

Introduction à la Statistique

Octobre 2023

Contents

1	Concepts généraux	2
1.1	Introduction: le pouvoir des données	2
1.2	Différence entre statistique descriptive, inférentielle, et bayésienne	3
1.3	Notion d'échantillon	3
1.4	Différence entre variables qualitative et quantitative, ou discrète et continue	4
1.5	Différence entre statistique unidimensionnelle et multidimensionnelle	4
2	Estimation ponctuelle	5
2.1	Paramètres et statistiques, modèle paramétrique	5
2.2	Notion de modèle statistique	5
2.3	Estimateur statistique	6
2.4	Estimateur par la méthode des moments (EMM)	6
2.5	Méthode du maximum de vraisemblance	7
2.5.1	Fonction de vraisemblance	8
2.5.2	Définition de l'estimateur du maximum de vraisemblance (EMV)	8
2.6	Qualité d'un estimateur	9
2.6.1	Biais d'un estimateur	10
2.6.2	Variance d'un estimateur	10
2.6.3	Risque (quadratique) et ESBVM	10
2.6.4	Consistance (convergence) d'un estimateur	11
2.6.5	Information de Fisher	11
2.6.6	Borne de Cramer-Rao et efficacité	12
2.6.7	Exemple de la loi de Bernoulli	13
2.7	Propriétés des estimateurs	13
2.7.1	Le cas des EMM	13
2.7.2	Le cas des EMV	15
2.7.3	La méthode Delta	15
3	Estimation par intervalle	15
3.1	Introduction aux intervalles de confiance (IC) et définition	16
3.2	Notion de pivot	16
3.3	Exemples de construction	17
3.3.1	IC pour la moyenne d'une loi normale de variance connue	17
3.3.2	IC pour la variance d'une loi normale	17
3.3.3	IC pour une proportion	18

4	Tests paramétriques	19
4.1	Introduction et concepts généraux	20
4.2	Formalisation des tests paramétriques	20
4.2.1	Test d'hypothèses simples	20
4.2.2	Test d'hypothèses composites	20
4.3	Test sur la moyenne d'une gaussienne	20
4.3.1	Exemple introductif et test naïf	20
4.3.2	Amélioration du test	20
4.4	p-valeur d'un test	20
4.5	Lien entre test d'hypothèses et intervalle de confiance	20
4.6	Construction d'un test : généralisation	20
4.7	Test sur la variance d'une gaussienne	20
4.8	Test sur une proportion	20
4.9	Test de comparaison de deux échantillons	20
4.9.1	Echantillons gaussiens indépendants	20
4.9.2	Comparaison de proportions	21
5	Tests non-paramétriques	21
5.1	Rappels sur la fonction de répartition empirique	21
5.2	Test du χ^2	21
5.2.1	Test sur les probabilités d'événement	21
5.2.2	Test d'adéquation à une loi	21
5.2.3	Test d'adéquation à une famille de lois	21
5.2.4	Test d'indépendance entre deux variables	21
5.3	Test de Kolmogorov	21
5.4	Test de Wilcoxon-Mann-Whitney	21
5.5	Test de Wilcoxon	21
5.6	Test de Kruskal-Wallis	21

1 Concepts généraux

1.1 Introduction: le pouvoir des données

La statistique est l'art et la science de collecter, d'analyser, d'interpréter et de présenter des données. Elle est omniprésente dans notre vie quotidienne, que nous en soyons conscients ou non. Des entreprises aux gouvernements en passant par les chercheurs, la statistique est un outil puissant pour prendre des décisions éclairées, résoudre des problèmes complexes et comprendre le monde qui nous entoure.

Au cœur de la statistique réside la capacité à extraire des informations significatives à partir d'un ensemble de données. Que vous soyez un économiste étudiant les tendances du marché, un épidémiologiste suivant la propagation d'une maladie, ou simplement un consommateur comparant les prix des produits en ligne, la statistique joue un rôle central dans la transformation des données en connaissances utiles.

La statistique se divise en plusieurs domaines, dont l'estimation, les tests d'hypothèses, l'analyse de régression, et bien d'autres. Elle englobe également des concepts fondamentaux tels que la théorie des probabilités, les indicateurs statistiques de variance, médiane, moyenne, qui nous aident à résumer, à interpréter et à tirer des conclusions à partir de données brutes. L'une des caractéristiques les plus remarquables de la statistique est son pouvoir prédictif. En utilisant des méthodes statistiques appropriées, nous pouvons faire des prévisions sur des événements futurs, identifier des tendances émergentes et même prendre des décisions stratégiques dans divers domaines.

En fin de compte, la statistique est une discipline polyvalente qui transcende les frontières de nombreuses professions. Que vous soyez un scientifique des données, un économiste, un psychologue, un biologiste ou un étudiant curieux, la statistique vous fournira des outils essentiels pour comprendre le monde complexe qui nous entoure. Ce cours d'introduction à la statistique vous aidera à acquérir les compétences et les connaissances nécessaires pour maîtriser cet art de la donnée et l'appliquer dans votre domaine d'intérêt.

1.2 Différence entre statistique descriptive, inférentielle, et bayésienne

Les statistiques descriptives, inférentielles et bayésiennes sont trois approches distinctes de l'analyse statistique des données, chacune ayant ses propres objectifs et méthodes. Voici la différence entre ces trois types de statistiques :

1. statistique descriptive : la statistique descriptive vise à décrire et résumer des données en utilisant des techniques simples, telles que des mesures de tendance centrale (moyenne, médiane, mode), des mesures de dispersion (écart-type, plage interquartile), des tableaux, des graphiques et des résumés numériques. Elle est principalement utilisée pour comprendre les caractéristiques fondamentales d'un ensemble de données et pour présenter des informations de manière concise et compréhensible.
2. statistique inférentielle : elle est utilisée pour tirer des conclusions générales à partir d'un échantillon de données et les appliquer à une population plus large. Elle repose sur des probabilités et des tests d'hypothèses. Elle est couramment utilisée pour faire des déclarations sur une population à partir d'un échantillon, comme l'estimation d'une moyenne, la comparaison de groupes, la vérification d'hypothèses, ...
3. statistique bayésienne : c'est une approche basée sur la théorie bayésienne en probabilité. Elle permet de mettre à jour des croyances ou des hypothèses à mesure que de nouvelles données deviennent disponibles, en utilisant une distribution de probabilité a priori et la mise à jour avec des données observées. Elle est souvent utilisée dans des domaines tels que la modélisation statistique, l'apprentissage automatique, la prise de décision, et d'autres situations où l'on souhaite tenir compte de l'incertitude et mettre à jour des informations en fonction de nouvelles données.

En résumé, la statistique descriptive se concentre sur la présentation des données, la statistique inférentielle sur l'inférence à partir d'un échantillon pour faire des généralisations sur une population, tandis que la statistique bayésienne repose sur la mise à jour des croyances ou des probabilités en utilisant des données observées. Les trois approches sont importantes dans des contextes différents et complémentaires en statistique.

1.3 Notion d'échantillon

En statistique, un échantillon est un sous-ensemble représentatif d'une population plus large. La population fait référence à l'ensemble complet de toutes les unités ou individus qui sont d'intérêt pour une étude ou une analyse statistique. Puisque l'étude de l'ensemble complet de la population est souvent difficile, coûteuse ou même impossible, on utilise des échantillons pour en tirer des conclusions générales sur la population entière. Voici quelques points clés concernant la notion d'échantillon :

- objectif : L'objectif de l'échantillonnage est de prélever un groupe plus restreint et gérable d'individus ou d'éléments de la population, de manière à ce qu'il soit représentatif de la population plus large. Cela permet d'extrapoler ou d'inférer des caractéristiques, des tendances ou des paramètres de la population entière à partir de l'analyse de l'échantillon.
- méthodes d'échantillonnage : Il existe différentes méthodes d'échantillonnage, notamment l'échantillonnage aléatoire simple, l'échantillonnage stratifié, l'échantillonnage par grappes, et d'autres méthodes spécifiques en fonction des objectifs de l'étude.
- taille de l'échantillon : La taille de l'échantillon doit être suffisamment grande pour obtenir des résultats fiables, mais elle dépend également de la variabilité des données et des objectifs de l'étude. Les statisticiens utilisent des calculs de puissance pour déterminer la taille de l'échantillon nécessaire.
- représentativité : Il est essentiel que l'échantillon soit représentatif de la population. Cela signifie que les caractéristiques clés de l'échantillon (âge, sexe, lieu, etc.) doivent refléter la répartition de ces caractéristiques dans la population.
- biais d'échantillonnage : Il est important de minimiser tout biais dans la sélection de l'échantillon, car un biais peut entraîner des conclusions inexactes. Les biais peuvent survenir si la méthode d'échantillonnage ou la collecte de données est défectueuse.

