Statistiques Descriptives

I- Notion de Variabilité

Toutes les données biologiques possèdent une variabilité. La connaissance de cette variabilité est nécessaire pour pouvoir classer nos données comme « normales » ou « anormales ».

- Une variabilité **maîtrisée** permet une **estimation**
- Une variabilité **non maîtrisée** conduit à des **biais**

Par exemple les valeurs normales de la glycémie sont comprises entre 0,75 et 1,25 g/L. Si on est en dessous de 0,75 g/L on a une valeur anormale, on est en hypoglycémie.

II- Estimations en statistiques

A) Définition :

Les études en biostatistique sont réalisées sur un **échantillon représentatif** de la population après « échantillonnage ». Après l'étude on doit réfléchir à la légitimité des résultats et à leur extrapolation potentielle à l'ensemble de la population. Pour ça on réalise une **estimation du résultat vrai** à partir des données obtenues sur l'échantillon : **On détermine des paramètres au niveau d'une population à partir d'observations réalisées sur un échantillon de cette population**.

On retrouve deux types d'estimations :

1. **L'estimation ponctuelle :** valeur **unique** jugée la meilleure à l'instant t *(peu fiable)*.
2. **L'estimation par intervalle :** il y a un intervalle de valeurs comprenant la valeur recherchée, c'est **l'Intervalle de Confiance** ou **IC** *(beaucoup plus fiable).*

Deux estimations ponctuelles d'une même variable réalisées sur 2 échantillons A & B donneront des valeurs ponctuelles proches mais pas nécessairement la même valeur.

Deux estimations par intervalle d'une même variable réalisées sur 2 échantillons A & B donneront des IC se recouvrant (car proches) mais pas nécessairement le même IC.

L'estimation par intervalle est moins précise. Cependant, si on refait la même estimation sur un autre échantillon, elle recouvrira la première, ce qui ne serait sûrement pas le cas avec des valeurs ponctuelles. Donc **l'estimation par intervalle est plus juste**, d'où son intérêt.

B) Estimation des données quantitatives :

Méthodologie :

1) Définition précise de la population étudiée = *Population cible*
2) Tirage au sort d'un échantillon représentatif
3) Calcul de l'intervalle de confiance

Pour les données quantitatives, on va estimer la moyenne !

L'estimation assure la correspondance entre ce qu'il se passe au niveau de l'échantillon et ce qu'il se passe au niveau de la population

ESTIMATION

Paramètre	Échantillon	Population
Moyenne	m = estimateur de la moyenne vraie μ au niveau de l'échantillon	μ = moyenne vraie
Ecart type	s = estimateur de l'écart type vrai σ au niveau de l'échantillon	σ = écart type vrai
Effectif	n	N

a) Écart-type :

Il mesure la **dispersion** d'un ensemble de données autour de la moyenne. C'est la variabilité des mesures entre elles et par rapport à la moyenne. **Plus l'écart-type est faible plus le caractère étudié est homogène** (plus les valeurs sont proches de la moyenne).

b) Degrés de liberté :

On définit « m » la moyenne, « x_i » les valeurs dont on veut faire la moyenne, « n » l'effectif, « $x_i - m$ » les écarts.

- ➔ Il y a **n écarts**
- ➔ Il y a **(n – 1) écarts indépendants à la moyenne**, ou **degrés de liberté**

Les degrés de liberté ou **ddl**, c'est le **nombre de valeurs nécessaires à connaître pour pouvoir résoudre** l'équation et connaître toutes les valeurs de la série. (Si je connais ma moyenne et toutes mes valeurs sauf une, je peux trouver la valeur manquante).

Exemple :

Paul a eu 3 notes mais une de ses évaluations est tachée. Il sait qu'il a eu 12 et 13 et il connait sa moyenne : 14.

m=14 ; x={12,13,y} ; n=3 Il peut donc avec n-1 valeurs, c'est à dire 2 valeurs, trouver la troisième, il y a différentes techniques, comme par exemple avec la moyenne:

$moyenne = 14 = \frac{12+13+y}{3}$ *donc y = 17*

On retrouve bien sa note à partir des autres, cependant s'il manquait deux notes on n'aurait pas pu déterminer la deuxième c'est pourquoi il y a n-1 et pas n-2 ddl.

c) Intervalle de Confiance :

L'IC c'est l'estimation de la moyenne vraie μ à partir de la moyenne calculée sur l'échantillon. On donne un intervalle auquel μ appartient.

$$\mu \in \left[m \mp \frac{\varepsilon s}{\sqrt{n}}\right]$$

L'IC est aussi appelé intervalle au risque α. Le risque α c'est le risque d'erreur dans l'estimation de μ (autrement dit le risque que notre intervalle de confiance ne comprenne pas la valeur vraie de μ). **On prend en général α = 5%** (donc on a 95% de chances que la moyenne vraie appartienne bien à l'IC).

