Visualisation des données & inférence statistique

UQAC | Université du Québec à Chicoutimi

Visualisation des données

Types de Variables	Types de graphiques
Une variable catégorique ou numérique discrète	Diagramme en secteurs
	Diagramme en bâtons
Une variable numérique	Histogramme
	Diagramme arborescent
	Diagramme en boîte
	Diagramme quantile-quantile théorique
	Estimateur à noyau
Deux variables catégoriques ou numériques discrètes	Diagramme en bâtons empilés
	Diagramme en bâtons groupés
Une variable catégorique et une variable numérique	Histogrammes groupés
	Diagrammes en boîte groupée
Deux variables numériques	Diagramme de dispersion
	Diagramme temporel
	Estimateur à noyau
Trois variables ou plus	Diagramme de dispersion groupé
	Diagramme temporel groupé
	Matrice de diagrammes de dispersion

UQAC

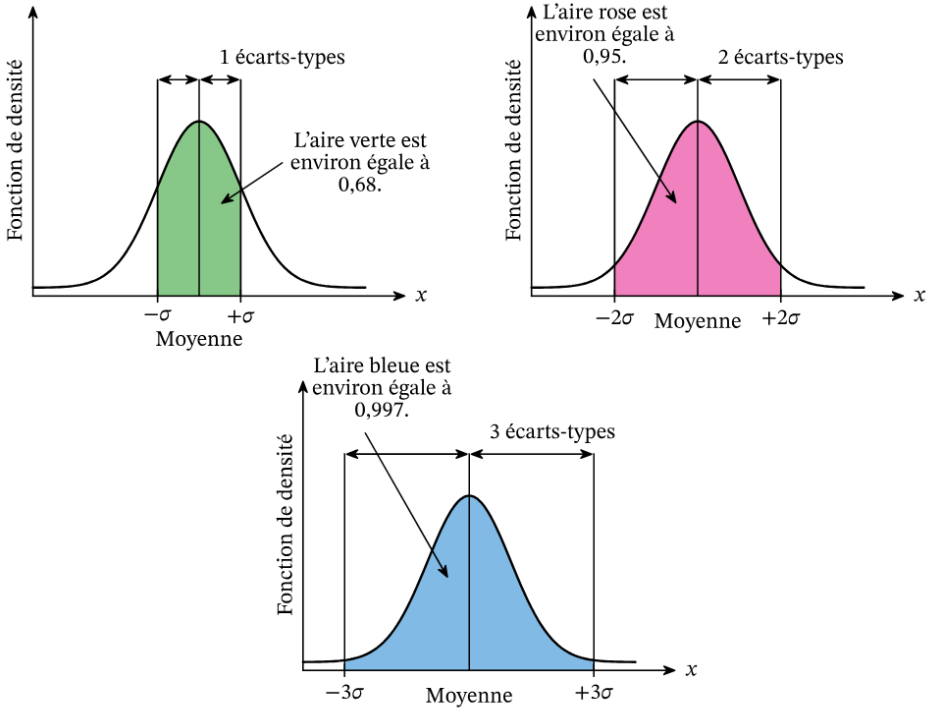
Inférence statistique

Diverses formes de distribution:

1. Normale
2. Khi-Deux
3. Binomiale
4. Poisson

Distribution normale

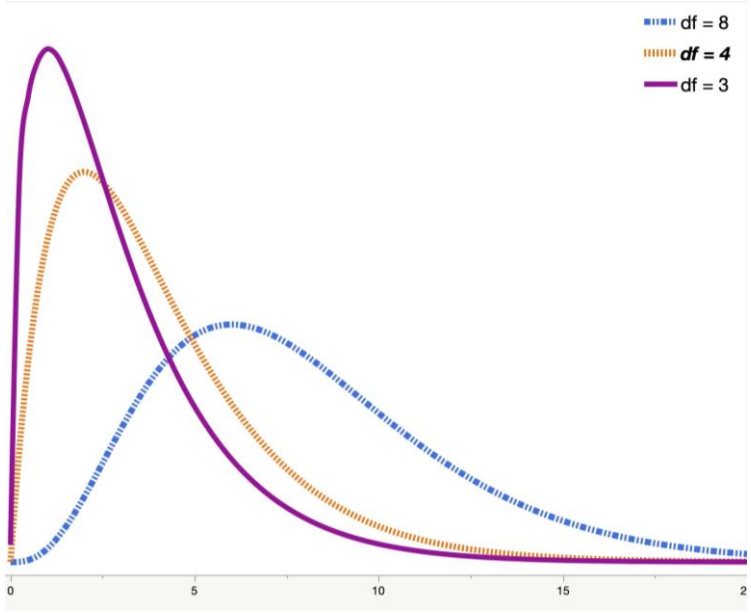
- Forme de cloche symmétrique
- Moyenne =Mode=Médiane
- Deux paramètres:
  - Moyenne (μ)
  - Écart-type (σ)



