Partie I

Introduction

Chapitre I : Définitions et généralités

1. Elément ou unité d'échantillonnage

2. La population statistique

3. L'échantillon

4. Les variables

Chapitre II : Statistique descriptive

Introduction

La biostatistique représente l'ensemble des méthodes mathématiques qui, à partir du recueil et de l'analyse de données réelles relatives à l'ensemble du vivant, permettent l'élaboration de modèles probabilistes autorisant les prévisions.

Les méthodes statistiques ne s'adaptent pas à n'importe quel mode de collecte de données et ne répondent pas à n'importe quelle question. La façon de recueillir les données a des implications sur le traitement statistique et vice – versa.

En outre, la cueillette et l'analyse des données sont toutes deux liées à un réseau complexe de décisions qu'il s'agit de prendre avant d'amorcer toute compagne de collecte de données. En effet, la logique et les contraintes des méthodes statistiques doivent s'articuler à toute la démarche scientifique du chercheur pour en arriver à un plan de recherche cohérent et à des résultats valides.

Chapitre I : Définitions et généralités

Avant d'amorcer toute étude statistique, il est important pour une meilleure compréhension de définir certains concepts fondamentaux.

1. Elément ou unité d'échantillonnage

L'élément ou l'unité d'échantillonnage est une entité concrète, comme un individu, un sujet, un objet, une colonie bactérienne etc., ou abstraite comme une association végétale. Définie clairement afin d'être identifiée sans difficultés. C'est l'élément de base sur lequel sont mesurés ou observés les descripteurs des populations. Par exemple, dans le cas de la mesure de la hauteur des arbres d'un peuplement forestier, l'arbre représente l'élément ou l'unité d'échantillonnage sur laquelle sont exécutées les mesures. Dans le cas d'un inventaire des types de bactéries dans un milieu donné, la petite partie choisi dans ce milieu sur laquelle sont dénombrés les taxons (analyse qualitative) ou les individus de chaque taxon (analyse quantitative) constitue l'unité d'échantillonnage.

2. La population statistique

Le but de la majorité des études statistiques est de formuler des lois valables pour un ensemble d'éléments appelé : population statistique, que l'on peut définir comme étant une collection d'éléments possédant au moins une caractéristique commune et exclusive, permettant de l'identifier et de la distinguer sans ambiguïté de toute autre, de laquelle on extrait un échantillon et sur laquelle porte les inférences, ou conclusions statistiques. Une inférence ou déduction statistique est une opération qui consiste à porter un jugement sur un ensemble vaste, la population statistique, à partir d'un sous–ensemble, l'échantillon.

3. L'échantillon

Pour des raisons techniques ou économiques, il n'est généralement pas possible de collecter des données sur tous les éléments de la population. En outre, si cette opération est possible, il est rarement utile de le faire car l'analyse d'un groupe restreint d'éléments, extrait de la population, fournit généralement des résultats d'une précision satisfaisante. Cette petite partie de la population que l'on va examiner s'appelle l'échantillon.

L'échantillon désigne un fragment d'un ensemble prélevé pour juger de cet ensemble. Dans la plupart des cas, il s'agit d'une collection d'éléments prélevés d'une façon particulière de la population statistique afin de tirer des conclusions sur cette dernière. Cependant, il arrive parfois que l'étude ou les observations s'étalent sur tous les éléments appartenant à la population, on parle alors d'échantillonnage exhaustif. Ce cas de figure se rencontre généralement au niveau de populations limités et de faible effectif. L'effort d'échantillonnage représenté par le rapport entre l'effectif de l'échantillon et celui de la population est très élevé dans le cas de populations restreintes et tend vers zéro pour les populations infinies.

4. Les variables

Une variable est une caractéristique mesurée ou observée sur chacun des éléments de l'échantillon. Cette caractéristique est sujette à des variations quantitatives ou qualitatives. Un caractère, ou une variable, est de nature qualitative s'il ne peut être mesuré tout en demeurant susceptible de classement, comme le sexe, la race, l'espèce etc. Une variable est de nature quantitative s'il peut être mesuré, comme la hauteur, la largeur le poids etc.

