BIOSTATISTIQUES

METHODE STATISTIQUE EN MEDECINE

Utilité pour décrire des populations/situations
évaluer des ttt, des techniques ou des coûts
quantifier des observations épidémiologiques

⇨ mise en place d'études comparatives, de sondages, de systèmes d'info

1- Variabilité

Qui peut être due au hasard ou physiologique, et donc nécessite d'être prise en compte lors de l'interprétation des résultats en statistique.
On définit 2 domaines de la statistique :

- descriptive = décrire une situation ou un ensemble de données à partir de paramètres précis
- déductive = tirer des ccl depuis des observations/mesures en déterminant la part de variabilité.

2- Concepts de base

- **Série statistique** : collection d'objets de mm nature avec des caractéristiques différentes (= variables) qui peuvent être qttives (mesures) ou qltives (classements en %).
- **Population** : série exhaustive qui correspond à l'ensbl des objets étudiés, pouvant être finie ou non.
- **Echantillon** : sous ensbl fini et d'effectif limité extrait de la population qui doit être tiré au hasard (randomisation, avec un échantillon connu) et représentatif (inférence statistique) de cette population afin de pouvoir lui extrapoler les résultats.
  Très utile car la population étudiée n'est pas tjrs accessible ou moyens limités pour réaliser l'étude.

STATISTIQUE DESCRIPTIVE

Va permettre la description des populations
le calcul d'estimateurs
l'introduction de sondage

La description des populations nécessite de maîtriser la variabilité de toute étude, qui est normale (entre différents individus ou différents types de mesures). Une mauvaise estimation de cette variabilité conduit à un biais.

1- Estimation statistique

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

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

2- Données qttives

- la **moyenne m** (quotient de la somme des valeurs 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, ac 2 paramètres de largeur = s et ε qui est en lien avec le risque α, devant être inférieur ou égal à 5% soit ε = 1,96 (référence). Plus il est grand, moins il sera précis // plus il est petit, plus le risque d'erreur sera élevé.
  Pour un échantillon de plus de 30 personnes, la répartition est gaussienne :
  l'intervalle **(μ - 1,96σ ; μ+1,96σ)** contient 95,4% de la population, d'où le risque d'erreur de 5%.

Il est possible de savoir combien de personnes sont nécessaires pour constituer un échantillon représentatif d'une population donnée. Plus la taille de l'échantillon est importante, plus le résultat sera précis.
Il permet le calcul de normes pour les différents pays, les différentes habitudes.

3- Données qltives

Se mesure 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

On peut également connaître le nombre de personnes nécessaires à la représentativité de l'échantillon, permettant de maîtriser la variabilité.
Dans ton les sondages, il faut donner les dates, la méthode et la taille de l'échantillon.

La **précision** varie comme √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 significative entre les groupes

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-α : si H0 est juste, la probabilité que Z soit compris dans l'intervalle de confiance est 1-α.
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 (*respectivement < ou > à ε*).
On peut alors interpréter les résultats en fonction des données de départ. Si H1 est retenue, il faut préciser le risque d'erreur (= valeur α qui correspond au résultat trouvé, dans les tables de référence).

	H0 vraie	H1 vraie
H0 acceptée	1-α	β
H0 rejetée	α	1-β

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-β = puissance de l'étude

1- Etude de liaison entre 2 caractères qltifs

De type comparaison de 2 pourcentages entre deux populations différentes

- **Comparaison de 2 pourcentages observés** : mettant en jeu la différence entre p_A et p_B qui peut-être proche de 0 = H0 (données quasiment identiques) ou importante = H1 (données éloignées l'une de l'autre).
- **Test du χ^2** = somme des quotients (observé – calculé) sur les résultats calculés. Il est fonction du degré de liberté = (nb de lignes-1) x (nb de colonnes-1)
  On compare avec le α de 5% correspondant, si < = H0 ; si > = H1.

Ex : étude de l'effet de 2 vaccins, càd réactions légère, moyenne ou forte.

2- Etude de liaison entre caractères qltifs et qttifs

- **Comparaison des moyennes** : pour de gds échantillons càd n_A et n_B > 30, pour laquelle on prend en compte les moyennes et le carré des écart-types.
- **Test t de Student** : lorsque l'un des échantillon (ou les deux) est < 30 personnes. On prend en compte un écart-type commun aux deux échantillons.
  Fonction également d'un degré de liberté : ddl = $(n_A-1)+(n_B-1)$

Ex : mesure du taux de T3 chez les femmes prenant la pilule ou non.

