Table des matières

Chapitre 1

Rappel

1.1 Terminologie et vocabulaire statistique

1.1.1 Définition et but des biostatistiques

La biostatistique est un champ scientifique constitué par l'application de la science statistique à la biologie. La statistique est la discipline qui étudie des phénomènes à travers la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre ces données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques.

But : les statistiques permet de décrire un état des lieus d'un phénomène à travers le calcul des différents paramètres statistiques ou la représentation graphique de phénomène en question, confirmer ou infirmer une hypothèse avec une marge d'erreur la plus petite possible et/ou prédire un événement à l'aide d'outils et paramètres statistiques. La démarche scientifique utilisée est basée sur l'application des formules mathématiques appelées ''tests statistiques". Ces derniers permettent d'une part l'interprétation des résultats, et d'autre part la prise d'une décision par rapport au phénomène étudié.

1.1.2 Notion de population et échantillon

— **La population** : un ensemble d'élément

— **L'échantillon** : (sample) est le fragment d'un ensemble prélevé pour juger de cet ensemble. Autrement dit, la fraction de la population statistique sur laquelle des mesures sont fait pour connaitre les propriétés de cette population.

— **L' élément** : ou individu est l'unité de base de l'échantillonnage, sur lequel on observe des phénomènes ou on mesure des performances.

— **La variable** : consiste à toute caractéristique mesurable ou observable sur un élément d'échantillonnage (variable propre) ou son environnement (variable associé). Il existe différents types de variables :

  — Variable qualitative : non mesurable correspond aux caractères appréciables. Elle peut être binaire (sexe des animaux), nominale (race) ou ordinale (classe d'âge).

  — Variable quantitative : mesurable correspond aux caractères quantifiables. Elle peut être continue (Poids) ou discrète (nombre de d'œufs pondus).

1.2 Statistiques descriptives univarié

1.2.1 Représentation des données

1.2.1.1 Représentation des données

Table de distribution des fréquences : Les données brutes et sommaires d'une série statistique est difficilement interprétable voir impossible. En effet, la constitution d'une table de distribution de fréquence permet de résumer les données sous forme de fréquences. Le nombre de classes ainsi établis est fonction de type de variable.

* En cas de variable qualitative chaque catégorie ou modalité d'observa-

tions forme une classe. Le nombre de classe correspond au nombre de modalités de la variable (binaire ou nominale).

* En cas de variable quantitative discrète à faible nombre de modalité, on suit la même règle de la précédente.

* En cas de variable quantitative continue on utilise des règles mathématiques permettant la production d'un nombre adéquat de classes.s. Parmi elles :

  * Règle de Sturge : Nombre de classes= $1 + (3.322 * log(n)$

  * Règle de Yule : Nombre de classes= $2,5 * n^{1/4}$

  Pour ces deux cas, nous arrondirons le nombre de classes à l'entier le plus proche, le nombre de classes étant un entier.
  Pour déterminer les intervalles et les bornes de classe :

  * La borne inférieure d'une classe est la plus petite valeur admise dans la classe.

  * La borne supérieure d'une classe est au contraire la plus grande valeur admise dans la classe

  * L'intervalle de classe se calcule approximativement avec la formule suivante :

$$\frac{x_{max} - x_{min}}{nombre\ de\ classes}$$

Les différentes fréquences représentées dans la table de distribution des fréquences sont :

* Fréquence absolue, notée n_i

* Fréquence relative, notée f_i où $f_i = \frac{n_i}{N}$

* Fréquence absolue cumulée d'une classe, notée N_i où : $N_i = \sum_{j=1}^{i} f_j$

* Fréquence relative cumulée d'une classe, notée F_i où : $F_i = \sum_{j=1}^{i} f_j$

1.2.1.2 Représentation graphique

Les graphiques sont des schémas représentatifs des variables statistiques et varient selon le type de variable aléatoire :

* Variable qualitative
  * Diagramme en bâtons
  * Camembert
* Variable quantitative continue
  * Histogramme
  * Polygone de fréquences
  * Courbe de fréquences

1.2.2 Paramètres de position

Il existe plusieurs paramètres permettant de décrire une distribution, avec en premier lieu les paramètres de position, correspondant aux valeurs centrales autour desquelles se groupent les valeurs observées.

