

XII - Statistique descriptive

I - Description statistique d'une population

I.1 - Statistique

Définition 1 - Population, Taille, Individus

- Une *étude statistique* est une étude visant à recueillir et analyser des informations sur un ensemble fini, généralement noté Ω , appelé *population*.
- Les éléments de la population Ω sont des *individus*.
- La *taille* de la population est le cardinal de l'ensemble Ω .

Exemple 1 - Populations

- L'ensemble des individus constituant la population française est une population.
- L'ensemble des élèves d'une classe est une population.

Définition 2 - Caractères, Modalités

Soit Ω une population.

- Un *caractère* sur Ω est une application définie sur Ω .
- Si X est une application, un élément $x \in X(\Omega)$ est une *modalité* du caractère.

Exemple 2

- Si Ω est la population française, on peut considérer les caractères :
 - ★ la taille en cm ; dont 150, 172.45, 190 sont des caractères.
 - ★ la couleur des yeux ; dont bleu, marron, vert sont des caractères.

- Si Ω est l'ensemble des élèves de la classe, la note obtenue lors du dernier devoir de mathématiques est un caractère.

Définition 3 - Quantitatif, Qualitatif

Soit Ω une population et X un caractère sur Ω .

- Si X est à valeurs réelles, X est un caractère *quantitatif*.
- Si X est à valeurs dans un ensemble quelconque non inclus dans \mathbb{R} , X est un caractère *qualitatif*.
- Si $X(\Omega)$ est dénombrable, le caractère est *discret*. Sinon, il est continu.

Exemple 3

- La taille des individus est un caractère quantitatif continu. La couleur des yeux est un caractère qualitatif discret.
- La note obtenue au dernier devoir de mathématiques est un caractère quantitatif discret (la note est généralement arrondie à une décimale).

I.2 - Statistiques univariées

Définition 4 - Effectif, Fréquence, Série

Soit Ω une population de cardinal N et X un caractère sur Ω . Posons

$$n_x = |\{\omega \in \Omega ; X(\omega) = x\}|.$$

- L'entier n_x est l'*effectif* de la modalité x . Il s'agit du nombre d'individus de la population qui présentent la modalité x .

- Le réel $f_x = \frac{n_x}{N}$ est la *fréquence* de la modalité x . Les fréquences sont souvent exprimées sous forme de pourcentages.
- Si l'effectif est discret, en notant $X(\Omega) = \{x_1, \dots, x_n\}$, la *série statistique* des effectifs associée à Ω est l'ensemble des couples $\{(x_i, n_i), i \in \llbracket 1, n \rrbracket\}$.

Exemple 4 - Notes au devoir

Supposons que la liste des notes au devoir est la suivante : 8.2, 5.5, 10, 4.7, 15.7, 12.1, 4.7, 12.1, 9.3, 4.7, 9.6, 14.5, 10, 18.9, 19.5, 7.5, 8.2, 18.9, 10, 11.1.

- Représentation sous forme de tableau des effectifs :

| | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|----|------|------|------|------|------|------|
| 4.7 | 5.5 | 7.5 | 8.2 | 9.3 | 9.6 | 10 | 11.1 | 12.1 | 14.5 | 15.7 | 18.9 | 19.5 |
| 3 | 1 | 1 | 2 | 1 | 1 | 3 | 1 | 2 | 1 | 1 | 2 | 1 |

- Représentation sous forme de tableau des fréquences :

| | | | | | | | | | | | | |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 4.7 | 5.5 | 7.5 | 8.2 | 9.3 | 9.6 | 10 | 11.1 | 12.1 | 14.5 | 15.7 | 18.9 | 19.5 |
| $\frac{3}{20}$ | $\frac{1}{20}$ | $\frac{1}{20}$ | $\frac{2}{20}$ | $\frac{1}{20}$ | $\frac{1}{20}$ | $\frac{3}{20}$ | $\frac{1}{20}$ | $\frac{2}{20}$ | $\frac{1}{20}$ | $\frac{1}{20}$ | $\frac{2}{20}$ | $\frac{1}{20}$ |

Proposition 1 - Somme des fréquences

Soit Ω une population et X un caractère sur Ω . Pour toute modalité x de X , on note f_x sa fréquence. Alors,

$$\sum_{x \in X(\Omega)} f_x = 1.$$

Définition 5 - Fréquences cumulées d'une modalité

Soit Ω une population de taille N et X un caractère discret dont les modalités peuvent être ordonnées. Posons $X(\Omega) = \{x_1, \dots, x_n\}$ où $x_1 < x_2 < \dots < x_n$.

