Faculté des sciences exactes et sciences de la nature et de la vie

Département : Sciences exactes et sciences de la nature et de la vie de

Module : Biostatistiques (2019-2020)

Chapitre 1 : Statistique descriptive à une variables

1. Introduction :

La statistique est l'ensemble des méthodes utilisées pour obtenir des renseignements sur une population à partir d'observations obtenues sur un échantillon supposé tiré au hasard de cette population.

2. Terminologie employée :

a) Population : c'est un ensemble dont les éléments s'appellent les individus (ou unités statistiques) et doivent être bien caractérisés du point de vue pratique.
Ex : ensemble des individus d'une espèce animale, végétale.

b) Echantillon :

C'est un sous ensemble connu et disponible d'individus d'une population, il est obtenu par un procédé faisant intervenir le hasard, en ayant recours au tirage au sort et en prenant des précautions pour que l'échantillon soit le plus représentatif possible de la population.

c) La variable :

La variable se présente avec une grandeur ou une intensité qui peut différer d'un individu à l'autre. Cette intensité est évaluée au moyen d'un processus de mesure.

Une variable est désignée par une lettre de l'alphabet comme : X, Y, Z, etc…Si on fait plusieurs observations sur une variable (cas d'un échantillon), on peut affecter la lettre associée à la variable d'un indice qui précise l'ordre de l'ordre de l'observation dans le paquet des valeurs obtenues. (Xi signifie valeur de X chez l'individu numéro i).

On peut classer les variables en deux groupes :

- **Les variables quantitatives (variables numériques):** Sont des variables mesurables : poids, taille, âge. Elles sont souvent accompagnées d'une unité de mesure (ex : poids = kg).

On distingue 2 sous – catégories :

- **Variables continues :** qui peuvent prendre un nombre infini de valeur dans un intervalle donné (taille, pression artérielle diastolique).

*** Variables discrètes :** ne peuvent prendre qu'un nombre fini de valeur : ex : nombre d'enfants d'une famille.

- **Les variables qualitatives ou catégorielles :**

Ce sont des variables non mesurables. Elles ont un certain nombre de catégories ou modalités.
Une variable catégorielle à 2 catégories est dite dichotomique ou (binaire).
Ex la variable fumeurs (fumeurs-non fumeurs) est une variable catégorielle à deux catégories.
En présence de plusieurs catégories, on distingue :

*** Les variables ordinales :** elles peuvent bénéficier d'un classement ordonné ou d'un ordre naturel.
Ex : l'intensité de douleur : nulle, légère, intense, insupportable.

*** Les variables nominales** : Il n'existe pas d'ordre naturel. Chaque classe désigne une catégorie (elle la nomme). Par exemple, pour la couleur des yeux : noir / marron / vert /bleu.

3. Statistique descriptive à une variable :

La statistique descriptive est la phase de la statistique qui se limite à décrire ou analyser une population donnée, sans tirer de conclusion pour une population plus grande. Elle englobe l'étude et l'élaboration de données au cours d'une série particulière (un échantillon) d'expériences ou d'observations. L'objectif de l'étude est d'extrapoler les résultats de la caractérisation de cet ensemble bien particulier d'observations, à la population toute entière.

4. Paramètres statistiques d'une série statistique :

Une série statistique peut se caractériser par 2 grands types de paramètres:

*** Paramètres de position :** ils donnent l'ordre de grandeur des observations et sont liés à la tendance centrale de la distribution.

*** Paramètres de dispersion :** ils montrent la manière dont les observations fluctuent autour de la tendance centrale.

4.1. Les paramètres de localisation (position) :

a) La médiane :

C'est la valeur de la variable telle qu'une moitié des valeurs lui soit supérieure ou égale et l'autre moitié des valeurs lui soit inférieure ou égale. Deux cas apparaissent suivant la parité de n.

n pair : Me = $\frac{x\left(\frac{n}{2}\right)+x\left(\frac{n}{2}+1\right)}{2}$

n impair : Me = x ($\frac{n+1}{2}$ *)*

Ex : considérons la série suivante : 2,50; 2,75; 3,20; 2,18; 1,85; 5,40; 3,65; 4.25; 5,65; 6,30.
Le tri par ordre croissant donne :1,85; 2,18; 2,50; 2,75; 3,20; 3,65; 4,25; 5,25; 5,40; 5,65; 6,30.
Le nombre d'observations est n = 10 (pair) donc:

La médiane est *Me* = $\frac{x\left(\frac{n}{2}\right)+x\left(\frac{n}{2}+1\right)}{2} = \frac{x(5)+x(6)}{2} = \frac{3.20+3.65}{2}$ = 3.425

b) La moyenne:

C'est la valeur de la variable qui occupe la position centrale parmi les valeurs de la série.

