Introduction

L'objectif de la biostatistique consiste à caractériser une population à partir d'une image plus ou moins floue constituée à l'aide d'un échantillon issu de cette population. *On peut alors chercher à extrapoler une information obtenue à partir de l'échantillon.*

La statistique constitue, en biologie, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence :

1. Quelle est la valeur normale d'une grandeur biologique, taille, poids… etc ?

2. Quelle est la fiabilité d'un examen complémentaire ?

3. Le traitement A est-il plus résistante que le traitement B ?

La bio-statistique propose des méthodes essentielles à la recherche dans toutes les branches de la biologie.

Il faut distinguer entre ses termes :

• **La statistique** = le domaine scientifique.

• **Une statistique** = une quantité (paramètre) statistique estimée (moyenne, variance, etc.)

• **Des statistiques** = des données, par exemple le nombre d'espèces animales dans une région donnée.

Objectifs du cours

- Connaître le vocabulaire particulier de la statistique ;
- Comprendre les principes du traitement des données ;
- Le choix de la méthode statistique convenable à chaque situation particulière ;
- la réalisation des calculs et des tests de base pour une et deux variables.

Chapitre I : Généralités et notions de base

Définitions

- C'est l'analyse statistique des données biologiques.
- La mathématique de l'expérimentation.
- Ensemble de méthodes à partir desquelles on recueille, organise, résume, présente et analyse des données afin d'en tirer des conclusions et de prendre des décisions avec prudence. Les conclusions, toujours entachées d'un certain pourcentage d'incertitude, nous permettent alors de prendre une décision.
- Étude scientifique des données numériques décrivant les variations naturelles.

Notions de base

Parmi les notions importantes nous avons :

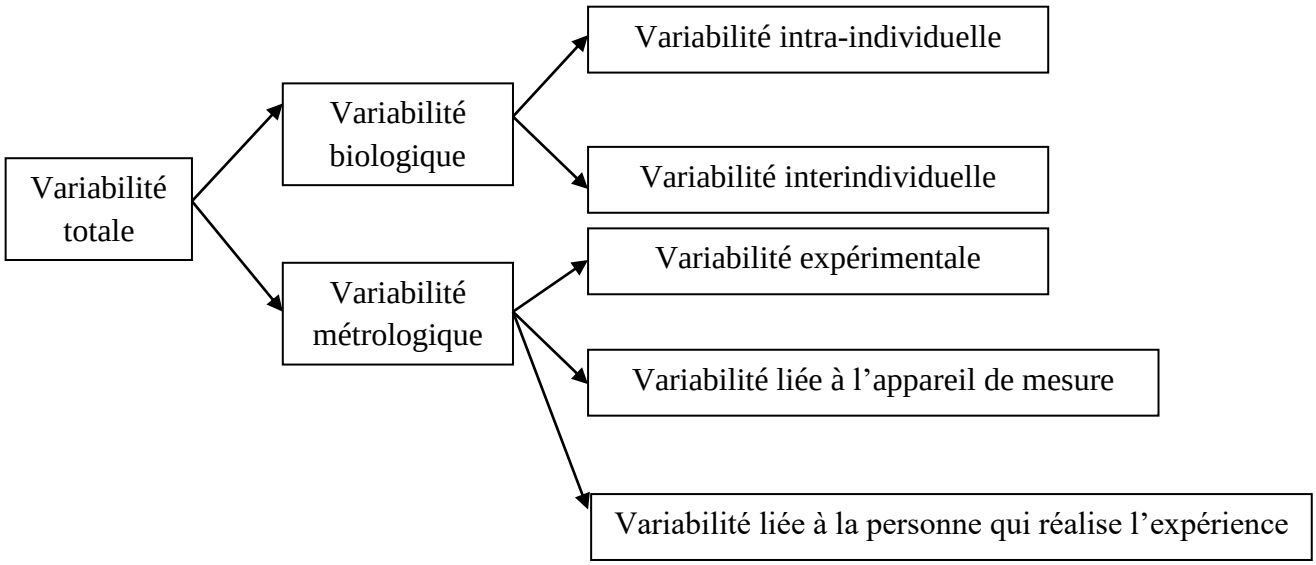
La variabilité

La variabilité est la somme d'une variabilité métrologique et d'une variabilité proprement biologique.

