un vocabulaire statistique essentiel, ainsi qu'une classification des types de caractères. Ensuite, nous aborderons les différentes représentations graphiques utiles pour visualiser les données, avant d'examiner les principaux paramètres statistiques permettant de décrire les tendances centrales et la dispersion des données.

Cette étape constitue le socle de toute analyse statistique rigoureuse, en préparant les données pour des traitements plus complexes comme l'inférence statistique ou les analyses multivariées.

Définitions

La statistique est un ensemble des méthodes qui servent à organiser les épreuves fournissant des observations, à analyser celles-ci et à interpréter les résultats.

L'analyse statistique se subdivise en deux parties :

Statistique descriptive

Elle a pour but de décrire, c'est-à-dire de résumer ou représenter les données.

Questions typiques :

— Représentation graphique

— Paramètres de position, de dispersion, de relation

Statistique inférentielle

L'ensemble des méthodes permettant de formuler un jugement. Elle nécessite des outils mathématiques plus pointus (théorie des probabilités).

Questions typiques :

— Estimation des paramètres
— Intervalle de confiance
— Tests d'hypothèses
— Modélisation (exemple : régression linéaire)

1.2 Vocabulaire statistiques

Population : la collection d'objets ou de personnes étudiées (élèves, habitants, voitures...)

Individu : élément de la population étudiée (un élève, un habitant, une voiture...)

Échantillon : partie de la population étudiée. Nombre d'individus dans un échantillon noté n, est appelé taille de l'échantillon.

Variable (caractère) : propriété commune aux individus de la population que l'on veut étudier.

Types de caractères

Qualitatif :

on ne peut associer ni valeur numérique ni un ordre naturel (type de voiture, couleur des cheveux...) Dans l'analyse statistique, les variables de type **qualitatif** (ou *catégorielles*) se divisent en deux types principaux :

1. **Qualitatif nominal**
   Ce sont des catégories sans ordre particulier.
   Exemples :
   — Couleur des yeux (bleu, vert, marron)
   — Genre (homme, femme)
   — Pays de naissance
2. **Qualitatif ordinal**
   Ce sont des catégories avec un ordre logique ou hiérarchique.
   Exemples :
   — Niveau d'éducation (primaire, secondaire, universitaire)
   — Niveau de satisfaction (faible, moyen, élevé)

Quantitatif :

peut prendre des valeurs numériques (poids,longueur).
un caractère quantitatif peut être :

— **Continue :** peut prendre toutes les valeurs numériques dans un intervalle déterminé (taille, poids...)
— **discontinue (Discrète) :** ne peut prendre que des valeurs numériques isolées (nombre de pièces d'habitations , nombre de fruits endommagés...)

1.3 Représentations graphiques

1.3.1 Diagramme en bâtons

si on porte en abscisse les valeurs des n_i et si on trace à partir de chacun de ces points,un segment à l'axe des ordonnées et de longueur l'effectif n_i on obtient un diagramme en bâton (si on joint les sommets des bâtons on obtient le polygone des fréquences)

1.3.2 Histogramme

lorsque le caractère étudié est continue on utilise un histogramme. chaque classe est représentée par un rectangle dont la base est égale à l'intervalle de la classe et dont la hauteur est égale à l'effectif correspondant le polygone des fréquences s'obtient en joignent les points d'abscisses les centres de classes et d'ordonnées les effectifs correspondants.

1.3.3 La courbe cumulative (signoide)

On considère les pointes dont Les abscisses sont les limites supérieures des classes et d'ordonnées n_i^c correspondants la limite inférieurs de La première classe à pour ordonnée le zéro. En reliant entre ces points par segments, on obtient la courbe cumulative.

1.3.4 Diagramme en boîte (la boîte à moustaches)

c'est un résumé visuel du sommaire d'une série de données, la médiane, les quartiles, la plus petite et la plus grande valeur de la série des valeurs aberrante (les valeurs qui s'écarte de façon marquée de l'ensemble des données) ce diagramme est utilise principalement pour comparer un même caractère dans deux populations de tailles différentes.

1.4 Paramètres à tendance centrale

1.4.1 Le mode M_0

Le mode d'un ensemble de nombres est la valeur qui y apparaît le plus, c'est-à-dire la valeur dominante. Le mode peut ne pas exister et, même s'il existe, peut ne pas être unique (dans le cas continu on parle de classe modale).

Si la variable discréte :

Exemple 1 : L'ensemble $2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18$ a comme mode 9.

Exemple 2 : L'ensemble $3, 5, 8, 10, 12, 15, 16$ n'a pas de mode.

Exemple 3 : L'ensemble $2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 9$ a deux modes 4 et 7. La série est appelée bimodale. Une série ayant un seul mode est appelée unimodale.

Dans le cas d'une variable continue : On applique la formule suivante :

$$M_0 = l_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot A_i.$$

- l_i : la limite inférieure de la classe modale
- Δ_1 : la différence entre la fréquence de la classe modale et celle d'avant
- Δ_2 : la différence entre la fréquence de la classe modale et celle d'après

1.4.2 La médiane M_e

La médiane d'un ensemble de nombres rangés par ordre croissant est :

— la valeur du milieu si le nombre des données est impair

— la moyenne arithmétique des deux valeurs du milieu si le nombre des données est pair

Exemple 1 : L'ensemble des nombres $3, 4, 4, 5, 6, 8, 8, 8, 10$ a comme médiane 6.

Exemple 2 : L'ensemble des données $5, 5, 7, 9, 11, 12, 15, 18$ a comme médiane 10 car $(9 + 11)/2 = 10$. Pour déterminer la médiane dans le cas continu, il est nécessaire de considérer les effectifs cumulés croissants ou décroissants, et de chercher le cas échéant par interpolation la valeur du caractère correspondant à 50% de l'effectif total.

Exemple :

$$M_e = \text{l'abscisse de } 32 \times 50\% = 32 \times 0.5 = 16.$$

La médiane dans le cas d'une variable continue

Pour une variable continue, présentée sous forme de tableau de fréquences par classes, la médiane se calcule à l'aide de la formule suivante :

$$\text{Médiane} = L + \left(\frac{\frac{N}{2} - F}{f_m} \right) \times h,$$

— L : borne inférieure de la classe médiane

— N : effectif total

— F : effectif cumulé avant la classe médiane

— f_m : effectif de la classe médiane

— h : amplitude de la classe médiane.

La classe médiane est déterminée comme la première classe dont l'effectif cumulé dépasse ou atteint $\frac{N}{2}$.

1.4.3 Moyenne arithmétique $\bar{X}$

Soit $x_1, x_2, x_3, \ldots, x_n$ une suite finie de nombres. La moyenne arithmétique est :

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Si chaque valeur x_i apparaît n_i fois dans la série, on peut encore écrire :

$$\bar{X} = \frac{1}{n}\sum n_i x_i.$$

En remarquant que $\frac{n_i}{n}$ est la fréquence relative f_i qui correspond à la valeur x_i, on a aussi :

$$\bar{X} = \sum f_i x_i.$$

Dans le cas de données groupées en classes, on prend pour valeur des x_i les centres de classes, on prend pour valeur des x_i les centres de classes.

— **Les percentiles** Le $k^{\text{ème}}$ percentile est la valeur du caractère :

— telle que l'ensemble des individus dont le caractère est au moins égal à C_k représente les $(100 - k)\%$ de l'effectif total.

