Statistiques descriptives

Données

- But : définir des indicateurs permettant de résumer, suivant les besoins, des informations sur la distribution des données collectées auprès d'une population étudiée.
- Les variables étudiées seront **quantitatives**.

I Vocabulaire

Le vocabulaire des statistiques est souvent polysémique, ce qui induit souvent des erreurs de compréhension : soyez donc très attentifs au sens des mots employés

A) Variable (Statistique) .....

[Variable - Constante] **Déf.1**

- **Variable** : Concept en lien à une « réalité » (physique ou non), et se manifestant sous différents aspects, appelés *valeurs* possibles de la variable.
- **Constante** : variable qui se manifeste toujours sous la même valeur.

Rem.1 | Les termes *variable*, et *constante* ont donc ici des acceptions différentes de leur acception mathématique, ou informatique,ou même physique.

[terminologie de base] **Déf.2**

a. L'ensemble $\mathscr{P}$ (au sens mathématique) des objets sur lesquels les informations seront collectées est appelé *population*.

b. Les éléments de la population $\mathscr{P}$ s'appellent les *individus*.

c. Le cardinal de $\mathscr{P}$ est appelé *effectif* de la population (si il est défini).

d. Une variable statistique attachée à $\mathscr{P}$ est aussi appelée *caractère* de la population.

e. Les valeurs prises par un caractère s'appellent les *modalités* du caractère ou de la variable.

f. Tout sous-ensemble fini $\mathscr{E}$ de $\mathscr{P}$ est appelé *échantillon* de la population.

g. Le cardinal de $\mathscr{E}$ est appelé *taille* de l'échantillon.

[Variable statistique catégorielle, ordinale, quantitative] **Déf.3**

Le caractère x est :

a. **qualitatif** ou **catégoriel** si les valeurs de x ne peuvent être ordonnées. Exemple : la couleur des yeux, le groupe sanguin.

b. **ordinal** si les valeurs peuvent être seulement ordonnées sans pouvoir donner de sens aux différences entre les valeurs (typiquement : un caractère mesuré sur une échelle de valeurs, telle que $0-1-2-3-4-5$ ou $\mathit{Satisfait}-\mathit{Moyen}-\mathit{Nul}$).

c. **quantitatif** quand les valeurs sont celles d'une grandeur numérique (typiquement : toute grandeur exprimée dans une certainte unité).

II Statistiques univariées

A) Série statistique .....

Usage.

a. Quitte à numéroter les n individus de l'échantillon $\mathscr{E}$, on peut supposer : $\mathscr{E} = \{1, 2, \dots, n\}$.

b. Dans ce cas : x_j est la valeur du caractère x observée sur l'individu $j \in \mathscr{E}$.

[Série statistique] **Déf.4**

La liste $\mathbf{x} = (x_1, \dots x_n)$ s'appelle *série statistique*.

Rem.2 |

a. Ainsi, si le caractère x est quantitatif, $\mathbf{x} \in \mathbf{R}^n$: on a alors à notre disposition tous les concepts de l'algèbre linéaire et du produit scalaire.

b. La série statistique $\mathbf{x}$ est constante (**Déf. 1**) ssi pour un certain réel a, on a : $\mathbf{x} = (a, a, \dots, a)$. Ce qui veut dire que $\mathbf{x}$ est colinéaire à la série : $\mathbf{1} = (1, \dots, 1)$.

c. Ainsi, Vect$(\mathbf{1})$ est le sev des séries statistiques constantes.

[Série chronologique] **Déf.5**

Soit $\mathbf{x}$ une série statistique. Si l'ordre dans lequel les mesures ont été prises est important, $\mathbf{x}$ s'appelle *série chronologique*, et la numérotation des éléments de $\mathscr{E}$ est choisie de sorte à respecter l'ordre d'acquisition des données.

[Effectifs, fréquences de la série $\mathbf{x}$] **Déf.6**

a. on note n_k le nombre de fois que cette valeur est observée, et on l'appelle effectif des cas où x prend la valeur x_k.

b. Le rapport $f_k = \dfrac{n_k}{n}$ est appelé fréquence de l'observation x_k.

B) Représentations graphiques

La visualisation des données donne une bonne synthèse des listes de nombres !

Représentations possibles :

a. Boites à moustache.

b. Diagramme en bâtons.

c. Histogramme (regroupement par classes dans ce cas).

d. Polygone des fréquences ou effectifs cumulés.

