

Chapitre 1 : Statistique descriptive univariée

1.1 Définitions

Variable statistique : Propriété qui varie d'un individu à l'autre.

Remarque : Les variables statistiques sont généralement notées par des lettres majuscules comme X , Y , Z , etc.

Individus : Éléments dont on étudie les propriétés : on étudie la valeur que prend, pour chaque individu la variable statistique.

Remarque : Les *individus* ne sont pas nécessairement des personnes. Par exemple dans l'exercice 10, ce sont des pays.

Population : Ensemble des individus que l'on considère.

Échantillon : Sous-ensemble d'individus, choisis au sein de la population.

Proportion d'individus qui ont une certaine propriété nombre entre 0 et 1, qui s'obtient en divisant le nombre d'individus qui ont cette propriété par le nombre total d'individus considérés.

Remarque : Il arrive de multiplier par cent ce nombre en 0 et 1, afin de l'exprimer en pourcentage (entre 0% et 100%).

Modalités : Valeurs que peut prendre la variable.

Exemple : Les modalités de la variable "*nombre de frères et soeurs*" sont 0, 1, 2, 3, ...

Variable quantitative : Variable dont les modalités sont des nombres (éventuellement munis d'une unité), pour lesquels l'addition a un sens

Exemples : La taille est une variable quantitative

Le numéro de téléphone n'en est pas une.

Variable quantitative discrète : Variable quantitative dont les modalités sont séparées par de nombreuses valeurs "interdites"

Exemple : Le nombre de frère et soeur peut valoir 1 ou 2, mais pas 1,5 ni 1,0356.

Variable quantitative continue : Variable dont les modalités ne sont séparées par aucune valeur interdite (elles forment un intervalle).

Exemples : La taille

Variable qualitative : Variable qui n'est pas quantitative

Variable qualitative ordinale : Variable qualitative dont les modalités sont ordonnées de manière claire et consensuelle.

Exemple : Mauvaise/Bonne/Très bonne humeur

Variable qualitative nominale : Variable qualitative qui n'est pas ordinale.

Exemple : Couleur des yeux

1.2 Regroupement des données

Données de l'exercice 24

A) Regroupement par modalités

effectif d'une modalité Nombre d'individus chez qui la variable prend cette valeur.

On écrit dans un tableau les effectifs de chaque modalité, comme ci-dessous :

	de mauvaise humeur	de relativement bonne humeur	de bonne humeur	de très bonne humeur
Modalité				
Effectif	2	1	1	7

Nb de frères/soeurs	0	1	2	3	4
Effectif	2	4	2	1	2

Notation :

- ▶ x_1 : première modalité
- ▶ x_2 : deuxième modalité
- ▶ etc
- ▶ r : nombre de modalités
- ▶ n_1 : Effectif de x_1
- ▶ n_2 : Effectif de x_2
- ▶ etc

Effectif total (taille de l'échantillon) : $n = \sum_{i=1}^r n_i$

B) Regroupement en classes

Classe Intervalle (généralement de la forme $[a; b[$, qui contient le minimum a mais pas le "maximum" b)

Effectif d'une classe Nombre d'individus chez qui la variable est dans cet intervalle

Classe	[1,60;1,65[[1,65;1,70[[1,70;1,75[[1,75;1,80[[1,80;1,85[[1,85;1,90[[1,90;1,95[
Effectif	2	3	2	1	2	0	1

Notation :

▶ a_1 : minimum de la première classe

▶ a_2 : minimum de la deuxième classe (et "maximum" de la première)

▶ etc

▶ n_1 : Effectif de la première classe

▶ n_2 : Effectif de la deuxième classe

▶ etc

C) Fréquences et fréquences cumulées

Fréquence d'une modalité (ou d'une classe) Proportion d'individus correspondant à cette modalité (ou cette classe)

$$f_i = \frac{n_i}{n}$$

Notation : $\mathbb{P}_r[X = \dots]$

Fréquence cumulée d'une modalité (ou d'une classe)

Sommes des fréquences de toutes les modalités (ou classes) jusqu'à celle là

Notation : $\mathbb{P}_r[X \leq \dots]$

Nb de frères/soeurs	0	1	2	3	4
Effectif	2	4	2	1	2
Fréquence	0,182	0,364	0,182	0,091	0,182
Fréquence cumulées	0,182	0,546	0,728	0,819	1,001

1.3 Représentations graphiques

A) Représentation des fréquences

- ▶ « Camembert » : Quartiers de surfaces proportionnelles aux fréquences.
Typiquement pour des variables qualitatives.
- ▶ Diagramme « en bâtons » (ou « en tuyaux d'orgue ») :
 - ▶ fréquences : hauteurs des bâtons
 - ▶ modalités : positions horizontale des bâtonsTypiquement pour des variables quantitatives discrètes.
- ▶ Histogramme (pour des données regroupées en classes ; typiquement quantitatives continues)
 - ▶ classes : position horizontale des rectangles
 - ▶ fréquences : surface des rectangles

