Tut Rentrée - Initiation à la biostatistique

I. La méthode statistique en médecine

La statistique est une méthode scientifique qui consiste en l'art de collecter, d'analyser et d'interpréter des données. Un des rôles fondamentaux de la statistique est de permettre de décider si une observation (ou une situation de santé publique) est due au hasard ou si elle a une explication concrète.

Les biostatistiques sont les *statistiques appliquées au domaine de la santé*. En santé publique, elles s'appliquent à 3 domaines :

- *Décrire* des populations par rapport à une maladie
- *Evaluer* des traitements, des techniques, des coûts
- *Mettre en place* des observations épidémiologiques et en *tirer des conclusions*

On distingue deux sortes d'analyse :

- ***Analyse descriptive* :** Cela correspond à la collecte de données sur une population. On décrit une situation grâce à des paramètres.
- **Analyse déductive :** Une observation est-elle due au hasard ou existe-t-il une explication ?

II. Quelques définitions à connaître

Population : Série exhaustive de tous les individus étudiés. Ex : Les professions médicales, la population française…

Echantillon : Ensemble fini et d'effectif limité extrait de la population le plus souvent randomisé constitué par tirage au sort. Il doit être représentatif de la population. S'il ne l'est pas, alors l'étude de la population est faussée.

➔ **ECHANTILLON CONNU, POPULATION INCONNUE !** On se sert de l'échantillon pour étudier la population.

Paramètre : Grandeur apportant une information résumée sur la variable étudiée

Données : Résultat de l'observation d'un individu, par l'utilisation d'un instrument de mesure ou par les sens de l'observateur. Ces données sont variables d'un individu à un autre (variabilité interindividuelle), on parle donc de variable.
Il existe une grande variabilité dans le domaine biologique, qui peut être due au hasard ou qui peut être physiologique. On distingue la variabilité intraindividuelle (ex : Selon les études, on est plus grand le matin que le soir) et la variabilité interindividuelle (ex : Paul est plus grand que Thomas).

Variable quantitative : Mesurable par un outil de mesure (taille, poids…). On distingue les variables quantitatives discrètes (Sophie a 6 ans, Lucien consomme 3 paquets de cigarettes par semaine (fumer tue), Sibille a perdu 500g en deux jours…) et les variables quantitatives continues (Rodolphe a un déficit auditif moyen de 8,5 dB…)
Une variable quantitative se représente sous forme de tableaux et d'histogrammes.

Variable qualitative : Non mesurable (couleur des cheveux, des yeux, forme de visage…). On sépare les variables qualitatives en 3 catégories :

- ☑ Binaire (Sexe : Féminin/Masculin…)
- ☑ Nominale (Couleur des cheveux : Brun, roux, blond ; couleur de peau…)
- ☑ Ordinale : Il y a un **ordre** dans la réponse (Degré de satisfaction du tutorat : pas content (personne !), peu satisfait, satisfait, très satisfait, J'ADOREEE (ouuui !!) ; mention au bac : pas de mention, assez bien, bien, très bien, félicitation du jury…)

Une variable qualitative se représente sous forme de pourcentages (tableau) et d'histogrammes.

III. Statistiques descriptives

Une variabilité non maîtrisée conduit à des **biais**, qui ne permettent pas d'étudier une population: résultats erronés, conclusions faussées. Une variabilité maîtrisée permet de conclure à des estimations.
Pour étudier une variable quantitative, on dispose de plusieurs paramètres :

- ♣ **Moyenne :** On distingue deux cas :

  -Pour une variable quantitative discrète : $m = \frac{\sum_{1}^{n} x_i}{n}$

  -Pour une variable qualitative continue : $m = \frac{\sum_{1}^{n} n_i x_i}{n}$

- ♣ **Variance= (Ecart-type)².** Elle indique la dispersion des données autour de la moyenne.

- ♣ **Médiane :** Valeur de l'observation centrale **si** rangement par ordre croissant.
  Ex : Eva a 6 poulets, Nicky en a 7, Lilo en a 4, Stitch 3 et Matthew 5. On range par ordre croissant : 3,4,**5**,6,7. La médiane est donc de 5 dans ce cas.

- ♣ **Quartile :** Ils partagent la série ordonnée en 4 groupes de même effectif. Ex : On veut calculer le premier quartile : Q1= 0,25x5=1,25. Le premier quartile est situé entre la première et la deuxième valeur, d'où : Q1= (3+4)/2= 7/2=3,5. De même, Q2=0,5x5=2,5, d'où Q2= 4,5. Q3= 0,75x5= 3,75, donc Q3= 5,5.

