COURS DE STATISTIQUES
3ÈME ANNÉE LICENCE
ISSEP du KEF

Présenté par:

MOHAMED BEN AISSA

Doctorant en STAPS : Sciences Humaines et Sociales

mbenaissa.hs@gmail.com

INTRODUCTION

La statistique est une discipline essentielle qui trouve une place de choix dans de nombreux domaines de la vie quotidienne. Lorsqu'elle est appliquée au sport et à l'éducation physique, elle devient un outil puissant pour analyser, comprendre et améliorer les performances athlétiques, ainsi que pour évaluer l'impact des programmes d'enseignement de l’éducation physique. Dans ce cours de statistique appliquée, nous explorerons comment les méthodes statistiques peuvent être utilisées pour obtenir des informations cruciales dans ces domaines.

INTRODUCTION

Nous explorerons les principes fondamentaux de la statistique. Nous aborderons les méthodes de collecte de données, les techniques d'analyse statistique.

De plus, nous examinerons des exemples concrets de l'application de la statistique dans des contextes sportifs et éducatifs, afin de mieux comprendre son importance et sa pertinence dans ces domaines.

La statistique est bien plus qu'une simple compilation de nombres ; elle est un outil puissant qui nous permet de comprendre, d'analyser et d'améliorer les performances sportives et les programmes d'éducation physique.

MÉTHODOLOGIE EXPÉRIMENTALE ET RECUEIL DES DONNÉES

La mesure est le processus de quantification ou d'évaluation d'une grandeur physique, d'une caractéristique ou d'une variable à l'aide d'unités de mesure standard. Elle permet de représenter de manière précise et objective une valeur numérique associée à une propriété ou à un phénomène.

la mesure n'est pas limitée aux grandeurs physiques, elle peut également être utilisée pour quantifier des concepts abstraits, des comportements et des attitudes. Voici quelques exemples supplémentaires pour illustrer cela :

- Niveau de satisfaction : La satisfaction d'un client à l'égard d'un produit ou d'un service peut être mesurée en utilisant des sondages ou des échelles de satisfaction, où les participants attribuent des scores à leur niveau de contentement.
- Niveau de compétence linguistique : La maîtrise d'une langue étrangère peut être mesurée à l'aide de tests de compétence linguistique qui évaluent la grammaire, la compréhension orale, la lecture et l'expression écrite.

MÉTHODOLOGIE EXPÉRIMENTALE ET RECUEIL DES DONNÉES

La mesure est essentielle dans de nombreux domaines de la science, de l'ingénierie, de la technologie, de la médecine, et elle joue un rôle crucial dans la prise de décision, la recherche et le développement, ainsi que dans la compréhension des phénomènes naturels. Elle permet de rendre des données quantitatives exploitables pour l'analyse, la comparaison et la communication.

La mesure empirique est une méthode de collecte de données basée sur l'observation directe ou l'expérience pratique, plutôt que sur des concepts théoriques. En d'autres termes, elle consiste à recueillir des informations en utilisant nos sens ou des instruments de mesure concrets plutôt que des modèles abstraits.

Les exemples suivants illustrent comment la mesure empirique permet de recueillir des données réelles et tangibles dans le domaine du sport pour évaluer la performance des athlètes et suivre leur progression.

MÉTHODOLOGIE EXPÉRIMENTALE ET RECUEIL DES DONNÉES

- **Force musculaire :** Utiliser un dynamomètre pour mesurer la force maximale qu'un athlète peut exercer dans un exercice spécifique, comme un soulevé de poids.
- **Fréquence cardiaque :** Utiliser un moniteur de fréquence cardiaque pour mesurer les battements cardiaques par minute pendant l'exercice physique.
- **Évaluation de la coordination :** Observer la capacité d'un joueur de football à dribbler le ballon autour des défenseurs lors d'un match.
- **Niveau de stress :** Le stress peut avoir un impact significatif sur la performance. Des échelles de mesure du stress, telles que l'inventaire de stress perçu (Perceived Stress Scale), peuvent être utilisées pour évaluer le niveau de stress des athlètes.
- **Niveau d'anxiété compétitive :** Les chercheurs et les entraîneurs utilisent des questionnaires standardisés pour mesurer le niveau d'anxiété avant une compétition. Par exemple, l'échelle CSAI-2 (Competitive State Anxiety Inventory-2) évalue l'anxiété somatique et cognitive chez les athlètes.

TYPOLOGIE DES VARIABLES

En statistique, une variable est une caractéristique, une propriété ou une quantité qui peut varier d'un individu, d'un objet ou d'une unité à l'autre au sein d'une population ou d'un échantillon (Age, Genre, Couleur des cheveux, Taille…etc.) .

Les variables sont utilisées pour collecter des données et étudier comment elles varient, ce qui permet d'effectuer des analyses statistiques et de tirer des conclusions.

les variables sont utilisées pour mesurer et étudier différents aspects de la performance sportive, de la condition physique et du comportement des athlètes.

En statistique, l'analyse de ces variables permet de mieux comprendre les relations et les tendances dans le domaine du sport.

TYPOLOGIE DES VARIABLES

Variables Qualitatives

Les variables qualitatives, également appelées catégorielles, sont des données qui décrivent des caractéristiques ou des attributs. Elles ne peuvent pas être mesurées numériquement, mais elles sont essentielles pour classer, comparer et caractériser des éléments.

Ces variables sont essentielles pour capturer des informations sur des aspects non numériques, tels que les caractéristiques personnelles, les préférences, les affiliations ou les identifications. Par exemple, elles permettent de distinguer entre des éléments tels que les couleurs, les genres, les équipes, les types d'animaux, les régions géographiques, les marques de produits, et bien d'autres encore. Les variables qualitatives facilitent la classification des données et l'organisation des éléments en groupes ou en catégories qui peuvent ensuite être analysés et comparés.

Elles sont généralement divisées en deux sous-catégories : les variables nominales et les variables ordinales.

TYPOLOGIE DES VARIABLES

Variables nominales: Une variable nominale, également appelée variable catégorielle nominale, est un type de variable qualitative qui représente des catégories distinctes ou des étiquettes sans ordre intrinsèque. Les catégories dans une variable nominale ne peuvent pas être classées ou ordonnées de manière significative. Elles sont utilisées pour classer et catégoriser des éléments en groupes distincts (Couleur de maillot d'équipe, Sport pratiqué, Pays d'origines des athlètes…etc.).

Les variables nominales sont principalement utilisées pour regrouper, classer et caractériser des éléments en fonction de leurs caractéristiques distinctes, sans établir d'ordre ou de hiérarchie entre ces catégories. Elles sont importantes pour effectuer des analyses descriptives et pour représenter des informations catégorielles dans les études et les recherches liées au sport.