- utilisation de l'échantillon : Une fois l'échantillon prélevé, des analyses statistiques sont effectuées pour tirer des conclusions sur la population plus large. Cela peut inclure des estimations de paramètres, des tests d'hypothèses et d'autres analyses.

Un échantillon en statistique est un sous-groupe soigneusement sélectionné de la population qui est utilisé pour effectuer des analyses statistiques et faire des généralisations sur l'ensemble de la population. Une sélection appropriée de l'échantillon est cruciale pour obtenir des résultats fiables et significatifs. D'un point de vue probabiliste, on parlera souvent en statistique d'échantillon lorsque les données sont issues de réalisations de variables aléatoires indépendantes et identiquement distribuées (iid).

1.4 Différence entre variables qualitative et quantitative, ou discrète et continue

Les variables en statistiques sont généralement classées en deux catégories principales : qualitatives/quantitatives et discrètes/continues. Voici la différence entre ces deux paires de catégories :

- Variables Qualitatives et Quantitatives :
 - Variable Qualitative (ou Catégorielle) : Une variable qualitative prend des valeurs qui sont des catégories ou des étiquettes, et ces catégories ne sont pas nécessairement numériques. Les variables qualitatives décrivent généralement des caractéristiques ou des attributs qui ne peuvent pas être mesurés numériquement. Par exemple, le genre (homme/femme), la couleur des yeux (bleu/marron/vert), la catégorie de produit (électronique/vêtements/produits alimentaires) sont des exemples de variables qualitatives.
 - Variable Quantitative : Une variable quantitative prend des valeurs numériques et mesure des quantités. Elle peut être continue ou discrète. Les variables quantitatives permettent d'effectuer des opérations mathématiques telles que l'addition, la soustraction, la multiplication, etc. Par exemple, l'âge, le poids, le revenu, le nombre d'heures de sommeil sont des exemples de variables quantitatives.
- Variables Discrètes et Continues :
 - Variable Discrète : Une variable discrète prend des valeurs distinctes et isolées, souvent entières, avec des sauts entre les valeurs. Par exemple, le nombre de voitures dans un parking, le nombre d'enfants dans une famille, le nombre de pièces de monnaie dans votre portefeuille sont des exemples de variables discrètes.
 - Variable Continue : Une variable continue peut prendre une gamme infinie de valeurs réelles dans un intervalle donné. Ces valeurs peuvent inclure des fractions ou des décimales, et il n'y a pas de sauts ou de discontinuités. Par exemple, la taille d'une personne, la température, le temps nécessaire pour effectuer une tâche sont des exemples de variables continues.

La principale distinction réside donc dans le fait que les variables qualitatives sont des catégories ou des étiquettes, tandis que les variables quantitatives sont des mesures numériques. De plus, les variables quantitatives peuvent être soit discrètes (valeurs distinctes) soit continues (valeurs sur une plage continue). Ces distinctions sont essentielles pour choisir les bonnes méthodes statistiques d'analyse en fonction du type de variable que vous traitez.

1.5 Différence entre statistique unidimensionnelle et multidimensionnelle

La différence entre la statistique unidimensionnelle et multidimensionnelle réside dans le nombre de variables prises en compte dans l'analyse statistique:

- statistique unidimensionnelle (ou statistique univariée) : elle concerne l'analyse de données comportant une seule variable. Elle se concentre sur la distribution, les caractéristiques et les propriétés d'une seule variable à la fois. Exemples : Une analyse unidimensionnelle pourrait impliquer l'examen de la distribution des notes d'un examen, de la fréquence des réponses "oui" et "non" à une question, ou de la moyenne des temps de réaction pour un groupe de personnes.

- statistique multidimensionnelle (ou statistique multivariée) : elle concerne l'analyse de données qui comportent plusieurs variables simultanément. Elle explore les relations, les corrélations et les interactions entre ces variables. Exemples : Une analyse multidimensionnelle pourrait impliquer l'examen des relations entre les variables telles que les revenus, l'éducation et l'âge dans une enquête sociodémographique. Elle peut également inclure des techniques d'analyse factorielle, de régression multivariée, etc.

La statistique unidimensionnelle se concentre ainsi sur une seule variable à la fois, tandis que la statistique multidimensionnelle explore les relations entre plusieurs variables simultanément. L'analyse multidimensionnelle est généralement plus complexe car elle permet de mieux comprendre les interactions entre les variables et de répondre à des questions plus complexes, mais elle nécessite des méthodes statistiques plus avancées pour être correctement appliquée.

2 Estimation ponctuelle

Dans ce cours, nous allons explorer particulièrement le domaine de l'estimation paramétrique, ainsi que certains tests statistiques. L'estimation paramétrique est une approche statistique qui vise à estimer les paramètres inconnus d'une distribution de probabilité à partir d'un échantillon de données observées.

2.1 Paramètres et statistiques, modèle paramétrique

Avant d'entrer dans les détails de l'estimation paramétrique, il est important de comprendre la distinction entre les paramètres et les statistiques. Les paramètres sont des caractéristiques de la distribution de probabilité sous-jacente, tandis que les statistiques sont des mesures calculées à partir des données observées.

Un modèle paramétrique est une représentation mathématique simplifiée d'un phénomène basée sur des hypothèses spécifiques concernant la distribution sous-jacente des données. Les modèles paramétriques jouent un rôle central dans l'analyse statistique, car ils permettent de réduire la complexité des données tout en fournissant des outils puissants pour la compréhension, l'estimation et la prise de décision. L'idée fondamentale derrière un modèle paramétrique est de supposer que les données observées suivent une distribution de probabilité spécifique, caractérisée par un ensemble de paramètres inconnus. Ces paramètres décrivent des propriétés clés de la distribution, telles que la moyenne, la variance, la médiane, etc. En d'autres termes, le modèle présume que les données sont générées à partir d'une famille de distributions bien définie, chaque membre de cette famille étant déterminé par des valeurs spécifiques des paramètres.

Plutôt que de traiter directement avec des observations individuelles, les statisticiens peuvent utiliser ces modèles pour résumer ces données en un ensemble gérable de paramètres. Cela facilite la description, la comparaison et l'interprétation des données. Un exemple courant de modèle paramétrique est le modèle de régression linéaire, qui suppose que la relation entre une variable dépendante et plusieurs variables indépendantes suit une forme linéaire. Dans ce modèle, les paramètres sont les coefficients de la régression, et ils déterminent la pente et l'ordonnée à l'origine de la droite de régression.

Les modèles paramétriques permettent également de réaliser des prévisions et des estimations. En ajustant les paramètres du modèle aux données observées, il est possible d'estimer les valeurs inconnues de ces paramètres, ce qui peut être utilisé pour faire des prédictions ou des inférences sur le phénomène sous-jacent. Cependant, il est essentiel de noter que les modèles paramétriques reposent sur des hypothèses spécifiques concernant la distribution des données. Si ces hypothèses ne sont pas correctes, les conclusions basées sur le modèle peuvent être biaisées ou incorrectes. Par conséquent, il est crucial de choisir un modèle approprié en fonction des données et de vérifier la validité des hypothèses sous-jacentes.