UQAC

Distribution du Khi-Deux

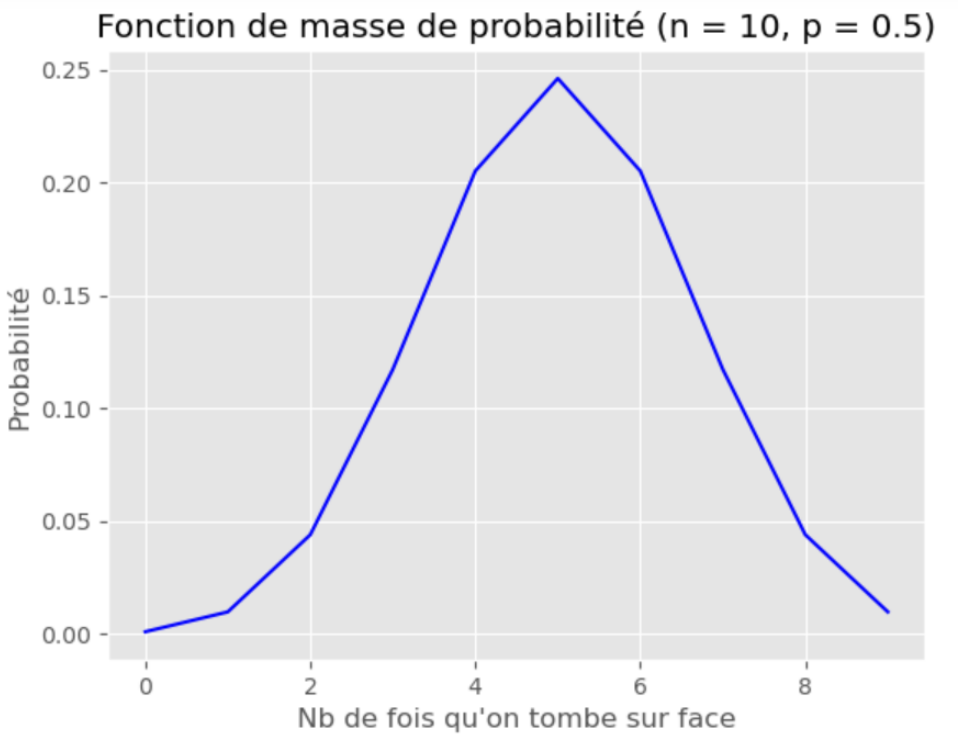
- Asymétrie vers la droite
- Un seul paramètre appelé le degré de liberté (df)
- Degré de liberté: nombre de valeurs dans l'étude qui sont libres de varier indépendamment les unes des autres
- Df=nb variables - 1



UQAC

Distribution Binomiale

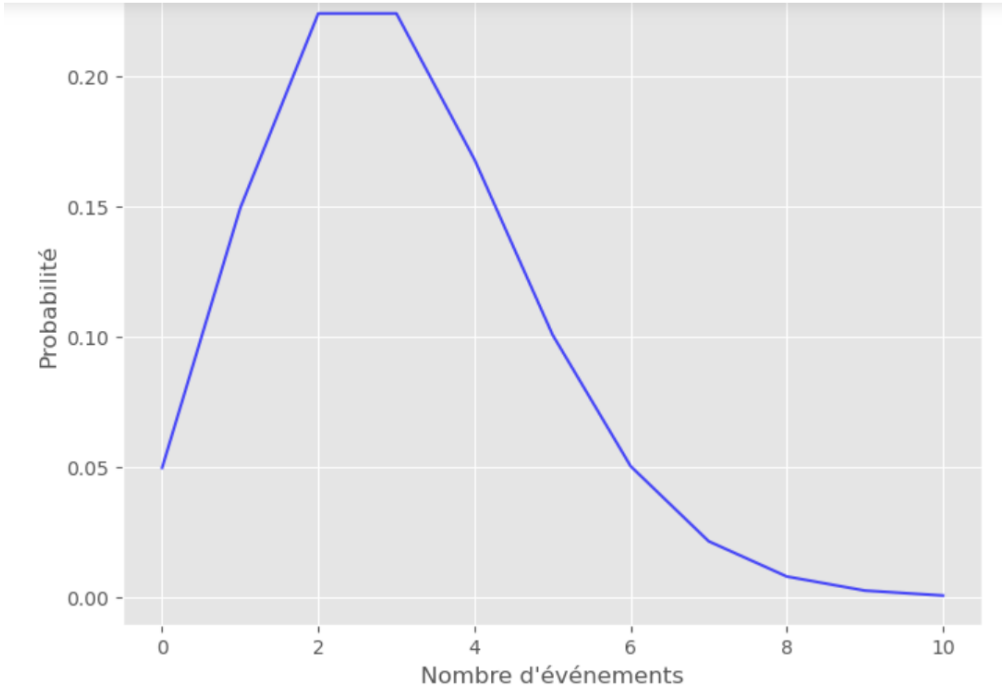
- Décrire le nombre de succès dans un nombre fixe d'expériences indépendantes
- Modéliser des événements binaire
- Deux paramètres:
  - le nombre d'essais (n)
  - la probabilité de succès dans chaque essai (p)



UQAC

Distribution de Poisson

- Décrit le nombre d'événements se produisant dans un intervalle de temps ou d'espace fixe
- Un seul paramètre:
  - λ (lambda): nombre moyen d'événements dans l'intervalle donné.