Elle est égale à $\bar{X} = \frac{\sum_{x=1}^{n} x_i}{n}$ où n : l'effectif total.

c) Le Mode (ou valeur dominante) :

C'est la valeur de la variable la plus souvent rencontrée. Dans la distribution d'une variable, le mode peut ne pas exister ou ne pas être unique.

X = (1, 2, 5, 2, 4, 2, 5) a pour mode 2

X = (1, 3, 5, 2, 4, 7) pas de mode

X = (2, 7, 5, 2, 5, 8, 9) a pour mode 2 et 5. On parle de distribution bimodale.

Sur un plan graphique, le mode est la valeur de x sur l'axe des abscisses dont l'ordonnée est la plus grande.

4.2. Paramètres (caractéristiques) de dispersion :

Ce sont des indices qui expriment le degré de dispersion ou de fluctuation des observations autour de la moyenne (valeur centrale) de la variable. Ce sont respectivement :

a) L'écart-type (erreur type, erreur standard, écart standard) :

Le plus utilisé des paramètres de dispersion, il est égal à la racine positive de V(X) soit : $\sigma=\sqrt{V(X)}$

La variance :

La variance est égale à la somme des carrés des écarts à la moyenne divisée par l'effectif total. On la note : V(X) et elle vaut :

$$V(X)=\frac{1}{n}\left(\sum_{i=}^{n}(x_i-\overline{X})^2\right)=\frac{1}{n}\left(\sum_{i=1}^{n}x_i^2\right)-(\overline{X})^2$$

La variance a l'unité de la variable au carré: si x est par exemple une longueur exprimée en cm, la variance est exprimée en cm^2

b) Les quartiles :

Sont des valeurs Q1, Q2 et Q3 de la grandeur mesurée qui partagent la série statistique en 4 parties de même effectif n/4 (ou 25%).

Le premier quartile (Q1) : qui est supérieur à 25% des premières (plus petites) valeurs de la variable, Q1 est donc inférieur à 75% des valeurs x.

Le deuxième quartile (Q2) est supérieur à la première moitié 50% des valeurs x, donc Q2 = Me.

Le troisième quartile (Q3) est supérieur à 75% des premières (plus petites) valeurs de la variable (Q3) est donc inférieur à 25% des valeurs X.

2) **série statistique simple <u>ordonnée</u>** (selon la valeur de n)

Les différents cas :

a) cas n = 4 p (multiple de 4, ex : n =24) $\Rightarrow\begin{cases} Q_1=\frac{X(p)+X(p+1)}{2} \\ Q_2=Me=\frac{X(2p)+X(2p+1)}{2} \\ Q_3=\frac{X(3p)+X(3p+1)}{2} \end{cases}$

b) cas n = 4p+1 (ex : n = 33) $\Rightarrow\begin{cases} Q_1=\frac{X(p)+X(p+1)}{2} \\ Q_2=Me=X(2p+1) \\ Q_3=\frac{X(3p+1)+X(3p+2)}{2} \end{cases}$

c) cas n = 4p+2 (ex : n = 18) $\Rightarrow\begin{cases} Q_1=X(p+1) \\ Q_2=Me=\frac{X(2p+1)+X(2p+2)}{2} \\ Q_3=X(3p+2) \end{cases}$

d) cas n = 4p +3 (ex : n =27) $\Rightarrow \begin{cases} Q_1 = X(p+1) \\ Q_2 = Me = X(2p+2) \\ Q_3 = X(3p+3) \end{cases}$

c) **L'intervalle interquartile :** c'est la longueur de l'intervalle (Q1, Q3) qui renferme 50 % des observations. On a : *IQ = Q3 – Q1*.