2.2 Notion de modèle statistique

Définition 1. *Un modèle statistique est généralement composé de trois éléments principaux, que l'on peut représenter sous forme d'un triplet :*

$$(\mathcal{E}, \Theta, \mathcal{P}) \tag{1}$$

où :

\mathcal{E} est l'espace des observations,

Θ est l'espace des paramètres du modèle

\mathcal{P} est la famille de lois de probabilité qui décrit le comportement des données dans l'espace des observations.

Dans ce triplet, \mathcal{E} représente l'espace dans lequel vos données existent, Θ contient les paramètres que vous cherchez à estimer, et \mathcal{P} spécifie comment les données sont distribuées dans cet espace en fonction des paramètres.

2.3 Estimateur statistique

Un estimateur statistique est une fonction des observations utilisée pour estimer le paramètre inconnu d'une distribution de probabilité, à partir d'un échantillon de données observées. Il permet d'inférer des informations sur des caractéristiques de populations ou de distributions à partir des données échantillonnées. Un estimateur statistique est utilisé lorsque nous ne connaissons pas la valeur réelle d'un paramètre (comme la moyenne, la variance, la proportion, ...) de la population ou de la distribution que nous voulons étudier. Son objectif est de fournir une meilleure approximation du paramètre inconnu à partir des données de l'échantillon.

Exemple 1. *Un exemple courant d'estimateur est la moyenne empirique pour estimer la moyenne d'une population. Si vous avez un échantillon de données, la moyenne de cet échantillon est souvent utilisée comme estimateur de la moyenne de la population.*

Définition 2. *Un estimateur statistique est défini comme suit :*

$$\hat{\theta} = g(X_1, X_2, \dots, X_n) \quad (2)$$

où :

$\hat{\theta}$ est l'estimateur du paramètre inconnu θ ,

g est une fonction (règle) mathématique appliquée aux données de l'échantillon (X_1, X_2, \dots, X_n) .

La fonction g prend en entrée les données de l'échantillon (X_1, X_2, \dots, X_n) et produit l'estimation $\hat{\theta}$.

2.4 Estimateur par la méthode des moments (EMM)

La méthode des moments est une technique couramment utilisée en statistiques et en théorie de la probabilité pour estimer les paramètres d'une distribution de probabilité. Elle repose sur le principe de faire correspondre les moments observés (les valeurs de données) aux moments théoriques d'une distribution probabiliste en ajustant les paramètres. Les étapes de mise en oeuvre de cette technique sont les suivantes:

1. collecte des données,
2. choix de la distribution : on suppose une distribution de probabilité,
3. calcul des moments empiriques : on calcule les moments empiriques à partir des données,
4. calcul des moments théoriques : on calcule les moments théoriques de la distribution choisie en fonction des paramètres qu'on souhaite estimer,
5. équation des moments : on égalise les moments empiriques aux moments théoriques par une transformation, en fonction des paramètres. Cela crée un système d'équations à résoudre pour estimer les paramètres inconnus,
6. résolution des équations : on résout le système d'équations pour obtenir les estimations des paramètres,

La méthode des moments est relativement simple à comprendre et à appliquer, en particulier pour les distributions de probabilité bien connues. Cependant, il est important de noter que la méthode des moments repose sur des suppositions sur la forme de la distribution, et elle peut ne pas être la meilleure méthode d'estimation dans tous les cas.

Exemple 2. Supposons que nous ayons un échantillon aléatoire de données X_1, X_2, \dots, X_n provenant d'une distribution exponentielle avec une fonction de densité de probabilité :

$$f(x|\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0, \lambda > 0$$

où λ est le paramètre d'échelle que nous souhaitons estimer.

Le moment d'ordre 1 de cette distribution est donné par :

$$E(X) = \int_0^\infty x \lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

Pour estimer le paramètre d'échelle λ par la méthode des moments, nous égalons le moment d'ordre 1 empirique (la moyenne de l'échantillon) au moment d'ordre 1 théorique :

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

En résolvant cette équation pour λ , nous obtenons l'estimateur par la méthode des moments du paramètre d'échelle de la distribution exponentielle :

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

où \bar{X} est la moyenne empirique de l'échantillon.

2.5 Méthode du maximum de vraisemblance

La méthode du maximum de vraisemblance (Maximum Likelihood Estimation, abrégée en MLE en anglais) est une technique couramment utilisée en statistiques pour estimer les paramètres d'une distribution de probabilité ou d'un modèle statistique en se basant sur les données observées. L'objectif est de trouver les valeurs des paramètres qui rendent les données observées les plus "vraisemblables" sous un modèle donné. Ses étapes se résument ainsi:

1. modèle statistique : on choisit un modèle statistique qui dépend de paramètres inconnus à estimer. Il décrit la distribution des données,
2. fonction de Vraisemblance : on écrit la fonction de vraisemblance (likelihood function) qui décrit la probabilité des données sous le modèle, en fonction des paramètres à estimer,
3. log-vraisemblance : on travaille souvent avec la log-vraisemblance (logarithme de la vraisemblance) plutôt qu'avec la vraisemblance elle-même, car cela simplifie les calculs,
4. estimation des paramètres : on cherche les valeurs des paramètres qui maximisent la log-vraisemblance. En d'autres termes, on trouve les paramètres pour lesquels les données sont les plus "vraisemblables" ou les plus probables,
5. optimisation : on utilise des techniques mathématiques telles que la dérivation, la résolution numérique ou d'autres méthodes d'optimisation. Le résultat est l'estimation des paramètres MLE.

Cette technique possède des propriétés statistiques souhaitables, notamment une efficacité asymptotique, ce qui signifie que pour de grands échantillons, les estimateurs MLE sont non biaisés et atteignent la limite de Cramér-Rao (cf plus loin dans le cours).

2.5.1 Fonction de vraisemblance

La fonction de vraisemblance est utilisée pour mesurer à quel point les données observées sont "vraisemblables" sous un modèle statistique avec un paramètre donné. Elle est une fonction du paramètre θ , et son rôle est de décrire comment la probabilité des données varie en fonction de θ .

Définition 3. La fonction de vraisemblance (likelihood function) est définie comme suit :

$$\mathcal{L}(\theta|x) = f(x|\theta) \quad (3)$$

où :

$\mathcal{L}(\theta|x)$ est la fonction de vraisemblance,

θ est le paramètre que nous cherchons à estimer,

x sont les données observées,

$f(x|\theta)$ est la fonction de densité (ou de probabilité) des données sous le modèle avec le paramètre θ .

Cette définition se décline différemment dans le cas d'un modèle discret ou d'un modèle continu. La probabilité apparaissant dans la définition est une probabilité multivariée puisqu'elle est liée à un échantillon de taille n . Ainsi, dans le cas discret, le terme $f(x|\theta)$ sera remplacé par

$$P(X = x|\theta) = P((X_1 = x_1, \dots, X_n = x_n)|\theta).$$

2.5.2 Définition de l'estimateur du maximum de vraisemblance (EMV)

L'estimateur du maximum de vraisemblance est la valeur du paramètre θ qui maximise la fonction de vraisemblance. En d'autres termes, il s'agit de l'estimation qui rend les données observées les plus "vraisemblables" sous un modèle probabiliste avec le paramètre θ .

Définition 4. L'estimateur du maximum de vraisemblance (EMV) est défini comme suit :

$$\hat{\theta}_{EMV} = \arg \max_{\theta} \mathcal{L}(\theta|X) \quad (4)$$

où :

$\hat{\theta}_{EMV}$ est l'estimateur du maximum de vraisemblance,

θ est le paramètre que nous cherchons à estimer,

$\mathcal{L}(\theta|X)$ est la fonction de vraisemblance,

X sont les données de l'échantillon $X = (X_1, \dots, X_n)$.

Comme évoqué précédemment, cet estimateur possède d'excellentes propriétés asymptotiques (lorsque $n \rightarrow +\infty$), qui seront vues ultérieurement dans ce cours.

Exemple 3. Supposons que nous ayons un échantillon de données binaires, notées X_1, X_2, \dots, X_n , où X_i vaut 1 en cas de succès et 0 en cas d'échec. Ces données suivent une distribution de Bernoulli avec une probabilité de succès p .