ε représente l'écart-réduit : C'est une valeur qui **dépend du risque α** *(ils varient en sens inverse, si α augmente, ε diminue)*. Un écart-réduit mesure de combien d'écarts-types une observation particulière est éloignée de la population.

$$pour\ \alpha = 5\%\ ;\ \varepsilon = 1{,}96$$

$$pour\ \alpha = 1\%\ ;\ \varepsilon = 2{,}60$$

d) Précision de l'estimation :

Les variations du risque α vont conditionner la précision de l'estimation et la largeur de l'intervalle de confiance.

IC Large	IC Resserré
Si **α**↘ alors **ε**↗ donc l'**IC**↗ ➔ On a plus de chances que μ soit comprise dans l'IC ➔ Par contre on perd en précision	Si **α**↗ alors **ε**↘ donc l'**IC**↘ ➔ On a moins de chance que μ soit dans l'IC ➔ Mais on diminue l'IC, on gagne en précision
 *La précision est mauvaise parce que les flèches ne sont pas au centre mais il n'y a pas de valeurs qui ne sont pas dans l'IC*	 *La précision est meilleure puisque les flèches sont + proches du centre mais les points verts ne sont pas dans l'IC*

Ici on visualise l'intervalle de confiance comme une cible

C'est assez logique : Si on prend moins de risque (α↘), on a un intervalle de confiance plus grand, on a plus de chances que la moyenne soit dedans. (et inversement).

Indice de précision « i » : Il permet de calculer la précision de l'estimation de μ. Cette valeur **représente la largeur de l'IC**.

$$i = \varepsilon \frac{s}{\sqrt{n}}$$

D'après la formule de l'IC vue juste avant, **l'IC est compris :**

entre $\left[m - \frac{\varepsilon s}{\sqrt{n}}\right]$ **et** $\left[m + \frac{\varepsilon s}{\sqrt{n}}\right]$ **donc entre** $[m - i]$ **et** $[m + i]$

D'après la formule de l'indice de précision,

Si **n**↗ alors **i**↘ donc l'**IC**↘ donc **la précision** ↗

Attention, quand l'indice de précision diminue la précision augmente ! On peut conclure que **plus la taille de l'échantillon augmente, plus la précision augmente**.

« n » le nombre de sujets nécessaires : $n = \varepsilon^2 \frac{s^2}{i^2}$

e) Loi de Gauss ou loi Normale :

En sciences humaines, on observe souvent des distributions des variables plutôt symétriques autour de la moyenne avec une forme de cloche : c'est la **courbe de Gauss**.

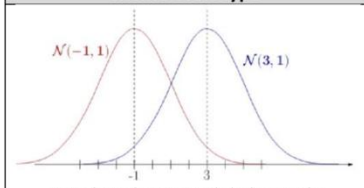
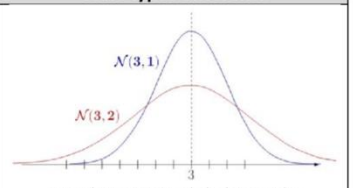
La représentation graphique de données par la loi de Gauss donne une courbe en cloche avec :

- ➔ En abscisse : [m +/- εs], donc l'IC
- ➔ En ordonnée : n
- ➔ L'aire sous la courbe : le % de la population concernée

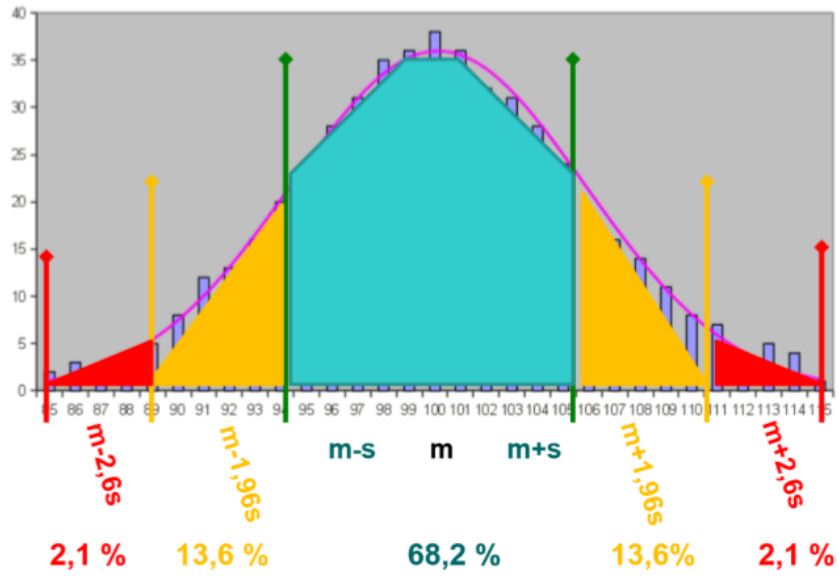
La loi de Gauss permet de visualiser **l'IC** autour de la moyenne, **l'écart type**, **la dispersion** autour de cette valeur moyenne et **la moyenne**.