Parmi les percentiles, on distingue :

— **Les déciles :** pour lesquels $k = 10, 20, 30, \ldots$

$$C_{10} = D_1, \quad C_{20} = D_2, \ldots$$

— **Les quartiles :** pour lesquels $k = 25, 50, 75$

$$C_{25} = Q_1, \quad C_{50} = Q_2, \quad C_{75} = Q_3.$$

Remarque :

La comparaison des trois permet de se faire une idée plus complète de la distribution (si les trois sont à peu près égales alors la série statistique a peu près symétrique).

1.5 Paramètres de dispersion (variabilité)

Un paramètre de dispersion se rapporte à la différence de deux valeurs du caractère. Alors qu'un paramètre de position représente une valeur du caractère.

1.5.1 L'étendue

L'étendue est une mesure de dispersion qui exprime la différence entre la plus grande et la plus petite valeur d'une série statistique.

$$\text{Étendue} = \text{Valeur maximale} - \text{Valeur minimale}$$

1.5.2 l'ecart absolu moyen

Elle est donnée par :

$$E_m = \frac{1}{n}\sum_i n_i \left|x_i - \bar{X}\right|.$$

1.5.3 Variance

La variance du caractère dans l'échantillon, notée V(X) est donnée par :

$$V(X) = \frac{1}{N}\sum_i n_i (x_i - \bar{X})^2.$$

La variance du caractère dans la population, notée σ^2 , est en général inconnue.

1.5.4 Écart-type

C'est la racine de la variance :

$$\sigma_X = \sqrt{V(X)}.$$

1.5.5 Moment d'une série statistique

On appelle moment d'ordre q par rapport à x_0 la quantité :

$$m_q = \frac{1}{n}\sum n_i (x_i - x_0)^q.$$

— Si $x_0 = 0$ et $q = 1$, on a $m_1 = \bar{X}$
— Si $x_0 = \bar{X}$ et $q = 2$, on a $m_2 = \mathrm{V(X)}$

Chapitre 2

Estimation de la statistique

2.1 Introduction

L'estimation statistique est une branche essentielle de l'inférence statistique. Elle a pour objectif de fournir des informations sur les paramètres inconnus d'une population à partir des données issues d'un échantillon. En pratique, il est souvent impossible d'étudier l'ensemble d'une population en raison de contraintes de temps, de coût ou d'accessibilité. C'est pourquoi l'analyse se base généralement sur un sous-ensemble représentatif de cette population.

À partir de cet échantillon, les statisticiens utilisent des méthodes rigoureuses pour estimer des paramètres tels que la moyenne, la variance ou la proportion. Ces estimations permettent de tirer des conclusions fiables sur l'ensemble de la population étudiée, tout en intégrant une certaine marge d'incertitude due au caractère aléatoire de l'échantillonnage.

Estimateur sans biais :

On dit qu'un estimateur est **sans biais** si, en moyenne, il donne la vraie valeur du paramètre.

Autrement dit : si on répétait l'expérience plusieurs fois (plusieurs échantillons), la moyenne des valeurs de l'estimateur serait égale à la vraie valeur du paramètre.

$$\mathbb{E}[\hat{\theta}] = \theta.$$

Exemple : la moyenne empirique $\overline{X}$ est un estimateur sans biais de la moyenne réelle μ.

Estimateur biaisé :

Un estimateur est dit **biaisé** si, en moyenne, il ne donne pas la vraie valeur du paramètre.

$$\mathbb{E}[\hat{\theta}] \neq \theta.$$

Exemple : l'estimateur de la variance $\frac{1}{n}\sum(x_i - \overline{X})^2$ est biaisé.

Estimateur convergent :

Un estimateur est dit **convergent** si, lorsque la taille de l'échantillon augmente $(n \to \infty)$, l'estimateur se rapproche de plus en plus de la vraie valeur du paramètre.

$$\hat{\theta} \xrightarrow[n\to\infty]{\mathbb{P}} \theta.$$

2.2 Estimation ponctuelle

Estimer un paramètre par exemple (une moyenne, une variance, ou une proportion) revient à calculer une valeur approchée à partir des données d'un échantillon.

Dans le cas où le paramètre est estimé par un seul nombre déduit des résultats de l'échantillon, on appelle ce nombre "une estimation ponctuelle du paramètre".

2.2.1 Estimation de la variance

Soit X une variable aléatoire qui suit une loi probabilité, on cherche estimer la variance σ^2 de X.

Définition :

La variance empirique de l'échantillon $(x_1, x_2, \ldots, x_n)$ de X (la statistique) est définie par :

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2,$$

qui est la variance de l'échantillon, aussi appelée variance empirique. Correspond donc à la moyenne des écarts, à la moyenne empirique.

Propriétés :

$$\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2.$$

Variance empirique modifiée : Soit s^2 la variance empirique modifiée, on la calcule comme suit :

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2\right).$$

On peut aisément prouver que :

$$\hat{S}^2 = \frac{n}{n-1}S^2.$$

Variance de la variance empirique :

L'expression de la variance de S^2 se présente comme :

$$\mathrm{Var}(S^2) = \frac{n-1}{n^2}\sigma^4.$$

Remarque :

Si $(n-1) > 30$: χ^2 remplacée $N(0,1)$:

$$IC(\sigma^2) = \left[\frac{2SCE}{\sqrt{2(n-3)} + u_{\alpha/2}}, \frac{2SCE}{\sqrt{2(n-3)} - u_{\alpha/2}}\right].$$

2.2.2 Estimation de la moyenne

Soit X une variable aléatoire, dont on cherche à estimer la moyenne (ou espérance) $\mu = \mathbb{E}(X)$ à partir d'un échantillon $(x_1, x_2, \ldots, x_n)$ de X
(on ne suppose rien sur la loi de X)

Définition :

La moyenne empirique de l'échantillon $(x_1, x_2, \ldots, x_n)$ de X (la statistique) est définie par :

$$m_n = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Théorème central limite 1 :

Si la variance σ^2 est connue et que l'échantillon prélevé est grand ($n \geq 30$), donc la moyenne échantillonnale vérifie :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \longrightarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \longrightarrow \mathcal{N}(0,1).$$

Remarque : le théorème précédent reste valide quand la variance est connue, l'échantillon est petit et que la variable aléatoire X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$

- Si la variable σ^2 de la population est inconnue et que l'échantillon prélevé est grand ($n \geq 30$), alors :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \longrightarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

c'est-à-dire

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \longrightarrow \mathcal{N}(0,1).$$

Théorème central limite 2 :

Si la variance de la population est inconnue, la variable X suit une distribution normale $\mathcal{N}(\mu, \sigma^2)$ et si la taille de l'échantillon est petite ($n < 30$)
Donc :

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \longrightarrow t_{n-1} \quad \text{loi de Student } (n-1) \text{ degrés de liberté (ddl).}$$

2.2.3 Estimation de la proportion

Considérons une variable aléatoire X qui peut prendre deux valeurs : 0 (échec) ou 1 (succès). Dans un contexte expérimental, on cherche à estimer la probabilité de succès p.

Lorsque l'expérience de Bernoulli est répétée n fois, on note K_n le nombre de succès observés. La **fréquence empirique** est alors définie par :

$$F = \frac{K_n}{n}.$$

Cette fréquence constitue un estimateur ponctuel du paramètre p.

