Faculté de Médecine et de Pharmacie de Casablanca

BIOSTATISTIQUES

OUSSAMA ESSAHILI©

Sommaire :

Oussama Essahili

Statistiques qui répondent à des questions comme

- *QUELLE EST LA VALEUR NORMALE D'UNE GRANDEUR BIOLOGIQUE ?*
- *QUEL EST LA FIABILITE D'UN EXAMEN COMPLEMENTAIRE ?*
- *TRAITEMENT A + EFFICACE QUE B ?*
- *RISQUE COMPLICATION D'UN ETAT PATHOLOGIQUE ?*

Comment ? Traiter la variabilité

BIOSTATISTIQUES

Définition

1. **Conception des expériences biologiques**
2. **Collecte, synthèse, analyse des données des expériences**
3. **Interprétation des résultats**

Rôles

1. **Organiser les données provenant des observations**
2. **Décrire les phénomènes par des paramètres résumant les observations**
3. **Estimer les valeurs de ces paramètres dans les populations d'où proviennent les échantillons observés**
4. **Comparer ces paramètres entre plusieurs populations**
5. **Prédire la probabilité de survenue des événements**

Données caractérisés par une **variabilité totale** (hasard)

- **Non pris en compte** par la biostatistique (On ne saura jamais l'erreur)

Causes de la variabilité :

1. Variabilités **métrologiques** (appareils de mesure) erreurs de mesure et expérimentales
2. Variabilité **biologique** (من عند الله), exemples : *moléculaire/génomique/cellulaire/fonctionnel/populationnel*

- **Pris en compte** par la biostatistique.
- Entraîne une variabilité dans les échantillons = **fluctuation d'échantillonnage**
- Est réservée aux mesures quantitatives **ET** qualitatives

Variabilité biologique

Intra individuelle : même individu dans des situations différentes.

Ex : variation de l'effet d'un médicament chez un même individu à des mesures différentes.

Inter individuelle : entre les individus au sein d'un groupe

Ex : variation de l'effet d'un médicament entre différents

VARIABILITÉ

- **Indépendante :**

- caractérisée par une variable **explicative** et manipulée par l'expérimentateur
- peut-être un **facteur de risque**

- **Dépendante :**

- caractérisée par une variable **qu'on cherche à expliquer**, qui subit l'effet de la variable indépendante => **variable d'intérêt**
- est généralement la **pathologie étudiée**

Les variabilités dépendantes et indépendantes peuvent être **qualitative ou quantitative. (Page 6)*

Ex : On compare **les femmes et les hommes** quant à **leur satisfaction au travail** dans une usine.

➢ **Question de recherche :** Quel est l'effet de l'appartenance à un sexe sur la satisfaction de travail ?

Variabilité **indépendante** : Le sexe, Variabilité **dépendante** : la satisfaction au travail

Ex : On compare **la satisfaction au travail des employés** quant à **leur sexe.**

➢ **Question de recherche :** Comment varie le sexe selon la satisfaction au travail ?

Variabilité **indépendante** : La satisfaction au travail, Variabilité **dépendante** : le sexe

CONSÉQUENCE DE LA VARIABILITE BIOLOGIQUE

Méthode Statistique

Enquêtes
(Maladies dans population)

Mesures
(sur un groupe)

- **GLOSSAIRE :**

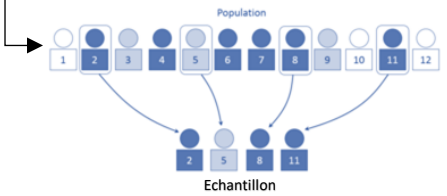
Unité statistique (U.S) : unité distincte dans laquelle on peut observer une ou plusieurs caractéristiques données. (Cela peut-être un arbre, une école, une personne etc.)

Population (P) : Taille finie ou infinie. Ensemble d'individus (ou unités statistiques) pour lequel on considère une ou plusieurs caractéristiques.

Taille de la population (N) : est le nombre d'individus constituant la population.

Echantillon (E) : Taille finie. Ensemble d'individus représentatif d'une population.

Obtenu grâce à l'échantillonnage : permet d'étudier la variable sur une partie jugé **représentatif**.
(Car la totalité est **difficile** et **coûteuse**)



Caractère ou variable statistique x_i : ce qui est observé ou mesuré sur les individus d'une population statistique.

Exemple : sexe, âge, taille, salaire d'un groupe.

Modalités : facettes sous lesquelles peut être étudié un caractère.