La moyenne

La moyenne est le centre de gravité d'une série statistique. Il existe plusieurs types de moyennes :

* **Moyenne arithmétique** : correspond à la somme des valeurs divisée par la taille de l'échantillon. Soit $y_1, y_2, y_3, \ldots, y_N$ la moyenne arithmétique est donnée par les formules suivantes :

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

* **Moyenne géométrique** : elle est utilisée dans le cas des données dont la distribution est déviée de la distribution normale.

$$\overline{Y}_G = (\prod_{i=1}^{N} y_i)^{1/N}$$

La médiane

C'est la valeur de la série statistique qui la divise en deux parties d'égal effectif. Le repérage se fait après un classement des valeurs de la série en ordre croissant. S'il ya un nombre impair d'observation, la médiane sera une observation de la série. Sinon, la médiane est située entre les deux observations centrales de la série statistique.

$$\begin{cases} Me = y_{\frac{N+1}{2}}, & \text{si } N \text{ est paire}\,; \\ Me = \frac{1}{2}(y_{\frac{N}{2}} + y_{\frac{N}{2}+1}), & \text{si } N \text{ est impaire.} \end{cases}$$

En Cas des données distribuées en classes dans un tableau de distribution de fréquences la médiane est donnée par la formule suivante après le repérage de la classe médiane :

$$Me = L_m + \frac{h}{n_m}(\frac{N+1}{2} - N_m)$$

où

- L_m : est la limite inférieure de la classe médiane.
- h : est l'étendu de la classe médiane.
- n_m : est la fréquence absolue (effectif) de la classe médiane.
- N : représente l'effectif global de l'échantillon.
- N_m : est la fréquence absolue cumulée jusqu'à la limite inférieure de la classe médiane.

Le mode

C'est la valeur de la série statistique (variable) qui présente la plus forte probabilité d'être observée. Dans une distribution continue, c'est la ''bosse" de la distribution.

Dans le cas des données distribuées en classes, le mode est calculé par la

formule mathématique suivante :

$$Mode = L + h\frac{D_i}{D_s + D_i}$$

Où :

— L : est la limite inférieure de la classe modale.

— h : est l'étendu de la classe modale.

— D_i : l'excédent d'effectif entre la classe modale et la classe précédente

— D_s : l'excédent d'effectif entre la classe modale et la classe suivant

1.2.3 Paramètres de dispersion

Les paramètres de dispersion nous renseignent sur l'étalement des valeurs observées de la série statistique et correspond donc au niveau de la variabilité de la série étudiée.

Etendu de variation

Est égale à la différence entre la valeur maximale et minimale de la série statistique $Etendu = Y_{max} - Y_{min}$

Variance

Permet d'estimer concrètement la variabilité d'une série statistique qui peut être d'ordre biologique, ou causée par la mauvaise qualité ou le faible nombre des mesures expérimentales. Mathématiquement la variance est la moyenne des écarts quadratique à la moyenne arithmétique. La variance peut être calculée dans la population ou estimée à partir d'un échantillon.

* Variance réelle de la population : $Var = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2$

* Variance estimée à partir d'un échantillon : $Var_{échantillon} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$ (où n est la taille de l'échantillon)

* Cas des données distribuées en classes : $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} n_i(y_i - \overline{y})^2$

Ecart type

Noté σ pour une population et S pour un échantillon présente la même unité que la moyenne (le plus utilisé dans la présentation des données), il correspond à la racine carré de la variance.

1.3 Statistique descriptives bi-varié

1.3.1 Coefficient de corrélation de Pearson

La Coefficient de corrélation de Pearson (r) est un paramètre qui mesure le degré de relation entre deux variables quantitatives. Le (r) varie entre (-1) et 1. La corrélation est dite positive si (r) et proche de 1 ; elle est négative si (r) est proche de (-1) ; elle est nulle si (r) est proche de (0).

$$\begin{aligned} r &= \frac{cov(X,Y)}{\sqrt{VAR(X)VAR(Y)}} \\ &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})^2\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{Y})^2}} \\ &= \frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{X})^2\sum_{i=1}^{n}(y_i-\overline{Y})^2}} \end{aligned}$$