La variabilité biologique peut être elle-même décomposée en deux termes : d'une part la variabilité intra-individuelle, qui fait que la même grandeur mesurée chez un sujet donné peut être soumise à des variations ; et d'autre part la variabilité interindividuelle (qui fait que cette même grandeur varie d'un individu à un autre).

En général, la variabilité intra est moindre que la variabilité inter.

La variabilité métrologique peut être elle aussi décomposée en deux termes : d'une part les conditions expérimentales ; et d'autre part les erreurs induites par l'appareil de mesure utilisé.



Exemple

La mesure de la glycémie peut grandement varier sur un individu donné suivant les conditions de cette mesure.

Population : Ensembles des ***éléments*** ou ***d'individus*** de même nature, visés par une problématique scientifique. Tous les étudiants de la faculté S.N.V. constituent une population.

Elément : Les éléments sont les unités qui composent une population.

Synonymes : Objet, individu, unité statistique, unité d'échantillonnage, sujet, événement, comportement, localité, parcelle, observation, prélèvement, entité …

Il est le plus souvent impossible, ou trop coûteux, d'étudier l'ensemble des individus constituant une population ; on travaille alors sur une partie de la population que l'on appelle ***échantillon***.

Echantillon : C'est un sous ensemble de la population considérée, prélevé pour juger de cet ensemble.

Echantillon représentatif : Échantillon qui reflète fidèlement la complexité et la composition de la population. Le tirage au sort ainsi que l'inventaire exhaustif (recensement), sont deux façons d'obtenir un échantillon représentatif d'une population.

Caractère : C'est la propriété ou l'aspect singulier que l'on se propose d'observer dans la population ou l'échantillon.

Synonymes : Variable statistique, descripteur, caractère, attribut, trait, profile (en géophysique), stimulus (en étude du comportement).

Ce variable peut être quantitative (numérique) ou qualitative (non numérique).

Variables quantitatives : Pouvant être classées en ***variables continues*** (taille, poids) ou ***discontinues*** (nombre d'enfants dans une famille, nombre d'œufs pondus par un oiseau).

Variables qualitatives : Pouvant être classées en ***variables catégorielles*** (couleurs des plumes des oiseaux) ou ***ordinales*** (résistance d'une plante vis-à-vis un ravageur classée en faible, moyenne, importante).

Notion d'hypothèse

L'hypothèse est une relation hypothétique (provisoire, postulée par le chercheur).

On distingue deux formes d'hypothèses :

Hypothèse nulle (H0) et *Hypothèse significative (H1)* ou *alternative.*

- **Hypothèse nulle (H0)** : $m_1 = m_2$ ou l'absence d'une différence significative entre les moyennes ;
- **Hypothèse alternative (H1)** : $m_1 \neq m_2$ ou l'existence d'une différence significative entre les moyennes.

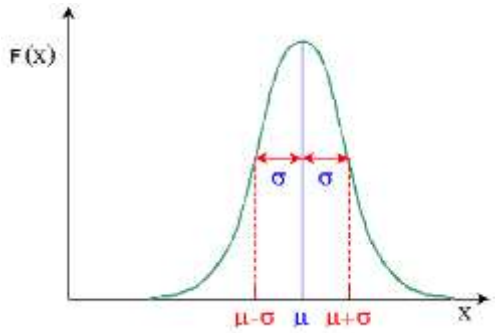
Seuil de signification ou marge d'erreur

En statistique, il n'existe pas de règle rigide permettant de tirer une conclusion concernant les hypothèses ; aucun test ne nous fournit une réponse en terme de oui ou non, mais indique dans quelle mesure nous pouvons être certain de tirer des conclusions ; cette mesure se nomme niveau ou seuil de signification, ou encore probabilité d'erreur.

La loi normale

Une distribution normale correspond à la distribution de probabilités d'une variable aléatoire continue dont la courbe est parfaitement symétrique et en forme de cloche.