Les différents types de variables ou de descripteurs peuvent-être rassemblés dans le tableau suivant :

Types de Variables	**Exemple**
Binaires (2 descriptions possibles) ⟶	Présence-absence
Multiples (plusieurs descriptions)	

*Descripteurs qualitatifs ⟶ Groupes géologiques

* Descripteurs semi – quantitatifs ⟶ Echelle d'abondance

* Descripteurs quantitatifs ⟶ Longueur, poids, hauteur…

Chapitre II : Statistique descriptive

Ce chapitre sera consacré uniquement à la description des données recueillies lors de l'échantillonnage ou de l'expérimentation. On ne cherche pas à tirer des conclusions sur la population statistique, mais seulement à analyser et à présenter sous forme synthétique les caractéristiques des données accumulées.

Lors de la collecte des données, les valeurs observées se trouvent sans ordre. Si l'effectif de l'échantillon est faible, il n'y a aucun inconvénient à conserver les données les unes à la suite des autres. Mais, sitôt que l'effectif s'élève, il devient peu avantageux de les laisser telles quelles. On procède alors, au regroupement ou au classement des observations en un tableau de distribution de fréquences à partir duquel il devient possible de les représenter graphiquement afin de faciliter leur description.

Un tableau de distribution de fréquences est un mode synthétique de présentation des données numériques, montrant comment les résultats enregistrés sur une variable se distribuent dans ses différentes classes. Pour préparer ce tableau, il faut déterminer des classes et dénombrer le nombre d'éléments appartenant à chacune d'elles.

1. Présentation des données d'une série statistique simple

1.1. Variable quantitative

Dans le cas des variables quantitatives (ex : le nombre d'enfants engendré par chaque couple), une classe correspond à une valeur précise de la variable (ex : 2 enfants, 3 enfants, 4 enfants etc.). Toutefois, dans la majorité des cas, une classe se rapporte à plusieurs valeurs de la variable (ex : enfants dont le poids est supérieur ou égal à 15 kg, mais inférieur à 20 kg). La plus petite valeur admise dans la classe s'appelle alors *la borne inférieure* de la classe (ex : 15 kg) et la plus grande valeur, *la borne supérieure* (ex : 19.99 kg). La gamme des valeurs admissibles constitue *l'intervalle de la classe* (ex : 15 à 20 kg, soit 5 kg). La valeur centrale de la classe représente *l'indice de classe* (ex : 17.5 kg).

Le nombre de classes à considérer peut être recherché à partir de l'une des formules empiriques suivantes :

- *Règle de Sturge*

  Nombre de classes = 1 + (3.3 log n)

Où log n représente le logarithme à base 10 de l'effectif n de l'échantillon.

- *Règle de Yule*

Nombre de classes = $2.5\sqrt[4]{n}$

Où n représente l'effectif de l'échantillon.

Suivant les deux formules, le nombre de classes obtenues est arrondi à l'entier le plus proche.

En divisant l'étendue de la variation (écart entre la valeur la plus élevée est la plus faible de la variable) par le nombre de classes ainsi trouvé, on obtient l'intervalle de classe.

$$\text{Intervalle de classe} = \frac{\text{Valeur maximum} - \text{Valeur minimum}}{\text{Nombre de classes}}$$

Les classes étant déterminées, il reste à dénombrer les éléments appartenant à chacune d'elles. Le nombre d'éléments constituant une classe s'appelle l'effectif de la classe ou la fréquence absolue de la classe.