TYPOLOGIE DES VARIABLES

Variables ordinales: Une variable ordinale est un type de variable statistique qui représente des catégories ou des groupes avec un ordre ou une hiérarchie intrinsèque. Contrairement à une variable nominale, une variable ordinale a des catégories qui ont une relation d'ordre définie,

- Niveau de compétence : Vous pourriez demander aux participants de classer leur niveau de compétence dans un sport de "débutant", "intermédiaire" et "avancé". Dans cet exemple, il existe un ordre défini, où "avancé" indique un niveau plus élevé que "intermédiaire" et "débutant".

- Classement des équipes : Dans une ligue sportive, vous pourriez avoir un classement des équipes en fonction de leurs performances. Les équipes peuvent être classées de la première à la dernière place. Il existe un ordre intrinsèque dans ce classement.

une variable ordinale classe les catégories en fonction d'un ordre spécifique ou d'une hiérarchie, ce qui signifie que les valeurs de cette variable peuvent être comparées numériquement de manière significative en tenant compte de l'ordre sous-jacent.

TYPOLOGIE DES VARIABLES

Variables Quantitatives

Une variable quantitative est une mesure numérique ou une quantité utilisée en statistiques pour représenter des données. Ces variables sont des valeurs numériques qui peuvent être soumises à des analyses mathématiques.

Elles sont essentielles pour quantifier des phénomènes, effectuer des calculs statistiques tels que la moyenne, la médiane, la variance, et elles jouent un rôle fondamental dans l'analyse statistique en permettant d'évaluer les relations et les tendances dans les données, ce qui en fait un élément central dans la prise de décision et la recherche dans divers domaines.

Elles sont généralement divisées en deux sous-catégories : les variables continues et les variables discrètes.

TYPOLOGIE DES VARIABLES

Variables discrètes: Une variable discrète est un concept en statistiques qui représente des données numériques avec des valeurs spécifiques et dénombrables. Cela signifie que les valeurs possibles d'une variable discrète sont distinctes et que l'on peut les compter.

Par exemple, si vous mesurez le nombre de buts marqués dans un match de football, les valeurs possibles seraient des nombres entiers tels que 0, 1, 2, 3, etc. Vous ne pouvez pas marquer un demi-but, ce qui signifie que les valeurs sont spécifiques et distinctes.

Les variables discrètes sont souvent utilisées pour représenter des données qui sont dénombrables par nature, comme le nombre d'éléments, d'événements ou de résultats dans divers domaines, y compris le sport.

TYPOLOGIE DES VARIABLES

Variables continues: Une variable continue est un concept important en statistiques qui se réfère à une caractéristique mesurée ou observée dans le monde réel qui peut théoriquement prendre n'importe quelle valeur au sein d'une plage spécifique. Contrairement aux variables discrètes, qui sont limitées à des valeurs spécifiques et distinctes, une variable continue peut varier de manière infiniment fine à l'intérieur de son intervalle.

Prenons un exemple simple : la taille des personnes. Si vous mesurez la taille de plusieurs individus, vous remarquerez qu'elle peut varier en continu, par exemple, de 150,5 cm à 150,6 cm, puis à 150,61 cm, et ainsi de suite, avec une infinité de valeurs possibles entre ces points. Cette continuité signifie que vous pouvez avoir une précision élevée lors de la mesure de la taille, sans limites strictes sur les valeurs possibles.

En statistiques, les variables continues sont souvent traitées à l'aide de fonctions de densité de probabilité, telles que la distribution normale, qui permettent de modéliser et d'analyser ces valeurs de manière précise.

LES NIVEAUX DE MESURE

En statistiques, les niveaux de mesure, également appelés échelles de mesure ou niveaux de mesure des données, définissent la nature des données que vous manipulez. Ils indiquent comment les données sont mesurées et classées, ce qui a des implications importantes pour le type d'analyses statistiques que vous pouvez effectuer. Il existe généralement quatre niveaux de mesure principaux :

1. Echelles nominales
2. Echelles ordinales
3. Echelles d’intervalles
4. Echelles de ratio

LES NIVEAUX DE MESURE

1. Echelles nominales: Les données nominales sont des catégories ou des étiquettes sans ordre intrinsèque. Elles sont utilisées pour regrouper des éléments similaires sans signification numérique. Par exemple, les couleurs des voitures (rouge, bleu, vert) sont nominales. Vous ne pouvez pas effectuer des opérations mathématiques telles que des moyennes ou des écarts types sur des données nominales.

2. Echelles ordinales: Les données ordinales représentent des catégories avec un ordre relatif, mais les écarts entre les catégories ne sont pas significatifs. Par exemple, un questionnaire de satisfaction peut utiliser une échelle de notation de 1 à 5, où 1 signifie "très insatisfait" et 5 signifie "très satisfait". Vous savez qu'un score de 3 est meilleur que 2, mais vous ne pouvez pas dire que c'est deux fois mieux.

LES NIVEAUX DE MESURE

1. Echelles d’intervalles: Les données d'intervalles ont un ordre, et les écarts entre les valeurs sont significatifs. De plus, les données d'intervalles ont un point zéro arbitraire. Un exemple courant est la température en degrés Celsius. Les valeurs intervalles peuvent être soumises à des opérations mathématiques telles que l'addition et la soustraction, mais il n'a pas de sens de dire qu'une valeur est "deux fois plus élevée" qu'une autre.

2. Echelles de ratio: Les données de ratio ont toutes les propriétés des données d'intervalles, mais elles ont également un point zéro absolu, ce qui signifie qu'il est possible de dire qu'une valeur est "deux fois plus élevée" qu'une autre. Les exemples courants de données de ratio incluent la longueur, le poids, le temps en secondes, etc.

LES NIVEAUX DE MESURE

Il est essentiel de comprendre le niveau de mesure de vos données car cela détermine quelles analyses statistiques sont appropriées. Par exemple, pour des données nominales, vous utiliseriez des statistiques telles que le mode et le chi-carré, tandis que pour des données de ratio, vous pouvez utiliser des statistiques telles que la moyenne, l'écart type, et effectuer des opérations mathématiques plus complexes. Le choix de l'analyse appropriée dépendra du niveau de mesure de vos données.

<table>
<tr><th>Variables</th><th>Structure</th><th>Modalités</th><th>Mesure</th><th>Exemple</th></tr>
<tr><td rowspan="2">Qualitatives</td><td>Nominale</td><td></td><td></td><td>Catégorie socio-pro</td></tr>
<tr><td>Ordinale</td><td></td><td></td><td>Réponse graduée</td></tr>
<tr><td rowspan="5">Quantitatives</td><td>Intervalle</td><td rowspan="4">Nombre infini</td><td rowspan="3">Continue</td><td>Différence de temps</td></tr>
<tr><td rowspan="4">Rapport</td><td>Temps de réaction</td></tr>
<tr><td>Pourcentage</td></tr>
<tr><td rowspan="2">Discrète</td><td>Nombre d'utilisateurs</td></tr>
<tr><td>Nombre fini</td><td>Nombre de pays</td></tr>
</table>

STATISTIQUES DESCRIPTIVES

STATISTIQUES DESCRIPTIVES

Les statistiques descriptives sont une branche de la statistique qui se concentre sur la collecte, la présentation, la synthèse et l'interprétation des données.