d) **L'étendue :** elle indique l'amplitude de variation de la variable X et c'est la distance ou longueur de la différence entre la plus grande valeur et la plus petite valeur de la variable x. on la note : *e = Xmax – Xmin*.

e) **Le coefficient de variation :** c'est une valeur (sans unité) qui exprime le rapport (donné en %) entre la valeur de tendance centrale (moyenne) et la valeur de la dispersion (écart type). Les fortes valeurs de ce coefficient indiquent une grande hétérogénéité de l'ensemble étudié par rapport à la variable. La formule pour ce coefficient est : $C.V = \frac{\sigma x}{\bar{X}} \times 100$ (%).

5. Cas d'une variable distribuée en classes :

Lorsque les observations de la variable étudiée sont réparties en classes dans un tableau de distribution, les caractéristiques présentées dans le paragraphe précédent, conservent la même signification mais se calculent autrement.

5.1. Caractéristiques de tendance :

a) **La moyenne pondérée** : $\overline{X} = (\sum_{i=1}^{k} n_i \times c_i)/n = \sum_{i=1}^{k} f_i \times c_i$ **(** c_i sont les centres de classes).

b) **Le mode :** c'est la valeur de la variable qui correspond à la plus grande fréquence dans la distribution. La classe à laquelle appartient le mode s'appelle : classe modale.

c) **La médiane :** c'est la valeur qui correspond à la fréquence relative cumulée F = 0.50 ou à l'effectif cumulé n/2, c'est à dire que 50 % des valeurs de X sont inférieures ou égales à Me, et les 50 % restantes sont supérieures à Me.

5.2. Caractéristiques (paramètres) de dispersion :

Ce sont des indices qui expriment le degré de dispersion ou de fluctuation des observations autour de la moyenne (valeur centrale) de la variable. Ce sont respectivement :

a) **La variance et l'écart type :**

On appelle variance pondérée d'une variable x distribuée en classes, la moyenne des carrés des différences des centres de classes Xi, avec la moyenne. On la note : V(X) et elle vaut :

$$V(X)=\frac{1}{n}\left(\sum_{i=}^{k} n_i \times (c_i - \overline{X})^2\right)=\frac{1}{n}\left(\sum_{i=1}^{k} n_i \times c_i^2\right)-(\overline{X})^2=\sum_{i=}^{k} f_i \times (c_i - \overline{X})^2=\sum_{i=1}^{n} f_i \times c_i^2-(\overline{X})^2$$

b) L'écart type (erreur type, erreur standard, écart standard) : il est égal à la racine positive de V(X) soit : $\sigma=\sqrt{V(X)}$

c) Les quartiles : Les trois valeurs Q1, Q2, Q3 sont reliées à la fonction cumulative par les relations : F(Q1) = 0.25 ; F(Q2) = 0.50 ; F(Q3) = 0.75

d) Intervalle interquartile : c'est la valeur *IQ = Q3 – Q1* et c'est la longueur de l'intervalle (Q1,Q3) qui renferme 50 % des observations.

e) L'étendue : c'est la distance ou longueur de la différence entre la plus grande valeur et la plus petite valeur de la variable x. on la note :*e = Xmax – Xmin* et elle mesure la longueur de l'intervalle de variation de X.

6. Représentation des données :

C'est la méthode qui englobe la présentation des résultats sous forme de schémas illustratifs. Elle doit être la plus claire et la plus concise possible

Les représentations graphiques sont très importantes en statistique descriptive. Elles ont l'avantage de renseigner immédiatement sur l'allure générale de la distribution. Elles facilitent l'interprétation des données recueillies.

La représentation des données se fait principalement sous deux formes :

*** Des tableaux :** qui représentent la liste des résultats des observations éventuellement regroupés et présentées de façon convenable,

*** Des graphiques :** qui peuvent être de forme très diverse selon la nature et le nombre de variables prises en considération.

6.1. Les variables quantitatives discrètes :

*** Tableaux de distribution de fréquences :**

L'effectif est le nombre de fois où la valeur xi a été observée.

La fréquence ou effectif relatif de la valeur d'un caractère quantitatif est le rapport entre l'effectif de cette valeur et l'effectif total de l'ensemble des valeurs. En général, elle est exprimée en %.