	MOYENNE	**MEDIANE**
AVANTAGES	-Facile à calculer et à manipuler -Significative si répartition symétrique des données et dispersion faible	-Facile à calculer -Peu sensible aux valeurs anormales (mini et maxi) -Utilisable pour les valeurs ordinales
INCONVENIENTS	-Sensible aux valeurs anormales (mini et maxi)	-Moins adéquat pour les calculs statistiques

A. Estimation statistique

Il s'agit de déterminer une grandeur définie sur une population à partir d'observations réalisées sur un échantillon de cette population. Pour pouvoir extrapoler à la population entière, l'échantillon doit être **REPRESENTATIF** de la population que l'on veut étudier.

Il existe deux types d'estimation :

- **Estimation ponctuelle** : C'est la valeur unique jugée la meilleure à l'instant t pour un échantillon donné unique.

- **Estimation par intervalle** : C'est un intervalle de valeurs contenant la valeur recherchée. On l'appelle également intervalle de confiance (noté IC) ou intervalle au risque α où α représente le risque d'erreur dans l'estimation de μ. L'estimation par intervalle est en général plus fiable que l'estimation ponctuelle.

Deux estimations ponctuelles d'une même variable réalisée sur deux échantillons A et B donnent des valeurs ponctuelles voisines, mais qui ne sont pas nécessairement la même valeur. Alors que deux estimations par intervalle d'une même variable réalisée sur deux échantillons A et B auront des IC qui se recouvreront.

Les techniques utilisées afin de constituer un échantillon représentatif de la population sont :

- ✓ La détermination précise des caractéristiques de la population
- ✓ **Le tirage au sort** (noté TAS) d'un grand nombre adapté d'individus. Le TAS permet d'éviter des biais d'échantillonnage (Ex de biais d'échantillonnage : On cherche le pourcentage d'hommes et de femmes ayant les yeux bleus dans le monde. Pour cela, on prend comme échantillon la population suédoise. Cela n'est pas représentatif, les suédois(es) ayant en moyenne beaucoup plus de personnes ayant les yeux bleus que par exemple en Afrique. Le TAS permet d'éviter ce type d'erreur).

Afin de réaliser une estimation statistique de données quantitatives, il y a trois étapes successives à respecter :

1. Constitution d'un échantillon représentatif par TAS
2. Calcul de la moyenne m et de l'écart-type s sur l'échantillon
3. Estimation de la valeur vraie de la moyenne μ et de l'écart-type σ au niveau de la population.

Les différents paramètres sont notés différemment selon l'étude de l'échantillon ou de la population :

	ECHANTILLON	**POPULATION**
MOYENNE	m	μ
ECART-TYPE	s	σ
EFFECTIF	n	N

On parle de **valeur vraie** lorsqu'on parle des paramètres de la population.

L'écart-type (tout comme la variance, un petit effort de mémoire svp) mesure la dispersion des données autour de la moyenne. (Les données sont-elles centrées autour de la moyenne ou au contraire groupées autour d'elle ? Les données sont-elles regroupées entre elles ou au contraire dispersées ?) Plus l'écart-type est FAIBLE, moins les données sont dispersées et donc les données sont réparties de façon HOMOGENE. Inversement, plus l'écart-type est GRAND, plus les données sont dispersées, elles sont donc réparties de façon HETEROGENE. Il permet également de calculer la précision de l'IC.

L'écart-type a pour formule : $\mathbf{s = \sqrt{\Sigma(xi-m)^2/n}}$.

L'intervalle de confiance : Il permet de délimiter un encadrement de valeurs dans lequel se trouvera la valeur vraie de la moyenne μ au niveau de la population. On a donc **μ qui appartient à IC :**

$$\mu \in IC$$
$$\mu \in [\, m +/- \varepsilon\, s/\sqrt{n}] \text{ avec } i = \varepsilon s/\sqrt{n} = \text{précision de l'IC}$$

On peut ainsi se rendre compte que plus la taille de l'effectif n augmente, plus i est petit et donc plus l'IC se resserre : Sa précision a donc augmenté. L'estimation tend donc de plus en plus vers la valeur vraie. Dans l'estimation de μ, on prend en compte le risque d'erreur α qui correspond au risque que la moyenne μ ne soit pas dans l'IC. On fixe généralement **α à 5%. ε** représente **l'écart réduit.** La valeur d'ε dépend de la valeur d' α.

Pour une valeur de **α = 5%,** ε vaut **1,96.**
Pour une valeur de **α= 10%,** ε vaut **2,6.**
NB : α et ε varient en sens inverse.

On distingue l'IC95 et l'IC99:
IC95 : Il y a 95% de chances que μ appartienne à l'IC.
IC99 : Il y a 99% de chances que μ appartienne à l'IC.