La *fréquence cumulée* de la modalité x_i est l'entier

$$F_i = \frac{|\{\omega \in \Omega ; X(\omega) \leq x_i\}|}{N}.$$

Le réel F_i est la proportion d'individus qui ont une modalité au plus égale à x_i .

Exemple 5 - Notes au devoir

Reprenons les données de l'exemple précédent.

- Le tableau des effectifs cumulés est le suivant :

| | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|----|------|------|------|------|------|------|
| 4.7 | 5.5 | 7.5 | 8.2 | 9.3 | 9.6 | 10 | 11.1 | 12.1 | 14.5 | 15.7 | 18.9 | 19.5 |
| 3 | 4 | 5 | 7 | 8 | 9 | 12 | 13 | 15 | 16 | 17 | 19 | 20 |

- Le tableau des fréquences est le suivant :

| | | | | | | | | | | | | |
|----------------|----------------|----------------|----------------|----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 4.7 | 5.5 | 7.5 | 8.2 | 9.3 | 9.6 | 10 | 11.1 | 12.1 | 14.5 | 15.7 | 18.9 | 19.5 |
| $\frac{3}{20}$ | $\frac{4}{20}$ | $\frac{5}{20}$ | $\frac{7}{20}$ | $\frac{8}{20}$ | $\frac{9}{20}$ | $\frac{12}{20}$ | $\frac{13}{20}$ | $\frac{15}{20}$ | $\frac{16}{20}$ | $\frac{17}{20}$ | $\frac{19}{20}$ | $\frac{20}{20}$ |

Définition 6 - Classe

Les valeurs prises par le caractère X peuvent être regroupées en sous-ensembles de $X(\Omega)$ appelées *classes*. On étend les notions d'effectifs et de fréquences aux classes.

Exemple 6 - Notes au devoir

Reprenons les données de l'exemple précédent. Les modalités peuvent être regroupées selon les classes suivantes :

| | | | | | | | | | | |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|----------|
| [0, 1[| [1, 2[| [2, 3[| [3, 4[| [4, 5[| [5, 6[| [6, 7[| [7, 8[| [8, 9[| [9, 10[| [10, 11[|
| 0 | 0 | 0 | 0 | 3 | 1 | 0 | 1 | 2 | 2 | 3 |

| | | | | | | | | |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| [11, 12[| [12, 13[| [13, 14[| [14, 15[| [15, 16[| [16, 17[| [17, 18[| [18, 19[| [19, 20] |
| 1 | 2 | 0 | 1 | 1 | 0 | 0 | 2 | 1 |

I.3 - Représentations graphiques

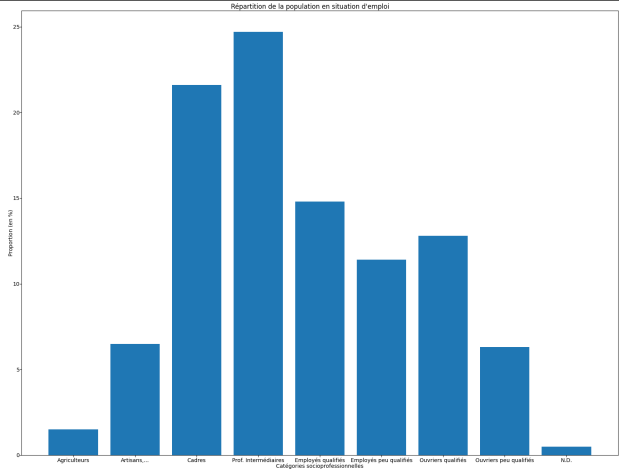
Définition 7 - Diagramme en bâtons

Dans un *diagramme en bâtons*, les modalités du caractère discret étudié sont indiquées en abscisses (il peut s'agir d'un caractère qualitatif). Les effectifs ou fréquences sont indiquées en ordonnées.