Exemple : sexe (mâle, femelle) poids (inférieur à 60KG, supérieur à 60KG, 60KG) couleur de la peau (noir, blanc, brun...)

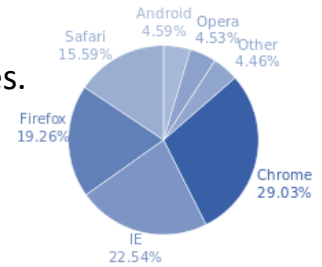
Sous-population : Ensemble d'individus de la même population ayant la même modalité

Exemple : Hommes atteint du VIH, Femmes fumeuses et ayant un cancer de poumons

Variabilité : c'est la fluctuation ou le hasard qui se présente dans la différence d'un individu d'un autre.

Stratification : partition de la population selon une variable à plusieurs modalités.

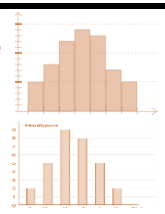
Exemple : Navigateur internet (Safari, Chrome, Firefox, Opéra...



Oussama Essahili

VARIABLE STATISTIQUE x_i

- Histogramme
- Diagramme



QUALITATIVE (Catégorielle)

non représenté par un nombre

POURCENTAGE
ECART-TYPE

VOIR PAGE &13

NOMINALE

plusieurs types (n)

Ex :

- Groupe sanguin : A,O,B,AB
- Nature du tabac fumé

Représenté par :

- Graphique à secteurs : Camembert
- Diagramme en secteurs (n<5)
- Diagramme en barres horizontales (n>5)

BINAIRE/DICHOTOMIQUE

deux types

Ex :

- Sexe : ♂ ou ♀
- Cancer : oui ou non
- Tabagisme : Fumeur ou non-fumeur
- Etat santé : sain ou malade

Représenté par :

- Diagramme en secteurs (n<5)

ORDINALE

types classés par ordre

Ex :

- Consommation de tabac : non > petite > grande
- Indice de douleur : peu, moyen, fort mal

Représenté par :

- Diagramme en bâtons

QUANTITATIVE

représenté par un nombre

MOYENNE
ECART-TYPE

VOIR PAGE 13

CONTINUE/MESURABLE

calculé par des appareils de mesures

Ex :

- Poids, Age, Taille
- Taux de Cholestérol
- Indice d'allergie : L'indice de risque allergique, concerne les allergiques de pollen qui souffrent de pollinose, il va de 0 (nul) à 5 (très élevé)

Représenté par :

- Polygone des effectifs ou des fréquences
- Histogramme des fréquences cumulées
- Box plot ou Boite à moustache

TEMPORELLE

en relation avec le temps

Types : durée, date ou heure

Ex :

- Age de grossesse (durée)
- Temps de sommeil (durée)
- Consommation de tabac journalière, hebdomadaire et mensuelle

DISCONTINUE/DISCRÈTE/DÉNOMBRABLE

0,1,2,3,4...

Ex :

- Nombre d'enfants
- Consommations de tabac par jour

Représenté par :

- Histogramme

- **2 Variables quantitatives** peuvent être représenté par **un nuage de points**.

BASE DE SONDAGE

- Liste qui **répertorie tous les individus** - *selon les critères d'inclusion et d'exclusion* – lors d'une **étude quantitative.**

➢ Former un échantillon lors d'une méthode **probabiliste** (choix aléatoire)

- Liste qui **ne maîtrise pas** la liste des éléments de la base de sondage.

➢ Utilisation méthode **non probabiliste** (choix non aléatoire et facile d'accès mais raisonné)

QCM Entraînement

1/ Quelles sont les propositions justes concernant les variables statistiques ?

A. Existence d'un cancer est une variable qualitative
B. Consommation journalière de tabac est une variable temporelle
C. Natrémie est une variable quantitative continue
D. Date d'accouchement est une variable temporelle
E. Durée de vie des cellules est une variable qualitative

2/- Quels types de représentations graphiques sont adaptés pour une variable quantitative continue ?

A. Diagramme en barres
B. Polygone de fréquences
C. Histogramme
D. Diagramme en secteur
E. Polygone des effectifs

3/ Quel est le graphique le plus adapté pour représenter la distribution du niveau d'étude?