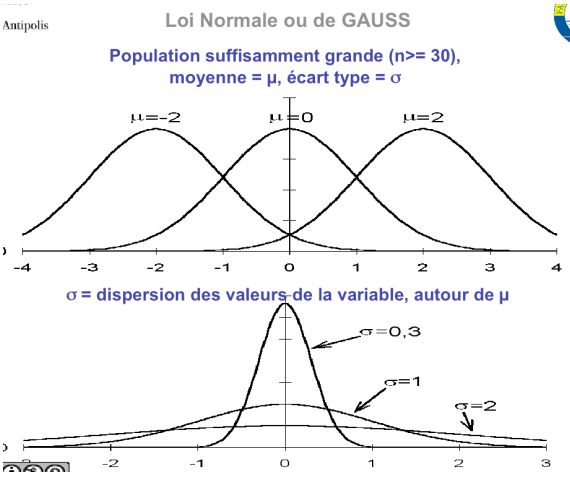
Les variations de α conditionnent la précision de l'estimation et la largeur de l'IC :

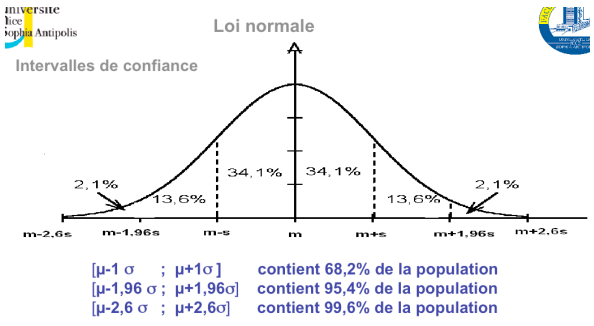
- Si **α est petit**, on ne prend pas beaucoup de risques dans l'estimation de l'IC, celui-ci est donc grand. α et ε variant inversement proportionnellement, si α est petit alors **ε est grand**. Dans ce cas, on ne rate pas la valeur vraie de μ, mais la **précision de l'IC est faible.**
- Si **α est grand**, on prend des risques dans l'estimation de l'IC, celui-ci est donc petit. α et ε variant inversement proportionnellement, si α est grand alors **ε est petit**. Dans ce cas, la **précision de l'IC est grande** mais on risque de rater la valeur vraie de μ.

B. LA LOI DE GAUSS = LOI NORMALE

La **loi de Gauss** est une loi qui permet, pour tout échantillon où **n ≥ 30**, de visualiser :

- ✓ La notion **d'IC** autour de la moyenne
- ✓ La notion **d'écart-type**
- ✓ La notion de **dispersion des données** autour de la moyenne





La **représentation graphique** de données par la loi de Gauss donne une courbe en cloche avec :
- ✓ Abscisse : m +/- εs, donc l'IC
- ✓ Ordonnée : n
- ✓ Aire sous la courbe : le % de la population concernée

La courbe de Gauss est toujours centrée sur la moyenne m de la population.
Pour α=5%, l'aire sous la courbe comprise dans l'intervalle [m – 1,96s ; m + 1,96s] correspond à 95% de la population.
Pour α=1%, l'aire sous la courbe comprise dans l'intervalle [m – 2,6s ; m + 2,6s] correspond à 99% de la population.
Plus l'écart-type est grand, plus les données sont dispersées et plus la courbe de Gauss a une forme aplatie. Plus l'écart-type est faible, plus les données sont rapprochées et plus la courbe de Gauss a une forme en cloche.

C. L'ESTIMATION DE DONNEES QUALITATIVES

Dans l'estimation de données **qualitatives**, on s'intéresse a la proportion de la population présentant une caractéristique quelconque A. Cette estimation se déroule en plusieurs étapes :

1. constitution d'un échantillon représentatif par TAS
2. calcul du pourcentage p_{obs} de l'échantillon présentant A et de l'écart-type s
3. Estimation de la valeur vraie p du pourcentage de la population présentant A et de l'écart-type σ

Comme pour l'estimation de données quantitatives, on utilise des paramètres que l'on nomme différemment selon l'étude de l'échantillon ou de la population :

	ECHANTILLON	**POPULATION**
POURCENTAGE (%)	$P_{obs}=p_{observé}$	p
ECART-TYPE	s	σ
EFFECTIF	n	N

L'écart-type a les mêmes caractéristiques pour une variable qualitative que pour une variable quantitative.

$$\mathbf{s = \sqrt{(p_{obs} \times q_{obs}/n)}}$$
$$\mathbf{avec\ q_{obs}=1-p_{obs}}$$

L'IC se calcule un peu différemment pour une variable qualitative :

$$\mathbf{p \in IC}$$
$$\mathbf{p \in [\ p_{obs} +/- \varepsilon\ s]}$$
$$\mathbf{avec\ i= \varepsilon\ s}$$

La précision de l'IC et l'influence d'α sur l'estimation de l'IC sont les mêmes que pour l'estimation des données quantitatives.
Le **sondage** est une application directe de l'IC calculée sur des données qualitatives. Les instituts de sondage fournissent toujours sous forme de pourcentage **la valeur centrale de l'IC calculé**. Un IC devrait accompagner tout résultat de sondage pour que le résultat soit fiable.