Exemple 7 - Catégories socioprofessionnelles

En 2021, la répartition (en pourcentages) de la population en situation d'emploi selon la catégorie socioprofessionnelle était la suivante (source INSEE) :

| Catégorie socioprofessionnelle | Proportion |
|---|------------|
| Agriculteurs | 1.5 |
| Artisans, commerçants, chefs d'entreprise | 6.5 |
| Cadres | 21.6 |
| Professions intermédiaires | 24.7 |
| Employés qualifiés | 14.8 |
| Employés peu qualifiés | 11.4 |
| Ouvriers qualifiés | 12.8 |
| Ouvriers peu qualifiés | 6.3 |
| Non déterminé | 0.5 |

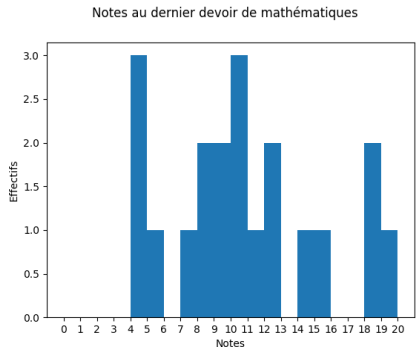


Définition 8 - Histogramme

Dans un *histogramme*, le caractère étudié est regroupé en classes. Les valeurs des classes sont indiquées en abscisses. Les effectifs ou fréquences sont indiquées en ordonnées.

Exemple 8 - Notes au devoir

En reprenant les notes obtenues lors du dernier devoir de mathématiques :

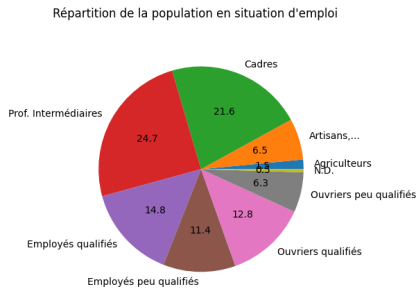


Définition 9 - Diagramme circulaire

Dans un *diagramme circulaire*, à chaque modalité du caractère (discret) ou chaque classe est dévolue une partie d'un disque dont l'aire est proportionnelle à sa fréquence.

Exemple 9 - Catégories socioprofessionnelles

En reprenant les données précédentes :



I.4 - Caractéristiques de position et de dispersion

Définition 10 - Mode

Soit Ω une population et X un caractère discret sur Ω . Un *mode* est une modalité $x \in X(\Omega)$ dont l'effectif est maximal.

Exemple 10 - Notes au devoir

En reprenant les notes au devoir, les modes sont 4.7 et 10.

Définition 11 - Médiane

Soit Ω une population de taille N et X un caractère discret ordonné sur Ω . On note $X(\Omega) = \{x_1, \dots, x_n\}$ où $x_1 < x_2 < \dots < x_n$.

La modalité x_i est la *médiane* de X si

$$|\{\omega \in \Omega ; X(\omega) < x_i\}| < \frac{N}{2} \text{ et } |\{\omega \in \Omega ; X(\omega) > x_i\}| < \frac{N}{2}.$$

La médiane est une modalité qui partage la population en deux parties contenant sensiblement les mêmes effectifs.

Plus généralement, les *quantiles* sont les modalités qui permettent de diviser la population en intervalles consécutifs contenant sensiblement les mêmes effectifs.

Exemple 11 - Notes au devoir

Rappelons que $N = 20$ soit $N/2 = 10$. En utilisant le tableau des effectifs cumulés,

- Le nombre d'étudiants ayant une note strictement inférieure à 10 vaut 9.
- Le nombre d'étudiants ayant une note strictement supérieure à 10 vaut $20 - 12 = 8$.