La fonction de vraisemblance pour un échantillon de données binaires est la suivante :

$$\mathcal{L}(p|x_1, x_2, \dots, x_n) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$$

La log-vraisemblance est donc :

$$\ln \mathcal{L}(p|x_1, x_2, \dots, x_n) = \sum_{i=1}^n [x_i \ln(p) + (1-x_i) \ln(1-p)]$$

Pour estimer la probabilité de succès p par la méthode du maximum de vraisemblance, nous maximisons la log-vraisemblance en trouvant la dérivée par rapport à p et en la mettant égale à zéro :

$$\frac{d}{dp} \ln \mathcal{L}(p|x_1, x_2, \dots, x_n) = \frac{1}{p} \sum_{i=1}^n x_i - \frac{1}{1-p} \sum_{i=1}^n (1-x_i) = 0$$

En résolvant cette équation, nous obtenons l'estimation MLE de p :

$$\hat{p}_{EMV} = \frac{1}{n} \sum_{i=1}^n x_i$$

C'est-à-dire que l'estimation MLE de la probabilité de succès est la moyenne empirique des données binaires.

Cet exemple nous fournit une estimation par maximum de vraisemblance du paramètre p .

Remarque 1. Il est fondamental de comprendre que l'estimateur du maximum de vraisemblance est une variable aléatoire, au contraire de l'estimation qui en fournit une valeur calculée. Dans l'exemple précédent, l'estimateur EMV s'écrit donc

$$\hat{p}_{EMV} = \frac{1}{n} \sum_{i=1}^n X_i.$$

2.6 Qualité d'un estimateur

La qualité d'un estimateur statistique fait référence à sa capacité à fournir des estimations précises et fiables des paramètres d'une distribution de probabilité ou d'un modèle statistique. Une bonne qualité d'estimateur repose sur plusieurs propriétés et critères essentiels :

- le biais,
- l'efficacité,
- l'efficacité asymptotique : certains estimateurs sont plus efficaces que d'autres lorsque la taille de l'échantillon est grande. L'efficacité asymptotique mesure la rapidité avec laquelle un estimateur converge vers la vraie valeur du paramètre.
- la consistance,
- la normalité asymptotique,
- robustesse : un estimateur robuste est celui qui est peu influencé par les valeurs aberrantes ou les erreurs de mesure. Les estimateurs robustes conservent leur qualité même en présence de données atypiques.
- la similitude : un estimateur est dit invariant s'il conserve sa qualité lorsqu'il est utilisé pour estimer un paramètre après une transformation des données.
- bonne performance en petite taille d'échantillon.

En fin de compte, la qualité d'un estimateur dépend des caractéristiques spécifiques de la distribution de probabilité sous-jacente et des objectifs de l'analyse. Dans de nombreuses situations, le choix d'un estimateur dépendra de l'équilibre entre ces différentes propriétés et de la pertinence par rapport au problème statistique particulier que vous essayez de résoudre. Nous allons approfondir dans ce cours les différentes notions évoquées ci-dessus.

2.6.1 Biais d'un estimateur

Le biais d'un estimateur statistique est une mesure de la tendance systématique de l'estimateur à s'éloigner de la vraie valeur du paramètre que vous essayez d'estimer. En d'autres termes, le biais quantifie l'erreur moyenne de l'estimateur. Si un estimateur est non biaisé, cela signifie que, en moyenne, il fournit des estimations qui sont exactes et ne s'écartent pas de la vraie valeur du paramètre.

Définition 5. *Le biais d'un estimateur est défini comme la différence entre l'espérance de cet estimateur et la vraie valeur du paramètre que vous essayez d'estimer.*

$$\text{Biais}(\hat{\theta}) = E(\hat{\theta}) - \theta \quad (5)$$

où :

$\text{Biais}(\hat{\theta})$ est le biais de l'estimateur $\hat{\theta}$,

$E(\hat{\theta})$ est l'espérance (moyenne) de l'estimateur $\hat{\theta}$,

θ est la vraie valeur du paramètre que vous cherchez à estimer.

2.6.2 Variance d'un estimateur

La variance mesure la dispersion des valeurs de l'estimateur autour de sa moyenne. Une variance faible indique que les estimations de l'estimateur sont regroupées étroitement autour de sa moyenne, ce qui suggère une plus grande précision. À l'inverse, une variance élevée indique une dispersion plus grande des estimations, ce qui suggère une moindre précision.

Définition 6. *La variance d'un estimateur est définie comme la mesure de la dispersion ou de la variabilité de cet estimateur. Mathématiquement, la variance s'exprime comme suit :*

$$\text{Var}(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}))^2] \quad (6)$$

où :

$\text{Var}(\hat{\theta})$ est la variance de l'estimateur $\hat{\theta}$,

$\hat{\theta}$ est l'estimateur que vous cherchez à évaluer,

$E(\hat{\theta})$ est l'espérance (moyenne) de l'estimateur $\hat{\theta}$.

2.6.3 Risque (quadratique) et ESBVM

Le risque quadratique d'un estimateur, également connu sous le nom d'erreur quadratique moyenne (EQM) ou mean squared error (MSE) en anglais, est une mesure de la précision de cet estimateur. Il combine deux aspects importants de la qualité d'un estimateur : le biais et la variance. Le risque quadratique est défini comme la moyenne de l'erreur quadratique, c'est-à-dire la moyenne des carrés des écarts entre les estimations de l'estimateur et les vraies valeurs du paramètre que vous essayez d'estimer.

Définition 7. *Le risque (quadratique) d'un estimateur vaut:*

$$\text{Risque Quadratique}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] \quad (7)$$

où :

$\text{Risque Quadratique}(\hat{\theta})$ est le risque quadratique de l'estimateur $\hat{\theta}$,

$\hat{\theta}$ est l'estimateur que vous cherchez à évaluer,

θ est la vraie valeur du paramètre que vous essayez d'estimer.

Le risque quadratique est une mesure globale de la précision de l'estimateur, prenant en compte à la fois le biais (tendance systématique) et la variance (variabilité aléatoire) de l'estimateur. Un risque quadratique plus faible indique une meilleure précision de l'estimateur. Un ESBVM est un estimateur sans biais et de variance minimale, c'est donc le meilleur estimateur que nous puissions obtenir, mais il n'existe pas toujours!

2.6.4 Consistance (convergence) d'un estimateur

La consistance d'un estimateur est une propriété importante en statistiques qui se réfère à son comportement lorsque la taille de l'échantillon augmente. En d'autres termes, un estimateur est considéré comme consistant s'il converge vers la vraie valeur du paramètre que vous essayez d'estimer à mesure que la taille de l'échantillon devient de plus en plus grande. Plus précisément, un estimateur est dit consistant si, pour toute vraie valeur du paramètre, la probabilité que l'estimateur s'éloigne de cette vraie valeur autant que possible tend vers zéro à mesure que la taille de l'échantillon augmente.

Définition 8. *Un estimateur est dit consistant s'il satisfait la condition suivante à mesure que la taille de l'échantillon augmente :*

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0 \quad (8)$$

où :

$\hat{\theta}$ est l'estimateur que vous cherchez à évaluer,

θ est la vraie valeur du paramètre que vous essayez d'estimer,

n est la taille de l'échantillon,

ϵ est une petite valeur positive,

$P(|\hat{\theta} - \theta| > \epsilon)$ est la probabilité que l'estimateur s'éloigne de la vraie valeur du paramètre de plus de ϵ .

La consistance signifie que, à mesure que la taille de l'échantillon augmente, la probabilité que l'estimateur s'éloigne de la vraie valeur du paramètre autant que possible tend vers zéro. En d'autres termes, l'estimateur devient de plus en plus précis et se rapproche de la vraie valeur du paramètre lorsque vous disposez de plus de données.

2.6.5 Information de Fisher

L'information de Fisher, également appelée information de Fisher-Rao, est une mesure fondamentale en statistiques qui quantifie la quantité d'information contenue dans une distribution de probabilité par rapport à un paramètre que vous essayez d'estimer. L'information de Fisher au point θ est définie comme la variance de la dérivée première du logarithme de la fonction de vraisemblance par rapport au paramètre θ .