IV. Statistique déductive

Dans les statistiques déductives, contrairement aux statistiques descriptives, on essaie, à partir des observations faites, de tirer des conclusions. Pour cela, les épidémiologistes utilisent des tests d'hypothèse.
La première étape d'une étude statistique est la **formulation d'hypothèses** qu'un test permettra ensuite de confirmer ou d'infirmer. On utilise les **hypothèses** lorsque l'on veut comparer deux groupes pour un caractère donné. Au début de chaque test, on définit deux hypothèses qui jouent un rôle symétrique :

- **H0= Hypothèse nulle=** Il n'y a **pas de différence** observée entre les deux groupes, il n'existe **pas de lien** entre les deux caractères étudiés, les fluctuations observées sont dues au seul hasard. (ex : Etre une fille et avoir le concours PAES n'a aucun lien)

- **H1= Hypothèse alternative=**Il y a une **différence significative** entre les deux groupes, il existe **un lien** entre les deux caractères étudiés, les fluctuations ne sont pas dues au hasard. (ex : Etre une fille et réussir le concours PAES du premier coup ont un lien significatif.)

Les tests utilisés le plus souvent en statistiques déductives sont les tests de comparaison. Il en existe deux types :

- entre 2 populations : on constitue alors 2 échantillons représentatifs et on essaie de déterminer s'il existe une **différence significative entre ces 2 échantillons.**

- Entre une population donnée A et la population générale de référence : on constitue alors un échantillon représentatif de la population A et on essaie de déterminer s'il existe une différence significative entre l'échantillon et la population générale et donc, in fine, entre A et la population générale.

Les tests sont des techniques permettant de décider si on **accepte ou si on rejette H0**, en ayant fixé le **risque d'erreur α** accompagnant cette décision.
NB : On choisit toujours pour **H0** l'hypothèse qu'il serait **le plus grave de rejeter à tort.**

	REJET H0/ACCEPT H1	**ACCEPT H0/REJET H1**
H0 VRAIE/ H1 FAUSSE	α=Risque de 1^{e} espèce	1-α
H0 FAUSSE/H1 VRAIE	1-β=Puissance du test	β=Risque de 2^{e} espèce

Ce tableau est à retenir pour le concours !! +++

Résumé du tableau:
-α est le **risque de 1^{e} espèce,** il représente le risque de **rejeter H0 si H0 est vraie.** Ce risque est maîtrisé et généralement **fixé à 5%.** Il est fixé **AVANT** l'application du test statistique (on y revient dans la suite)
-β est le **risque de 2^{e} espèce**, il représente le risque **d'accepter H0 si H0 est fausse.** Ce risque est négligé et peut être assez important.
-**1-α est** la probabilité **d'accepter H0 si H0 est vraie.**
-**1-β** représente la **puissance du test.** Il s'agit de la **probabilité de rejeter H0 si H0 est fausse.**

Les étapes d'un test statistique :
Pour réaliser un test d'hypothèse, il faut **TOUJOURS respecter DANS L'ORDRE les étapes suivantes** :

1. Définir les **hypothèses H0 et H1**
2. **Choisir le test en fonction du type de données** (Qualitatif/Qualitatif ; Quantitatif/Qualitatif ; Quantitatif/Quantitatif, on étudiera ces tests dans les cours à venir). On nomme **Z le paramètre calculé.**
3. Choisir le **risque α**
4. **Recueil de données**
5. **Interprétation** des résultats (au niveau de l'échantillon puis de la population)

V. Le mot de la fin

Vous arrivez à la fin de la fiche du deuxième cours de Biostat de la tut rentrée. Sur les deux cours, on a réussi à couvrir 4 cours sur 16 du programme de Biostat. Donc si vous les comprenez bien et que vous les apprenez bien, vous avez fait ¼ du programme ☺. Maintenant concernant ce cours-ci, il n'est pas très compliqué en soit. Il faut d'abord bien comprendre les notions et surtout ne pas s'embrouiller entre elles, car beaucoup se ressemblent. Une fois bien comprises, quand vous devrez le revoir par la suite il vous paraitra assez facile, du moins on l'espère. Je vous conseille de bien vous entraîner à vous réciter le cours, comme ça vous saurez si vous arrivez bien à « jouer » avec les notions. Au Concours Blanc, il y aura 10 questions sur 19 en Biostat sur ce cours-ci. Les QCMs seront bien répartis sur tout le cours, certains simples, d'autres avec quelques pièges, mais si vous apprenez bien et que vous faites attention le jour du Concours Blanc, ça devrait marcher ;)
Bon courage !

Cleaar et attention83