A. Polygone de fréquence
B. Histogramme
C. Diagramme en secteur
D. Nuage de points
E. Diagramme en bâton

4/ Choisissez la ou les réponses justes concernant la classification des variables suivantes :

A. Groupe sanguin : variable qualitative ordinale
B. Créatinémie : variable quantitative discrète
C. Age de la grossesse : variable temporelle type durée
D. Intensité de la douleur : variable quantitative continue
E. Nature du tabac fumé : variable qualitative nominale

5/ Même question :

A. Créatinémie : variable quantitative continue
B. Evolution d'un cancer en différents stades : variable qualitative nominale
C. Groupe sanguin : variable qualitative ordinale
D. Durée de sommeil : variable temporelle type durée
E. Localisation de la douleur : variable qualitative nominale

6/ Quelles sont les propositions qui s'appliquent à la variabilité biologique ?

A. La variabilité biologique existe uniquement pour les mesures quantitatives
B. La variabilité biologique correspond à la variabilité métrologique
C. La méthode statistique élimine la variabilité biologique
D. La méthode statistique n'a aucun lien avec la variabilité biologique
E. La méthode statistique prend en considération la variabilité biologique

1	ABCD
2	BCE
3	E
4	CE
5	ADE
6	E

Oussama Essahili

Pourquoi l'échantillonnage ?

- à cause des contraintes, coûts et erreurs de la base de sondage.

Les choix des individus selon : (Facteurs)

- disponibilité ou non des données
- ressources financières et humaines
- La durée
- La nature de question de recherche

Erreurs de l'échantillonnage :

- Biais de sélection
- Variations aléatoires

ECHANTILLONNAGE

***Inférence statistique :** est la capacité de généraliser les résultats à partir d'un échantillon.

▪ PROBABILISTE

- vise pratiquement la couverture de l'ensemble de la population même si la non-réponse reste importante
  - ➢ **Inférence statistique :** OUI
- ○ Représentatif

→ **Aléatoire simple :** on prélève les individus de la population sans remise.

→ **Aléatoire systématique/Systématique :** on prélève les individus de la population de façon à ce qu'il y'a un intervalle ou un écart entre chaque individu sélectionné.

Ex : Population : 2000 / Echantillon : 200

$\frac{2000}{200}$ = 10 (1ère personne dans chaque dizaine de 2000 est choisie jusqu'à obtenir 200 personnes)

→ **En grappe :** prélèvement en plusieurs groupes similaires dont les individus sont différents.

→ **Stratifié :** conserver le % dichotomique après prélèvement d'une population.

Ex : P : 200 (84♂ [42%] et 116♀ [58%])

⇒ E : 50 (21♂ [42%] et 29♀ [58%])

▪ NON PROBABILISTE (EMPIRIQUE)

- c'est la personne qui sélectionne l'échantillon et ses unités statistiques et non le hasard (voisins, proches non représentatifs)
  - ➢ **Inférence statistique :** NON
- ○ Peu coûteuse, rapide, facile à appliquer.

→ **Accidentel/De convenance:** on prélève les individus de la population sans remise.

→ **Volontaire :** individus se sélectionnent eux-mêmes.

→ **Quotas (+++) :** individus pris au hasard qui seront échantillonné au choix (stimule l'échantillonnage stratifié)

→ **Boule de neige :** individus de l'échantillon connaissant d'autres profils similaires pour participer à la recherche.

→ **Choix résonné :** selon le choix du caractère d'intérêt.

ORGANISATION DES DONNÉES

1)- Tri des données
2)- Regroupement en classes
3)- Transformation de variable
4)- Effectifs et fréquences
5)- Distribution : Plus les distributions sont **normales**, plus les **données se rapprochent du centre**.

Valeur et effectif : Exemple pour comprendre

- La valeur X_i peut-être assimilé à la matière
- L'effectif f_i peut-être assimilé au coefficient de cette matière

Fréquence f_i :

$f_i = \frac{n_i}{N} (0 \leq f_i \leq 1)$

- n_i : $effectif\ partiel$
- N : effectif total

$\sum f_i$ = Distribution de fréquence

Centre de la classe :

$C_i = \frac{X_i + X_J}{2}$

Ex : Poids **(40 – 44 KG)**

$C_i = \frac{44 + 40}{2} = 42$

Limites de la classe :

$X_i \leq t \leq X_J$

Ex : Poids **(40 – 44 KG)**

$40 \leq t \leq 44$

- Mesure de tendance centrale (de position) : *servent à résumer*

∑ des données d'une variable en un seul paramètre.

- sont identiques pour chaque deux groupes.