UQAC

Z-score

Le score Z représente combien d'écart-type sépare une observation d'un jeu de données à la moyenne de celle-ci. L'utilisation du z-score est courante dans les analyses statistiques, en particulier lors de la détection des valeurs extrêmes dans un ensemble de données. Les points de données ayant un z-score élevé (en valeur absolue) sont généralement considérés comme des valeurs extrêmes et peuvent nécessiter un examen plus approfondi.

Il peut être calculé à l'aide de la formule suivante :

$$Z = \frac{X - \mu}{\sigma}$$

Où :

- X est l'observation.
- μ est la moyenne de l'échantillon.
- σ est l'écart type de l'échantillon.

UQAC

Intervalle de confiance

Un intervalle de confiance est une plage de valeurs dans laquelle on estime qu'un paramètre statistique, comme la moyenne, est susceptible de se trouver, avec un certain niveau de confiance. En d'autres termes, c'est une méthode statistique utilisée pour estimer l'incertitude autour d'une mesure statistique d’une population basée sur un échantillon de celle-ci.

L'intervalle de confiance (IC) à 95 % est un intervalle calculé à partir d'un échantillon de données provenant d'une séquence infinie, dont 95 % comprennent le paramètre de population. À long terme, 95 % de ces intervalles incluent la véritable moyenne.

Plus le niveau de confiance est élevé (par exemple, 95% au lieu de 90%), plus l'intervalle de confiance sera large pour accueillir une plus grande incertitude.

UQAC

Les tests statistiques

Les tests statistiques permettent de valider une hypothèse par rapport à des données.

Voici les étapes afin de réaliser un test statistique:

1. Construire les hypothèses H0 et H1
2. Déterminer le risque d'erreur acceptable
3. Choisir le test adapté
4. Calculer la *p-value* et l'interpréter

L'hypothèse nulle et l'hypothèse alternative

L'hypothèse nulle (H0) représente le statu quo tandis que l'hypothèse alternative (H1) est l'hypothèse que l'on souhaite démontrer.

Exemple:

Supposons que nous ayons deux médicaments (A et B), et nous voulons savoir si le médicament A est plus efficace que le médicament B en termes de réduction de la pression artérielle. Nous prélevons des échantillons de patients sous chaque traitement et mesurons leur pression artérielle

Voici ce que pourrait être les deux hypothèses:

Hypothèse Nulle (H0) : Les moyennes de la pression artérielle des patients sous les deux médicaments sont égales.

Hypothèse Alternative (H1): La moyenne de la pression artérielle des patients sous le médicament A est inférieure à la moyenne de la pression artérielle des patients sous le médicament B.

Risque d'erreur

Il existe deux types d'erreur:

- Type 1: On suppose qu'H0 est fausse alors qu'en réalité H0 est vraie
- Type 2: On suppose qu'H0 est vraie alors qu'en réalité H0 est fausse

Par défaut, le pourcentage acceptable de commettre une erreur de type 1 (aussi appelé alpha) est fixée à 5%. Ainsi, on accepte qu'on a 5% de chance de rejeter H0 si elle est vraie.

Dans le cas de notre exemple, cela reviendrait à dire que le traitement A est efficace à faire diminuer la pression artérielle.

L'erreur de type 2 dans notre exemple reviendrait à dire qu'il n'y a pas de différence au niveau de la pression artérielle en fonction du traitement même si le traitement A est en fait plus efficace pour faire diminuer la pression artérielle que le traitement B.

Risque d’erreur

Relation avec le % de chance de faire une erreur de type 1 & l’intervalle de confiance

$$IC= (1-alpha) x100\%$$

UQAC

Les tests unilatéraux et bilatéraux

La différence réside au niveau de l'hypothèse alternative (H1).

Il existe deux types de tests:

1. Unilatéraux
   - il existe une différence entre les deux groupes dont le sens est connu
   - ***lower-tailed test:*** La moyenne sous le médicament A < la moyenne sous le médicament B
   - ***upper-tailed test:***La moyenne sous le médicament A > la moyenne sous le médicament B.
2. Bilatéraux.
   - il existe une différence entre les deux groupes mais on ne mentionne pas le sens.