Rem.3 |

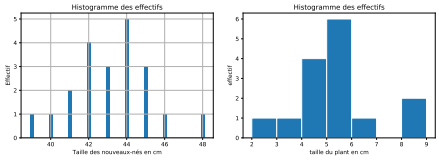
a. Le choix de l'amplitude des classes ne va pas de soi.

b. En effet, si les classes sont trop amples, la synthèse des données risque de refléter assez peu les tendances de la série.

c. *A contrario*, des classes trop étroites brouillent la lisibilité.

d. On peut toujours essayer de regrouper les classes en classes de largeur de l'ordre d' un écart-type de la série de données : ceci devrait faire ressortir

une distribution normale d'après le TCL, si ses conditions sont vérifiées. Toutefois, il n'y a pas de règle absolue.



c) Caractéristiques de position

[Mode d'une série statistique, classe modale] **Déf.7**

a. Le mode d'une série statistique est la valeur observée le plus grand nombre de fois.

b. En cas de regroupement par classe, la classe modale est la classe d'effectif le plus élevé.

Rem.4

a. S'il y a plusieurs modes ou plusieurs classes modales, n'importe laquelle des valeurs la plus observée constitue un mode.

b. En cas de regroupement par classes, les classes modales sont celles dont les rectangles sont d'aire maximale.

[Moyenne ou moyenne empirique de la série $\mathbf{x}$] **Déf.8**

Soit $\mathbf{x} = (x_1, \dots, x_n)$ une série statistique. Alors :

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j.$$

Rem.5

a. On a donc : $\bar{x} = \dfrac{\mathbf{x}\cdot\mathbf{1}}{\|\mathbf{1}\|^2}$.

b. La série $\mathbf{k} = (\bar{x}, \dots, \bar{x}) = \dfrac{\mathbf{x}\cdot\mathbf{1}}{\|\mathbf{1}\|^2}\mathbf{1}$ est donc le projeté orthogonal de $\mathbf{x}$ sur le sev des séries constantes.

[Stabilité de la moyenne par changement affine] **Prop.1**

Si $(a,b) \in \mathbf{R}^2$, $\mathbf{x} = (x_j)_{1\leq j\leq n}$ et $y = (ax_j + b)_{1\leq j\leq n}$. Alors :

$$\overline{y} = a\overline{x} + b$$

[Médiane d'une variable non qualitative] **Déf.9**

Soit $(x_j)_{1\leq j\leq n}$ une série statistique. Quitte à renuméroter les modalités, supposons que ces dernières sont rangées dans l'ordre croissant :

$$x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n.$$

La médiane de la série est le nombre noté Q_2 défini par :

a. $Q_2 = x_{\frac{n+1}{2}}$ si n est impair.

b. $Q_2 = \dfrac{1}{2}\left(x_{\frac{n}{2}} + x_{1+\frac{n}{2}}\right)$ si n est pair.

Rem.6 La détermination de Q_2 passe par le classement des données dans l'ordre croissant. Voir la fiche `Méthodes de tri`.

d) Caractéristiques de dispersion

[Étendue d'une série $\mathbf{x}$] **Déf.10**

Soit $x = (x_j)_{1\leq j\leq n}$ une variable statistique quantitative.

$$\max_{1\leq j\ n}(x_j) - \min_{1\leq j\ n}(x_j).$$

Rem.7 Donne une idée de la «variabilité» des données, mais est extrêmement sensible aux données aberrantes, puisqu'une valeur anormalement petite par exemple (obtenue par erreur de manipulation par exemple) modifie complètement la valeur de l'étendue des données.

[variance/écart-type d'une série] **Déf.11**

- Variance des données :

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2 = \overline{(x-\overline{x})^2}.$$

- Écart-type des données :

$$s_x = \sqrt{s_x^2}.$$

Rem.8

a. On peut remarquer, avec les notations de **Rem. 5 b.**, que

$$s_x^2 = \|\mathbf{x} - \mathbf{k}\|^2$$

b. Ainsi, par définition de projeté orthogonal réalisant le minimum de distance, σ_x est la distance de $\mathbf{x}$ au sev des variables constantes.

[Koenig pour les séries stats.] **Prop.2**

$$s_x^2 = \overline{x^2} - \overline{x}^2$$

Rem.9 C'est simplement le théorème de Pythagore (cf. fiche `Produit scalaire`, **Prop. 4**), puisque :

$$\|\mathbf{x}\|^2 = \|\mathbf{x} - \mathbf{k}\|^2 + \|\mathbf{k}\|^2$$

[Propriétés de la variance] **Prop.3**

a. Si x est en unités u, s_x^2 est en unités u^2.

b. $\forall a \in \mathbf{R}\ s_{ax}^2 = a^2 s_x^2$

c. $\forall b \in \mathbf{R}\ s_{x+b}^2 = s_x^2$.