I. **Moyenne :** (« m » pour échantillon et « M » pour population) somme des valeurs des données divisées par le nombre des valeurs qui existent **(5 valeurs -> on divise par 5)**

Moyenne pondérée $\bar{X}$: valeurs multipliés par leur effectif, l'ensemble divisée par le nombre totale d'effectifs.

$$\bar{X} = \frac{n_1 x_1 + n_2 x_2 + ..}{N\ (nb\ effectifs)}$$

II. **Médiane (Quartile 2) :** valeur qui coupe les séries des données en 2 parts égales.

- **Quartile 1 :** divise les séries des données en 2 parts (La première 25% et la deuxième 75%)
- **Quartile 3 :** divise les séries des données en 2 parts (La première 75% et la deuxième 25%)

III. **Mode :** Valeur dont effectif est le plus grand.

- **Distribution unimodale :** 1 seule valeur MAX
- **Distribution bimodale :** Plusieurs valeurs MAX

Pour une meilleure description : 1 tendance + 1 position

Médiane + Etendue : Si valeurs très éloignés

Moyenne + Ecart-type : Si valeurs très rapprochés

Mesure de dispersion : *indices qui permettent de savoir si les données sont fortes serrées autour d'un paramètre de tendance centrale ou dispersées.*

- sont **toujours différents entre deux groupes.**
- **↗ Dispersion des données ➞ ↗ Valeurs des paramètres**

I. **Etendue :** $X_{max} - X_{min}$

II. Variance (S^2) et écart-type (σ) : (Echantillon)

2 paramètres reliés car l'écart-type S est égal à la racine carrée de la variance S^2

$S = \sqrt{V_x}$

Echantillon : $$S^2 = \frac{n_1(\bar{x}_1 - \mu)^2 + n_2(\bar{x}_2 - \mu)^2 + \dots}{n-1}$$

Population : $$S^2 = \frac{n_1(x_1 - \mu)^2 + n_2(x_1 - \mu)^2 + \dots}{N}$$

n-1 = Nombre d'individus, n = nombre d'individus de l'échantillon, N = nombre d'individus de la population

Variable quantitative binaire :

Si la variance augmente -> Pmax diminue.

Variance : $\sigma^2 = P(1-P)$

Ecart-type : $\sigma = \sqrt{P(1-P)}$

- Exemple :

Pour P(homme) = 50% et P(femme) = 50% -> $\sigma^2 = 0{,}5(1-0{,}5) = 0{,}25\ (MAX)$

P(homme) = 90% et P(homme) = 10% -> $\sigma^2 = 0{,}1(1-0{,}1) = 0{,}9(1-0{,}9) = 0{,}09$

III. Coefficient de variance (CV) : (%)

Mesure la distribution de la moyenne

Est relatif, ne dépend pas des valeurs, compare les variabilités de plusieurs variables même si elles n'ont pas la même unité.

$CV = \frac{\sigma}{\bar{x}} \times 100$ **= …… %**

IV. Intervalle interquartile (IIQ) et semi-interquartile (SIQ) : Ne permet pas d'estimer

IIQ = Q3 – Q2 et **SIQ = IIQ/2**

Etendue + IIQ : sont des paramètres de dispersion

Etendue + IIQ + Médiane : permettent la description de distribution de variable quantitative

QCM Entraînement

1/ Les méthodes d'échantillonnage permettant l'inférence statistique sont :

A. Accidentel
B. En grappes
C. Choix raisonné
D. Aléatoire simple
E. Stratifié

2/ On dose un médicament chez des sujets normaux avec les résultats suivants : (unité arbitraire U; classe avec borne inférieure comprise et borne supérieure exclue)
Classe A (6 à 10 U) : 55 sujets
Classe B (18 à 22 U) : 95 sujets
Classe C (10 à 14 U) : 80 sujets
Classe D (22 à 26 U) : 80 sujets
La classe modale est :

A. La classe D
B. La classe A
C. Égale à 95
D. La classe qui divise en 2 parties égales la série des données
E. La classe B

3/ Parmi les mesures de tendance centrale on trouve

A. Le mode
B. L'écart type
C. La différence entre la valeur maximale et la valeur minimale
D. La moyenne arithmétique
E. La valeur de la variable qui divise la série en deux parties égales

4/ Les échantillonnages permettant l'inférence statistique sont :

A. Aléatoire simple
B. En boule de neige
C. En grappes
D. L'échantillonnage accidentel
E. Stratifié

5/ Les éléments de choix d'un plan d'échantillonnage donné sont :

A. Le nombre d'unités statistiques
B. Les ressources humaines nécessaires
C. Le temps disponible pour l'étude
D. Les ressources financières disponibles
E. La question de recherche