Son objectif principal est de résumer et de décrire les caractéristiques essentielles d'un ensemble de données, en utilisant des techniques graphiques et numériques.

la statistique descriptive permet de donner une image concise et informative d'un ensemble de données, facilitant ainsi la compréhension et la communication des informations contenues dans ces données.

STATISTIQUES DESCRIPTIVES

Les statistiques descriptives sont un ensemble de techniques et de méthodes utilisées pour résumer, organiser et présenter les données d'une manière qui permette une compréhension plus aisée.

Les statistiques descriptives visent à simplifier des données complexes pour en extraire des informations utiles, telles que des tendances, des distributions, des mesures de centralité (comme la moyenne, la médiane et le mode), des mesures de dispersion (comme la variance et l'écart type), des graphiques et des tableaux statistiques.

STATISTIQUES DESCRIPTIVES

Les concepts et le vocabulaire de base:

1. Minimum (min) ou la valeur minimale est la plus petite valeur observée dans un ensemble de données. Il représente le point le plus bas de l'ensemble.

2. Maximum (max) ou la valeur maximale est la plus grande valeur observée dans un ensemble de données. Il représente le point le plus élevé de l'ensemble.

3. Etendue (R) ou l'écart est une mesure de la dispersion qui quantifie la variation entre la valeur maximale et la valeur minimale dans un ensemble de données. Elle indique l'amplitude totale des valeurs dans l'ensemble.

4. Taille (n) ou le nombre d'observations est simplement le nombre total de valeurs présentes dans un ensemble de données. Il représente la taille de l'ensemble de données.

5. Somme (sum) ou la somme totale est le résultat de l'addition de toutes les valeurs d'un ensemble de données. Elle représente la somme totale des observations dans l'ensemble.

LES MESURES DE TENDANCE CENTRALE

Les mesures de tendance centrale sont des statistiques qui permettent de résumer l'emplacement central ou typique d'un ensemble de données. Elles sont utilisées pour décrire où la plupart des valeurs d'un ensemble de données se situent. Les trois principales mesures de tendance centrale sont les suivantes :

1. Moyenne (Moyenne arithmétique)
2. Médiane
3. Mode

➡ Chacune de ces mesures de tendance centrale présente des avantages et des inconvénients en fonction de la nature des données et des objectifs de l'analyse.

MOYENNE ARITHMÉTIQUE

La moyenne arithmétique, également appelée moyenne simple, est une mesure de tendance centrale largement utilisée pour résumer un ensemble de données. La moyenne arithmétique est la **"valeur moyenne"** d'un ensemble de données. Elle permet de résumer ces données en une seule valeur, qui représente le point central autour duquel les autres valeurs gravitent.

- La formule de la moyenne arithmétique : vous additionnez toutes les valeurs de l'ensemble de données (la somme des valeurs) puis vous divisez cette somme par le nombre total de valeurs pour obtenir la moyenne.

Moyenne (m) = (Somme des valeurs) / (Nombre de valeurs)

$$m = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Notation Mathématique:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

MOYENNE ARITHMÉTIQUE

La moyenne arithmétique est sensible aux valeurs extrêmes (valeurs aberrantes) dans les données. Une seule valeur extrême peut considérablement influencer la moyenne, la rendant moins robuste que d'autres mesures de tendance centrale.

On calcule généralement deux types de moyennes:

- La moyenne obtenue sur un groupe d’individus dans la même situation.
- La moyenne obtenue par le même individu dans des situations différentes.

En résumé, la moyenne arithmétique est une mesure de tendance centrale qui représente la "valeur moyenne" d'un ensemble de données en additionnant toutes les valeurs et en les divisant par le nombre total de valeurs. Elle est largement utilisée pour résumer et comprendre des données dans divers domaines.

MÉDIANE

La médiane est une mesure de tendance centrale essentielle qui nous permet de trouver une valeur de référence au sein d'un ensemble de données. Lorsque les données sont arrangées dans un ordre croissant, la médiane est la valeur située exactement au milieu.

En d'autres termes, elle divise l'ensemble de données en deux parties égales. Cela signifie que la moitié des valeurs sont inférieures à la médiane, tandis que l'autre moitié est supérieure à la médiane.

Cette caractéristique en fait une mesure robuste, particulièrement utile lorsque les données peuvent contenir des valeurs extrêmes ou des valeurs aberrantes, car elle n'est pas influencée par ces valeurs comme le serait la moyenne.

La médiane nous donne une perspective précieuse sur la valeur "centrale" de nos données, ce qui est essentiel pour comprendre leur distribution et leur tendance.

MÉDIANE

Comment trouver la médiane

La médiane md ou $\tilde{x}$ est la valeur des données qui sépare la moitié supérieure d'un ensemble de données de la moitié inférieure.

1. Triez les valeurs des données du plus bas au plus élevé.

2. La médiane est la valeur des données au milieu de l'ensemble.

3. Si deux valeurs de données se trouvent au milieu, la médiane est la moyenne de ces deux valeurs.

Exemple de Médiane:

- Pour l'ensemble de données 1, 1, 2, **5**, 6, 6, 9, la médiane est 5.
- Pour l'ensemble de données 1, 1, **2**, **6**, 6, 9, la médiane est 4. Prenez la moyenne de 2 et 6, c'est-à-dire (2+6)/2 = 4.

MÉDIANE

Formule de la Médiane

En ordonnant un ensemble de données $x_1 \leq x_2 \leq x_3 \leq ... \leq x_n$ du plus bas au plus élevé, la médiane $\tilde{x}$ est le point de données qui sépare la moitié supérieure des valeurs des données de la moitié inférieure.

Si la taille de l'ensemble de données n est impair, la médiane est la valeur à la position p où

$$\mathbf{p = (n + 1)/2}$$

$$\mathbf{md = x_p}$$

Si n est pair, la médiane est la moyenne des valeurs aux positions p et p + 1 où

$$\mathbf{p = n/2}$$

$$\mathbf{md = (x_p + x_{p+1})/2}$$

MODE

Le mode est la valeur qui apparaît le plus fréquemment dans un ensemble de données. En d'autres termes, c'est la valeur qui a la plus grande fréquence d'occurrence.

Formule : Il n'y a pas de formule mathématique complexe pour calculer le mode. Vous examinez simplement les données pour identifier la valeur qui se répète le plus souvent.

Utilisation :

- Le mode est utile pour identifier la valeur la plus courante ou la catégorie la plus fréquente dans un ensemble de données.
- Il est souvent utilisé dans des contextes où vous souhaitez déterminer la préférence des gens, comme les couleurs préférées, les marques de produits les plus populaires, etc.
- Dans les statistiques descriptives, le mode est l'une des mesures de tendance centrale qui permet de décrire la concentration des données autour d'une valeur spécifique.

