Chapitre 1

Statistique descriptive

Ce chapitre est une introduction à la description de séries statistiques. Une série statistique s'obtient en observant une ou plusieurs variables statistiques dans une population. Par exemple, la taille de chaque étudiant d'une classe de lycée forme une série statistique issue de la variable taille, observée dans la population des élèves de la classe de lycée considérée.

La statistique descriptive précède la statistique inférentielle dont l'objet est la modélisation des variables statistiques étudiées sur une population en vue de prévoir le résultat d'une étude portant sur une population plus vaste. La statistique inférentielle est introduite dans la seconde partie de ce cours.

Le formalisme de la statistique descriptive, et de la statistique en général, est celui des mathématiques. Dans ce chapitre, l'emploi du formalisme mathématique est dosé de façon à fournir au lecteur un exposé rigoureux, sans être trop théorique.

Chaque notion ou résultat est appliqué à un problème issu de la psychologie ou de la psychopharmacologie.

La Section 1.1 est consacrée aux séries statistiques issues d'une variable qualitative ou quantitative. La description de données issues d'une variable quantitative passe par des représentations graphiques, mais également par le calcul et l'interprétation d'indicateurs de position, de dispersion et de forme.

A la Section 1.1, des représentations graphiques classiques (diagramme en bâtons, histogramme et boîte à moustaches) et des indicateurs classiques (mode et classe modale, moyenne, écart-type, quantiles et coefficient de Yule) sont présentés, puis appliqués en psychologie.

La Section 1.2 est consacrée aux séries statistiques bivariées. Après avoir introduit les notions de distributions jointes, marginales et conditionnelles d'un couple de variables statistiques, le coefficient de corrélation linéaire est défini pour un couple de variables quantitatives. Des applications en psychologie et en psychopharmacologie sont proposées.

Sur la statistique descriptive, le lecteur pourra également se référer à Dodge [6] ou Tenenhaus [19]. En particulier, les éléments de statistique descriptive présentés dans ce chapitre constituent la base pour l'analyse de données dont une bonne introduction est proposée dans [19].

1.1 Séries univariées

Cette section est consacrée à l'étude des variables statistiques unidimensionnelles.

Définition 1.1.1 *Soient $\mathscr{P}$ une population,* E *un ensemble et* X *une application de $\mathscr{P}$ dans* E *prolongeable en un certain sens à toute population contenant $\mathscr{P}$.*

1. *Si* E *est un ensemble fini et non numérique, alors* X *est une variable (statistique) qualitative.*
2. *Si* $\mathrm{E} \subset \mathbb{R}$ *et que l'expérience statistique suggère que* $\mathrm{X}(\mathscr{P}) = \mathrm{X}(\mathscr{P}')$ *pour toute population $\mathscr{P}'$ contenant $\mathscr{P}$, alors* X *est une variable quantitative discrète.*
3. *Si* $\mathrm{E} = [a,b]$ *($a, b \in \mathbb{R}$ tels que $a < b$) et que l'expérience statistique suggère que pour des populations de plus en plus vastes les images de celles-ci par* X *tendent à valoir* E*, alors* X *est une variable quantitative continue.*

Remarque. A la Définition 1.1.1.(3), parler de continuité est abusif car l'application X n'est pas continue. Néanmoins, c'est une façon commode de faire référence au fait que $\mathrm{X}(\mathscr{P}) \approx [a,b]$ pour une population $\mathscr{P}$ très vaste.

Exemples :

1. Le groupe sanguin est une variable qualitative à valeurs dans
$$\mathrm{E} = \{\mathrm{A}, \mathrm{B}, \mathrm{AB}, \mathrm{O}\}.$$
2. Le QI est une variable quantitative discrète à valeurs dans
$$\mathrm{E} = \{0, \ldots, 200\}.$$
3. En année(s), l'espérance de vie d'un humain est une variable quantitative continue à valeurs dans $[0, 130]$.

Selon son type (qualitative, quantitative discrète ou quantitative continue), l'étude d'une variable statistique dans une population donnée nécessite l'emploi de méthodes spécifiques.

La Sous-Section 1.1.1 est consacrée aux variables qualitatives et quantitatives discrètes, tandis que la Sous-Section 1.1.2 est consacrée aux variables quantitatives continues.

1.1.1 Variables qualitatives et quantitatives discrètes

Soient $\mathscr{P}$ une population formée de $n \in \mathbb{N}^*$ individus, $\mathrm{E} := \{x_1, \ldots, x_m\}$ un ensemble fini de cardinal $m \in \mathbb{N}^*$ et $\mathrm{X} : \mathscr{P} \to \mathrm{E}$ une variable statistique. Les éléments $x_1, \ldots, x_m$ de E sont les modalités de la variable X.