Choix du test

Il existe beaucoup de tests statistiques différents. Chaque test a ses propres conditions d'application.

Afin de déterminer le test adapté, on a besoin d'identifier certains critères :

- Nature des variables (catégorielle, continues etc.)
- Nombre de groupes
- Normalité de la distribution des variables

Dans les prochaines sections, nous allons couvrir les tests suivants:

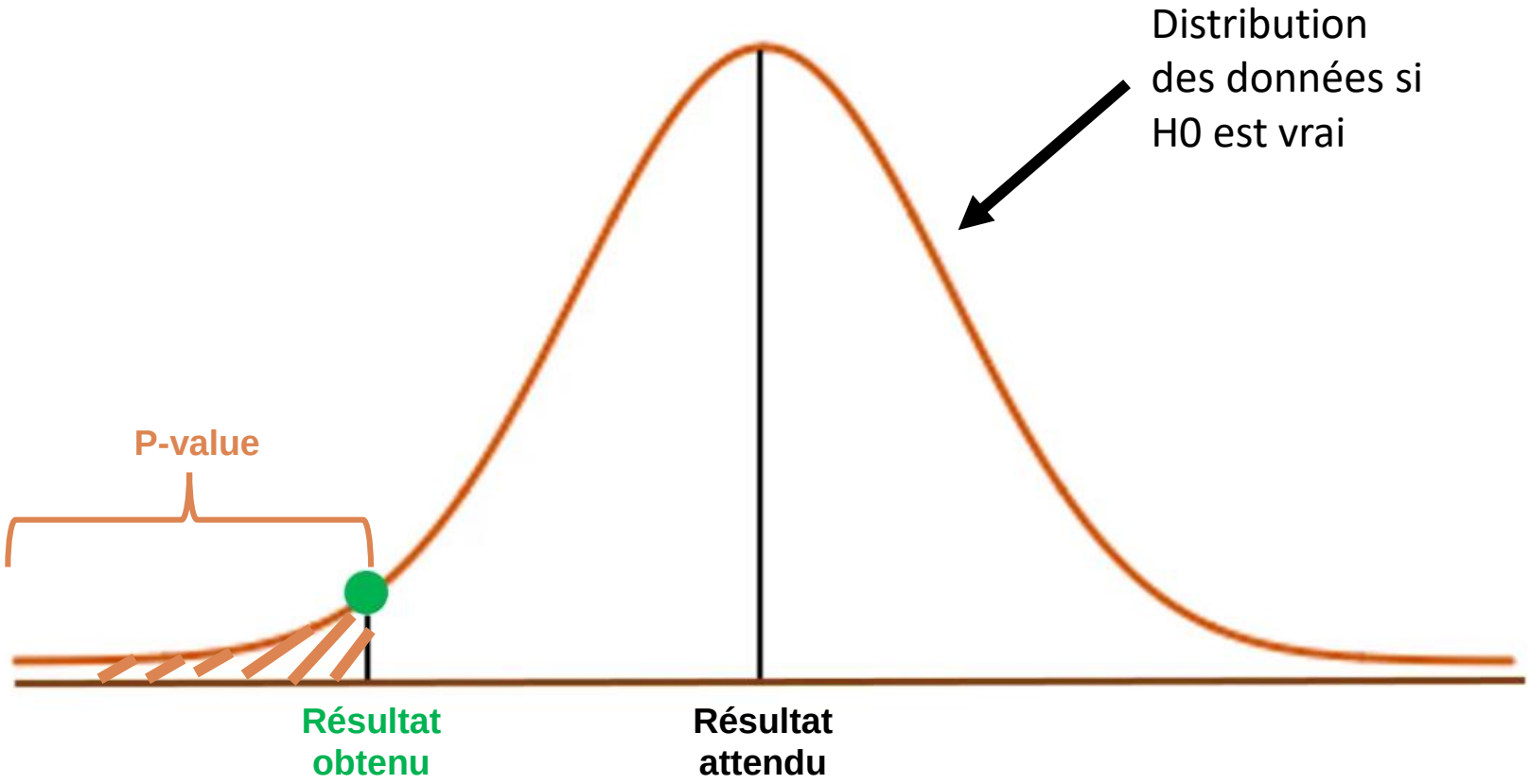
- Z-test
- T-test
- Test du Khi-Deux
- ANOVA

P-Value

- La *p-value* quantifie la probabilité d'observer des données similaires à celles observées, supposant que l'hypothèse nulle est vraie.
- **Si la *p-value* < alpha (5%),** cela suggère que les données qu'on a observées sont peu probables si l'hypothèse nulle est vraie. Cela signifie qu'il y a suffisamment de preuves pour rejeter l'hypothèse nulle au profit de l'hypothèse alternative.
- **Si la *p-value* > alpha (5%),** cela indique que les données qu'on a observées sont assez probables même si l'hypothèse nulle est vraie. Cela signifie qu'il n'y a pas suffisamment de preuves pour rejeter l'hypothèse nulle.