Définition 9. *Statistiquement, l'information de Fisher est définie comme suit:*

$$I(\theta) = \text{Var} \left[\frac{d}{d\theta} \ln L(\theta) \right] \quad (9)$$

où :

$I(\theta)$ est l'information de Fisher au point θ ,

E représente l'espérance (moyenne),

$\frac{d^2}{d\theta^2} \ln L(\theta)$ est la deuxième dérivée par rapport à θ du logarithme de la fonction de vraisemblance $L(\theta)$.

Elle mesure la courbure de la fonction de vraisemblance autour du paramètre θ et est utilisée pour calculer la variance asymptotique des estimateurs du maximum de vraisemblance. Elle joue un rôle essentiel en statistiques pour évaluer la précision des estimateurs, calculer des intervalles de confiance, et effectuer des tests d'hypothèses. En bref, l'information de Fisher est un concept fondamental dans la théorie statistique et l'inférence statistique. L'information de Fisher a plusieurs propriétés:

- elle est toujours positive ou nulle : $I(\theta) \geq 0$,
- elle est maximale lorsque la fonction de vraisemblance est la plus informative à propos du paramètre θ .

- elle est utilisée pour calculer la variance asymptotique des estimateurs EMV (en anglais MLE (Maximum Likelihood Estimators)),
- il existe deux autres expressions de l'information de Fisher, qui découlent de sa définition:

$$I(\theta) = E \left[\left(\frac{d}{d\theta} \ln L(\theta) \right)^2 \right], \quad (10)$$

et

$$I(\theta) = -E \left[\frac{d^2}{d\theta^2} \ln L(\theta) \right]. \quad (11)$$

Cette dernière expression est souvent utilisée dans les calculs. Dans le cas de données indépendantes, il est facile de montrer que

$$I(\theta) = nI_1(\theta),$$

où $I(\theta)$ désigne l'information de Fisher sur l'échantillon alors que $I_1(\theta)$ désigne celle d'une observation de l'échantillon.

2.6.6 Borne de Cramer-Rao et efficacité

La borne de Cramér-Rao est un concept fondamental en statistiques. Cette borne établit une limite inférieure théorique à la variance de n'importe quel estimateur non biaisé d'un paramètre. Elle a été nommée d'après le statisticien Harald Cramér et le mathématicien Cuthbert C. Rao.

Supposons que nous ayons un échantillon aléatoire de données X suivant une loi de probabilité dépendant d'un paramètre inconnu θ .

Définition 10. Si $\hat{\theta}$ est un estimateur non biaisé de θ , la variance de cet estimateur $Var(\hat{\theta})$ doit satisfaire la borne de Cramér-Rao :

$$Var(\hat{\theta}) \geq \frac{1}{I(\theta)} \quad (12)$$

où :

$$\begin{aligned} Var(\hat{\theta}) &\text{ est la variance de l'estimateur } \hat{\theta}, \\ I(\theta) &\text{ est l'information de Fisher au point } \theta. \end{aligned}$$

En d'autres termes, la variance de tout estimateur non biaisé ne peut pas être plus faible que l'inverse de l'information de Fisher. Cette borne est un outil essentiel en statistiques pour évaluer la performance des estimateurs.

L'efficacité d'un estimateur fait référence à sa capacité à fournir des estimations précises et fiables du paramètre. En d'autres termes, un estimateur est dit efficace s'il atteint la borne de Cramér-Rao pour la variance de l'estimateur. L'efficacité est un critère important pour évaluer la qualité d'un estimateur, car elle détermine si l'estimateur est le meilleur parmi tous les estimateurs non biaisés en termes de précision.

Pour qu'un estimateur soit considéré comme efficace, il doit satisfaire les critères suivants :

- non biaisé,
- efficace asymptotiquement : L'estimateur doit atteindre la borne de Cramér-Rao pour la variance de l'estimateur lorsque la taille de l'échantillon augmente. En d'autres termes, sa variance doit tendre vers l'inverse de l'information de Fisher à mesure que la taille de l'échantillon devient grande. Cela signifie que l'estimateur est le plus précis possible parmi tous les estimateurs non biaisés.

L'efficacité est souvent associée aux estimateurs du maximum de vraisemblance (EMV), car ils sont connus pour être asymptotiquement efficaces et atteindre la borne de Cramér-Rao dans de nombreuses situations. Cependant, il est important de noter que l'efficacité dépend du modèle statistique et de la distribution de probabilité sous-jacente. Dans certaines situations, d'autres estimateurs peuvent être plus efficaces que les EMV.

2.6.7 Exemple de la loi de Bernoulli

L'efficacité de l'estimateur du maximum de vraisemblance (EMV) pour une distribution de Bernoulli peut être démontrée en utilisant la borne de Cramér-Rao. Pour une distribution de Bernoulli, la variance de l'estimateur de la probabilité de succès (paramètre de la distribution) est égale à l'inverse de l'information de Fisher. En effet, l'efficacité de l'estimateur EMV pour une distribution de Bernoulli est déterminée en utilisant la borne de Cramér-Rao.

Pour une distribution de Bernoulli, le paramètre d'intérêt est la probabilité de succès, notée θ . L'estimateur EMV de θ est la proportion de succès observée dans l'échantillon. La variance de l'estimateur EMV de θ est donnée par :

$$\text{Var}(\hat{\theta}) = \frac{\theta(1-\theta)}{n} \quad (13)$$

où $\hat{\theta}$ est l'estimateur EMV de θ et n est la taille de l'échantillon.

L'information de Fisher pour une distribution de Bernoulli est donnée par :

$$I_1(\theta) = \frac{1}{\theta(1-\theta)} \quad (14)$$

Maintenant, pour démontrer l'efficacité de l'estimateur EMV, nous pouvons utiliser la borne de Cramér-Rao. La borne de Cramér-Rao stipule que la variance de tout estimateur non biaisé ne peut pas être plus faible que l'inverse de l'information de Fisher, c'est-à-dire :

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)} \quad (15)$$

En substituant les expressions précédentes pour la variance et l'information de Fisher, nous obtenons :

$$\frac{\theta(1-\theta)}{n} \geq \frac{1}{n \frac{1}{\theta(1-\theta)}} \quad (16)$$

Ce qui conduit à une égalité, donc la variance de l'estimateur MLE atteint la borne de Cramér-Rao, ce qui signifie que l'estimateur EMV est asymptotiquement efficace pour une distribution de Bernoulli.

2.7 Propriétés des estimateurs

Nous allons ici étudier les propriétés des estimateurs issus des deux précédentes méthodes.

2.7.1 Le cas des EMM

Si $\theta = E(X)$, alors l'EMM de θ est $\hat{\theta} = \bar{X}_n$. La justification de cette méthode est la loi des grands nombres (\bar{X}_n converge presque sûrement vers $E(X)$). Donc, si $\theta = E(X)$, \bar{X}_n est un estimateur de θ convergent presque sûrement. Autrement dit, si on a beaucoup d'observations, on peut estimer une espérance par une moyenne empirique.

Même sans la loi des grands nombres, on peut calculer:

$$E[\bar{X}_n] = E\left[\frac{1}{n} \sum_i X_i\right] = \frac{1}{n} E\left[\sum_i X_i\right] = \frac{1}{n} \sum_i E[X_i] = \frac{1}{n} n E[X_i] = \theta.$$

Ainsi, \bar{X}_n est un estimateur non biaisé de θ .

On peut également en calculer la variance:

$$\text{Var}(\bar{X}_n) = \text{Var}\left(\frac{1}{n} \sum_i X_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_i X_i\right) = \frac{1}{n^2} \sum_i \text{Var}(X_i) = \frac{1}{n^2} n \text{Var}(X_i) = \frac{\text{Var}(X_i)}{n},$$

en utilisant l'hypothèse iid des X_i .