Définition :

- **Fréquence empirique :** La variable aléatoire F représente la proportion de succès dans l'échantillon et sert d'estimation pour la probabilité p.
- **Propriétés :**

$$\mathbb{E}(F) = p \quad \text{et} \quad \text{Var}(F) = \frac{p(1-p)}{n}.$$

2.3 Estimation par intervalle de confiance :

Les estimations ponctuelles, bien qu'utiles, ne fournissent aucune information concernant la précision des estimations.
Autrement dit, elles ne tiennent pas compte de l'erreur possible dans l'estimation due aux fluctuations d'échantillonnage. La théorie des intervalles de confiance (IC) vise à construire autour de l'estimation ponctuelle un intervalle qui aura une grande probabilité $(1-\alpha)$ de contenir la vraie valeur du paramètre.

Soit X une variable aléatoire dont la loi dépend d'un paramètre θ inconnu, soit $(x_1, \ldots, x_n)$ un échantillon issu de X et $\alpha \in [0, 1[$

Définition :

On appelle intervalle de confiance pour θ de niveau $1-\alpha$ (où α est un seuil), un intervalle $[t_1, t_2]$ qui a la probabilité $1-\alpha$ de contenir la vraie valeur de θ :

$$\mathbb{P}(t_1 < \theta < t_2) = 1 - \alpha.$$

Plus le niveau de confiance est élevé, plus la certitude est grande et que la méthode d'estimation produira une estimation contenant la vraie valeur de θ.
Lorsque l'on augmente le niveau de confiance $1-\alpha$, on augmente la longueur de l'intervalle.

2.3.1 Intervalle de confiance pour la variance :

Supposons que $X \sim \mathcal{N}(\mu, \sigma^2)$ et soit S^2 la variance empirique observée sur un échantillon de taille n.

La statistique suivante :

$$\frac{(n-1)S^2}{\sigma^2},$$

suit une loi du $\chi^2(n-1)$ à $n-1$ degrés de liberté.

En fixant un niveau de confiance α, les bornes de l'intervalle de confiance peuvent être obtenues à partir des quantiles $\chi^2_{1-\alpha/2}$ et $\chi^2_{\alpha/2}$:

$$I_C(\sigma^2) = \left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}},\ \frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right].$$

Pour obtenir l'intervalle de confiance pour l'écart-type (σ), il suffit de prendre les racines carrées des bornes obtenues.

Remarque :

Si $(n-1) > 30$: χ^2 remplacée par $N(0,1)$:

$$IC(\sigma^2) = \left[\frac{(n-1)s^2}{\sqrt{2(n-3)} + U_{1-\alpha/2}},\ \frac{(n-1)s^2}{\sqrt{2(n-3)^2} - U_{1-\alpha/2}}\right].$$

2.3.2 Intervalle de confiance pour la moyenne :

Cas où , la taille de l'échantillon est petite ($n < 30$), on suppose que $X \sim \mathcal{N}(\mu, \sigma)$

Cas 1 : si la variance inconnue, on a :

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

la loi de Student à $(n-1)$ ddl (degrés de liberté), on cherche dans la table de la loi de Student, α étant fixé la valeur $t_{n-1;1-\alpha/2}$. que

$$P\left(-t_{n-1;1-\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{n-1;1-\alpha/2}\right) = 1 - \alpha.$$

On a :

$$P\left(\bar{X} - t_{n-1;1-\alpha/2}\hat{S}\sqrt{n} < \mu < \bar{X} + t_{n-1;1-\alpha/2}\hat{S}\sqrt{n}\right) = 1 - \alpha.$$

Cas 2 : si la variance de la population de X est connue :

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{ou bien} \quad \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1).$$

On se fixe le risque α et on cherche dans la table de la loi normale la valeur $U_{1-\frac{\alpha}{2}}$ (c'est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite) telle que :

$$P\left(-U_{1-\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < U_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

$$P\left(\bar{X} - U_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + U_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Cas où, la taille de l'échantillon est grande $n > 30$, il n'est plus nécessaire de supposer que X est gaussienne.

Cas 1 : si la variance connue, on a :

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Cas 2 : si la variance inconnue

On a :

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim \mathcal{N}(0,1),$$

On se fixe l'erreur α et on cherche dans la table de la loi normale la valeur $U_{1-\frac{\alpha}{2}}$ telle que :

$$P\left(-U_{1-\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < U_{1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

$$\left(\bar{X} - U_{1-\frac{\alpha}{2}} \cdot \frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + U_{1-\frac{\alpha}{2}} \cdot \hat{S}\sqrt{n}\right) = 1-\alpha.$$

2.3.3 Intervalle de confiance pour la proportion

On a $f = \frac{k}{n}$ est le meilleur estimateur de F est la proportion de la population possédant une caractéristique considérée.

On cherche dans la table de $\mathcal{N}(0,1)$ la valeur $U_{1-\frac{\alpha}{2}}$ telle que :

$$P\left(-U_{1-\frac{\alpha}{2}} < \frac{F-f}{\sqrt{f(1-f)/n}} < U_{1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

$$P\left(f - U_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} < F < f + U_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}\right) = 1-\alpha.$$

L'intervalle de confiance au seuil $1-\alpha$ pour la proportion f de la population est de la forme :

$$IC_f = \left[f - U_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}; f + U_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}\right].$$

Chapitre 3

Tests d'hypothèses statistiques

3.1 Introduction

Le test d'hypothèses statistiques est l'un des outils fondamentaux de l'inférence statistique. Il est utilisé pour vérifier la validité d'une ou plusieurs hypothèses concernant un ou plusieurs paramètres d'une population statistique à partir des données d'un échantillon. Le test repose sur la comparaison d'une statistique calculée à partir de l'échantillon avec une valeur théorique supposée, et sur un niveau de signification fixé à l'avance pour prendre une décision.

3.2 Test de conformité

3.2.1 Construction générale

Le test de conformité permet de vérifier si un paramètre observé (moyenne, variance, proportion) est conforme à une valeur théorique supposée dans la population.

- Formulation des hypothèses H_0 et H_1
- Définir la statistique de test.
- Choisir le niveau de signification.
- Déterminer les zones d'acceptation et de rejet.
- Calculer la statistique expérimentale.
- Prendre la décision.

3.2.2 Comparaison d'une moyenne observée à une moyenne théorique

Hypothèses :

$$H_0 : \mu = \mu_0 \quad \text{et} \quad H_1 : \mu \neq \mu_0.$$

Cas 1 : Variance connue (courbe Z)

Statistique de test :

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Conditions : Population normale ou grand échantillon, variance connue.
Courbe : Courbe Z : loi normale centrée réduite ($\mu = 0$, $\sigma = 1$) avec zones d'acceptation et de rejet.

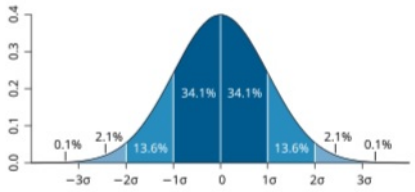

FIGURE 3.1 – Courbe de Z

Décision : on rejet de H_0 quand

$$Z > U_{1-\frac{\alpha}{2}}.$$

Cas 2 : Variance inconnue (courbe T)

Statistique de test :

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}.$$

avec S l'écart-type de l'échantillon, degrés de liberté : $n - 1$.
Conditions : Population normale, variance inconnue, petit échantillon.
Courbe : Courbe T : similaire à Z mais plus aplatie pour petits degrés de liberté.