On note v1, : : : , vk les k valeurs différentes que peut prendre la variable.

Pour $1 \leq j \leq n$, on note nj l'effectif des individus pour lesquels la variable prend la valeur vj .

On note fj la fréquence relative ou proportion pour la valeur vj et φj = f1 + ··· + fj la j-ème fréquence relative cumulée.

Valeurs prises par la variable	V1	V2	…	vk	Total
Effectif ni	n_1	n_2	…	n_k	n
Fréquence relative *fi*	n_1/n	n_2/n	…	n_k/n	1
Fréquence relative cumulée Fi	f1	f1+f2	…	f1+f2+…+fk	Pads de sens

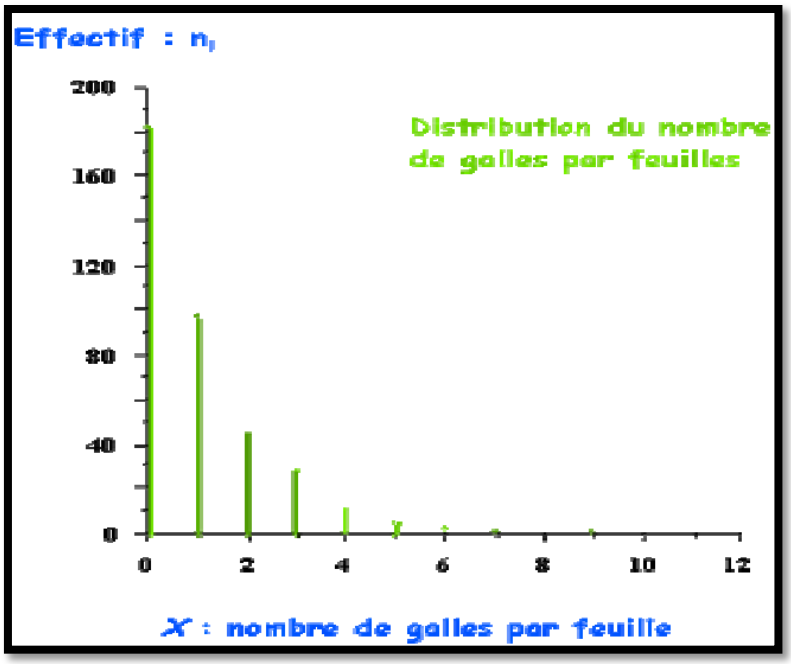
Exemple : Tableau de distribution des valeurs de la glycémie :

Xi (g/l)	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
Fréquence absolue	8	12	24	26	32	32	28	26
Fréquence relative	0,04255	0,0638	0,1276	0,1382	0,1702	0,1702	0,1489	0,1382
Fréquence relative cumulée	0,04255	0,1063	0,2340	0,3723	0,5425	0,7127	0,8617	1

*** Représentation graphique :**

*** Diagramme en bâtons :** est un graphique où l'axe des abscisses représente les différentes valeurs prises par la variable, placées en respectant une échelle, et l'ordonnée représente les fréquences relatives ou les fréquences absolues.

Exemple : l'exemple de la **cécidomyie** du hêtre, la distribution des fréquences observées du nombre de galles par feuille peut être représentée par **un diagramme en bâtons** avec en ordonnée les **effectifs *ni*** et en abscisse les différentes **modalités** de la variable étudiée.



6.2. Les variables quantitatives continues :

*** Tableau des effectifs:**

C'est un tableau qui est construit par regroupement en classes du nombre total n des valeurs de la variable.

Dans ce tableau on définit des intervalles semi-ouverts ou semi-fermés disjoints (sans intersection commune ou éléments communs) de la forme [ai,ai+1[sur le domaine de variation de X.

On note ni **l'effectif** ou **fréquence absolue** ou nombre de valeurs de la variable X appartiennent à cet intervalle. Cette répartition en classes fournit un tableau de distribution de la forme :

Classe	] a_0, a_1]	] a_1, a_2]	….	] a_i, a_{i+1}]	….	] a_{k-1}, a_k]	Total
Effectif	n1	n2	….	ni	….	nk	n

Remarquons que : n = n1 + n2 + …. + nk = $\sum_{i=1}^{k} n_i$

*** Tableaux de distribution de fréquences - fréquences cumulées :**

– L'intervalle où la variable prend ses valeurs est divisé en k classes : [b0; b1[, [b1; b2[, . . ., [bk-1; bk[(il est possible d'avoir des bornes infinies).