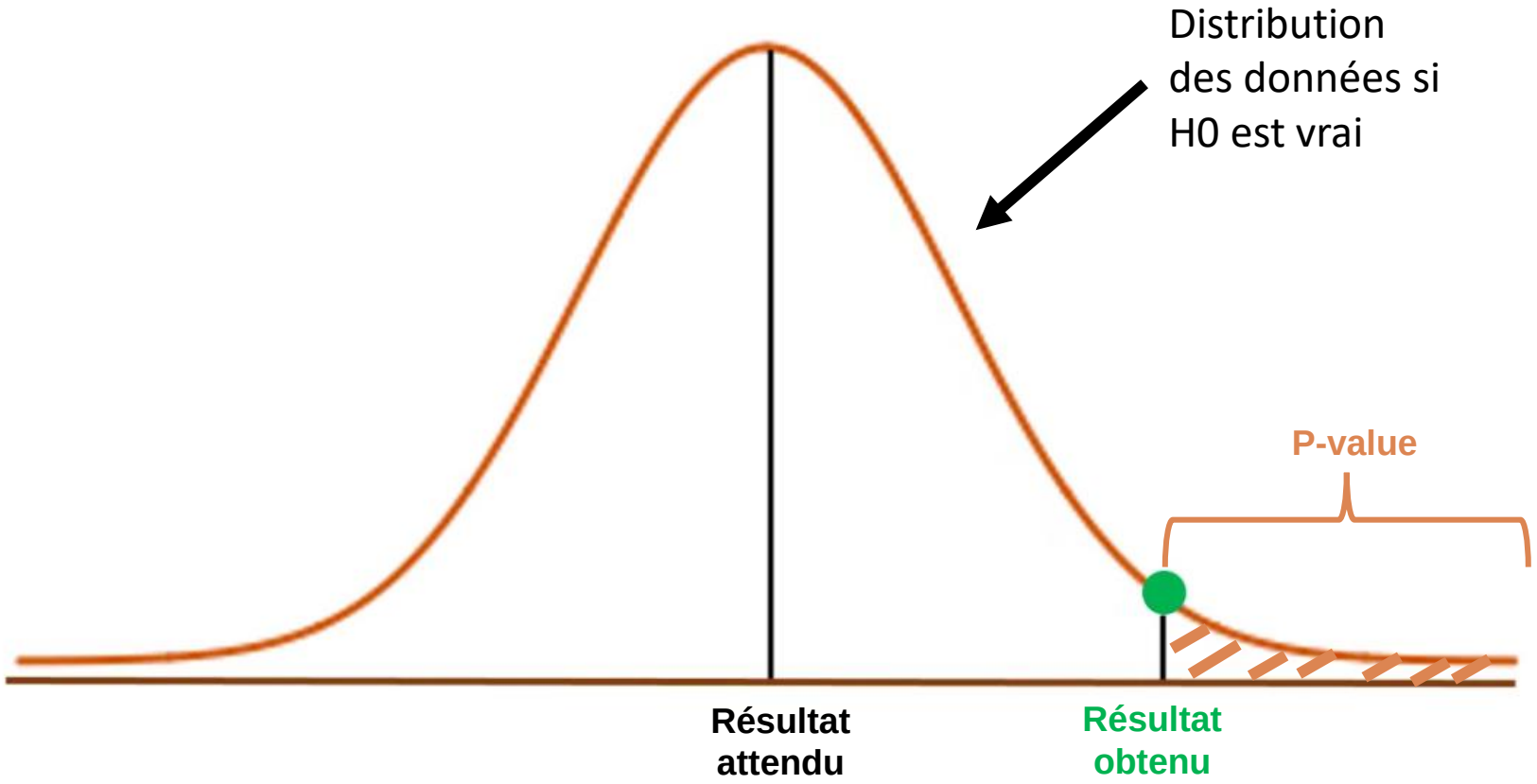
Il est important de noter que la p-value n'indique pas la probabilité que l'hypothèse nulle soit vraie ou fausse. Elle mesure la probabilité d'obtenir les données observées, compte tenu de l'hypothèse nulle.

P-Value (*lower-tailed)*



UQAC

P-Value (*upper-tailed)*



UQAC

Z-Test

Le Z-test est utilisé pour évaluer les différences entre les moyennes de deux groupes lorsqu'on connaît l'écart-type de la population de notre échantillon de données ou que la taille de l'échantillon de données est supérieure à 30.

Il existe trois principaux types de Z-test:

- Z-test pour échantillon unique (comparaison d'une moyenne avec une moyenne connue)
- Z-test indépendant (2 groupes indépendants)
- Z-test apparié (2 groupes dépendants)

Le test s'effectue en calculant le score Z (légèrement modifié) et en calculant la p-value associée à celle-ci grâce à un tableau de distribution normale.