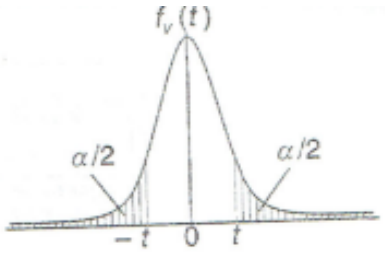

FIGURE 3.2 – Courbe de T

Décision : on rejet de H_0 quand

$$T > t_{n-1\left(1-\frac{\alpha}{2}\right)}.$$

3.2.3 Comparaison d'une variance observée à une variance théorique

Hypothèses :

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{et} \quad H_1 : \sigma^2 \neq \sigma_0^2.$$

Cas 1 : Grand échantillon

Statistique de test : loin normale centrée réduite N(0,1)

$$Z = \frac{\left|2(n-1)S^2 - \sqrt{2n-3}\right|}{\sigma^2}.$$

Conditions : Population normale.
Décision : on rejet H_0 quand

$$Z > U_{1-\frac{\alpha}{2}}.$$

Cas 2 : Petit échantillon

Statistique de test :

$$\chi^2 = \frac{(n-1)s^2}{\hat{\sigma}^2}.$$

Conditions : Population normale.
Décision : on rejet H_0 quand

$$\chi^2 > \chi^2_{1-\frac{\alpha}{2}}.$$

3.2.4 Comparaison d'une proportion observée à une proportion théorique

Hypothèses :

$$H_0 : p = p_0 \quad \text{et} \quad H_1 : p \neq p_0.$$

Statistique de test :

$$Z = \frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Conditions : Échantillon grand : $np > 5$ et $n(1-p) > 5$.
Décision : on rejet H_0 si

$$|Z| > U_{\alpha/2}.$$

3.3 Test d'homogénéité

3.3.1 Construction générale

Vérifier si deux ou plusieurs échantillons proviennent d'une même population ou de populations homogènes par rapport à un paramètre donné.

3.3.2 Comparaison de deux moyennes

Hypothèses :

$$H_0 : \mu_1 = \mu_2 \quad \text{et} \quad H_1 : \mu_1 \neq \mu_2.$$

Cas 1 : Variances connues et égales (courbe Z)

Statistique de test :

$$Z = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Conditions : Populations normales, variances connues.
Décision : on rejet H_0 si

$$|Z| > U_{\alpha/2}.$$

Cas 2 : Variances inconnues ou inégales (courbe T)

Statistique de test :

$$T = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\left(\frac{(n-1)S_1^2+(n-1)S_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

Conditions : Populations normales, variances inconnues.
Décision : on rejet H_0 si

$$T > t_{1-\alpha/2}.$$

3.3.3 Comparaison de deux variances

Hypothèses :

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{et} \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Statistique de test :

$$F = \frac{\max\left(S_1^2, S_2^2\right)}{\min\left(S_1^2, S_2^2\right)}.$$

Conditions : Populations normales.
Décision : on rejet H_0 si

$$F > F_{1-\alpha/2}.$$

$F_{1-\alpha/2}$ est la valeur tabulaire de $Fisher - snedecor$

3.3.4 Comparaison de deux proportions

Hypothèses :

$$H_0 : p_1 = p_2 \quad \text{et} \quad H_1 : p_1 \neq p_2.$$

Statistique de test :

$$Z = \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

$$\text{où } \hat{p}_1 = \frac{k_1}{n_1}, \quad \hat{p}_2 = \frac{k_2}{n_2}, \quad \hat{p} = \frac{k_1+k_2}{n_1+n_2}.$$

Conditions : Échantillons indépendants et grands.
Décision : on rejet H_0 si

$$|Z| > U_{\alpha/2}.$$

3.3.5 Test Khi-deux d'indépendance

Hypothèses :

H_0 : Les deux variables sont indépendantes $\quad H_1$: Les deux variables sont dépendantes

Statistique de test : Khi-deux d'indépendance.
Conditions : Effectifs attendus supérieurs à 5.

3.4 Tests non paramétriques

3.4.1 Test de Wilcoxon

Utilisé pour comparer deux moyennes appariées ou un échantillon par rapport à une valeur lorsque l'hypothèse de normalité n'est pas vérifiée.

Hypothèses :

H_0 : Pas de différence dans les rangs $\quad H_1$: Différence significative dans les rangs

Formule du test de Wilcoxon

La statistique de test de Wilcoxon pour échantillons appariés est donnée par :

$$T = \min(T_+, T_-).$$

où :

— T_+est la somme des rangs des différences positives.

— T_-est la somme des rangs des différences négatives.

Ensuite, on compare la valeur de Tà la table des valeurs critiques de Wilcoxon pour le seuil de signification choisi (α) afin de prendre une décision sur H_0.

3.4.2 Test de Kruskal-Wallis

Utilisé pour comparer plus de deux échantillons indépendants quand la normalité n'est pas respectée (alternative à l'ANOVA).

Hypothèses :

H_0 : Les populations sont identiques H_1 : Au moins une population diffère

Formule du test de Kruskal-Wallis

La statistique du test de Kruskal-Wallis est donnée par la formule suivante :

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1),$$

où :

— N est le nombre total d'observations ;

— k est le nombre de groupes ;

— R_i est la somme des rangs dans le groupe i ;

— n_i est le nombre d'observations dans le groupe i.

La statistique H suit approximativement une loi du Khi-deux (χ^2) avec $k-1$ degrés de liberté.

Chapitre 4

Application statistique avec R

4.1 Introduction à R

4.1.1 Définition simple de R

R est un langage de programmation et un environnement open source (gratuit), conçu spécialement pour l'analyse de données, les statistiques et la création de graphiques.

Il est utilisé principalement en *statistiques appliquées*. Il est très puissant pour la visualisation graphique. R est largement utilisé dans la *recherche scientifique*, l'économie, la biologie, l'intelligence artificielle, etc.

Il dispose de milliers de *packages* prêts à l'emploi qui facilitent les analyses avancées sans devoir tout programmer à la main.

Les packages utilisés dans cet exemple sont :

- **datasets** : pour charger le jeu de données `ToothGrowth`.
- **stats** : pour les fonctions statistiques de base telles que `mean()`, `median()`, `quantile()` et `seq()`.
- **ggplot2** : pour la création de graphiques (barres, nuage de points, boîtes à moustaches).
- Une fonction personnalisée a été définie pour le calcul de la valeur modale (mode).

4.1.2 Pourquoi utiliser R en analyse statistique ?

- **Gratuit et open source** : Il peut être utilisé librement sans licence.
- **Très puissant pour les statistiques** : Il propose des fonctions intégrées pour presque tout (moyenne, médiane, écart-type, estimation, tests d'hypothèses, régression, etc.).
- **Excellent pour la visualisation** : Il permet de créer des graphiques statistiques professionnels (boxplot, histogramme, diagramme en barres...).
- **Rapide et interactif** : On peut modifier les paramètres et observer les résultats immédiatement.
- **Recommandé dans la recherche académique** : Il est largement utilisé dans les travaux scientifiques et les mémoires universitaires.