MODE

Valeurs modales multiples :

Un ensemble de données peut avoir:

- un mode « mode unimodal »
- plusieurs modes « mode multimodal »
- aucun mode si toutes les valeurs sont uniques et aucune ne se répète.

Par exemple: dans l'ensemble de données [1, 2, 2, 3, 4, 4, 4, 5], le mode est 4 car il apparaît plus fréquemment que les autres valeurs,

Sensibilité aux valeurs extrêmes : Contrairement à la moyenne, le mode n'est pas affecté par les valeurs extrêmes. Il se concentre uniquement sur la fréquence des valeurs.

LES MESURES DE DISPERSION

Les mesures de dispersion, en statistiques, sont des indicateurs qui quantifient la répartition ou la variabilité des données au sein d'un ensemble de données. Elles permettent de comprendre à quel point les valeurs sont dispersées autour de la moyenne ou de la médiane. Les principales mesures de dispersion sont les suivantes :

1. Variance
2. Ecart-type (Standard deviation)
3. Etendue (Range)
4. Ecart interquartile (Interquartile Range, IQR)
5. Ecart moyen Absolu (Mean Absolute Deviation, MAD)

VARIANCE

La variance mesure à quel point les valeurs individuelles dans un ensemble de données diffèrent de la moyenne de l'ensemble. Plus la variance est élevée, plus les valeurs sont dispersées autour de la moyenne, indiquant une plus grande variabilité. En d'autres termes, la variance quantifie l'écart moyen entre chaque valeur de données et la moyenne.

Unité de mesure : L'unité de mesure de la variance est le carré de l'unité originale. Par exemple, si les données sont exprimées en unités de longueur, la variance sera en unités de longueur au carré.

Sensibilité aux valeurs extrêmes : La variance est sensible aux valeurs extrêmes ou aberrantes. Une seule valeur extrême peut augmenter considérablement la variance, car elle peut être très éloignée de la moyenne.

VARIANCE

Formule :

La formule de la variance (σ^2 pour une population et s^2 pour un échantillon) est la somme des carrés des différences par rapport à la moyenne, divisée par la taille de l'ensemble de données.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}$$

x_i : Chaque valeur individuelle dans l'ensemble de données.

m : La moyenne de l'ensemble de données.

n : Le nombre total de valeurs.

ÉCART-TYPE

L'écart-type mesure à quel point les valeurs individuelles dans un ensemble de données diffèrent de la moyenne de l'ensemble. Il indique la dispersion des valeurs autour de la moyenne. Un écart-type élevé signifie que les valeurs sont plus dispersées, tandis qu'un écart-type faible indique que les valeurs sont plus regroupées autour de la moyenne.

Unité de mesure : L'unité de mesure de l'écart-type est la même que celle des valeurs originales, contrairement à la variance qui est mesurée en unités au carré.

Sensibilité aux valeurs extrêmes : Comme la variance, l'écart-type est sensible aux valeurs extrêmes ou aberrantes. Une seule valeur aberrante peut augmenter considérablement l'écart-type.

ÉCART-TYPE

Formule :

La formule pour l'écart type est la racine carrée de la somme des carrés des différences par rapport à la moyenne, divisée par la taille de l'ensemble de données.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}}$$

x_i : Chaque valeur individuelle dans l'ensemble de données.

m : La moyenne de l'ensemble de données.

n : Le nombre total de valeurs.

√ : La racine carrée de la somme.

ÉCART-TYPE VS VARIANCE

La variance est une mesure de la dispersion des données en utilisant des carrés, tandis que l'écart-type est une mesure de la dispersion en utilisant les mêmes unités que les données d'origine, ce qui le rend plus intuitif à interpréter. Les deux mesures sont importantes en statistiques pour comprendre la variabilité des données.

Mesure	Ecart-type	Variance
Qu'est-ce que c'est ?	La racine carrée de la variance	La moyenne des carrés des différences par rapport à la moyenne.
Indication	L'écart entre les nombres dans un ensemble de données.	Le degré moyen selon lequel chaque point diffère de la moyenne.
Unité de mesure	La même chose que les unités dans l'ensemble de données.	En unités carrées ou en pourcentage.
Signification	Une faible déviation standard (écart) signifie une faible volatilité, tandis qu'une déviation standard élevée (écart) signifie une volatilité plus élevée.	Le degré de variation ou de changement des rendements au fil du temps.

INTERVALLE DE CONFIANCE DE LA MOYENNE

L'intervalle de confiance de la moyenne est une plage de valeurs dans laquelle on estime que la vraie moyenne d'une population se trouve avec un certain degré de confiance. Il est souvent utilisé en statistiques pour quantifier l'incertitude associée à l'estimation d'une moyenne à partir d'un échantillon de données.

L'intervalle de confiance pour la moyenne peut être calculé en utilisant la formule générale suivante :

$$\bar{X} - Z\frac{s}{\sqrt{n}} < IC < \bar{X} + Z\frac{s}{\sqrt{n}}$$

ou

$$IC = \bar{X} \pm Z\frac{s}{\sqrt{n}}$$

INTERVALLE DE CONFIANCE DE LA MOYENNE

- ***Explication :***

➢ $\bar{X}$ est la moyenne de l'échantillon, qui est une estimation de la moyenne de la population.

➢ s est l'écart type de l'échantillon, qui mesure la dispersion des données dans l'échantillon. Il est utilisé pour tenir compte de la variabilité des données.

➢ n est la racine carrée de la taille de l'échantillon, et il est utilisé pour ajuster la largeur de l'intervalle en fonction de la taille de l'échantillon.

➢ Z est basé sur le niveau de confiance choisi. Par exemple, pour un niveau de confiance de 95%, Z est égal au score Z correspondant à un niveau de confiance de 95%. En général, pour un niveau de confiance de C% (par exemple, 95%), vous utilisez le score Z correspondant à la probabilité: $(C/100+(1-C/100)/2)(C/100+(1-C/100)/2)$. Ce score Z est souvent noté $Z_{\alpha/2}$.

Confidence Level	α (level of significance)	$Z_{\alpha/2}$
99%	1%	2.575
95%	5%	1.96
90%	10%	1.645

LA DISTRIBUTION

La distribution est un concept fondamental en statistiques qui décrit la façon dont les valeurs d'une variable particulière sont réparties ou distribuées dans un ensemble de données. En d'autres termes, elle nous donne une idée de la fréquence ou de la probabilité de chaque valeur possible pour cette variable.

Les distributions sont essentielles pour comprendre les caractéristiques d'un ensemble de données, analyser des phénomènes naturels, prendre des décisions éclairées et effectuer des prédictions. Deux types de distributions couramment utilisés sont les distributions discrètes, qui s'appliquent aux variables qui prennent des valeurs distinctes et isolées, et les distributions continues, qui s'appliquent aux variables avec une plage infinie de valeurs possibles.