Pour pouvoir faire des calculs on va supposer que notre variable X (quantitative continue) suit une distribution « modèle » : **la loi Normale**. Ainsi, Pour chaque (μ,σ) il existe une loi normale de moyenne μ et d'écart type σ : on la note **N (μ, σ)**

Exemples de lois normales pour bien comprendre le fonctionnement de la courbe

Moyennes différentes, même écart type	Même moyenne, écart types différents
	
Décalage horizontal de la courbe	*Décalage vertical de la courbe (la courbe s'étale = dispersion↗)*

A partir de la loi Normale ou loi de Gauss on peut trouver des IC ++++



[m - 1 s ; m + 1 s] contient 68,2% de la population
[m - 1,96 s ; m + 1,96 s] contient 95,4% de la population (95%)
[m - 2,6 s ; m + 2,6 s] contient 99,6% de la population (99%)

C) Estimation des données qualitatives :

Méthodologie :

1) Constitution d'un échantillon représentatif par TAS
2) Calcul du pourcentage p_{obs} au de l'échantillon présentant un caractère 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 »

Pour les données qualitatives, on va estimer un pourcentage !

Comme précédemment, l'estimation assure la correspondance entre ce qui se passe au niveau de l'échantillon et au niveau de la population.

	Echantillon	Population
Proportion (%)	$p_{obs} = p_{observé}$ = estimateur du pourcentage inconnu p	p = pourcentage vrai
Ecart type	s = estimateur de l'écart type vrai σ au niveau de l'échantillon	σ = écart type vrai
Effectif	n	N

Tout ce qui va suivre sera le même procédé que pour les données qualitatives, seuls changeront les paramètres utilisés et donc les formules qui en découlent.

a) Écart-type :

Il a les mêmes caractéristiques que la variable soit qualitative ou quantitative. Il est donné par $s = \sqrt{(\frac{p_{obs} \cdot q_{obs}}{n})}$ avec $q_{obs} = 1 - p_{obs}$

b) Intervalle de Confiance :

L'IC c'est l'estimation de la moyenne vraie μ à partir de la moyenne calculée sur l'échantillon. On donne un intervalle auquel μ appartient.

$$p \in [p_{obs} \mp \varepsilon s]$$

c) Précision de l'estimation :

Indice de précision « i » : Il représente toujours la largeur de l'IC.

$$i = \varepsilon \sqrt{\frac{pq}{n}} = \varepsilon s$$

Si n est multiplié par 100, alors s est divisé par 10 et donc la précision augmente d'un facteur 10

On peut aussi conclure sans problème la même chose :

Si **n**↗ alors **i**↘ donc l'**IC**↘ donc **la précision** ↗

On peut conclure que **plus la taille de l'échantillon augmente, plus la précision augmente**. La précision dépend de la taille de l'échantillon, et de l'écart-type « s ».

« n » le nombre de sujets nécessaires : $$n = \varepsilon^2 \frac{pq}{i^2}$$

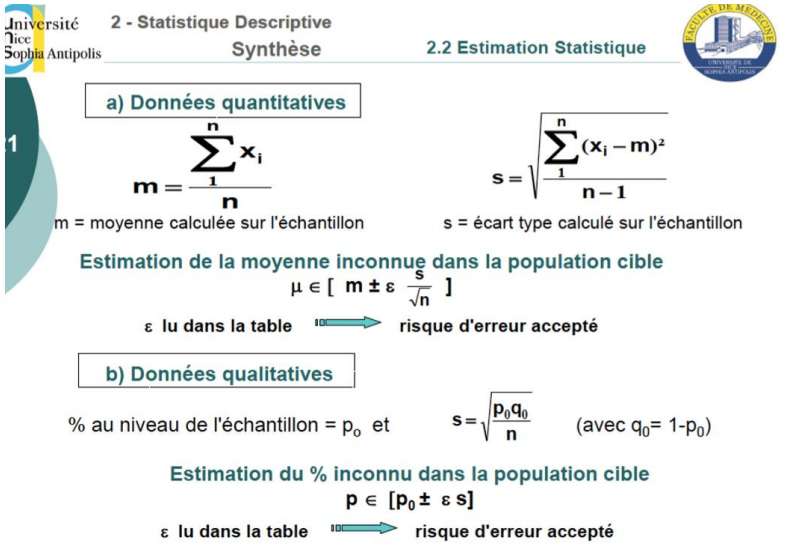
d) Sondages :

Le sondage est une **application directe de l'IC calculée sur des données qualitatives**. Tout résultat de sondage doit être accompagné d'un IC.

Pour une bonne estimation il nous faut :

- ➔ Un échantillon représentatif constitué par TAS
- ➔ Pas de biais pendant la sélection
- ➔ Un IC qui accompagne toujours l'estimation *(il montre la variabilité des données)*
- ➔ Une taille importante de l'échantillon : Si n↗ la précision ↗

La Biostats vous zem <3



Dédi du soir bonsoir

À Grohl & Léaccouchement parceque j'vous zem

À mes ptites vieilles toutes meugnonnes

À mon arrière-grand-père Guerric et mon arrière-grand-mère Mayanne

Antho continue tout droit j'ai grave confiance en toi

Mes fillots qui ne sont plus malades (maintenant c'est moi) mais qui gèrent la fougère : Lamia, Inès, Elodie, Timotey, Andréa, Teresa <3 continuezzz

L'UE13 en baaaambe, Marioscopie & Email <3

La team de l'Ambro'sProject kesk'on fout là aucune idée

À l'espoir