Ainsi, la médiane vaut 10.

Définition 12 - Moyenne

Soit Ω une population de taille N et X un caractère sur Ω . On note $X(\Omega) = \{x_1, \dots, x_n\}$. Pour toute modalité x_i , on note n_i son effectif et f_i sa fréquence. La moyenne \bar{X} du caractère X est le réel

$$\bar{X} = \frac{1}{N} \sum_{i=1}^n n_i x_i = \sum_{i=1}^n f_i x_i.$$

Exemple 12 - Notes au devoir

En utilisant un logiciel de calcul, la moyenne vaut

$$\frac{3 \cdot 4.7 + 5.5 + 7.5 + 2 \cdot 8.2 + 9.3 + 9.6 + 3 \cdot 10 + 11.1 + 2 \cdot 12.1 + 14.5 + 15.7 + 2 \cdot 18.9 + 19.5}{20} = 10.76$$

Définition 13 - Étendue

Soit Ω une population de taille N et X un caractère discret sur Ω à valeurs réelles. On note $X(\Omega) = \{x_1, \dots, x_n\}$.

- L'*étendue* est la différence entre la plus grande et la plus petite modalité du caractère X .
- Si \bar{X} est la moyenne du caractère x , la *variance* du caractère X est le réel :

$$V(X) = \frac{1}{N} \sum_{i=1}^n (x_i - \bar{X})^2.$$

- L'*écart-type* est le réel $\sigma(X) = \sqrt{V(X)}$.
- Si $x_p < x_q < x_r$ sont trois modalités qui divisent la population en 4 intervalles successifs ayant sensiblement les

mêmes effectifs :

$$\begin{aligned} |\{\omega \in \Omega ; X(\omega) < x_p\}| &< \frac{N}{4}, \\ |\{\omega \in \Omega ; x_p < X(\omega) < x_q\}| &< \frac{N}{4}, \\ |\{\omega \in \Omega ; x_q < X(\omega) < x_r\}| &< \frac{N}{4} \\ \text{et } |\{\omega \in \Omega ; x_r < X(\omega)\}| &< \frac{N}{4}, \end{aligned}$$

La modalité x_p est le premier *quartile*, la modalité x_q est le second *quartile* et la modalité x_r est le troisième *quartile*. La différence $x_r - x_p$ entre le plus grand et le plus petit quartile est l'*écart interquartile*.

- Le *coefficient de variation* de X est égal au rapport $\frac{\sigma(X)}{\bar{X}}$.

Exemple 13 - Notes au devoir

En utilisant un logiciel de calcul,

- l'étendue vaut $19.5 - 4.7 = 14.8$.
- la variance vaut environ 20.9.
- l'écart-type vaut environ 4.6.

L'avantage de l'écart-type est qu'il s'exprime dans la même unité que les données initiales (contrairement à la variance qui s'exprime en fonction de cette unité au carré). Le coefficient de variation est un nombre sans dimension.

En reprenant le tableau des effectifs,

- le nombre de notes strictement inférieures à 7.5 vaut 4,
- le nombre de notes strictement comprises entre 7.5 et 10 vaut 4,
- le nombre de notes strictement comprises entre 10 et 14.5 vaut 4,
- le nombre de notes strictement supérieures à 14.5 vaut 4.

Nous avons ainsi déterminé les quatre quartiles.

II - Statistiques bivariées

Définition 14 - Effectifs, Marginales

Soit X, Y deux caractères définis sur une population Ω de taille N . On note x_1, \dots, x_n les modalités de X et y_1, \dots, y_p les modalités de Y .