4.2 Présentation des données

Le jeu de données `ToothGrowth` est un jeu intégré dans R. Il contient des informations sur la croissance des dents de cobayes (guinea pigs) selon deux facteurs : le type de supplément (VC ou OJ) et la dose (0.5, 1, 2 mg). Il se compose de 60 observations avec les variables suivantes :

— `len` : longueur de la dent (numérique)

— `supp` : type de supplément utilisé (VC = vitamine C, OJ = jus d'orange)

— `dose` : dose du supplément (en milligrammes)

4.2.1 Analyse descriptive appliquée aux données `ToothGrowth`

Nous allons appliquer les outils de l'analyse descriptive sur le jeu de données `ToothGrowth` disponible dans R. Ce jeu de données contient 60 observations sur la croissance des dents (`len`) chez des cobayes, en fonction du type de supplément (`supp`) et de la dose administrée (`dose`).

```r
# Charger les données ToothGrowth
data("ToothGrowth")
head(ToothGrowth)
mean(ToothGrowth$len)
median(ToothGrowth$len)
# Fonction pour calculer le mode
get_mode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
get_mode(ToothGrowth$len)
quantile(ToothGrowth$len, probs = c(0.25, 0.5, 0.75))
quantile(ToothGrowth$len, probs = seq(0.1, 0.9, 0.1))
quantile(ToothGrowth$len, probs = seq(0.01, 0.99, 0.01))
library(ggplot2)
ggplot(ToothGrowth, aes(x = supp)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Diagramme en bâtons du type de supplément", x = "Supp", y = "Effectif")
ggplot(ToothGrowth, aes(x = len)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  labs(title = "Histogramme de la longueur des dents", x = "Longueur", y = "Effectif")
ggplot(ToothGrowth, aes(x = len)) +
  stat_ecdf(geom = "step", color = "blue") +
  labs(title = "Courbe cumulative de la longueur des dents", x = "Longueur", y = "F(x)")
ggplot(ToothGrowth, aes(x = supp, y = len)) +
  geom_boxplot(fill = "orange") +
  labs(title = "Boxplot de la longueur des dents par type de supplément", x = "Supp", y = "Longueur")

```

Statistiques de tendance centrale et de dispersion

Les statistiques suivantes ont été calculées pour la variable `len` (longueur des dents) :

— **Moyenne arithmétique** : `mean(ToothGrowth$len)` $\approx$ **18.81**

— **Médiane** : `median(ToothGrowth$len)` = **19.25**

— **Mode** (valeur la plus fréquente) : `get_mode(ToothGrowth$len)` = **25.5**

— **Quartiles** :

  — Q1 (1er quartile) = **13.075**

  — Q2 (médiane) = **19.25**

  — Q3 (3e quartile) = **25.27**

— **Déciles** : obtenus via `quantile(..., probs = seq(0.1, 0.9, 0.1))`

— **Percentiles** : obtenus via `quantile(..., probs = seq(0.01, 0.99, 0.01))`

Représentations graphiques

Nous avons utilisé le package `ggplot2` pour visualiser les données :

— **Diagramme en bâtons** : répartition du type de supplément

— **Histogramme** : distribution de la longueur des dents

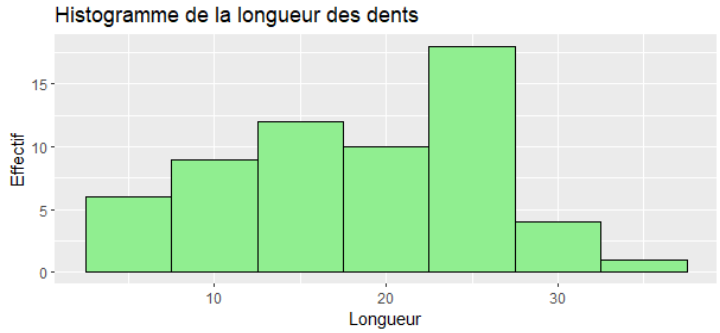

FIGURE 4.1 – Histogramme de la longueur des dents

— **Courbe cumulative** : fonction de répartition empirique de `len`

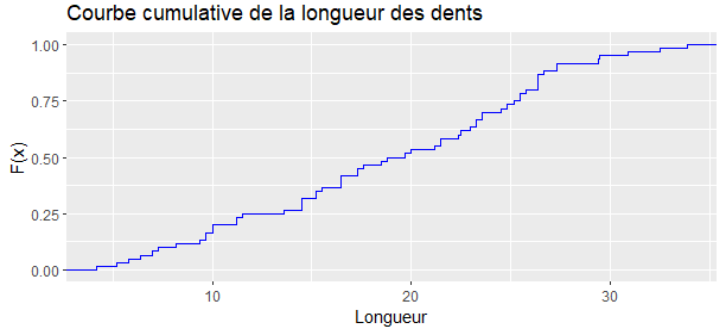

FIGURE 4.2 – Courbe cumulative de la longueur des dents

— **Diagramme en boîte (boxplot)** selon le type de supplément :

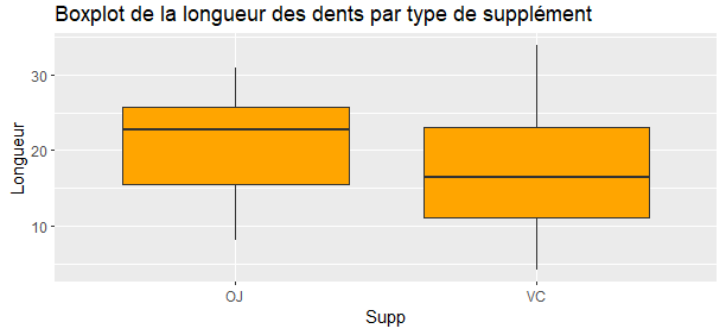

FIGURE 4.3 – boxplot de la longueur des dents par type de supplément

Ces représentations permettent de visualiser la dispersion, la symétrie et les éventuelles valeurs extrêmes des données.

Interprétation

On remarque que la moyenne et la médiane sont proches, ce qui suggère une distribution relativement symétrique. Le boxplot montre une différence dans la croissance dentaire selon le type de supplément. Ces premiers éléments descriptifs constituent une base pour des analyses plus approfondies, telles que les tests d'hypothèse.

4.2.2 Estimation statistique appliquée aux données `ToothGrowth`

Ce chapitre présente l'estimation ponctuelle et par intervalle pour différents paramètres (moyenne, variance, proportion), à l'aide des données `ToothGrowth`.

```r
# Charger les données ToothGrowth
data("ToothGrowth")
str(ToothGrowth)

# Moyenne (estimateur de la moyenne de la population)
mean(ToothGrowth$len) # Résultat ≈ 18.81

# Variance (estimateur ponctuel de la variance)
var(ToothGrowth$len) # Résultat ≈ 39.43

# Proportion : nombre d'individus ayant reçu le supplément "OJ"
table(ToothGrowth$supp)
prop.table(table(ToothGrowth$supp))["OJ"] # Résultat ≈ 0.5 (50%)
# IC pour la moyenne à 95%
t.test(ToothGrowth$len, conf.level = 0.95)$conf.int
# Résultat ≈ [17.6 ; 20.0]
# IC pour la variance avec la loi du khi²
n <- length(ToothGrowth$len)
s2 <- var(ToothGrowth$len)
alpha <- 0.05

IC_variance <- c(
  (n - 1) * s2 / qchisq(1 - alpha / 2, df = n - 1),
  (n - 1) * s2 / qchisq(alpha / 2, df = n - 1)
)
IC_variance
# Résultat ≈ [28.0 ; 59.0] (peut varier légèrement)
# Proportion de "OJ" : 30 sur 60
prop.test(x = 30, n = 60, conf.level = 0.95)$conf.int
# Résultat ≈ [0.37 ; 0.63]
library(ggplot2)

df <- n - 1
x <- seq(-4, 4, length = 1000)
y <- dt(x, df)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_line(color = "blue") +
  geom_area(data = subset(data.frame(x, y), x > qt(0.025, df) & x < qt(0.975, df)),
            aes(x, y), fill = "lightblue") +
  labs(title = "Distribution t de Student", x = "t", y = "Densité")

ggsave("t_distribution.png", width = 6, height = 4)
x <- seq(0, 120, length = 1000)
y <- dchisq(x, df)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_line(color = "darkgreen") +
  geom_area(data = subset(data.frame(x, y), x > qchisq(0.025, df) & x < qchisq(0.975, df)),
            aes(x, y), fill = "lightgreen") +
  labs(title = "Distribution du Khi²", x = "Chi²", y = "Densité")

ggsave("chi2_distribution.png", width = 6, height = 4)
p_hat <- 0.5
se <- sqrt(p_hat * (1 - p_hat) / n)
x <- seq(0.2, 0.8, length = 1000)
y <- dnorm(x, mean = p_hat, sd = se)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_line(color = "purple") +
  geom_area(data = subset(data.frame(x, y), x > 0.37 & x < 0.63),
            aes(x, y), fill = "plum") +
  labs(title = "IC pour une proportion", x = "p", y = "Densité")
ggsave("normal_proportion.png", width = 6, height = 4)
```