On obtient ainsi le résultat suivant:

Propriété 1. \bar{X}_n est un estimateur sans biais en convergent en moyenne quadratique de $E[X]$. En effet,

$$\lim_{n \rightarrow +\infty} E[(\bar{X}_n - X)^2] = 0.$$

Qu'en est-il de l'estimateur de la variance, mesure de la dispersion des données dans un échantillon? Nous considérons la variance empirique, notée S_n^2 :

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \quad (17)$$

Cet estimateur est-il biaisé? L'espérance de la variance empirique est donnée par :

$$E(S_n^2) = E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right]$$

Utilisons les propriétés de l'espérance pour développer cette expression :

$$E(S_n^2) = E \left[\frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X}_n)^2 \right] = \frac{1}{n} \sum_{i=1}^n E[X_i^2] - E[\bar{X}_n^2] = E[X_i^2] - E[\bar{X}_n^2]$$

Maintenant, nous allons utiliser la propriété $Var(X) = E(X^2) - (E(X))^2$:

$$E(S_n^2) = Var(X_i) + E[X_i]^2 - Var(\bar{X}_n) - E[\bar{X}_n]^2 = Var(X_i) + E[X_i]^2 - \frac{Var(X_i)}{n} - E[X_i]^2$$

Ainsi, nous obtenons

$$E(S_n^2) = \left(1 - \frac{1}{n}\right) Var(X_i) = \frac{n-1}{n} Var(X_i)$$

La variance empirique est donc un estimateur biaisé de la variance (mais asymptotiquement sans biais)! En revanche, l'estimateur

$$S_n'^2 = \frac{n}{n-1} S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (18)$$

est un estimateur non-biaisé de $Var(X_i)$.

Par ailleurs,

$$Var(S_n'^2) = Var \left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right) = \frac{1}{n(n-1)} [(n-1)E[(X_i - E[X_i])^4] - (n-3)Var(X_i)^2]$$

tend vers 0 lorsque n tend vers l'infini. En utilisant le fait que le risque quadratique résulte du biais au carré auquel est ajouté la variance de l'estimateur, on obtient:

Propriété 2. $S_n'^2$ est un estimateur sans biais en convergent en moyenne quadratique de $Var(X)$.

On peut aussi montrer que $S_n'^2$ converge presque sûrement vers $Var(X_i)$. Puisque \bar{X}_n et $S_n'^2$ convergent presque sûrement respectivement vers $E[X_i]$ et $Var(X_i)$, alors ils convergent également en probabilité vers ces quantités. On obtient ainsi:

Propriété 3. \bar{X}_n est un estimateur consistant de $E[X_i] = E[X]$.

$S_n'^2$ est un estimateur consistant de $Var(X_i) = Var(X)$.

Remarque 2. $Cov(\bar{X}_n, S_n'^2) = E[(X_i - E[X_i])^3]$, donc les deux estimateurs ne sont pas indépendants en toute généralité. Ils sont corrélés, bien qu'asymptotiquement non-corrélés. En revanche, on peut montrer qu'ils sont indépendants si les observations sont de loi normale.

2.7.2 Le cas des EMV

Un estimateur de maximum de vraisemblance n'est pas forcément unique (la vraisemblance peut avoir plusieurs maxima), ni sans biais, ni de variance minimale. Mais il a d'excellentes propriétés asymptotiques si la loi des observations satisfait certaines conditions de régularité (les mêmes que celles nécessaires à l'existence de la quantité d'information). Lorsque la taille de l'échantillon augmente, l'estimateur du maximum de vraisemblance (EMV) tend à suivre une distribution gaussienne. Cela est dû au théorème central limite, qui stipule que la somme d'un grand nombre de variables aléatoires indépendantes et identiquement distribuées suit approximativement une distribution normale, quelle que soit la distribution d'origine des variables.

Théorème 1. *Si les X_i sont identiquement distribués et indépendants, issus d'une même loi dépendante d'un paramètre θ , et que les conditions de régularité sur la loi des observations sont satisfaites, alors*

$$\hat{\theta}_{EMV} \stackrel{\mathcal{L}}{\sim} \mathcal{N}(\theta, I^{-1}(\theta))$$

où :

$\hat{\theta}$ est l'estimateur du maximum de vraisemblance (EMV) du paramètre θ ,

θ est la vraie valeur du paramètre,

$I(\theta)$ est l'information de Fisher,

$\stackrel{\mathcal{L}}{\sim}$ indique que l'estimateur est asymptotiquement équivalent en distribution,

$\mathcal{N}(\theta, I^{-1}(\theta))$ est une distribution gaussienne (normale) avec une moyenne θ et une variance $I^{-1}(\theta)$.

Cette propriété est essentielle pour comprendre le comportement asymptotique de l'estimateur du maximum de vraisemblance, en particulier lorsqu'il s'agit de dériver des intervalles de confiance et des tests d'hypothèses basés sur la distribution asymptotique de l'estimateur. On peut également écrire:

$$\hat{\theta}_{EMV} \stackrel{\mathcal{L}}{\sim} \mathcal{N}\left(\theta, \frac{1}{nI_1(\theta)}\right) \Leftrightarrow \sqrt{n}(\hat{\theta}_{EMV} - \theta) \stackrel{\mathcal{L}}{\sim} \mathcal{N}\left(0, \frac{1}{I_1(\theta)}\right)$$

2.7.3 La méthode Delta

On peut s'intéresser à la transformation par une application de l'estimateur EMV, et à sa nouvelle loi. C'est ce qu'on appelle la méthode Delta. Considérons une fonction g dérivable. On a les deux résultats suivants.

Propriété 4. *En notant $\hat{\theta}_{EMV}$ l'estimateur du maximum de vraisemblance de θ , alors:*

- $g(\hat{\theta}_{EMV})$ est l'EMV de $g(\theta)$;
- avec g dérivable, on obtient également que

$$\sqrt{n}(g(\hat{\theta}_{EMV}) - g(\theta)) \stackrel{\mathcal{L}}{\sim} \mathcal{N}\left(0, \frac{g'(\theta)^2}{I_1(\theta)}\right).$$

Le fait que l'EMV soit asymptotiquement sans biais et efficace fait que, si on a beaucoup de données, on est pratiquement certains que l'EMV donnera la meilleure estimation possible. C'est la raison pour laquelle cette méthode est utilisée en priorité à toute autre méthode, y compris celle des moments.

3 Estimation par intervalle

Les intervalles de confiance (IC) sont des outils fondamentaux en statistique qui permettent de mesurer l'incertitude entourant les estimations de paramètres. Ils offrent une approche plus informative que de simplement rapporter un seul chiffre en tant qu'estimation d'un paramètre, car ils tiennent compte de la variabilité des données et de la taille de l'échantillon.

3.1 Introduction aux intervalles de confiance (IC) et définition

Les IC sont largement utilisés dans la prise de décision, la recherche, et la communication des résultats statistiques. Ils rendent compte de l'incertitude statistique: lorsque nous estimons un paramètre à partir d'un échantillon de données, il existe une incertitude associée à cette estimation en raison de la variabilité naturelle des données. Deux échantillons différents de la même population donneront généralement des estimations légèrement différentes. L'incertitude statistique est une mesure de cette variation.

Les IC sont construits autour du principe que, pour une estimation donnée, nous pouvons définir un intervalle à l'intérieur duquel le véritable paramètre inconnu est susceptible de se situer avec une certaine probabilité. Par exemple, un IC à 95% signifie que, dans 95% des cas où nous répétons l'échantillonnage et l'estimation, le paramètre réel sera contenu dans l'intervalle.

La construction d'un IC implique généralement deux étapes:

1. tout d'abord, nous calculons un estimateur ponctuel du paramètre à partir de l'échantillon. Ensuite, nous utilisons des méthodes statistiques pour déterminer les limites de l'intervalle de confiance autour de cet estimateur. La largeur de l'intervalle est généralement basée sur la distribution de probabilité de l'estimateur.
2. niveau de confiance : c'est la probabilité que l'intervalle de confiance contienne le véritable paramètre. Les niveaux de confiance couramment utilisés sont 90%, 95% et 99%. Plus le niveau de confiance est élevé, plus l'intervalle de confiance sera large, reflétant une plus grande incertitude.