B) Représentation des fréquences cumulées

i) Fonction de répartition

Notation :

$$\underbrace{\mathbb{P}_r[X \leq 2]}_{F_X(2)} \simeq 0,728$$

$$\underbrace{\mathbb{P}_r[Y \leq 1,85]}_{F_Y(1,85)} \simeq 0,91$$

F_X s'appelle la « fonction de répartition » de X

ii) Polygone des fréquences cumulées

Polygone des fréquences cumulées Représentation graphique (approchée) de la fonction de répartition

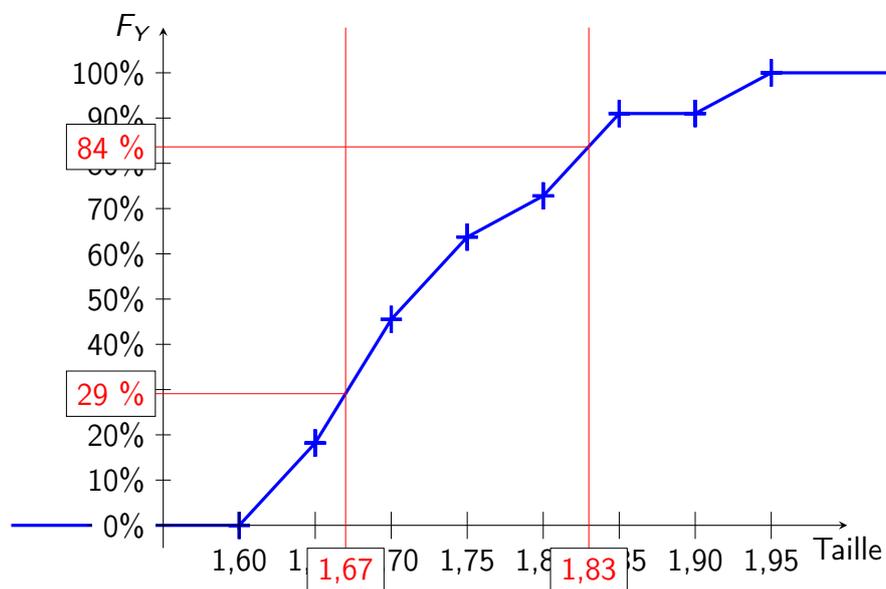
Exemple de lecture graphique : Calcul de

$$\mathbb{P}_r[1,67 < Y \leq 1,83]$$

$$= \mathbb{P}_r[Y \leq 1,83] - \mathbb{P}_r[Y \leq 1,67]$$

$$\simeq 0,84 - 0,29$$

$$\simeq 0,55$$



1.4 Calcul d'indicateurs

A) Médiane d'une variable X , quantitative ou "qualitative ordinale"

Médiane modalité notée Med telle que

$$\mathbb{P}_r[X \geq \text{Med}] \geq 0,5 \text{ et que } \mathbb{P}_r[X \leq \text{Med}] \geq 0,5.$$

Concrètement : Une fois les valeurs classées par ordre croissant, la médiane est la valeur du milieu plus précisément (page 3 du formulaire) : c'est la

$\left(\frac{n+1}{2}\right)^{\text{ième}}$ valeur. Si $\frac{n+1}{2}$ n'est pas entier, on prend le milieu entre la $\left(\frac{n}{2}\right)^{\text{ième}}$ et la $\left(\frac{n}{2} + 1\right)^{\text{ième}}$.

Exemples : ● Nombre de frères/soeurs :

0 0 1 1 1 1 2 2 3 4 4

La médiane est 1

● humeur :

de très bonne humeur

de très bonne humeur ← Médiane

de très bonne humeur

de bonne humeur

de relativement bonne humeur

de mauvaise humeur

de mauvaise humeur

La médiane est donc "de très bonne humeur"

● taille des hommes de l'échantillon:

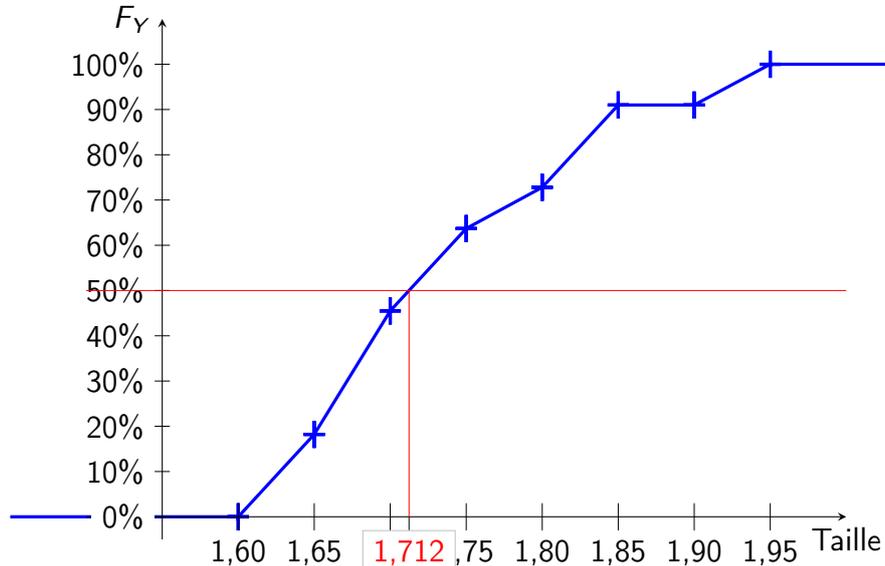
1,74 1,74 1,78 1,81 1,82 1,91

La médiane est $\frac{1,78+1,81}{2} = 1,795m$

Cas de données regroupées en classes

i) lecture graphique

La médiane est le moment où la fonction de répartition atteint 50%



La taille médiane dans cet échantillon est d'environ 1,712m.

ii) formule du formulaire

- ▶ on appelle a_i et a_{i+1} le minimum et le maximum de la première classe dont la fréquence cumulée est supérieure à 0,5.

- ▶ la médiane est alors donnée par la formule

$$\text{Med} \simeq a_i + \frac{a_{i+1} - a_i}{F_X(a_{i+1}) - F_X(a_i)} (0,5 - F_X(a_i)).$$

Exemple Taille médiane dans l'échantillon :

$a_i = 1,7$ et $a_{i+1} = 1,75$. On a donc $F_Y(a_i) = 0,455$ et $F_Y(a_{i+1}) = 0,637$, d'où

$$\text{Med} \simeq 1,7 + \frac{1,75 - 1,7}{0,637 - 0,455} (0,5 - 0,455) \simeq 1,712m.$$

B) Quartiles (*dans ce cours : uniquement pour des données regroupées en classes*)

- ▶ premier quartile Q_1 : là où la fonction de répartition atteint 25%.
- ▶ troisième quartile Q_3 : là où la fonction de répartition atteint 75%.
- ▶ on peut encore utiliser la formule du formulaire en remplaçant 0,5 par 0,25 (pour Q_1) ou par 0,75 (pour Q_3).

C) Moyenne d'une variable X , quantitative

Moyenne somme des valeurs que prend la variable pour chaque individu, divisée par l'effectif total

► Données brutes : $m(X) = \frac{1}{n} \sum_{i=1}^n x_i$

► Données regroupées par modalités :

$$m(X) = \frac{1}{n} \sum_{i=1}^r n_i x_i$$

► Données regroupées en classes :

$$m(X) = \frac{1}{n} \sum_{i=1}^r n_i c_i$$

Exemples :

• $m(Y) = \frac{1,69+1,82+1,65+1,74+1,91+1,74+1,62+1,78+1,81+1,6+1,69}{11} = \frac{19,05}{11} \simeq 1,732m$

• $m(X) = \frac{0 \times 2 + 1 \times 4 + 2 \times 2 + 3 \times 1 + 4 \times 2}{11} = \frac{19}{11} \simeq 1,727$

• $m(Y) \simeq \frac{1,625 \times 2 + 1,675 \times 3 + 1,725 \times 2 + 1,775 \times 1 + 1,825 \times 2 + 1,875 \times 0 + 1,925 \times 1}{11} = \frac{19,075}{11} \simeq 1,734m$

D) Écart type d'une variable X , quantitative

Variance $\text{Var}(X) = m(X^2) - (m(X))^2$

Écart type racine carré de la variance (notée $s(X) = \sqrt{\text{Var}(X)}$).

Exemples :

• $m(Y^2) =$

$$\frac{1,69^2 + 1,82^2 + 1,65^2 + 1,74^2 + 1,91^2 + 1,74^2 + 1,62^2 + 1,78^2 + 1,81^2 + 1,6^2 + 1,69^2}{11} = \frac{33,0793}{11}$$

D'où $\text{Var}(Y) = \frac{33,0793}{11} - \left(\frac{19,05}{11}\right)^2 \simeq 0,008015$

Et $s(Y) \simeq \sqrt{0,008015} \simeq 0,09m$

• $m(X^2) = \frac{0^2 \times 2 + 1^2 \times 4 + 2^2 \times 2 + 3^2 \times 1 + 4^2 \times 2}{11} = \frac{53}{11}$

D'où $\text{Var}(X) = \frac{53}{11} - \left(\frac{19}{11}\right)^2 \simeq 1,835$

Et $s(X) \simeq \sqrt{1,835} \simeq 1,355$

• $m(Y^2) \approx$

$$\frac{1,625^2 \times 2 + 1,675^2 \times 3 + 1,725^2 \times 2 + 1,775^2 \times 1 + 1,825^2 \times 2 + 1,875^2 \times 0 + 1,925^2 \times 1}{11} = \frac{33,166875}{11}$$

D'où $\text{Var}(Y) \approx \frac{33,166875}{11} - \left(\frac{19,075}{11}\right)^2 \simeq 0,0081$

Et $s(Y) \simeq \sqrt{0,0081} \simeq 0,09m$