Comprendre les propriétés d'une distribution, telles que sa forme, sa moyenne, son écart type et ses moments, est essentiel pour les statisticiens, les chercheurs et les décideurs afin de mieux appréhender le comportement des données et d'en tirer des conclusions significatives.

TYPES DE DISTRIBUTIONS

Il existe de nombreux types de distributions en statistiques. Certains des types les plus couramment rencontrés :

1. La distribution normale (ou gaussienne) : C'est la distribution la plus célèbre et est caractérisée par une forme en cloche symétrique. Elle est souvent utilisée pour modéliser des phénomènes naturels tels que la taille, le poids, et le QI. La distribution normale est définie par sa moyenne et son écart type.

2. La distribution uniforme : Dans cette distribution, toutes les valeurs possibles ont la même probabilité d'occurrence. Par exemple, lorsqu'on lance un dé équilibré, les valeurs de 1 à 6 ont une probabilité égale d'apparaître.

3. La distribution exponentielle : Elle est souvent utilisée pour modéliser le temps entre les occurrences d'événements rares et aléatoires. Elle est caractérisée par sa fonction de survie décroissante.

4. La distribution de Poisson : Elle est utilisée pour modéliser le nombre d'événements rares dans un intervalle de temps ou d'espace donné, comme le nombre d'appels à un centre d'urgence en une heure.

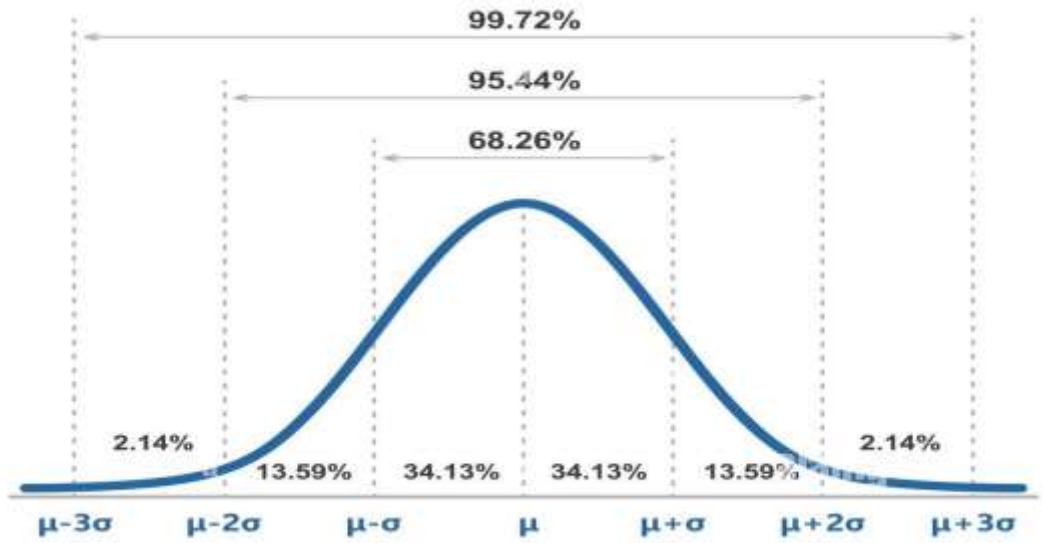
5. La distribution binomiale : Elle modélise le nombre de succès dans un certain nombre d'essais indépendants, chaque essai ayant deux résultats possibles (succès ou échec). C'est couramment utilisé pour des problèmes de probabilité binaire, comme le lancer d'une pièce de monnaie.

6. La distribution de Bernoulli : C'est une distribution spéciale de la distribution binomiale qui modélise un seul essai avec deux résultats possibles (succès ou échec).

LA LOI NORMALE

La loi normale, également connue sous le nom de distribution normale ou distribution gaussienne, est l'une des distributions de probabilité les plus fondamentales en statistiques et en mathématiques. Elle est caractérisée par une courbe en forme de cloche symétrique. La distribution normale est essentielle dans de nombreuses applications statistiques en raison de ses propriétés bien comprises et de sa grande importance dans la modélisation de phénomènes naturels.

La distribution normale est largement utilisée pour modéliser de nombreux phénomènes naturels, tels que la taille des individus dans une population, les scores aux tests standardisés, les erreurs de mesure, et de nombreux autres. Elle joue un rôle clé en statistiques, notamment dans l'estimation des probabilités, la prise de décision statistique, et la réalisation d'inférences statistiques. De plus, de nombreux tests statistiques et techniques d'analyse des données sont basés sur l'hypothèse que les données suivent une distribution normale.