Plutôt que de simplement obtenir une estimation ponctuelle à une hypothèse nulle, utiliser un IC permet d'évaluer si un paramètre est significativement différent de certaines valeurs de référence. En termes de communication de résultats, il est généralement préférable d'inclure des IC pour indiquer la précision de l'estimation. Cela permet aux autres personnes de comprendre l'incertitude associée aux résultats (pensez par exemple à la météo...).

Définition 11. Noté $IC_{1-\alpha}(\theta)$, un intervalle de confiance (IC) de seuil α pour le paramètre θ , avec $0 < \alpha \leq 1$, est un intervalle avec bornes aléatoires tel que

$$P(\theta \in IC_{1-\alpha}(\theta)) = 1 - \alpha.$$

Ainsi, α est la probabilité que le paramètre inconnu θ n'appartienne pas à l'intervalle. Si l'on suggère que le paramètre appartient à cet intervalle, on a donc $\alpha\%$ de chance de se tromper. Le niveau α est donc une probabilité d'erreur, en général fixée relativement petite en fonction des applications ($\alpha = 0.1$, $\alpha = 0.05$, $\alpha = 0.01$, $\alpha = 0.001$).

Remarque 3. Il est important de noter que les bornes de l'intervalle de confiance sont des variables aléatoires. En effet, nous verrons qu'elles sont basées sur une transformation d'un estimateur du paramètre, qui est lui-même une variable aléatoire.

3.2 Notion de pivot

Le procédé logique de détermination d'un intervalle de confiance pour un paramètre θ est de proposer un intervalle centré sur un estimateur performant (noté $\hat{\theta}$). Ainsi, cet intervalle de confiance aura la forme $IC = [\hat{\theta} - \eta, \hat{\theta} + \eta]$. Il faut ensuite trouver η de sorte que

$$P(\theta \in IC) = P(\hat{\theta} - \eta \leq \theta \leq \hat{\theta} + \eta) = P(|\hat{\theta} - \theta| \leq \eta) = \alpha.$$

Ici, α est un réel fixé à l'avance qui ne doit pas dépendre de θ . Le terme η ne doit pas non plus dépendre de θ pour que l'intervalle soit utilisable. Du coup, on ne peut déterminer un η vérifiant cette égalité que si la loi de probabilité de $\hat{\theta} - \theta$ ne dépend pas de θ . C'est ce qu'on appelle un pivot!

Définition 12. Un pivot pour l'estimation de θ est une fonction des observations X_i et du paramètre θ dont la loi de probabilité ne dépend pas de θ .

3.3 Exemples de construction

3.3.1 IC pour la moyenne d'une loi normale de variance connue

On suppose que les observations $x = (x_1, \dots, x_n)$ sont des réalisations de v.a. iid de loi $\mathcal{N}(m, \sigma^2)$, où σ^2 est considéré connu. On cherche à proposer un intervalle de confiance de niveau α pour le paramètre m , inconnu.

Nous avons vu précédemment que \bar{X}_n est un bon estimateur de $E[X_i] = m$ en toute généralité. Ainsi, il apparaît logique de chercher un intervalle de la forme

$$IC_{1-\alpha}(m) = [\bar{X}_n - \eta, \bar{X}_n + \eta].$$

Pour α fixé, on cherche donc η tel que

$$P(|\bar{X}_n - m| \leq \eta) = 1 - \alpha.$$

Or, on sait que $\bar{X}_n \sim \mathcal{N}(m, \sigma^2/n)$, donc on obtient de suite que

$$U = \frac{\bar{X}_n - m}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0, 1).$$

Nous venons de déterminer un pivot! En effet, la loi normale centrée réduite étant bien connue, nous disposons d'une table numérique nous permettant d'évaluer une probabilité pour un quantile donné. Ainsi, la densité d'une gaussienne étant symétrique,

$$P(|U| \leq q_{1-\alpha/2}^{\mathcal{N}(0,1)}) = 1 - \alpha,$$

où $q_{1-\alpha/2}^{\mathcal{N}(0,1)}$ désigne le quantile d'ordre $(1 - \alpha/2)$ de la loi normale centrée réduite. Soit en remplaçant,

$$P(|\bar{X}_n - m| \leq \sqrt{\sigma^2/n} q_{1-\alpha/2}^{\mathcal{N}(0,1)}) = 1 - \alpha,$$

et donc,

$$P(\bar{X}_n - \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2}^{\mathcal{N}(0,1)} \leq m \leq \bar{X}_n + \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2}^{\mathcal{N}(0,1)}) = 1 - \alpha.$$

Nous venons donc de déterminer l'intervalle en question.

Propriété 5. *Un intervalle de confiance de seuil α pour la moyenne m d'une loi normale $\mathcal{N}(m, \sigma^2)$ avec σ^2 connu vaut:*

$$IC_{1-\alpha}(m) = \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2}^{\mathcal{N}(0,1)}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2}^{\mathcal{N}(0,1)} \right].$$

Remarque 4. *Dans la vraie vie, on ne connaît pas la valeur de σ^2 . Il faut donc proposer un estimateur de ce paramètre et l'injecter dans la recherche d'un pivot. Evidemment, le pivot changera, et donc l'intervalle de confiance également.*

Remarque 5. *Il est intéressant de remarquer que:*

- plus le nombre d'observations n augmente, plus l'intervalle se rétrécit (on gagne en précision),
- plus la variance σ^2 augmente, plus l'intervalle s'agrandit. Ceci est logique puisqu'une variable aléatoire avec une plus grande variance induit plus de difficulté dans l'estimation de la moyenne (cf loi de \bar{X}_n).

3.3.2 IC pour la variance d'une loi normale

Comme pour la partie précédente, on recherche un pivot, c'est à dire une fonction des observations X_i et du paramètre σ^2 dont la loi ne dépend pas de σ^2 (et pas de m bien entendu!). A la fin de la partie de cours sur les probabilités, nous avons vu un résultat important. Il s'agit du théorème de Fisher.

Théorème 2. La statistique $\frac{nS_n^2}{\sigma^2}$ suit une loi du χ^2 à $(n-1)$ degrés de liberté. Autrement dit, $\frac{nS_n^2}{\sigma^2} \sim \chi_{n-1}^2$.

De ce résultat, il est aisé d'en déduire un pivot afin de déterminer un intervalle de confiance. En effet, nous savons donc que quelque soit $(a, b) \in \mathbb{R}^{+2}$, avec $0 < a < b$,

$$P\left(a \leq \frac{nS_n^2}{\sigma^2} \leq b\right) = F_{\chi_{n-1}^2}(b) - F_{\chi_{n-1}^2}(a).$$

Mais on peut aussi écrire

$$P\left(a \leq \frac{nS_n^2}{\sigma^2} \leq b\right) = P\left(\frac{nS_n^2}{b} \leq \sigma^2 \leq \frac{nS_n^2}{a}\right).$$

Il y a donc une infinité de choix possibles pour a et b afin d'égaliser cette probabilité à $(1 - \alpha)$. D'ordinaire, on répartit l'erreur symétriquement (à gauche de la densité et à droite), et on choisit donc a et b de sorte que $F_{\chi_{n-1}^2}(b) = 1 - \frac{\alpha}{2}$ et $F_{\chi_{n-1}^2}(a) = \frac{\alpha}{2}$.

On en déduit immédiatement le résultat.

Propriété 6. Un intervalle de confiance de niveau (ou seuil) α pour le paramètre σ^2 de la loi $\mathcal{N}(m, \sigma^2)$ est donné par:

$$IC_{1-\alpha}(\sigma^2) = \left[\frac{nS_n^2}{z_{n-1, 1-\alpha/2}}, \frac{nS_n^2}{z_{n-1, \alpha/2}} \right] = \left[\frac{(n-1)S_n'^2}{z_{n-1, 1-\alpha/2}}, \frac{(n-1)S_n'^2}{z_{n-1, \alpha/2}} \right],$$

où $z_{n-1, \alpha}$ désigne le quantile d'ordre α de la loi du χ^2 à $(n-1)$ degrés de liberté.