Z-test pour échantillon unique

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Où :

- $\bar{X}$ est la moyenne.
- μ_0 est la moyenne de la population
- σ est l'écart-type de la population.
- n est la taille de l'échantillon.

UQAC

Z-test indépendant

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Où :

- $\bar{X}_1$ et $\bar{X}_2$ sont les moyennes des deux groupes.
- σ_1^2 et σ_1^2 sont les écart-types des populations des 2 groupes.
- n_1 et n_2 sont les tailles des échantillons des deux groupes.

UQAC

Z-test apparié

$$Z = \frac{\bar{X}_d}{\frac{\sigma_d}{\sqrt{n}}}$$

Où :

- $\bar{X}_d$ est la moyenne des différences entre les paires.
- σ est l'écart type des différences entre les paires de la population.
- n est la taille de l'échantillon.

UQAC

T-test

Le T-test est également une statistique inférentielle utilisée pour évaluer les différences entre les moyennes de deux groupes.

La différence avec le Z-test est qu'il peut être utilisé lorsqu'on ne connait pas l'écart-type de la population et pour des échantillons qui peuvent être de plus petite taille que le Z-test.

Il existe trois principaux types de T-test:

- T-test pour échantillon unique
- T-test indépendant
- T-test apparié

T-test pour échantillon unique

Le T-test pour échantillon unique est utilisé pour comparer la moyenne d’un échantillon à une moyenne standard connue.

Hypothèses:

- H0: La moyenne de l'échantillon est égale à la moyenne standard connue
- H1: La moyenne n'est pas égale à la moyenne standard connue (test bilatéral)

Voici les conditions à remplir pour pouvoir l'utiliser:

- le jeu de données provient d'un échantillon aléatoire
- Les données sont numériques et continues
- Le jeu de données doit avoir une distribution relativement normale ou l'échantillon est suffisamment grand
- Pas de valeurs extrêmes

T-test indépendant

Ce test est utilisé pour comparer les moyennes de deux groupes et déterminer s'ils sont statistiquement différents l'un de l'autre.

Hypothèses:

- Hypothèse Nulle (H0) : Les moyennes des deux groupes sont égales
- Hypothèse Alternative (H1): Les moyennes des deux groupes sont différentes.

Voici les conditions à remplir pour pouvoir l'utiliser:

1. La variable dépendante doit être continue.
2. La variable indépendante doit avoir 2 groupes (catégories).
3. Les observations doivent être indépendantes.
4. Pas de valeurs extrêmes dans la variable dépendante
5. La variable dépendante doit avoir une distribution relativement normale au sein de chacun des groupes de la variable indépendante ou l'échantillon est suffisamment grand.

T-test indépendant

Il existe 2 versions du T-test indépendant qui peut être utilisé en fonction de la condition suivante:

1. Si les écarts-types des valeurs de la variable dépendante sont relativement égale pour les deux groupes, on peut utiliser le **Student T-test**.
2. Sinon, on peut utiliser le **Welch T-test**.

T-test apparié

Le test t apparié est utilisé pour comparer les moyennes de deux groupes appariés ou dépendants, comme des mesures avant et après un traitement sur le même groupe de personnes.

Hypothèses:

- Hypothèse Nulle (H0) : Les moyennes des deux groupes appariés sont égales.
- Hypothèse Alternative (H1) : Les moyennes des deux groupes appariés sont différentes.

T-test apparié

Voici les conditions à remplir pour pouvoir l'utiliser:

1. La variable dépendante doit être continue.
2. La variable indépendante doit avoir 2 groupes (catégories).
3. Les observations doivent être indépendantes.
4. Pas de valeurs extrêmes dans la variable dépendante
5. La variable dépendante doit avoir une distribution relativement normale au sein de chacun des groupes de la variable indépendante ou l'échantillon est suffisamment grand.
6. Les données doivent être appariées, ce qui signifie qu'il y a une correspondance entre les observations des deux groupes. Par exemple, les mesures prises avant et après un traitement pour chaque individu constituent des données appariées.
7. Variances des valeurs de la variable dépendante sont relativement égale pour les deux groupes

Z-test ou T-test

Voici des règles du pouce afin de déterminer quel test prendre:

1. Si vous disposez d'un grand échantillon avec des données normalement distribuées, utilisez le test Z.
2. Si vous avez un petit échantillon et que la population est normale :
    1. Utilisez le test Z si vous connaissez l'écart-type de la population.
    2. Utilisez le test t si vous ne connaissez pas l'écart-type de la population.

En réalité, il est souvent rare que nous connaissons l'écart-type de la population. Pour cette raison, le T-test est souvent préféré au Z-test.