Application

La teneur en acide ascorbique des boissons douces, varie selon des paramètres divers (variété du fruit utilisé, conditions climatiques et culturales, stade de maturité à la récolte, temps d'entreposage du fruit avant traitement à l'usine etc.). Afin de déterminer la teneur en vitamine C de la boisson Ngaous, 18 bouteilles représentant un échantillon représentatif de la production journalière ont été analysées. Les résultats exprimés en mg/l sont les suivants :

Tableau 1 : Teneur en vitamine C de la boisson Ngaous

79	86	105	96	99	90	100	97	106
98	100	114	87	94	98	90	110	92

Nombre de classes :

Selon la formule de Sturge

N. cl. = 1 + (3.3 log 18) = 1 + 4.14 = 5.14 résultat que l'on arrondi à 5 classes.

Selon la formule de Yule

N. cl. = $2.5\sqrt[4]{n}$ = $2.5\sqrt[4]{18}$ = 2.5 * 2.06 = 5.15 résultat que l'on arrondi à 5 classes.

Sachant que la valeur maximum de la série est 114 et la valeur minimum 79, l'intervalle de classe est alors égal à : Int. cl. = $\frac{114-79}{5}$ = 7

En prenant comme borne inférieure de la première classe la valeur de 79 mg/l, les limites de classes successives sont les suivantes : 79 ; 86 ; 93 ; 100 ; 107 ; 114.

On obtient ainsi 5 classes : [79 – 86[; [86 – 93[; [93 – 100[; [100 – 107[; [107 – 114].

Les indices de classes correspondent aux valeurs centrales de chaque classe et sont successivement : 82.5 ; 89.5 ; 96.5 ; 103.5 ; 110.5

Enfin pour avoir la distribution de fréquences correspondante, il suffit de dénombrer les éléments appartenant à chaque classe. Après comptage, on construit le tableau de distribution de fréquences en portant dans une colonne (x) les indices de classes et dans une deuxième (f) la fréquence.

Tableau 2 : Tableau de distribution de fréquences

x	f
82.5	1
89.5	5
96.5	6
103.5	4
110.5	2

$\sum fi$ = n (effectif de l'échantillon)

1.3. Les fréquences

La fréquence absolue d'une classe (f_i) est simplement l'effectif ou le nombre d'éléments appartenant à la classe. La fréquence relative d'une classe ($f_{rel.}$) est le rapport de son effectif à l'effectif total de l'échantillon ($f_{rel.} = f_i /n$). La fréquence cumulée d'une classe ($f_{cum.}$) correspond à l'effectif total des valeurs plus petites que la borne supérieure d'une classe ; c'est donc la somme des fréquences d'une classe et de toutes celles qui la précédent.

Tableau 4 : Distribution des teneurs en acide ascorbique de la boisson Ngaous

x	fi	$f_{rel.}$	$f_{cum.}$
82.5	1	0.06	1
89.5	5	0.28	6
96.5	6	0.33	12
103.5	4	0.22	16
110.5	2	0.11	18

2. Description des séries statistiques (la réduction des données)

La réduction des données a pour objet le calcul de paramètres permettant de caractériser les séries statistiques et les distributions de fréquences.

Les paramètres les plus couramment utilisés sont les suivants :

2.1. Paramètres de position : Ils servent à caractériser l'ordre de grandeur des observations.

2.1.1. Moyenne arithmétique : La moyenne arithmétique que nous appellerons tout simplement moyenne et que nous désignerons par le symbole $\bar{x}$ (x barre) ou simplement par la lettre m, est égale à la somme des valeurs observées x_1, x_2, ……, x_i, ……, x_n, divisée par le nombre d'observations :

$$\bar{x} = \frac{1}{n} (x_1 + x_2 + \ldots + x_n) = \frac{1}{n} \sum_{i=1}^{n} xi$$

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

$$\overline{X} = \frac{12 + 3 + 24 + 1 + 5 + 8 + 7}{7} = 8.57$$

2.1.2. Médiane : C'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).

+ Lorsque le nombre d'observations **n** est **impair**, la médiane correspond à l'observation de rang (n + 1)/2.

$$\boldsymbol{Me = Valeur\ (n + 1)/2}$$

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

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

Me = Valeur (7 + 1)/2 = 4^{éme} valeur donc 7 est la médiane de la série

+ Lorsque le nombre d'observations **n** est **pair**, tout nombre compris entre la valeur (n/2) et la valeur (n/2) + 1 répond à la définition et on convient de prendre comme valeur de la médiane la moyenne arithmétique de ces deux observations.