3.3.3 IC pour une proportion

Le contexte ici est celui de la détermination d'un intervalle de confiance pour le paramètre p d'une loi de Bernoulli, grâce à un échantillon X_1, \dots, X_n . On suppose donc que les observations X_i sont iid, distribuées selon une loi de Bernoulli: $X_i \sim \mathcal{B}(p)$. Nous avons déjà vu que $\hat{p} = \bar{X}_n$ est un ESBVM de p .

IC exact On cherche donc un pivot basé sur \bar{X}_n et p , dont la loi ne dépende pas de p . Nous savons que $T = n\bar{X}_n$ suit une loi binomiale car $n\bar{X}_n = \sum_i X_i \sim \mathcal{B}(n, p)$. Mais à ce stade nous ne connaissons pas de transformation simple permettant de faire disparaître le paramètre p de la loi binomiale... On admet donc le résultat suivant.

Propriété 7. Un intervalle de confiance exact sur p , de seuil α , est donné par

$$IC_{1-\alpha}(p) = \left[\frac{1}{1 + \frac{n-T+1}{T} q_{2(n-T+1), 2T, 1-\alpha/2}}, \frac{1}{1 + \frac{n-T}{T+1} q_{2(n-T), 2(T+1), \alpha/2}} \right],$$

avec $q_{2(n-T+1), 2T, 1-\alpha/2}$ le quantile d'ordre $(1 - \alpha/2)$ de la loi de Fisher-Snedecor.

Cet intervalle de confiance est cependant difficilement exploitable, et on lui préfère souvent un intervalle de confiance asymptotique.

IC asymptotique Le principe de l'intervalle de confiance asymptotique repose ici sur l'approximation de la loi binomiale par la loi normale. En effet, comme n est sensé être relativement grand, on peut appliquer le théorème central limite à la loi des X_i . Ainsi on a

$$T = \sum_i X_i \sim \mathcal{N}(nE[X_i], nVar(X_i)) \Leftrightarrow \frac{T - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0, 1).$$

Nous obtenons ainsi un pivot, ce qui nous permet d'aboutir immédiatement à l'intervalle de confiance asymptotique. En effet, pour tout $(a, b) \in \mathbb{R}^2$,

$$P\left(a \leq \frac{T - np}{\sqrt{np(1-p)}} \leq b\right) = F_{\mathcal{N}(0,1)}(b) - F_{\mathcal{N}(0,1)}(a).$$

Pour que cette probabilité vaille $(1 - \alpha)$, on peut choisir $F_{\mathcal{N}(0,1)}(b) = 1 - \alpha/2$ et $F_{\mathcal{N}(0,1)}(a) = \alpha/2$. Ainsi $b = F_{\mathcal{N}(0,1)}^{-1}(1 - \alpha/2) = q_{1-\alpha/2}^{\mathcal{N}(0,1)}$ et $a = F_{\mathcal{N}(0,1)}^{-1}(\alpha/2) = q_{\alpha/2}^{\mathcal{N}(0,1)}$. Ce qui donne:

$$P\left(q_{\alpha/2}^{\mathcal{N}(0,1)} \leq \frac{T - np}{\sqrt{np(1-p)}} \leq q_{1-\alpha/2}^{\mathcal{N}(0,1)}\right) = 1 - \alpha.$$

En développant le calcul pour isoler le paramètre p , on obtient un résultat un peu compliqué... Mais qui se simplifie au prix d'une approximation supplémentaire (le quantile au carré est négligeable par rapport à n). Au final, on obtient le résultat très simple suivant:

Propriété 8. *L'intervalle de confiance asymptotique sur p , de seuil α , pour des observations suivant une loi de Bernoulli de paramètre p , vaut:*

$$IC_{1-\alpha}(p) = \left[\bar{X}_n - q_{1-\alpha/2}^{\mathcal{N}(0,1)} \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}}, \bar{X}_n + q_{1-\alpha/2}^{\mathcal{N}(0,1)} \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \right],$$

Notons que \bar{X}_n représente ici la proportion observée de fois que X_i vaut 1 dans l'échantillon. On remarque également que cet intervalle de confiance est très proche de celui de la moyenne dans le cas d'un échantillon gaussien à variance connue, ce qui semble logique étant donné l'approximation faite initialement.

4 Tests paramétriques

Les tests statistiques sont des outils essentiels pour nous aider à prendre des décisions éclairées en nous permettant de tirer des conclusions à partir d'échantillons de données, tout en prenant en compte l'incertitude inhérente à la variabilité des données. Une introduction aux tests statistiques nécessite de comprendre les concepts fondamentaux suivants :

- population et échantillon : la population est l'ensemble complet de toutes les observations possibles, tandis qu'un échantillon est un sous-ensemble de la population que nous examinons. Les tests statistiques sont souvent utilisés pour faire des déclarations sur la population en se basant sur les caractéristiques observées dans l'échantillon;
- hypothèses : les tests statistiques reposent sur des hypothèses. L'hypothèse nulle H_0 est une affirmation que nous souhaitons tester, tandis que l'hypothèse alternative H_1 est ce que nous essayons de prouver. Les tests statistiques évaluent la probabilité que les données observées soient compatibles avec l'hypothèse nulle.
- statistique de test : une statistique de test est un nombre calculé à partir des données de l'échantillon qui nous aide à prendre une décision sur l'hypothèse nulle. La sélection de la statistique de test dépend du type de données et de la question;
- niveau de signification (souvent noté α): il est le seuil de tolérance pour l'erreur que nous sommes prêts à accepter. Cela détermine le seuil de probabilité en dessous duquel nous rejeterons l'hypothèse nulle;
- p-valeur : c'est la probabilité que la statistique de test soit observée, ou plus extrême, si l'hypothèse nulle est vraie. Une p-valeur faible (inférieure à α) suggère que les données fournissent des preuves solides contre l'hypothèse nulle;
- décision : en fonction de la p-valeur par rapport au niveau de signification, on peut décider de rejeter ou de ne pas rejeter l'hypothèse nulle;
- interprétation des résultats d'un test statistique: elle dépend de la décision prise. Si l'hypothèse nulle est rejetée, cela indique que les données fournissent des preuves pour l'hypothèse alternative, tandis que si elle n'est pas rejetée, cela signifie que les données ne fournissent pas suffisamment de preuves pour conclure en faveur de l'hypothèse alternative. En aucun cas on ne peut conclure sur la véracité de H_0 .

Les tests statistiques peuvent être classés en différentes catégories en fonction de la nature des données et des objectifs de l'analyse. Ils sont utilisés pour comparer des groupes, évaluer des relations, mesurer des différences, prédire des résultats, et bien plus encore. En résumé, les tests statistiques sont des outils puissants pour transformer des données brutes en informations significatives et parlantes.

4.1 Introduction et concepts généraux

4.2 Formalisation des tests paramétriques

4.2.1 Test d'hypothèses simples

4.2.2 Test d'hypothèses composites

4.3 Test sur la moyenne d'une gaussienne

4.3.1 Exemple introductif et test naïf

4.3.2 Amélioration du test

Remarque: si la variance est inconnue, ce test n'est pas applicable...

4.4 p-valeur d'un test

4.5 Lien entre test d'hypothèses et intervalle de confiance

4.6 Construction d'un test : généralisation

4.7 Test sur la variance d'une gaussienne

4.8 Test sur une proportion

4.9 Test de comparaison de deux échantillons

4.9.1 Echantillons gaussiens indépendants

Comparaison des variances: test de Fisher

Comparaison des moyennes: test de Student

4.9.2 Comparaison de proportions

5 Tests non-paramétriques

5.1 Rappels sur la fonction de répartition empirique

5.2 Test du χ^2

5.2.1 Test sur les probabilités d'événement

5.2.2 Test d'adéquation à une loi

5.2.3 Test d'adéquation à une famille de lois

5.2.4 Test d'indépendance entre deux variables

5.3 Test de Kolmogorov

5.4 Test de Wilcoxon-Mann-Whitney

5.5 Test de Wilcoxon

5.6 Test de Kruskal-Wallis