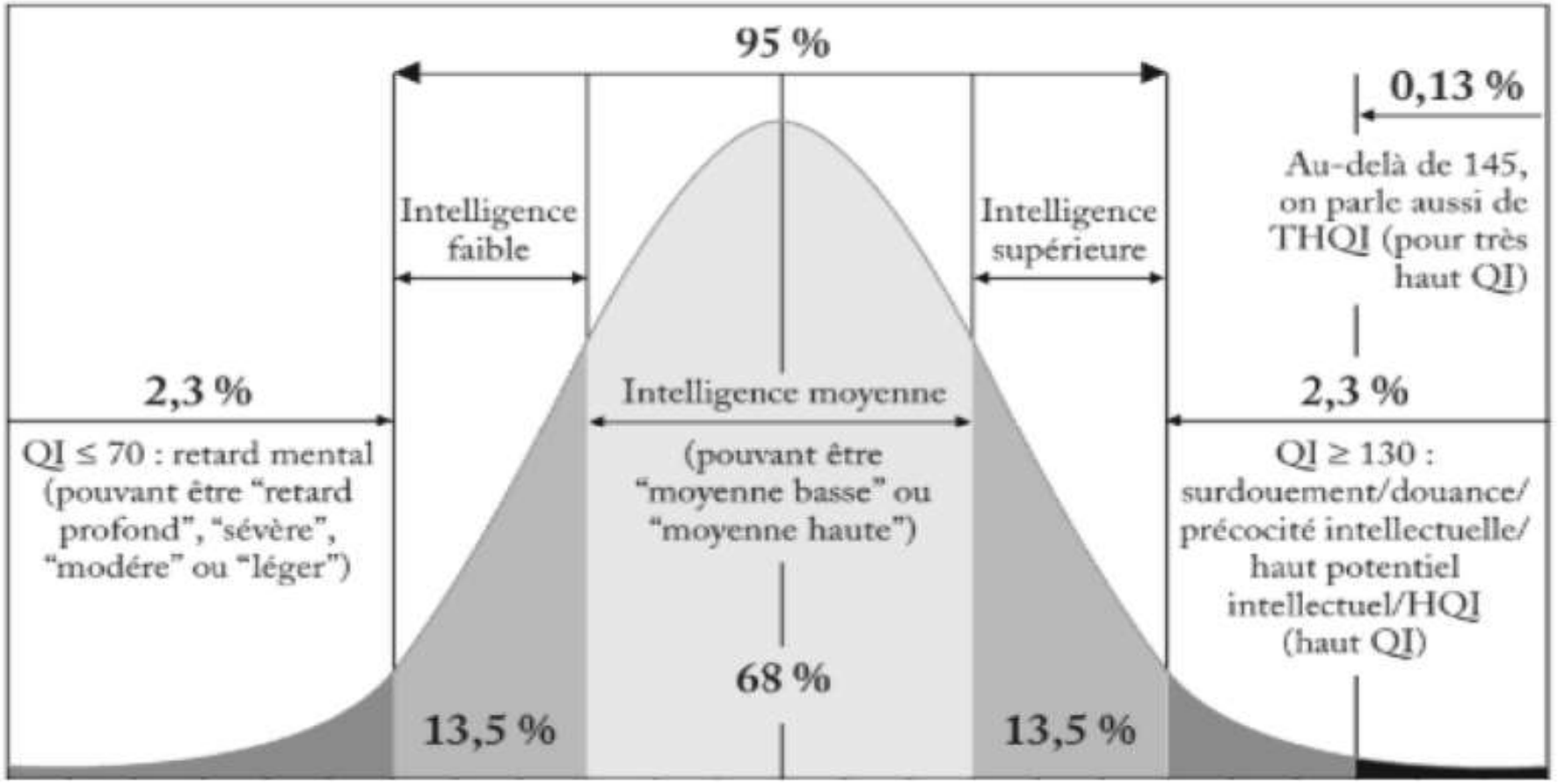


LA LOI NORMALE

- Les caractéristiques principales de la distribution normale sont les suivantes :

1. **Symétrie :** La courbe est symétrique par rapport à sa moyenne, ce qui signifie que la moitié des valeurs se trouvent à gauche de la moyenne, et l'autre moitié à droite.
2. **Moyenne :** La moyenne est le point central de la distribution normale. C'est également le point où la courbe atteint son maximum.
3. **Écart type :** L'écart type, noté σ (sigma), mesure la dispersion des valeurs autour de la moyenne. Plus l'écart type est grand, plus la courbe est étalée, et plus il est petit, plus la courbe est resserrée.
4. **Forme en cloche** : La courbe est étroitement centrée autour de la moyenne, et sa forme ressemble à une cloche. La majorité des valeurs se trouvent près de la moyenne, et à mesure que l'on s'éloigne de la moyenne, la probabilité de rencontrer une valeur diminue.
5. **Distribution continue** : La distribution normale est continue, ce qui signifie qu'elle peut prendre n'importe quelle valeur réelle. Il n'y a pas de "sauts" entre les valeurs.
6. **Asymptotique** : Les queues de la courbe s'étendent à l'infini dans les deux directions, bien qu'elles deviennent rapidement proches de zéro à mesure que l'on s'éloigne de la moyenne.

EXEMPLE : DISTRIBUTION « NORMALE » DE L'INTELLIGENCE



STATISTIQUES INFÉRENTIELLES

STATISTIQUES INFÉRENTIELLES

La statistique inférentielle est une branche de la statistique qui vise à tirer des conclusions et à faire des prédictions sur une population ou un ensemble de données à partir d'un échantillon représentatif de cette population.

Elle repose sur l'idée que l'on peut généraliser à une population plus large en se basant sur des observations faites sur un échantillon plus restreint, à condition que l'échantillon soit sélectionné de manière aléatoire et que des méthodes statistiques appropriées soient utilisées.

Les méthodes couramment utilisées en statistique inférentielle comprennent l'utilisation de tests de signification, de l'intervalle de confiance, de la régression, de l'analyse de variance, et d'autres techniques statistiques avancées.

la statistique inférentielle est un outil essentiel dans la prise de décisions basées sur des données, en permettant d'extrapoler à partir d'un échantillon pour faire des généralisations et des prédictions sur une population plus vaste.

STATISTIQUES INFÉRENTIELLES

Les principaux objectifs de la statistique inférentielle incluent :

- **Estimation :** L'estimation consiste à estimer les paramètres d'une population à partir des données de l'échantillon. Par exemple, on peut estimer la moyenne, la variance, ou d'autres caractéristiques de la population en se basant sur les données de l'échantillon.
- **Test d'hypothèses :** Les tests d'hypothèses permettent de vérifier des affirmations ou des hypothèses sur une population en se basant sur les données de l'échantillon. Par exemple, on peut tester si une nouvelle méthode de traitement est plus efficace qu'une méthode existante.
- **Prédiction :** La statistique inférentielle permet également de faire des prédictions sur des événements futurs en se basant sur des données passées. Par exemple, on peut prédire les ventes futures d'un produit en se basant sur les ventes passées.

TEST D’HYPOTHÈSE

Un test d'hypothèse est une procédure statistique utilisée pour évaluer une affirmation ou une hypothèse concernant une population à partir d'un échantillon de données. Il permet de déterminer si les données de l'échantillon fournissent des preuves suffisantes pour accepter ou rejeter une hypothèse concernant une caractéristique ou un paramètre de la population sous-jacente. Les tests d'hypothèses sont couramment utilisés en statistique pour prendre des décisions basées sur des données.

Les tests d'hypothèses sont largement utilisés dans de nombreux domaines de la recherche et de la prise de décisions pour évaluer des hypothèses, que ce soit en sciences, en économie, en médecine, ou dans d'autres domaines où des données sont collectées et analysées.

TEST D'HYPOTHÈSE

Un test d'hypothèse suit généralement un processus en deux étapes :

1. ***Formulation des hypothèses :***

   - **Hypothèse nulle (H_0)** : C'est l'hypothèse de départ ou l'hypothèse à tester. Elle stipule généralement qu'il n'y a pas de différence, pas d'effet, ou pas de relation dans la population.

   - **Hypothèse alternative (H_1 ou H_a) :** C'est l'hypothèse que l'on cherche à prouver, et elle indique généralement que la population présente une différence, un effet, ou une relation spécifique.

2. ***Collecte et analyse des données :***

   - On collecte des données à partir de l'échantillon.

   - On utilise des techniques statistiques pour calculer une statistique de test, qui mesure à quel point les données de l'échantillon sont compatibles ou non avec l'hypothèse nulle.

TEST D'HYPOTHÈSE

En fonction de la statistique de test calculée, on peut prendre l'une des deux décisions suivantes:

- **Rejeter l'hypothèse nulle (H0)** : Cela signifie que les données de l'échantillon fournissent suffisamment de preuves pour soutenir l'hypothèse alternative (H1). En d'autres termes, il y a une différence, un effet, ou une relation significative dans la population.

- **Ne pas rejeter l'hypothèse nulle (H0)** : Cela signifie que les données de l'échantillon ne fournissent pas suffisamment de preuves pour soutenir l'hypothèse alternative (H1). On ne peut pas conclure qu'il y a une différence, un effet, ou une relation significative dans la population.