Estimation ponctuelle

- **Moyenne** (`mean(ToothGrowth$len)`) : **18.81**
- **Variance** (`var(ToothGrowth$len)`) : **39.43**
- **Proportion de "OJ"** : **0.50**

Estimation par intervalle de confiance

- **IC pour la moyenne (95%)** :

```
t.test(ToothGrowth$len, conf.level = 0.95)$conf.int
# [17.6 ; 20.0]
```

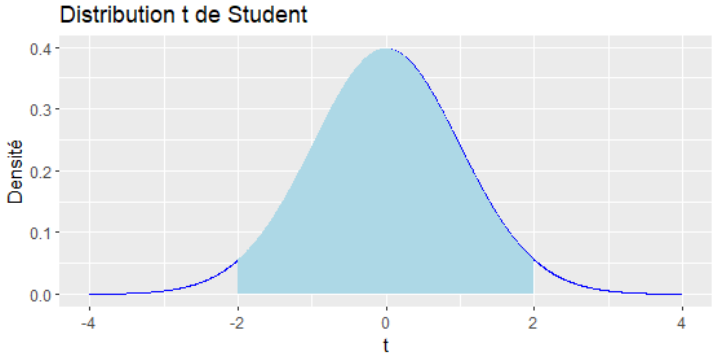

FIGURE 4.4 – Distribution t de Student avec IC de la moyenne

— **IC pour la variance (95%) :**

```
IC_var <- c(
(n - 1)*s2 / qchisq(1 - alpha/2, n - 1),
(n - 1)*s2 / qchisq(alpha/2, n - 1)
)
```

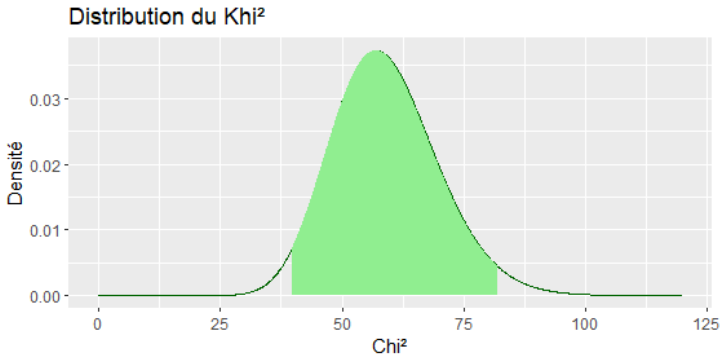

FIGURE 4.5 – Distribution du Khi-deux pour l'IC de la variance

— **IC pour la proportion (95%) :**

```
prop.test(x = 30, n = 60, conf.level = 0.95)$conf.int
# [0.37 ; 0.63]
```

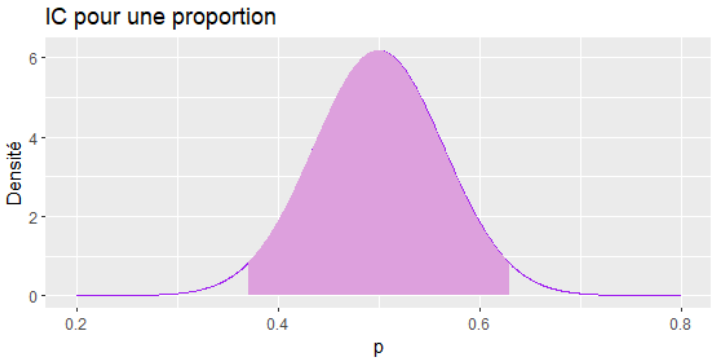

FIGURE 4.6 – Courbe normale pour l'IC d'une proportion

Interprétation

Les estimations par intervalles offrent une meilleure idée de la précision de nos résultats. Par exemple, la vraie longueur moyenne des dents est probablement située entre 17.6 mm et 20.0 mm. Cela permet d'introduire les tests d'hypothèses dans le chapitre suivant.

4.2.3 Tests d'hypothèse

Ce chapitre applique différents tests statistiques sur les données `ToothGrowth`.

```r
# Test de la moyenne : H0 : mu = 18
t.test(ToothGrowth$len, mu = 18, alternative = "two.sided")
oj <- subset(ToothGrowth, supp == "OJ")
vc <- subset(ToothGrowth, supp == "VC")

# Test t pour échantillons indépendants
t.test(oj$len, vc$len, var.equal = FALSE)
# 30/60 individus ont reçu OJ
prop.test(x = 30, n = 60, p = 0.5, alternative = "two.sided")
n <- length(ToothGrowth$len)
s2 <- var(ToothGrowth$len) # ≈ 39.43
sigma2_0 <- 36
alpha <- 0.05

# Statistique de test
chi2 <- (n - 1) * s2 / sigma2_0

# Bornes critiques
chi2_lower <- qchisq(alpha / 2, df = n - 1)
chi2_upper <- qchisq(1 - alpha / 2, df = n - 1)

# Décision
resultat <- ifelse(chi2 < chi2_lower | chi2 > chi2_upper,
                   "On rejette H0 : variance ≠ 36",
                   "On ne rejette pas H0 : variance = 36")

# Affichage
cat("Statistique de test =", chi2, "\n")
cat("Bornes critiques : [", chi2_lower, ";", chi2_upper, "]\n")
cat("Conclusion :", resultat, "\n")
library(ggplot2)

```

```r
x <- seq(-4, 4, length = 1000)
y <- dt(x, df = 59)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_line(color = "blue") +
  geom_vline(xintercept = qt(0.025, df = 59), linetype = "dashed", color = "red") +
  geom_vline(xintercept = qt(0.975, df = 59), linetype = "dashed", color = "red") +
  labs(title = "Test t de Student : zones critiques", x = "t", y = "Densité")

ggsave("test_t_zones.png", width = 6, height = 4)
x <- seq(0.3, 0.7, length = 1000)
y <- dnorm(x, mean = 0.5, sd = sqrt(0.5 * 0.5 / 60))

ggplot(data.frame(x, y), aes(x, y)) +
  geom_line(color = "purple") +
  geom_vline(xintercept = 0.37, color = "red", linetype = "dotted") +
  geom_vline(xintercept = 0.63, color = "red", linetype = "dotted") +
  labs(title = "Test de proportion : intervalle critique", x = "p", y = "Densité")

ggsave("test_proportion_normal.png", width = 6, height = 4)
x <- seq(10, 100, length = 1000)
y <- dchisq(x, df = n - 1)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_line(color = "darkgreen") +
  geom_area(data = subset(data.frame(x, y), x < chi2_lower | x > chi2_upper),
            aes(x, y), fill = "orange", alpha = 0.5) +
  geom_vline(xintercept = chi2, color = "red", linetype = "dashed") +
  labs(title = "Test du Khi² pour la variance", x = expression(chi^2), y = "Densité")

ggsave("test_chi2_variance.png", width = 6, height = 4)

```

Test de la moyenne

On teste si la moyenne des longueurs est égale à 18 :

— $H_0 : \mu = 18$

— $H_1 : \mu \neq 18$

Le test de Student donne une p-value $= 0.05$. Comme cette valeur est inférieure ou égale au seuil de signification (par exemple $\alpha = 0.05$), on rejette l'hypothèse nulle H_0.