1	BDE
2	E
3	ADE
4	ACE
5	BCDE

Les lois de distribution

Loi binomiale

$$P(X = k) = \frac{n!}{k!(n-k)!} P^k (1-P)^{n-k}$$

Conditions :

1. Variable de type binaire

2. Tentatives indépendantes

3. Évènement avec même probabilité de succès + Individus avec même chance d'être tiré au sort

Loi de poisson :

$$P(X = K) = \frac{e^{-\mu} \times \mu^k}{k!}$$

Conditions :

1. Dénombrables

2. Indépendants les uns des autres

3. S'applique aux évènements rares (P<0,05)

Au cas où P>0,05 -> Loi binomiale

Loi centrée réduite :

Transformation d'une variable aléatoire de telle sorte que :
sa moyenne soit nulle et son écartype soit égal à 1.

Permet de rendre les variables comparables.

Pour centrer la distribution : $X' = X - \mu$

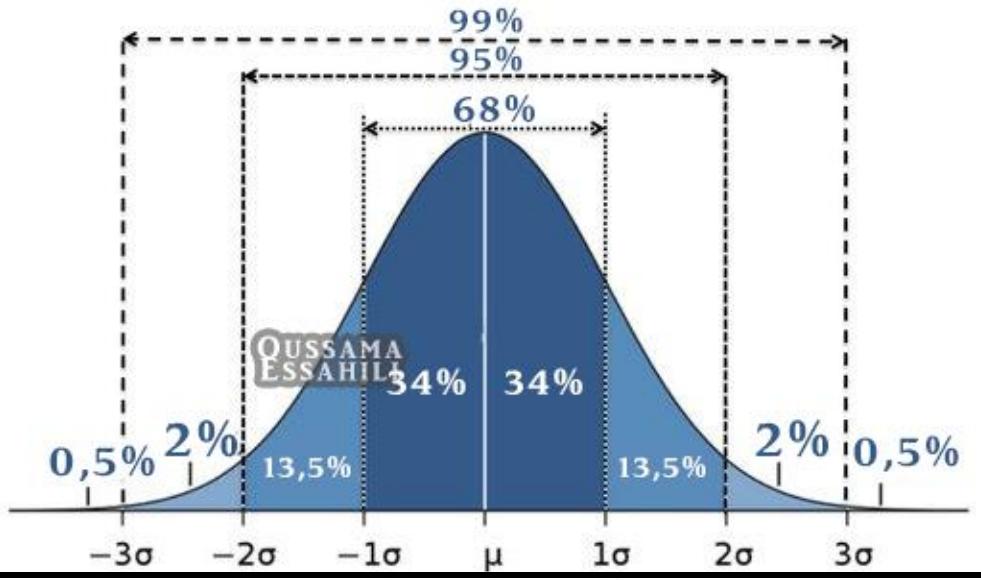
Pour réduire la distribution : Z = $\frac{X-\mu}{\sigma}$

Loi de normale :

S'applique aux variables quantitatives continues

La distribution des valeurs d'une variable est autour d'une moyenne μ

Les valeurs décroissent de part et d'autres.



	ESTIMATION D'UNE MOYENNE INCONNUE μ	ESTIMATION D'UN POURCENTAGE INCONNU P
Fluctuation d'échantillonnage	∑ des moyennes des échantillons donnent + de précision sur $\boldsymbol{\mu}$	∑ des pourcentages des échantillons donnent + de précision sur P
Ecart-type	$S_m = \frac{S_e}{\sqrt{n}}$ S_m : *Ecart type de la moyenne de l'échantillon* S_e : *Ecart type des valeurs de l'échantillon* n : *Taille de l'échantillon*	$S_p = \sqrt{\frac{p_e(1-p_e)}{n}}$ S_p : *Ecart type du pourcentage de l'échantillon* p_e : *Pourcentage de l'échantillon* n : *Taille de l'échantillon*
Intervalle de confiance	$[\boldsymbol{\mu_i}\ ;\ \boldsymbol{\mu_s}]$ $\quad \boldsymbol{\mu} = m_e \pm 1{,}96S_m$ μ_i : *Borne inférieur de l'intervalle* μ_s : *Borne supérieur de l'intervalle*	$[\boldsymbol{P_i}\ ;\ \boldsymbol{P_s}]$ $\quad \boldsymbol{P} = p_e \pm 1{,}96S_p$ $\begin{cases} P_i = p_e - 1{,}96S_p \\ P_s = p_e + 1{,}96S_p \end{cases}$ P_i : *Borne inférieur de l'intervalle* P_s : *Borne supérieur de l'intervalle*
Condition pour respecter la loi normale	Il faut vérifier **avant** les calculs. $n \leq 10$ *de Population* $n \geq 30$	Il faut vérifier **après** les calculs. $\begin{cases} nP_i \geq 5 \\ nP_s \geq 5 \end{cases}$ $\begin{cases} n(1-P_i) \geq 5 \\ n(1-P_s) \geq 5 \end{cases}$