Test d'indépendance du Khi-deux

Le test d'indépendance du Khi-deux (χ^2) est utilisé pour analyser s'il existe une association significative entre deux variables catégorielles.

Hypothèses :

- Hypothèse nulle (H0): les variables sont indépendantes, il n'y a pas de relation entre les deux variables catégorielles. Connaître la valeur d'une variable n'aide pas à prédire la valeur de l'autre variable
- Hypothèse alternative (H1): les variables sont dépendantes, il existe une relation entre les deux variables catégorielles. Connaître la valeur d'une variable permet de prédire la valeur de l'autre variable

Le test d'indépendance du Khi-deux fonctionne en comparant les fréquences observées dans l'échantillon (table de contingence) aux fréquences attendues s'il n'y avait pas de relation entre les deux variables catégorielles (donc les fréquences attendues si l'hypothèse nulle était vraie).

Test d'indépendance du Khi-deux

La formule pour calculer la valeur de khi-deux est:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Où:

Oi est la fréquence observée pour chaque cellule de la table de contingence

Ei est la fréquence attendue pour chaque cellule de la table de contingence

Test d'indépendance du Khi-deux

La fréquence attendue (Ei) se calcule ainsi:

$$E_i = \frac{(\text{Total de la ligne} \times \text{Total de la colonne})}{\text{Total de l'échantillon}}$$

Le nombre de données doit être suffisamment grand pour que les valeurs des fréquence attendues dans la table de contingence soient supérieures à un certain seuil (souvent 5)

Test d'adéquation du Khi-deux

Le test d'adéquation du khi-deux (χ^2) est utilisé pour déterminer si un échantillon de données suit une distribution théorique spécifique. Il compare les fréquences observées dans l'échantillon aux fréquences attendues sous l'hypothèse que les données suivent la distribution. Si la variable est continue, on peut la séparée en intervalle (binning) afin d'utiliser ce test.

Hypothèses:

- Hypothèse Nulle (H0) : Les données suivent la distribution théorique spécifiée.
- Hypothèse Alternative (H1) : Les données ne suivent pas la distribution théorique spécifiée.

Le test ANOVA

L’utilisation du test statistique ANOVA permet d’estimer la force d’association une variable numérique continue et une variable catégorielle. Ce test est utile quand il y a plus de 2 groupes dans la variable catégorielle.

Ce test compare la moyenne observée de la variable continue au sein de chacun des groupes (catégories) de la variable catégorielle.

Voici les deux hypothèses du test:

- H0: Toutes les moyennes des groupes sont égales
- H1: Au moins, une moyenne est différente des autres

Les corrélations (Pearson, Spearman)

Corrélation de Pearson

La corrélation de Pearson est une relation linéaire entre deux variables continues telle que les variations de leurs valeurs soient toujours de même sens (corrélation positive) ou de sens opposé (corrélation négative).

Le coefficient de corrélation nommé “r” sert à mesurer la force de la relation linéaire qui existe entre deux variable : à quel point les observations sont alignées autour d’une droite.

Une corrélation proche de 0 indique seulement qu’il n’y a pas de relation linéaire, mais il peut y avoir une relation quadratique.

Les corrélations (Pearson, Spearman)

Corrélation de Pearson

Les conditions pour utiliser ce test:

- Les deux variables que vous comparez doivent suivrent une distribution normale
- Il n'y a pas présence de valeurs extrêmes

Les hypothèses de ce test sont:

- H0: Il n'y a pas de relation linéaire entre les 2 variables
- H1: Il y a une relation linéaire entre les 2 variables

Les corrélations (Pearson, Spearman)

Corrélation de Spearman

La corrélation de Spearman est une mesure statistique qui évalue la force et la direction de la relation entre deux variables. Contrairement à la corrélation de Pearson, qui mesure la corrélation linéaire entre deux variables continues, la corrélation de Spearman évalue la corrélation monotone, c'est-à-dire toute relation systématique, croissante ou décroissante, entre les valeurs des deux variables. Cette mesure peut être utile quand il n'existe pas de relation linéaire entre les variables.

Ce test est utile, entre autres, lorsque vous voulez comparer des variables et que l'une d'entre elles est ordinales (il y a un ordre dans les valeurs (ex: 1-faible, 2-moyen, 3-élevé)). Vous pouvez utiliser ce test lorsqu'il y a présence de valeurs extrêmes et que vos variables ne sont pas distribuées normalement.