Lorsqu'une variable (x) se distribue de telle sorte que les fréquences de ses différentes éventualités suivent la loi normale, alors elle est dite variable normale.



Types de test

On parle de ***tests paramétriques*** lorsque l'on stipule que les données sont issues d'une distribution paramétrée. Dans ce cas, les caractéristiques des données peuvent être résumées à l'aide de paramètres estimés sur l'échantillon (moyenne, mode et médiane). La distribution des données suit la loi normale.

Les ***tests non paramétriques*** ne font aucune hypothèse sur la distribution sous-jacente des données (la distribution des données ne suit pas la loi normale). On les qualifie souvent de tests pour distribution libre.

Chapitre II : Statistiques descriptives à une dimension

Le but de simplification de la statistique descriptive peut être atteint en condensation les données d'observations sous formes :

- des tableaux statistiques ;
- des représentations graphiques.

1- Les paramètres de position

1-1 Le mode

Le mode, désigné par Mo est la valeur de la variable statistique la plus fréquente.

Remarque : Le mode ou la classe modale n'est pas obligatoirement unique.

1-2 La médiane

La médiane, désignée par Me, est la valeur qui partage l'échantillon en deux groupes de même effectif ; pour la calculer, il faut commencer par ordonner les valeurs (les ranger par ordre croissant par exemple).

Exemple : soit la série 12 3 24 1 5 8 7

On l'ordonne : 1 3 5 7 8 12 24

7 est la médiane de la série

1-3 la moyenne

Lorsque x désigne la variable statistique, la valeur moyenne, ou ***moyenne*** de la série se note m ou x. Elle est l'analogue d'un centre de gravité.

$$m = 1/n \sum x_i$$ n = effectif total

Avec la série précédente, qui comporte $n = 7$ valeurs, on obtient :

$x = 1/7 \sum(1+3+5+7+8+12+24) = 8,57.$

1-4 Les valeurs extrêmes

Ce sont la valeur minimale x_{min} (x_1) et valeur maximale x_{max} ou (x_n).

2- Les paramètres de dispersion

2-1 L'étendue

L'étendu, noté e, représente la différence entre les valeurs extrêmes de la distribution : $e = x_n - x_1$.

Exemple : Soit la série suivante : 15 14 18 17 19 12 56 48 47 59

e = xn – x1 = 59 – 12 = 47.

2-2 La variance

C'est la moyenne arithmétique des carrés des écarts par rapport à la moyenne. Symbolisé par le signe (s^2) ou dans la littérature par (σ^2), et elle est donnée par la relation suivante :

$Sx^2 = 1/n-1 \sum(x_i - m)^2$

Exemple : Avec la série précédente

m = 30,5

$Sx^2 = 1/9 \sum (15-30,5)^2 + (14-30,5)^2 + \ldots + (59-30,5)^2$

$Sx^2 = 374,05$

2-3 L'écart-type

Aussi appelé déviation standard, c'est la racine carrée de la variance, symbolisé par le signe (S) ou (σ) et donné par : Sx = √variance

Avec l'exemple précédent on obtient

Sx = √374,05 = 19,34

2-4 Le coefficient de variation

Il permet d'apprécier l'homogénéité des observations dans un échantillon. Symbolisé par le signe *V* ou C.V. et donné par la relation suivante :

C.V. ou $V = Sx / m \times 100$

Si le $V < 5\%$ il s'agit d'observations très homogènes ;

Si $5\% < V < 10\%$ il s'agit d'observations homogènes ;

Si $10\% < V < 15\%$ il s'agit d'observations moyennement homogènes ;

Si $15\% < V < 30\%$ il s'agit d'observations hétérogènes ;

Si le $V > 30\%$ il s'agit d'observations très hétérogènes.

Exemple

Avec la série précédente, *quel est le degré d'homogénéité entre les observations ?*

La solution

$V = Sx / m \times 100$ implique C.V. = 19,34 / 30,5 × 100 = 63,41%

Cette dernière valeur indique que les observations de la série sont très hétérogènes.