– Pour 1≤ j≤ k, on note nj l'effectif associé à la classe [bj-1; bj [, fj = nj/n la fréquence relative associée à cette classe et φj = f1 + _ _ _ + fj la j-ième fréquence cumulée

– On note aj = bj – bj-1 l'amplitude de la classe [bj-1; bj [.

– On note dj = fj/aj la densité de proportion pour la classe [bj-1; bj [.

Variable X	[b0; b1[	[b1; b2[	. . .	[bk-1; bk[	Total
Fréq. absolue	n1	n2	. . .		**n**
Fréq. relative	f1 = n1/n	f2 = n2/n	. . .	fk = nk/n	**1**
Fréq. relative cumulée	Φ1 = f1	Φ2 = f2	. . .	Φk = fk	
Amplitude	a1 = b1 - b0	a2 = b2 - b1	. . .	ak = bk – bk- 1	
Densité de proportion	d1 = f1/ a1	d2 = f2/a2	. . .	dk = fk/ak	

Remarques :

– la densité de proportion permet de comparer les effectifs dans chaque classe en tenant compte de la taille de ces classes (cf. la notion de densité de population en géographie).

– Dans le cas de classes qui ont toutes la même longueur, il n'est pas nécessaire de calculer la densité de proportion, il est suffisant d'étudier les fréquences relatives ou absolues (qui sont directement proportionnelles à la densité de proportion).

Exemple : on s'intéresse à la taille, notée T et exprimée en mètres, de 20 individus. On a obtenu la série statistique suivante :

1,72 ; 1,87 ; 1,66 ; 1,73 ; 1,64 ; 1,77 ; 1,80 ; 1,81 ; 1,60 ; 1,78 ; 1,83 ; 1,75 ; 1,70 ; 1,58 ; 1, 68 ; 1,66 ; 1,93 ; 1,75 ; 1,80 ; 1,85.

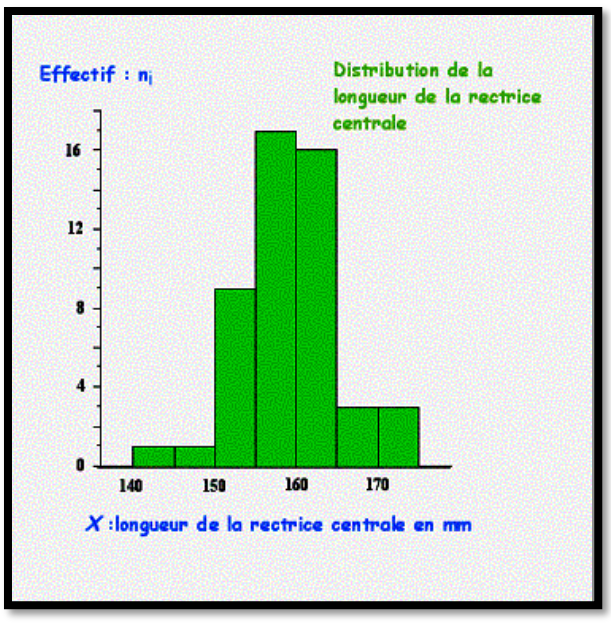
Taille	[1,50 ; 1,65[	[1,65 ; 1,75[	[1,75 ; 1,85[	[1,85 ; 2,00[
Fréq. Absolue	3	6	8	3
Fréq. relative	0,15	0,3	0,4	0,15
Fréq. relative cumulée	0.15	0.45	0.85	1
Amplitude	0.15	0.10	0.10	0.15
Densité de proportion	1	3	4	1

*** L'histogramme :**

Un histogramme est une représentation graphique dans laquelle les rectangles représentés ont des largeurs proportionnelles aux amplitudes des classes et des aires proportionnelles aux fréquences de ces classes. L'histogramme permet de visualiser rapidement des données.