- Pour tout $i \in \llbracket 1, n \rrbracket$ et $j \in \llbracket 1, p \rrbracket$, on note $n_{i,j}$ l'*effectif* du couple de modalités (x_i, y_j) et $f_{i,j} = \frac{n_{i,j}}{N}$ sa fréquence.
- Les *effectifs marginaux* en X sont, pour tout $i \in \llbracket 1, n \rrbracket$, les entiers $n_i = \sum_{j=1}^p n_{i,j}$.
- Les *effectifs marginaux* en Y sont, pour tout $j \in \llbracket 1, p \rrbracket$, les entiers $m_j = \sum_{i=1}^n n_{i,j}$.

On définit de manière analogue les fréquences et fréquences marginales. Ces quantités peuvent être représentées dans un tableau. On définit, comme dans le cadre des variables aléatoires, les valeurs de X conditionnellement à $Y = y_j$ ainsi que les valeurs de Y conditionnellement à $X = x_i$.

Exemple 14 - Mesures d'arbres

On liste ci-dessous la circonférence X (en cm) et le volume Y (en cm^3) de différents arbres : (9, 10), (9, 10), (9, 10), (11, 15), (11, 20), (11, 20), (11, 15), (11, 15), (11, 20), (11, 20), (11, 25), (11, 20), (11, 20), (11, 20), (13, 15), (13, 20), (13, 35), (13, 20), (15, 30), (15, 30), (15, 35), (15, 35), (15, 35), (15, 40), (15, 45), (19, 55), (19, 55), (19, 60), (19, 50), (19, 50), (21, 60).

Les marginales selon Y (resp. X) sont indiquées dans la dernière colonne (resp. ligne)

| $\begin{matrix} Y \backslash X \\ \end{matrix}$ | 9 | 11 | 13 | 15 | 19 | 21 | |
|---|---|----|----|----|----|----|----|
| 10 | 3 | 0 | 0 | 0 | 0 | 0 | 3 |
| 15 | 0 | 3 | 1 | 0 | 0 | 0 | 4 |
| 20 | 0 | 7 | 2 | 0 | 0 | 0 | 9 |
| 25 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 30 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| 35 | 0 | 0 | 1 | 3 | 0 | 0 | 4 |
| 40 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 45 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 50 | 0 | 0 | 0 | 0 | 2 | 0 | 2 |
| 55 | 0 | 0 | 0 | 0 | 2 | 0 | 2 |
| 60 | 0 | 0 | 0 | 0 | 1 | 1 | 2 |
| | 3 | 11 | 4 | 7 | 5 | 1 | 31 |

Définition 15 - Covariance, Coefficient de corrélation

On reprend les notations de la définition précédente.

- La *covariance* de X et de Y est le réel

$$\sigma_{X,Y} = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^p n_{i,j} (x_i - \bar{X})(y_j - \bar{Y}).$$

- Le *coefficient de corrélation* de X et Y est le réel

$$r = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.$$

Exemple 15 - Mesures d'arbres

Un logiciel de calcul permet d'obtenir

$$\bar{X} = 13.6$$

$$\bar{Y} = 29.4$$

$$\sigma_X \simeq 3.4$$

$$\sigma_Y \simeq 15.3$$

$$\sigma_{X,Y} \simeq 51.8$$

$$r \simeq 0.95$$

Théorème 1 - Méthode des moindres carrés

On reprend les notations des questions précédentes. La *droite de régression linéaire* de Y par rapport à X a pour équation :

$$y = ax + b = \frac{\sigma_{X,Y}}{\sigma_X^2} (x - \bar{X}) + \bar{Y}.$$

Si le coefficient de corrélation r^2 est proche de 1, l'approximation du nuage de points par la droite de régression est bonne. Les caractères X et Y sont *corrélés*. On peut alors effectuer une prédiction sur la valeur de y connaissant une valeur de x .

Exemple 16 - Mesures d'arbres

Un logiciel de calcul permet d'obtenir

$$a \simeq 4.3 \text{ et } b \simeq -29.5$$

On peut représenter le nuage de points ainsi que la droite de régression sur un même graphique. On a mis également en évidence le point de coordonnées (\bar{X}, \bar{Y}) . Comme r est proche de 1, l'approximation par une droite de régression est bonne.