Remarques :

- Marge d'erreur α (5%) -> IZ_αI = 1,96 selon Table de Z
- α *diminue* $\rightarrow$ *précision diminue* $\rightarrow$ *intervalle augmente en largeur* (*Il faut diminuer la taille de l'échantillon*)
- La taille de l'échantillon est inversement proportionnelle à l'intervalle de confiance
- α : le risque pour que la moyenne ou le pourcentage se trouve, en dehors de l'intervalle
- Si les conditions ne sont pas vérifiées « % » -> Loi binomiale

La précision : 1,96 S_m ou 1,96S_p

COMPARAISON STATISTIQUE

- Echantillon d'une Pop inconnue X Pop de référence → **La population inconnue :** Même **ou # distribution que la population de référence ?**
- Echantillons entre eux des populations

- Cas de distribution # entre les populations :

- Hypothèse 1 : La différence est due aux fluctuations.
- Hypothèse 2 : La différence est réelle.

- Application d'un test d'hypothèse : *(Pour s'assurer si la différence est due aux fluctuations)*

Étapes :

1/ Formation des hypothèses
- **H0 : Hypothèse nulle** $P_A = P_B$
- **H1 : Hypothèse alternative** $P_A \neq P_B$ (*Unilatéral*) ou $P_A > P_B$; $P_A < P_B$ (*Bilatéral*)

2/ Sélection échantillon + collecte des données

3/ Calcul de la probabilité P (Test de comparaison)

4/ Règle de décision : H0 est à rejetter ou à conserver ?

Conclusion : Différence significative ou non ?

Quantitée calculée à l'extérieur du modèle théorique → Rejet de H0 → Accepter H1 → La différence est réelle → Différence significative

Quantitée calculée rejoint celle du modèle théorique → Différence n'est pas significative → La différence est due aux fluctuations → Pas de rejet de H0

5/ Choix du risque d'erreur :

- Risque α ou Risque de 1ère espèce : probabilité de rejeter H0 si H0 est vraie.

- Risque β :

- Probabilité de ne pas rejeter H0 si H1 est vraie.
- On ne le choisit pas.
- $(1 - \beta)$ est la puissance du test.

6/ Interprétation finale :

CAS 1 : H0 est rejetée

- H1 est vraie
- Degré significatif au risque p
- Degré de signification p : p<α toujours, valeur minimale pour α

Signification :

1)- Statistique (ex : % cancer de poumons selon le nombre de personnes ayant un briquet)

2)- Clinique (en fonction des tests et mesures)

✓ *Dans le cadre de la santé, contrairement à d'autres, on ne se suffit pas de la signification statistique (présente des troubles) mais aussi de la signification clinique. On ne tolère pas la fluctuation.*

CAS 2 : H0 n'est pas rejetée

- On ne peut pas affirmer que les paramètres comparés sont différents. (Pas de réponse à la question)

➢ *TEST DE COMPARAISON (Récapitulatif) :*

TEST	CONDITIONS	VARIABLE QUALITATIF	VARIABLE QUANTITATIVE
Ecart-réduit	$\begin{cases} nP \geq 5 \\ nQ \geq 5 \ / \ Q = 1 - P \end{cases}$ **%**	**Cas 1 :** % observé + % théorique **Cas 2 :** % observés	
X^2	$C_i \geq 5$ $(chaque\ effectif\ th)$ **Effectifs Ci**	**Cas 1 :** 1 variable **Cas 2 :** plusieurs variables	
Student	- Distribution normale et homogénéité des variances. $n < 30$		**Cas 1 :** 2 échantillons **Cas 2 :** 2 études sur un même groupe
Z	$n \geq 30$		**Cas 1 :** 2 échantillons **Cas 2 :** 1 échantillon et une population de référence

Situation de liaison : étude liaison entre 2 variables (2 qualitatives/2 quantitatives/une chacune) si elle se rapproche de la valeur théorique.

- Grâce aux : Tests de tendance, Tests de corrélation, Tests de régression **- Exemple :** La créatinémie entraîne le cancer de foie ?