$$\boldsymbol{Me = \frac{[Valeur\ (n/2)\ + Valeur\ (n/2) + 1]}{2}}$$

Exemple : soit la série 12 ; 3 ; 24 ; 1 ; 5 ; 8 ; 7 ; 30

On l'ordonne : 1 ; 3 ; 5 ; **7** ; 8 ; 12 ; 24 ; 30

$$Me = \frac{[\text{Valeur } (n/2) + \text{Valeur } (n/2)+1]}{2} = \frac{[\text{Valeur } (8/2) + \text{Valeur } (8/2)+1]}{2} =$$

$$\frac{[4 \text{ ème valeur} + 5 \text{ème valeur}]}{2} = \frac{7+8}{2} = 7.5$$

2.2. Paramètres de dispersion : ils permettent de chiffrer la variabilité ou la dispersion des valeurs observées autour d'un paramètre de position.

2.2.1. Variance : La variance d'une série statistique est la moyenne arithmétique des carrés des écarts par rapport à la moyenne.

$$\boldsymbol{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (Xi - \overline{X})^2 = \frac{SCE}{n}$$

$\boldsymbol{\sigma^2}$ représentent la variance calculée ou variance de l'échantillon, nous verrons plus loin, qu'il est possible d'avoir la variance de la population d'où l'échantillon a été extrait en faisant une estimation. Cette variance notée $\mathbf{S^2}$ représente la variance de la population ou variance estimée. Elle est donnée par la formula suivante :

$$\mathbf{S^2} = \frac{1}{n-1} \sum_{i=1}^{n} (Xi - \overline{X})^2 = \frac{SCE}{n-1}$$

La valeur de $\boldsymbol{\sigma}^2$ s'accroît au fur et à mesure que la variabilité ou la dispersion augmente.

2.2.2. Ecart – type : ou déviation standard est la racine carrée de la variance.

$$\boldsymbol{\sigma} = \sqrt{\frac{1}{n}\ \sum_{i=1}^{n}(\text{xi} - \overline{\text{X}})2} = \sqrt{\frac{SCE}{n}}$$

σ représentent l'écart type calculé ou écart type de l'échantillon, nous verrons plus loin, qu'il est possible d'avoir l'écart type de la population d'où l'échantillon a été extrait en faisant une estimation. Cet écart type noté **s** représente l'écart type de la population ou écart type estimé. Il est donné par la formula suivante :

$$\mathbf{S} = \sqrt{\frac{1}{n-1}\ \sum_{i=1}^{n}(\text{xi} - \overline{\text{X}})2} = \sqrt{\frac{SCE}{n-1}}$$

2.2.3. Coefficient de variation : ou encore coefficient de variabilité est obtenu en exprimant l'écart – type en valeur relative ou en pourcent de la moyenne.

$$\mathbf{CV} = \frac{\boldsymbol{\sigma}}{\overline{\text{X}}} \text{ ou } 100\ \frac{\boldsymbol{\sigma}}{\overline{\text{X}}} \text{ (en \%)}$$

Exemple : Reprenons les données de la teneur en acide ascorbique de la boisson N'gaous :

$$\overline{\text{X}} = \frac{1}{n}\sum_{i=1}^{18} xi = \frac{1}{18}(79 + 88 + \ldots\ldots\ldots + 114) = 97.38$$

$$\boldsymbol{\sigma^2} = \frac{1}{n}\ \sum_{i=1}^{18}(\text{xi} - \overline{\text{X}})^2 = \frac{(79-97.38)2+ \ldots..+(114-97.38)2}{18} = \frac{1628.27}{18} = 90.46$$

$$\boldsymbol{\sigma} = 9.51$$

$$\mathbf{S^2} = \frac{1628.27}{17} = 95.78$$

$$\mathbf{S} = 9.78$$

$$\mathbf{CV} = \frac{\boldsymbol{\sigma}}{\overline{\text{X}}} = \frac{9.51}{97.38} = 0.09 \text{ ou } 9\%$$