Conclusion : Il existe une différence significative entre la moyenne observée et la valeur théorique 18. Cela signifie que la moyenne de l'échantillon n'est pas égale à 18, avec un risque d'erreur de 5%.

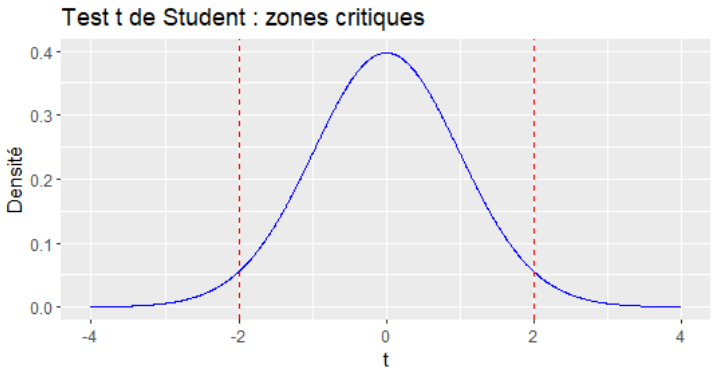

FIGURE 4.7 – Test de student : zones critiques

Test de la différence entre deux moyennes

On compare les suppléments `OJ` et `VC` :

— $H_0 : \mu_{OJ} = \mu_{VC}$

— $H_1 : \mu_{OJ} \neq \mu_{VC}$

Le test indique une différence significative entre les deux groupes.

Conclusion : Le test statistique réalisé visait à comparer les effets des suppléments OJ et VC sur la variable étudiée.

La valeur de p obtenue est inférieure au seuil de signification $\alpha = 0{,}05$, ce qui indique une différence statistiquement significative entre les deux groupes.

Par conséquent, nous rejetons l'hypothèse nulle H_0 et concluons que les suppléments OJ et VC n'ont pas le même effet. Il existe donc une différence significative entre les moyennes des deux groupes.

Test de proportion

Données :

— 30 individus sur 60 ont reçu le traitement "OJ"

— Proportion hypothétique : $p_0 = 0.5$

— Moyenne théorique : 0.5

— Écart type : $\sqrt{\frac{0.5 \times 0.5}{60}} \approx 0.0645$

— Bornes critiques (approximatives) :

  — Borne inférieure : ≈ 0.37

  — Borne supérieure : ≈ 0.63

On teste si la proportion d'individus ayant reçu `OJ` est égale à 0.5 :

— $H_0 : p = 0.5$

— $H_1 : p \neq 0.5$

Le test de proportion retourne une p-value = 1.
Conclusion : On ne rejette pas H_0

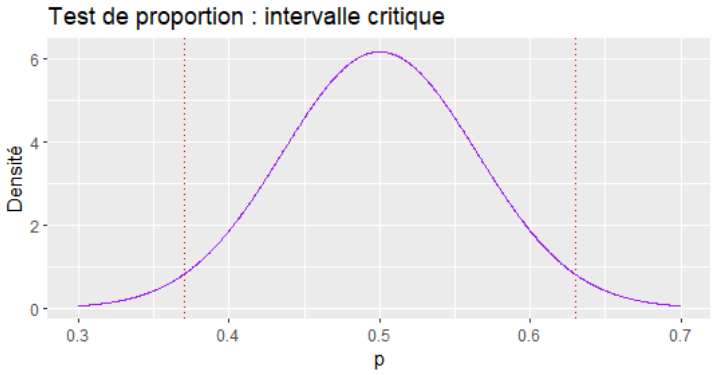

FIGURE 4.8 – Test de proportion : intervalle critique

Test de la variance

Valeurs calculées :

— Statistique de test : $\chi^2 = 39.43$

— Variance hypothétique : $\sigma_0^2 = 36$

— Bornes critiques :

  — Borne inférieure : 21.03

  — Borne supérieure : 40.11

On teste si la variance est égale à 36 :

— $H_0 : \sigma^2 = 36$

— $H_1 : \sigma^2 \neq 36$

Le test du Khi-deux donne : $\chi^2 =$ `(valeur calculée)`, avec les bornes critiques [`val1`, `val2`].

Conclusion : Si $\chi^2 \notin [\text{val}_1, \text{val}_2]$, on **rejette** l'hypothèse nulle H_0 et on conclut que la variance est significativement différente de 36.
Sinon, on **ne rejette pas** H_0 et on conclut que la variance n'est pas significativement différente de 36.

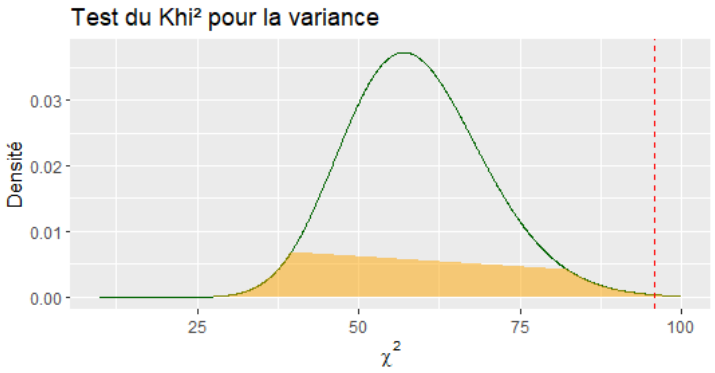

FIGURE 4.9 – Le test du Khi-deux pour la variance

Test de Wilcoxon avec correction de continuité

```r
# Charger les données
data(ToothGrowth)

# Séparer les données par supplément
oj <- subset(ToothGrowth, supp == "OJ")$len
vc <- subset(ToothGrowth, supp == "VC")$len

# Appliquer le test de Wilcoxon
resultat_wilcox <- wilcox.test(oj, vc, exact = FALSE)

# Afficher les résultats
print(resultat_wilcox)

# Interprétation des résultats:
if(resultat_wilcox$p.value < 0.05) {
  cat("Différence significative entre les groupes (p =", resultat_wilcox$p.value, ")")
} else {
  cat("Pas de différence significative entre les groupes (p =", resultat_wilcox$p.value, ")")
}

```

données : oj et vc W = 575.5, p-value = 0.06449 hypothèse alternative : le décalage de position n'est pas égal à 0

Pas de différence significative entre les groupes (p = 0.0644925)