- Leur finalité : Vérifier si la relation observée entre les variables étudiées de l'échantillon se rapprochent suffisamment du modèle théorique

Oussama Essahili

Test de Khi2

=> Comparaison des effectifs observés et des effectifs calculés (les effectifs peuvent être transformés en pourcentages : l'écart-réduit)

1)- Condition à vérifier :

Les effectifs calculés : $Ci \geq 5$

2)- Khi 2 calculé : $\chi_c^2 = \frac{(\theta_1 - C_1)^2}{C_1} + \frac{(\theta_1 - C_1)^2}{C_1}$ ..

3)- Poser α et ddl pour obtenir Khi 2 table χ_t^2

Avec ddl = K-1 avec K : le nombre de modalités

4)- Comparer χ_c^2 et χ_t^2

Si $\chi_c^2 < \chi_t^2$: non significative

Si $\chi_c^2 > \chi_t^2$: significative à p (p est la valeur α pour $|Z\alpha|$ = χ_c^2 selon le tableau de Khi 2)

Pour $\alpha = 5\%$: DDL = 1 et χ_t^2 = 3,841

Tableau de contingence :

Pour un échantillon : (Efficacité d'un médicament)

	OUI	NON
Ci		
θi		

Pour deux échantillons : (Efficacité d'un médicament chez 2 groupes)

	OUI	NON	Total
Groupe A			X
Groupe B			Y
Total	W	Z	X+Y = W+Z

Cas particulier :

I. Etude de + d'une variable :

DDL = (L-1) (C-1)

L : Nombre de lignes

C : Nombre de colonnes

Pour calculer χ_c^2 , il faudrait d'abord calculer Ci pour chaque case du tableau de contingence

$$Ci = \frac{Total\ ligne \times Total\ colonne}{Total\ général}$$

			= Total ligne
			= Total ligne
= Total colonne	= Total colonne	= Total colonne	= Total général

II. Test χ_c^2 corrigé – Correction de Yates

Condition : $3 < np,nq \leq 3$ / q = 1-p

$$\chi_c^2 = \frac{(|\theta_1 - C_1| - 1/2)^2}{C_1} + \frac{(|\theta_2 - C_2| - 1/2)^2}{C_1}$$

- Méthode valable pour la **variable dichotomique (2 classes)**

III. : χ_c^2 non valable – Test exact de Fisher

- Lorsqu'on ne peut pas conclure si la différence est significative ou pas (car l'effectif théorique est trop **faible Ci < 3**)

Ecart-réduit

• Comparaison de 2 pourcentages observées ou un pourcentage observé et un pourcentage supposé théorique

1/ Calcul du pourcentage unique

$$P(\%) = \frac{n_1 \times p_{O1} + n_2 \times p_{02}}{n_1 + n_2}$$

% observée et % théorique

$$SP = \sqrt{P(1-P)/n}$$

*Dans ce cas, le pourcentage unique p est égale au pourcentage théorique

2/ Ecart-type de la différence de la distribution des 2 pourcentages (SDP)

2 % observée

$$SDP = \sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

3/ Ecart-réduit

$$\varepsilon = \frac{|p - p_O|}{SP}$$

$$\varepsilon = \frac{|p_{O1} - p_{O2}|}{SDP}$$

4/ Comparaison de l'écart réduit avec $Z_\propto$

Pour $\propto = 5\%$ -> $Z_\propto = 1,96$ (2)

$\boldsymbol{\varepsilon < 2 : différence\ non\ significative}$

$\boldsymbol{\varepsilon > 2 : différence\ significative\ à\ p}$

(p : est la valeur $\propto$ pour que $\boldsymbol{Z_\propto = \varepsilon}$ selon la table de l'écart réduit)

P n'est pas calculée, il est constaté.

5/ Conditions à vérifier

$np \geq 5$

$n(1-p) \geq 5$

< ! > Test exact de Fisher pour les petits échantillons :

np, nq < 5 / $Ci \leq 3$

- Utilisée quand les conditions de l'écart-réduit et Khi2 ne sont pas vérifiés.

Comparaison des moyennes

Les distributions normales grâce à :
- **Histogramme**
- **Boîte à Moustache**
- **Test de Kolmogorov-Smirnov**

Test de Student

Conditions :

- n < 30
- Les deux groupes d'échantillons suivent des lois normales (distribution normale des variances) et sont de variances égales ou homogènes.