Exemple :

Dans l'exemple de la longueur de la rectrice centrale des individus mâles de la gélinotte huppée, la distribution des fréquences observées est représentée par un **histogramme** avec en ordonnée les **effectifs** n_i et en abscisse **les limites de classe** de la variable étudiée.



*** Le polygone des fréquences:**

Il s'obtient en joignant les milieux des cotés supérieures de chaque rectangle. La surface ainsi obtenue est identique à celle de l'histogramme. La représentation graphique d'une distribution des effectifs cumulés ou des fréquences cumulées est illustrée par un graphique appelé: **courbe cumulative (diagramme intégral)** dans le cas où la variable est discrète) dans le cas où la variable étudiée est continue et les classes sont représentées par des intervalles.

6.3. Les variables qualitatives :

*** Tableaux de fréquence :**

Exemple : On s'intéresse à la variable "couleur des yeux" sur un groupe de 20 personnes.

On code chaque modalité de la manière suivante : M=marron, V=vert, N=noir, B=bleu.

On obtient la série statistique suivante : M, V, M, M, M, N, M, B, M, B.

Couleur des yeux	M	V	N	B	Total
Effectif	7	3	8	2	20
Proportion	35%	15%	40%	10%	100%

*** Représentation graphique :**

*** Le diagramme à barres :**

Permet de donner la fréquence (ou le nombre) de chaque catégorie, un espace est laissé entre chaque barre.

*** Diagrammes en secteurs :**

Dans le diagramme en secteurs chaque modalité est représentée par un secteur d'un disque dont l'angle est proportionnel à la fréquence de la modalité (ou au pourcentage), l'angle 360 degrés équivalant à la fréquence relative 1 (ou au pourcentage 100%).

6.4. Les graphiques explorateurs:

Les données générales dans certaines expériences peuvent être déformées, et peuvent même contaminées par des valeurs aberrantes. Les graphiques explorateurs permettent d'étudier graphiquement la distribution d'un ensemble de mesures.

*** Les nuages de points:**

Un graphique de dispersion ou nuage de points est un graphique qui met en relation les valeurs de deux variables sur un repère de coordonnées cartésiennes. Dans le cas d'un graphe avec un nuage de points, l'abscisse n'est pas divisée en catégories, mais elle représente une mesure continue comme la taille le poids ou toute autre variable. La paire (x,y) de chaque sujet est tracée comme une donnée à deux variables sur deux échelles de mesures continues. Ce type de graphe est habituellement utilisé pour réaliser une corrélation.

*** La boîte à moustaches de Tukey (Box Plot) ou box-and-whisker plot**

La boîte à moustaches est un outil explorateur qui utilise les percentiles d'un ensemble de mesures pour décrire la forme et l'intervalle de la distribution. Ce graphique est en particulier utile pour comparer les distributions de différents groupes (traitement et témoin).

La bordure inférieure de la boîte représente le premier quartile (Q_1) et la bordure supérieure représente le troisième quartile (Q_3). La portion du diagramme comprise dans la boîte représente donc l'étendue interquartile ou la moitié centrale (50 %) des observations.

La ligne horizontale qui traverse la boîte représente la médiane des données.

Les lignes qui sortent de la boîte sont appelées moustaches. Les moustaches s'étendent vers l'extérieur pour indiquer à, leurs extrémités la valeur la plus basse (Q_1 - 1.5IQ) et la valeur la plus haute (Q_3 + 1.5IQ) dans la série (à l'exception des valeurs aberrantes).

La boîte à moustaches permet aussi d'évaluer la symétrie des données : lorsque les données sont symétriques, la ligne médiane se situe à peu près au milieu de la boîte interquartile et les moustaches sont de la même longueur.

Si les données sont asymétriques, il se peut que la médiane ne tombe pas au milieu de la boîte interquartile et une moustache peut être nettement plus longue que l'autre.

Les boîtes à moustaches peuvent également aider à repérer les valeurs aberrantes' douteuses'' ou ''suspectes'' (en anglais : ''outliers'') (une valeur est considérée comme aberrante si la valeur absolue de l'écart avec Q_1 ou Q_3 est supérieure à plus de 1,5 fois l'étendue interquartile.