

INTRODUCTION AUX BIOSTATISTIQUES

NOTIONS DE BASE

Afin de simplifier les études, on définit les individus en fonction de caractéristiques d'intérêt.

1- Types de variables

La variabilité existant entre les individus peut acquise ou due au hasard et doit être prise en compte lors de toute étude pour éviter l'introduction de biais. Une variable peut être :

- **qualitative** : se présentant sous forme de modalités de réponse sans prépondérance d'une modalité sur l'autre = variable qualitative nominale ex : *couleur de cheveux*
avec une ordre entre les modalités = variable qualitative ordinale
- **quantitative** : se présentant sous forme numérique et qui est
continue = peut prendre n'importe quelle valeur, dans un intervalle donné ex *taille, poids*
discrète = les valeurs possibles sont dénombrables ex : *nb d'enfants dans une fratrie* . En pratique, tous les instruments de mesure donnent lieu à des variables discrètes !

2- Groupes étudiés

- **Population** : ensemble exhaustif des objets sur qui portent l'étude. La population source est celle dont on extrait l'échantillon // population cible qui est celle à qui on extrapolera les résultats obtenus à partir de l'échantillon.
- **Série statistique** : collection d'objets de même nature possédant des caractéristiques différentes
- **Echantillon** : sous-ensemble fini, d'effectif limité, extrait de la population source. Il doit être représentatif de cette dernière (cad que les individus sont sélectionnés au hasard depuis la population source dans l'échantillon) pour permettre une généralisation des résultats.

3- Paramètres de variables quantitatives

Il s'agit de grandeurs apportant des infos sur la variable d'intérêt. Ces paramètres sont mesurés sur l'échantillon, mais ne peuvent être qu'estimés dans une population !

- de position : moyenne, médiane, mode...
- de dispersion : écart-type, variance, extrema...

STATISTIQUE DESCRIPTIVE

1- Estimation statistique

Permet la détermination d'une grandeur définie sur une population à partir d'observations réalisées sur un échantillon de celle-ci, et peut-être de 2 types :

- ponctuelle = valeur qui semble la meilleure à un instant donné
- par intervalle, qui contient lui-même la valeur recherchée = intervalle de confiance
Elle se fait en 3 temps : détermination de la population source
échantillonnage
calcul de l'intervalle de confiance

2- Données quantitatives

- la **moyenne m** (quotient de la somme des valeurs de la série statistique sur l'effectif n de l'échantillon) qui va permettre d'estimer la moyenne vraie μ de la population
- l'**écart type s** (variabilité des valeurs entre elles et avec la moyenne m), estimateur de σ
- **intervalle de confiance** de μ , avec 2 paramètres de largeur = s et ε qui est en lien avec le risque d'erreur de première espèce α , pris à 5% soit $\varepsilon = 1,96$ (référence). Plus cet intervalle est grand, moins il sera précis // plus il est petit, plus le risque d'erreur α sera élevé.

3- Données qualitatives

Se mesurent uniquement en %, on ne peut faire que des catégories. Il faut que l'échantillon soit représentatif, randomisé, nécessitant le calcul d'un intervalle de confiance :

- le **pourcentage p_{obs}** de l'échantillon qui permet d'estimer le % vrai **p**
- l'**écart type s** avec p_o et q_o
- l'**intervalle de confiance** de p , dont la largeur dépend de s et de ε

La **précision** varie comme $\frac{1}{\sqrt{n}}$ et comme l'inverse de σ => imp de la taille de l'échantillon. Plus la précision est gde, plus l'intervalle de confiance est réduit.

STATISTIQUE DEDUCTIVE

On tire des conclusions à partir d'observations en tranchant entre 2 hypothèses :

- **H0 = hypothèse nulle**, càd aucune différence observée entre les groupes
- **H1 = hypothèse alternative** càd différence entre les groupes. Elle peut être de 3 types : non égalité, infériorité ou supériorité. Lorsqu'on réalise un test, soit on cherche juste à voir s'il existe une différence (situation bilatérale) soit on cherche on teste une différence supérieure ou inférieure à 0 (situation unilatérale)

Après avoir défini les 2 hypothèses, on définit le test en fonction des données et du paramètre calculé Z. On choisit le risque α (*en pratique, 5%*) puis on construit un intervalle de pari $1-\alpha$: si H0 est vraie, la probabilité que Z soit compris dans l'intervalle de confiance est $1-\alpha$.

On recueille les données en calculant le paramètre observé, puis on détermine s'il est compris dans l'intervalle de pari (H0) ou non (H1) en comparant avec les tables de référence.

On peut alors interpréter les résultats en fonction des données de départ.

	H0 vraie	H1 vraie
H0 acceptée	$1-\alpha$	β
H0 rejetée	α	$1-\beta$

α : risque de trouver une différence où il n'y en a pas = risque de première espèce
 β : risque de ne pas trouver de différence qd il y en a une = risque de 2° espèce
 $1-\beta$ = puissance de l'étude

Donc en fonction des effectifs, on aura :

Effectif	données quantitatives	données qualitatives	données qttives + qltive
n<12	coeff r' de Spearman	comparaison de % ou χ^2	U de Mann et Withney
12<n<30	coeff de corrélation r		test t Student
n>30			comparaison de moyenne

Les tests du coefficient r' et du U de Mann et Whithney sont des tests dits non paramétriques : ce sont des tests de rang pour la plupart, ne préjugant pas de la distributions des valeurs dans la série statistique. Lorsque l'effectif de l'échantillon est trop faible, la moyenne, l'écart type etc. ne sont plus pertinents d'où le recours à d'autres tests.