1/ $S^2 = \frac{(n_1-1){s_1}^2+(n_2-1){s_2}^2}{n_1+n_2-2}$

2/ Calcul de t_0 valeur seuil du test Student

*t (ddl ; α) : à déterminer de la Table de loi Student

Cas 1 : Une étude sur deux échantillons

$$t_0 = \frac{m_1 - m_2}{\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}}$$

n_1, n_2 : *taille des 2 échantillons*

m_1, m_2 : *moyenne des 2 échantillons*

s_1, s_2 : *les 2 écartypes*

3/ Vérifier hypothèse :

$t_0 \geq t(n_1 + n_2 - 2\,;\,\alpha) \rightarrow$ ***Rejet de*** H_0

$t_0 < t(n_1 + n_2 - 2\,;\,\alpha) \rightarrow$ ***Non rejet de*** H_0

Cas 2 : 2 études sur un même échantillon

(Série appariée, ex : Avant/Après)

$$t_0 = \frac{m \times \sqrt{n}}{S}$$

m : moyenne de la différence

S : écart type de la différence

3/ Vérifier hypothèse :

$t_0 \geq t(n-1\,;\,\alpha) \rightarrow$ ***Rejet de*** H_0

$t_0 < t(n-1\,;\,\alpha) \rightarrow$ ***Non rejet de*** H_0

***Pour : $\alpha = 5\%$ (*bilatéral*) / $\frac{\alpha}{2} = 97,5\%$ (*unilatéral*) :**

t = 2,00

Test de Z (Test de l'écart réduit)

Condition : $n \geq 30$

Cas 1 : Deux échantillons

$$Z_0 = \frac{m_1 - m_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Cas 2 : échantillon d'une population et population de référence (théorie)

$z_0 = \frac{n - \mu_2}{SP}$ avec SP = $\frac{S}{\sqrt{n}}$

μ_2 : *la moyenne théorique*

$|z_0| \geq 1,96 \rightarrow$ ***Rejet de*** H_0 (signification)

$|z_0| < 1,96 \rightarrow$ ***Non rejet de*** H_0 (non signification)

***Sur la table normale centrée réduite :**

$Z_{\alpha/2} = 1,96$ ***pour*** $\alpha = 5\%$

Test de Shapiro : vérifier si une série de données suit une loi normale.

N.B : Pour n'importe quel test, p<0,05 veut dire que la différence est statistiquement significative.

QCM Entraînement

1/ Sur un groupe supposé représentatif de 520 malades atteints de maladies cardiaques, on en a observé 162 cardiopathies ischémiques
L'intervalle de confiance à 95% (ε = 2) du pourcentage de cardiopathies ischémiques dans la population des malades cardiaques (valeurs arrondies à l'entier) est :

A. Les conditions de validité ne sont pas remplies
B. 68% - 84%
C. 39% - 49%
D. 21% - 38%
E. 27% - 35%

2/ Quelle est la probabilité d'observer moins de 4 malades dans un échantillon de 10 sujets choisis au hasard dans une population où la fréquence de la maladie est de 17% ?

A. 0,160
B. 0,293
C. 0,155
D. 0,318
E. 0,926

3/ Pour connaître la fréquence d'une parasitose dans une région de 350000 habitants, on pratique une enquête sur un échantillon de 5775 personnes, on dépiste parmi eux 1155 sujets atteints de cette parasitose.
Pour un risque d'erreur de 5%, la précision est de :

A. 1,03%
B. 2,3%
C. 0,3%
D. 3,2%
E. 5%

4/ 4 étudiants parmi 10 ont réussi à un examen de passage, sachant que le pourcentage théorique de réussite à cet examen est de 10%.

Avec un risque d'erreur de 5%, on veut savoir si le pourcentage de réussite observé diffère du pourcentage théorique ?

A. En utilisant le test de Khi2, la différence est significative
B. En utilisant tous les tests, la différence est non significative
C. En utilisant le test de l'écart réduit, la différence est significative
D. En utilisant le test de khi2 corrigé ou la correction de Yates, la différence est significative
E. On utilise le test exact de Fisher.

5/ Le test de Student

A. A la même signification que le test de Khi2
B. Permet de comparer 2 pourcentages observés
C. Permet de comparer 2 moyennes
D. Permet de comparer un pourcentage observé et un pourcentage théorique
E. Permet de comparer 2 effectifs

2/ P(X<4) = P(0) + P(1) + P(2) +P(3)

= 0,155+0,318+0,293+0,160 = 92,6% (0,926)

1	E
2	E
3	A
4	E
5	C