Les corrélations (Pearson, Spearman)

Corrélation de Spearman

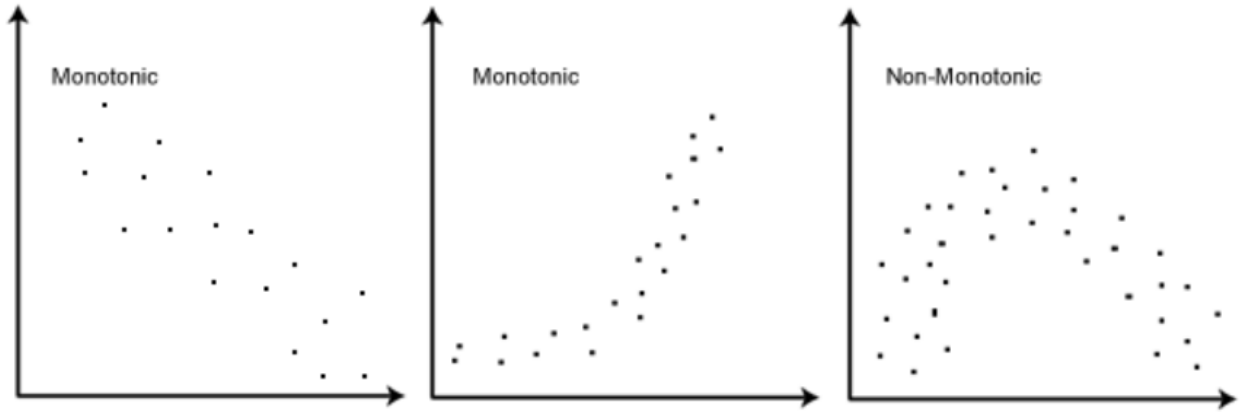
Les hypothèses de ce test sont:

- H0: Il n'y a pas de relation monotone entre les 2 variables
- H1: Il y a une relation monotone entre les 2 variables

UQAC

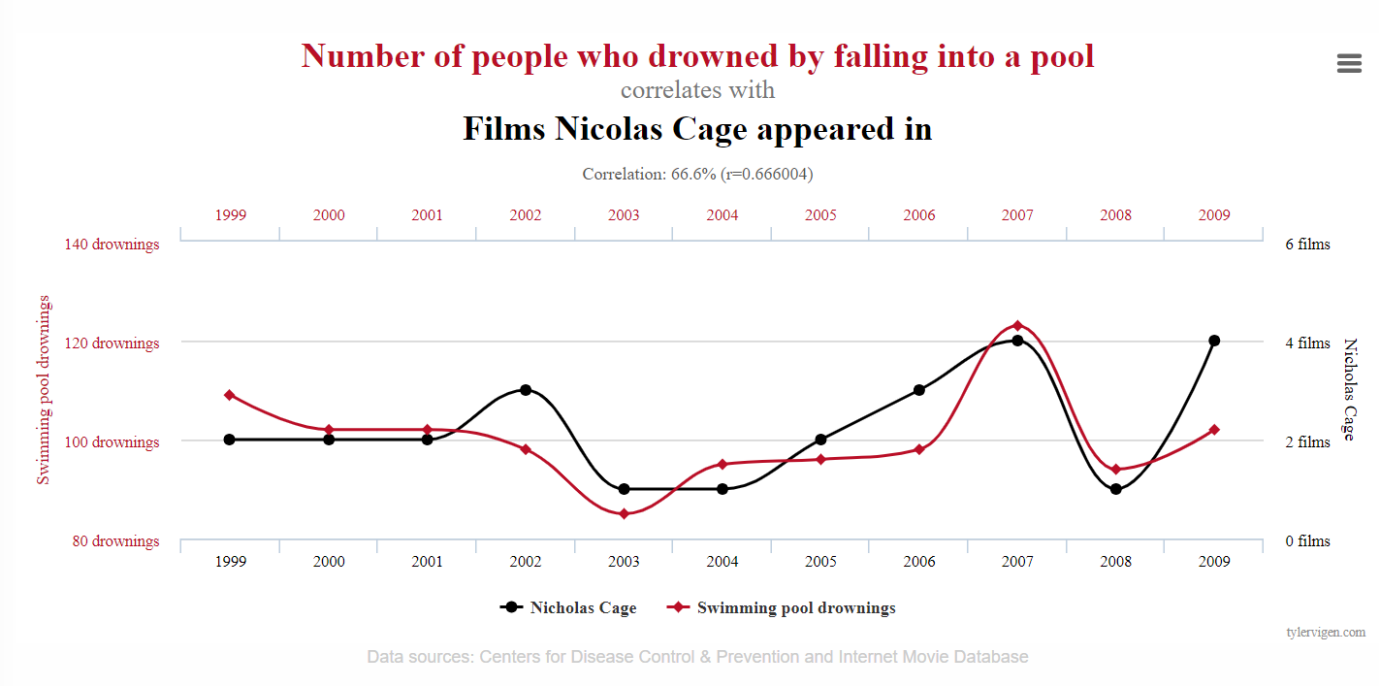
Les corrélations (Pearson, Spearman)

Corrélation de Spearman



UQAC

Corrélation n’est pas causalité !



https://www.tylervigen.com/spurious-correlations

UQAC