3- Etude de liaison entre 2 caractères qttifs

Par le test du **coefficient de corrélation** : pente de la droite de régression passant au plus près de chaque point, dont les coordonnées correspondent à 2 séries numériques (caractères étudiés). On calcule le coefficient r en regardant :

- la valeur absolue : > ou < à celle de α = 5% (regarder le ddl = nb de personnes -1)
- le signe, positif ou négatif selon la relation entre les caractères (dans le mm sens ou inversement).

Ex : existence d'un lien entre l'efficacité d'un ttt anti HTA (valeur de la baisse) et l'âge du patient.

4- Tests non paramétriques

Ils sont utilisés obligatoirement pour des effectifs n<5 et privilégiés pour de petits effectifs, où les populations ne se distribuent pas de façon normale

⇨ les tests pour les <u>données qttives</u> ne s'appliquent plus

Il faut donc transformer ces données en ordinales (*ex : le patient du plus âgé au plus vieux*).

Donc en fonction des effectifs, on aura :

Effectif	données qttives	données qltives	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

- **<u>Coefficient r' (ρ) de Spearman</u>** = indice statistique compris entre -1 et +1, exprimant l'intensité et le sens de la relation entre 2 variables ordinales (*ex : comparer les classements de 2 juges*)
- **<u>U de Mann-Withney</u>** = permet de déterminer si 2 gpes indépendants possèdent les mêmes caractéristiques (dc viennent de la même population). On range les chiffres des 2 séries (croissant ou décroissant) puis on compte combien de chiffres de la série A arrivent avt ceux de B (U_{AB}) et inversement (U_{BA}). On calcule ensuite le <u>ddl : n_B-n_A</u> puis on compare aux tables de référence :
  - si le plus petit des **U est inférieur** à la valeur indiquée => différence significative entre les groupes (**H1)**
  - si les 2 U sont supérieurs, alors pas de différence significative (H0)

LES INDICES EN MEDECINE

Le diagnostic constitue une variable binaire : le patient souffre de la maladie ou non. On va donc rechercher un signe chez les individus.

	M	nM	Total
S	VP	FP	VP+FP
nS	FN	VN	VN+FN
Total	VP+FN	VN+FP	N

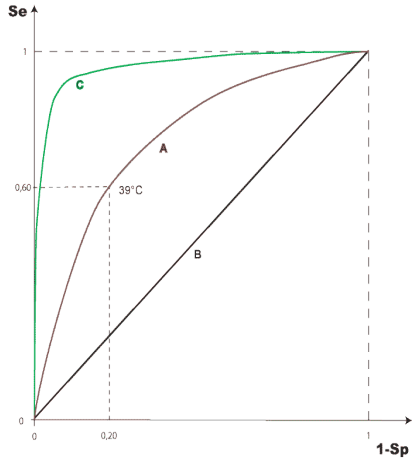
- **<u>Prévalence</u>** = nb de malade dans la population totale (VP+FN)
- **<u>Sensibilité</u>** = fqce du signe chez les malades (d'autant sensible que les M pstent le signe S) **VP/(VP+FN)**
- **<u>Spécificité</u>** = fqce de non-M ne pstant pas le signe : **VN/(VN+FP)**

Sensibilité et spécificité varient en sens inverse, et un examen est d'autant plus performant que ces deux paramètres se rapprochent de 1.

Les **<u>valeurs prédictives</u>** rendent comptent de la probabilité d'être malade ou non en fonction de la psce ou non du signe :

- VPPositive = **VP/(VP+FP)**
- VPNégative = **VN/(VN+FN)**

Pour fixer un seuil entre variables pathologiques et normales, on réalise une <u>courbe ROC</u> qui exprime la sensibilité en fonction de 1- spécificité :



Le point de la courbe qui se rapproche le plus du coin supérieur G (valeurs maximales de la Se et de la Sp) correspond au meilleur seuil possible.

Si on obtient une droite, le signe et la maladie sont indépendants, donc l'examen n'a aucun intérêt diagnostique.