Il est important de noter que le terme "ne pas rejeter l'hypothèse nulle" ne signifie pas que l'hypothèse nulle est prouvée ou que l'effet est nul, mais seulement que l'on n'a pas suffisamment de preuves pour la rejeter.

TEST D'HYPOTHÈSE

Pour rejeter l'hypothèse nulle (H_0) dans un test d'hypothèse :

1. **Collecte et analyse des données :** Collectez un échantillon de données et effectuez les calculs ou tests statistiques nécessaires. Le choix de la méthode et de la statistique de test dépend du type de données et de la question de recherche.

2. **Calculez la statistique de test :** Calculez une statistique de test qui résume l'information dans vos données d'échantillon. Le choix de la statistique de test dépend du test d'hypothèse spécifique que vous effectuez. Les statistiques de test courantes comprennent les t-scores, les z-scores, les statistiques du chi-carré et les statistiques F, entre autres.

3. **Déterminez le niveau de signification (Alpha) :** Choisissez un niveau de signification (souvent noté α), qui représente le seuil de signification statistique. Les choix courants pour alpha incluent 0,05 ou 0,01, ce qui correspond à un niveau de signification de 5 % ou 1 %, respectivement.

TEST D'HYPOTHÈSE

4. **Comparez la statistique de test à la valeur critique ou à la valeur p :** Il existe deux approches courantes pour décider de rejeter ou non l'hypothèse nulle :

- ***Approche de la valeur critique :*** Si votre statistique de test est plus extrême (par exemple, plus élevée ou plus basse) que la valeur critique provenant d'un tableau ou calculée pour votre niveau de signification choisi, vous pouvez rejeter l'hypothèse nulle. Dans ce cas, la valeur p (probabilité) est souvent utilisée pour déterminer la valeur critique.

- ***Approche de la valeur p :*** Calculez la valeur p associée à votre statistique de test. Si la valeur p est inférieure ou égale à votre niveau de signification choisi (α), vous pouvez rejeter l'hypothèse nulle. Plus la valeur p est faible, plus les preuves contre l'hypothèse nulle sont fortes.

TESTS PARAMÉTRIQUES

Un test paramétrique est une approche statistique employée pour évaluer des hypothèses concernant les caractéristiques d'une population en se basant sur certaines conditions préalables relatives à cette population, comme la distribution des données, qui doivent être respectées.

Les tests paramétriques sont basés sur des suppositions spécifiques concernant les propriétés des données, en particulier la distribution des données, généralement supposée être une distribution normale.

Exemples courants de tests paramétriques comprennent le test t de Student, l'analyse de la variance (ANOVA), la régression linéaire, et le test de corrélation de Pearson. Ces tests sont utilisés pour comparer des moyennes, des variances, ou évaluer des relations linéaires entre variables.

TESTS PARAMÉTRIQUES

caractéristiques des tests paramétriques :

Distribution normale : Les tests paramétriques supposent souvent que les données suivent une distribution normale (ou gaussienne). Cela signifie que les données sont symétriques et se répartissent de manière spécifique autour de la moyenne. Les tests paramétriques peuvent être moins fiables si cette supposition n'est pas satisfaite.

Homoscédasticité : Les tests paramétriques supposent également une variance égale dans les groupes ou échantillons comparés. Cela signifie que la dispersion des données est constante à travers les groupes.

Niveau d'échelle : Les tests paramétriques sont conçus pour les données qui sont à un niveau d'échelle, c'est-à-dire des données quantitatives qui ont un sens en termes de valeurs numériques, telles que les mesures de longueur, de poids, de temps, etc.

TEST T DE STUDENT

Le test t de Student, souvent appelé simplement test t, est une méthode statistique couramment utilisée pour comparer les moyennes de deux groupes d'échantillons afin de déterminer s'il y a une différence significative entre eux. Il a été développé par le statisticien William Sealy Gosset, qui a utilisé le pseudonyme "Student" lors de sa publication initiale en 1908.

Le test t de Student est généralement utilisé dans ***la Comparaison de deux groupes*** : Lorsque vous souhaitez comparer deux groupes (par exemple, un groupe de contrôle et un groupe expérimental) pour déterminer s'il existe une différence significative entre les moyennes des deux groupes.

TESTS T DE STUDENT

Il existe 3 variantes du test t, chacune adaptée à des situations spécifiques :

- ***Le test t de Student pour un échantillon unique :*** Utilisé pour comparer la moyenne d'un seul groupe à une valeur connue ou attendue.
- ***Le test t de Student pour 2 groupes indépendants :*** Utilisé pour comparer les moyennes de deux groupes indépendants.
- ***Le test t de Student pour 2 groupes dépendants ou appariés :*** Utilisé pour comparer les moyennes de deux groupes appariés ou dépendants, où les mêmes individus sont mesurés avant et après un traitement, par exemple.

LE TEST T DE STUDENT POUR UN ÉCHANTILLON UNIQUE

Le test t pour un échantillon unique (ou test t unilatéral) est utilisé pour déterminer si la moyenne d'un échantillon provient d'une population avec une moyenne connue ou attendue. Il repose sur la comparaison de la moyenne de l'échantillon à cette valeur de référence.

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

$\bar{x}$ est la moyenne de l'échantillon.

μ_0 est la moyenne attendue ou la valeur de référence sous l'hypothèse nulle.

s^2 est la variance de l'échantillon (la mesure de la dispersion des données).

n est la taille de l'échantillon (le nombre d'observations dans l'échantillon).

LE TEST T DE STUDENT POUR UN ÉCHANTILLON UNIQUE

Exercice:

Un athlète de haut niveau prétend que sa performance en saut en hauteur est en moyenne supérieure à 2 mètres. Pour tester cette affirmation, une équipe de recherche a collecté des données de 10 de ses sauts en hauteur. Les hauteurs en mètres sont les suivantes :

2.05, 2.10, 1.95, 2.15, 2.20, 2.25, 2.05, 2.10, 2.18, 2.12

Question 1 : Formulez l'hypothèse nulle (H0) et l'hypothèse alternative (H1) pour ce test t unilatéral.

Question 2 : Calculez la moyenne, l'écart-type et la taille de l'échantillon pour les données de sauts en hauteur.

Question 3 : Effectuez un test t unilatéral pour déterminer si la performance moyenne en saut en hauteur est statistiquement supérieure à 2 mètres. Utilisez un niveau de signification de 0,05.

Question 4 : Interprétez les résultats du test t. Quelle conclusion pouvez-vous tirer concernant la performance en saut en hauteur de l'athlète ?

LE TEST T DE STUDENT POUR 2 GROUPES INDÉPENDANTS

Le test t de Student pour deux groupes indépendants (également appelé test t indépendant) est une méthode statistique utilisée pour comparer les moyennes, **d'une variable numérique mesurée en valeurs continues**, de deux groupes distincts afin de déterminer s'il existe une différence significative entre ces moyennes.