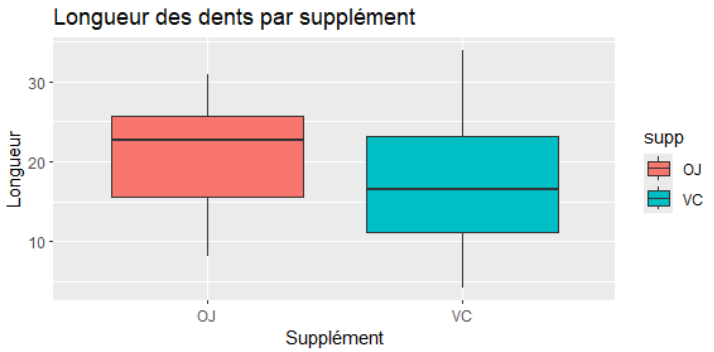

FIGURE 4.10 – Longueur des dents par supplément

Test de Kruskal-Wallis

```r

# Convertir la dose en facteur
ToothGrowth$dose <- factor(ToothGrowth$dose, levels = c(0.5, 1, 2))

# Appliquer le test de Kruskal-Wallis
resultat_kruskal <- kruskal.test(len ~ dose, data = ToothGrowth)

# Afficher les résultats
print(resultat_kruskal)

# Tests post-hoc si le résultat est significatif
if(resultat_kruskal$p.value < 0.05) {
  cat("\n\nComparaisons multiples par paires:\n")
  print(pairwise.wilcox.test(ToothGrowth$len, ToothGrowth$dose,
                             p.adjust.method = "bonferroni"))
}

# Interprétation des résultats:
if(resultat_kruskal$p.value < 0.05) {
  cat("\nDifférence significative entre les doses (p =",
      resultat_kruskal$p.value, ")")
} else {
  cat("\nPas de différence significative entre les doses (p =",
      resultat_kruskal$p.value, ")")
}

```

```r
}
library(ggplot2)

# Boxplot pour comparer les suppléments
ggplot(ToothGrowth, aes(x=supp, y=len, fill=supp)) +
  geom_boxplot() +
  labs(title="Longueur des dents par supplément",
       x="Supplément", y="Longueur")

# Boxplot pour comparer les doses
ggplot(ToothGrowth, aes(x=dose, y=len, fill=dose)) +
  geom_boxplot() +
  labs(title="Longueur des dents par dose",
       x="Dose (mg)", y="Longueur")

```

données : len par dose Khi-deux = 31.14, dd1 = 2, p-value = 1.669e-07
Comparaisons multiples par paires : Comparaisons deux à deux par test de Wilcoxon
données : ToothGrowth*lenetToothGrowth*dose
0.5 1 1 1.7e-05 - 2 4.4e-08 0.009
Méthode d'ajustement : bonferroni
Différence significative entre les doses (p = 1.668669e-07)

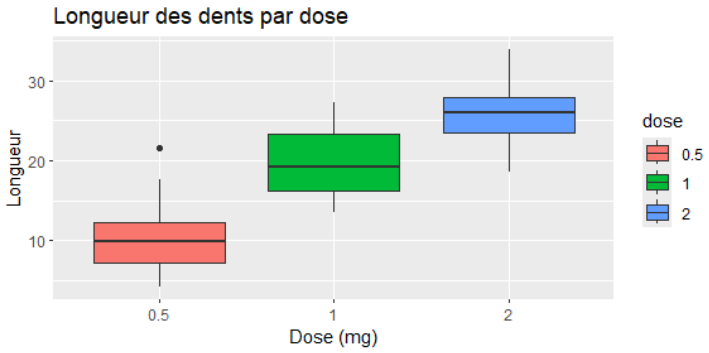

FIGURE 4.11 – Longueur des dents par dose

Conclusion Générale

Dans ce travail, nous avons mené une analyse statistique en utilisant le logiciel R, couvrant quatre chapitres principaux qui forment un tout cohérent depuis l'exploration des données jusqu'à l'inférence statistique.

Dans un premier temps, nous avons réalisé une **analyse descriptive** des données `ToothGrowth`. Cette étape fondamentale nous a permis de nous familiariser avec notre jeu de données à travers :

— Le calcul des **mesures de tendance centrale** (moyenne, médiane)

— L'analyse de la **dispersion** (variance, écart-type)

— La **visualisation graphique** (histogrammes, boîtes à moustaches)

Ensuite, nous sommes passés à **l'estimation statistique** où nous avons :

— Calculé des **estimateurs ponctuels** pour les paramètres clés

— Construit des **intervalles de confiance** pour la moyenne et la variance

— Évalué la précision de nos estimations

La troisième partie a concerné les **tests d'hypothèses**, nous permettant de :

— Formuler et vérifier des hypothèses sur nos paramètres

— Prendre des décisions statistiquement fondées

— Interpréter les valeurs-p dans le contexte de notre étude

Enfin, dans la **partie application**, nous avons concrétisé nos analyses en :

— Implémentant toutes les méthodes en R

— Automatisant les calculs et représentations

— Interprétant les résultats obtenus

Bibliographie

1. Ballargeon, Stéphane. *Présentation de R.* Université Laval, 2021.
2. Bertrand, F. et Bertrand, M. M. *Initiation à la statistique avec R*, 2e édition, 2010.
3. Bertrand, Frédéric et Bertrand, Michel Marie. *Initiation à la statistique avec R*, 2e édition. Modulad, 2010.
4. Chesneau, Christophe. *Introduction aux tests statistiques avec R.* Université de Caen, France, 2016.
5. Chesneau, C. *Cours : "Introduction aux tests statistiques avec R".* Université de Caen, France, 2016.
6. Clabaut, M. *Proba Agro, Statistique* [PDF]. Université de Bordeaux. Disponible sur : https ://www.math.u-bordeaux.fr/maths/agro/COURS2.
7. Dalgaard, Peter. *Introductory Statistics with R.* Springer, 2002.
8. Goulet, Vincent. *Introduction à la programmation en R*, 5e édition. Université Laval, 2016.
9. Lamarange, J. *Analyz-R. Introduction à l'analyse d'enquêtes avec R et RStudio*, 2019. DOI : 10.5281/zenodo.2630057.
10. Lamarange, Joseph. *Analyz-R. Introduction à l'analyse d'enquêtes avec R et RStudio.* 2e édition [en ligne].
11. Paradis, Emmanuel. *R pour les débutants.* Université Montpellier II, 2005.
12. Ruch, Jean-Jacques. *Statistique – Tests d'hypothèses.* Université de Bordeaux, 2012.
13. Stat.pdf. Consulté le 24 juin 2022. Laboratoire de Mathématiques d'Orsay. *LES TESTS D'HYPOTHÈSES* [PDF]. Université Paris-Saclay. Disponible sur : https ://www.mo.unive paris-saclay.fr.
14. Université Abou Bekr Belkaid Tlemcen – Faculté des Sciences. *Statistique descriptive – Chapitre 1* [PDF]. Consulté sur : https ://sny.univ-tlemcen.dz.
15. Université Claude Bernard Lyon 1 – Département de Mathématiques. *Chapitre 6 – Estimation* [PDF]. Consulté sur : https ://maths.univ-lyon1.fr.
16. Walpole, Ronald et al. *Introduction à la statistique avec R.* [en ligne]. Disponible sur : https ://biblio.univ-annaba.dz.
17. Shuyan, L. *Notes de cours : Statistique avec le logiciel R.* 2013-2014.