❖ Les variance des deux groupe ***sont égales***, La formule est la suivante:

$$t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\left(\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

LE TEST T DE STUDENT POUR 2 GROUPES INDÉPENDANTS

- Les variances des deux groupes ***ne sont pas égales***, La formule est la suivante:

$$t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

LE TEST T DE STUDENT POUR 2 GROUPES INDÉPENDANTS

$\overline{x_1}$ Moyenne du Groupe 1

$\overline{x_2}$ Moyenne du Groupe 2

n_1 Taille de l'effectif du groupe 1

n_2 Taille de l'effectif du groupe 2

s_1^2 Ecart-type du Groupe 1

s_2^2 Ecart-type du Groupe 2

LE TEST T DE STUDENT POUR 2 GROUPE APPARIÉS OU DÉPENDANTS

Le test t de Student pour deux groupes dépendants (aussi appelé test t apparié ou test t pour échantillons appariés) est une méthode statistique utilisée pour comparer les moyennes de deux groupes qui sont appariés ou reliés d'une manière ou d'une autre. Cette méthode est couramment utilisée lorsque les mêmes individus ou éléments sont mesurés à deux moments différents, ou lorsqu'il y a une correspondance entre les observations dans les deux groupes.

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

$\bar{d}$ est la moyenne des différences entre les paires de mesures après-avant.

s_d est l'Ecart type de la moyenne des différences.

n est la taille de l'échantillon (le nombre d'observations dans l'échantillon).

LE TEST T DE STUDENT POUR 2 GROUPE APPARIÉS OU DÉPENDANTS

$\bar{d}$ **:** La notation "d-barre" représente la moyenne des différences entre deux groupes ou deux mesures appariées. Elle est couramment utilisé dans les tests t appariés (tests t pour deux groupes dépendants), où vous mesurez une variable avant et après une intervention ou une modification. C'est la moyenne de ces différences entre les deux moments.

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

$$d_i = après_i - avant_i$$

Exemple:

Si n = 4, la formule pour calculer la moyenne de la différence $\bar{d}$ entre les 2 groupes est la suivante:

$$\bar{d} = \frac{(après_1 - avant_1) + (après_2 - avant_1) + (après_3 - avant_3) + (après_4 - avant_4)}{4}$$

INTERPRÉTATION DE LA VALEUR DE T

L'interprétation du test t de Student en utilisant la méthode de valeur critique implique de comparer la statistique de test t calculée (généralement notée t) à une valeur critique t associée à un certain niveau de signification (généralement noté α). Voici les étapes de la procédure d'interprétation en utilisant la méthode de valeur critique :

1. **Formule du test t :** Avant de commencer, assurez-vous que vous avez calculé correctement la statistique de test t à partir de vos données, en utilisant la formule appropriée pour votre test t spécifique (par exemple, test t unilatéral ou bilatéral, test t pour échantillons indépendants ou appariés, etc.).

2. **Définissez l'hypothèse nulle et l'hypothèse alternative :** Clarifiez vos hypothèses avant de poursuivre. L'hypothèse nulle (H_0) est généralement formulée comme l'absence de différence ou d'effet, tandis que l'hypothèse alternative (H_1 ou H_a) exprime la différence ou l'effet que vous souhaitez prouver.

INTERPRÉTATION DE LA VALEUR DE T

3. **Choisissez un niveau de signification (α) :** Le niveau de signification, souvent fixé à 0,05 (5%), représente le seuil de probabilité au-delà duquel vous rejetez l'hypothèse nulle. Vous pouvez également choisir d'autres niveaux de signification, en fonction de la précision requise pour votre étude.

4. **Déterminez les degrés de liberté *(df)* :** Les degrés de liberté dépendent du type de test t que vous effectuez. Assurez-vous de connaître le nombre de degrés de liberté associé à votre test t, car cela affectera la valeur critique t.

df pour un seul échantillon

$$df = n - 1$$

n : nombre d'observations

df pour 2 groupes appariés ou dépendants

$$df = n - 1$$

n : nombre de paires

INTERPRÉTATION DE LA VALEUR DE T

- ***df pour 2 Groupes indépendants avec variances égales:***

$$df = n_1 + n_2 - 2$$

n_1 : nombre d'observations du groupe 1 ; n_2 : nombre d'observations du groupe 2

- ***df pour 2 Groupes indépendants avec des variances inégales:***

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2\,(n_1 - 1)} + \frac{s_2^4}{n_2^2\,(n_2 - 1)}}$$

s_1^2 : variance du groupe 1 ; s_2^2 : variance du groupe 2

n_1 : nombre d'observations du groupe 1 : n_2 : nombre d'observations du groupe 2

INTERPRÉTATION DE LA VALEUR DE T

5. **Consultez une table de valeurs critiques t :** Trouver des tables de valeurs critiques t en fonction de différents niveaux de signification (α) et de degrés de liberté. Vous pouvez également utiliser des logiciels statistiques ou des calculateurs en ligne pour trouver la valeur critique t. La table vous fournira une valeur t critique correspondant à votre α et à vos degrés de liberté.

6. **Comparez la statistique de test t à la valeur critique t :** Comparez la statistique de test t calculée avec la valeur t critique à partir de la table. Voici comment interpréter le résultat :

- Si la statistique de test t est supérieure à la valeur t critique ***(t_calculé > t_critique)***, vous pouvez rejeter l'hypothèse nulle. Cela suggère que vous avez des preuves statistiques d'une différence significative ou de l'effet que vous vouliez prouver.

- Si la statistique de test t est inférieure ou égale à la valeur t critique ***(t_calculé ≤ t_critique)***, vous ne pouvez pas rejeter l'hypothèse nulle. Cela signifie que vous n'avez pas suffisamment de preuves statistiques pour conclure à une différence significative.

INTERPRÉTATION DE LA VALEUR DE T

7. **Rédigez une conclusion :** Sur la base de la comparaison entre la statistique de test t et la valeur critique t, rédigez une conclusion en rapport avec vos hypothèses. Indiquez si vous avez réussi à prouver l'effet ou la différence que vous cherchiez.

- *Il est essentiel de respecter la méthode de valeur critique pour interpréter correctement les résultats du test t de Student. Cela vous aidera à prendre des décisions informées en fonction des preuves statistiques que vous avez recueillies.*

INTERPRÉTATION DE LA VALEUR DE T

Table de distribution t unilatérale (à 95% de niveau de confiance) pour les degrés de liberté (df) de 1 à 25 avec les valeurs critiques de t.

Degrés de liberté (df)	*Valeur critique t*
1	6.314
2	2.92
3	2.353
4	2.132
5	2.015
6	1.943
7	1.895
8	1.86
9	1.833
10	1.812
11	1.796
12	1.782
13	1.771
14	1.761
15	1.753
16	1.746
17	1.739
18	1.734
19	1.729
20	1.725
21	1.721
22	1.717
23	1.714
24	1.711
25	1.708

MERCI POUR VOTRE ATTENTION

"Que la réussite accompagne chacun de vos pas et que le chemin que vous empruntez mène vers la concrétisation de vos rêves"