Définition 1.1.2 *Pour tout $i \in \{1, \ldots, m\}$, le nombre $n(\mathrm{X} = x_i)$ d'individus $\lambda \in \mathscr{P}$ tels que $\mathrm{X}(\lambda) = x_i$ est l'effectif de la modalité x_i dans la population $\mathscr{P}$.*

Définition 1.1.3 *Pour tout $i \in \{1, \ldots, m\}$,*
$$p(\mathrm{X} = x_i) := \frac{n(\mathrm{X} = x_i)}{n}$$
est la proportion d'individus $\lambda \in \mathscr{P}$ tels que
$$\mathrm{X}(\lambda) = x_i.$$
Ces proportions, également appelées fréquences, forment la distribution de la variable X *dans la population $\mathscr{P}$.*

Remarque. La somme des proportions des modalités $x_1, \dots, x_m$ dans la population $\mathscr{P}$ vaut 1 :

$$\sum_{i=1}^{m} p(\mathrm{X} = x_i) = \frac{1}{n} \sum_{i=1}^{m} n(\mathrm{X} = x_i) = 1.$$

Définition 1.1.4 *Pour tous $m^* \in \{1, \dots, m\}$ et $i_1, \dots, i_{m^*} \in \{1, \dots, m\}$ deux à deux distincts,*

$$p(\mathrm{X} \in \{x_{i_1}, \dots, x_{i_{m^*}}\}) := \sum_{j=1}^{m^*} p(\mathrm{X} = x_{i_j})$$

est la proportion d'individus $\lambda \in \mathscr{P}$ tels que $\mathrm{X}(\lambda) \in \{x_{i_1}, \dots, x_{i_{m^*}}\}$.

Notation. Pour tous $i_1, i_2 \in \{1, \dots, m\}$ tels que $i_1 \leqslant i_2$,

$$p(x_{i_1} \leqslant \mathrm{X} \leqslant x_{i_2}) := p(\mathrm{X} \in \{x_{i_1}, \dots, x_{i_2}\}).$$

Définition 1.1.5 *Le mode $\mathrm{M}_{\mathscr{P}}(\mathrm{X})$ de la variable X dans la population $\mathscr{P}$ est la modalité de proportion la plus élevée (si cette dernière est unique).*

En pratique, la série statistique (i.e. la donnée des modalités et des effectifs associés dans la population $\mathscr{P}$) est présentée dans un tableau :

X	x_1	…	x_m
Effectifs	$n(\mathrm{X} = x_1)$	…	$n(\mathrm{X} = x_m)$
Proportions	$p(\mathrm{X} = x_1)$	…	$p(\mathrm{X} = x_m)$

Il est commode de représenter la série statistique par un diagramme en bâtons. Il se construit comme suit :

1. Les modalités $x_1, \dots, x_m$ sont placées sur l'axe des abscisses d'un repère orthogonal à égales distances les unes des autres.
2. L'axe des ordonnées est gradué de 0 à la proportion la plus élevée de la série.
3. Un segment issu de x_i, parallèle à l'axe des ordonnées et de longueur $p(\mathrm{X} = x_i)$, est tracé pour tout $i \in \{1, \dots, m\}$.

Exemple. Les colonnes de Beck constituent un outil de décentration introduit par le psychiatre Aaron T. Beck, père de la thérapie cognitive. Lorsqu'une situation difficile s'impose au sujet, ce dernier doit décrire cette situation, préciser une émotion associée et lui attribuer un score de 1, 2, 3, 4

ou 5 selon son intensité, décrire cinq pensées automatiques, chercher cinq à dix pensées alternatives, puis réattribuer un score à l'émotion. Lorsqu'il recherche des pensées alternatives, le sujet décentre et l'intensité de l'émotion doit diminuer à l'issue de cette tâche.

Le score S de l'émotion après la recherche de pensées alternatives a été enregistré chez 91 patients (population $\mathscr{P}$) d'un psychiatre exerçant en cabinet libéral, tous formés à l'utilisation des colonnes de Beck depuis au plus deux ans :

S	1	2	3	4	5
Effectifs	8	14	28	22	19
Proportions	0.088	0.154	0.308	0.242	0.209

Le mode de la variable S dans la population $\mathscr{P}$ est 3.

Désormais, E est un sous-ensemble fini de $\mathbb{R}$. Ainsi, X est une variable quantitative discrète.

Définition 1.1.6 *La moyenne (arithmétique) de la variable* X *dans la population* $\mathscr{P}$ *est*

$$\mu_{\mathscr{P}}(\mathrm{X}) := \sum_{i=1}^{m} p(\mathrm{X} = x_i)\, x_i.$$

La moyenne est un indicateur de position de la série statistique.

Proposition 1.1.7 *Soient* $m^* \in \{1,\dots,m\}$ *et* $\{\mathscr{P}_1,\dots,\mathscr{P}_{m^*}\}$ *une partition de la population* $\mathscr{P}$ *(i.e.* $\mathscr{P}_1 \cup \cdots \cup \mathscr{P}_{m^*} = \mathscr{P}$ *et* $\mathscr{P}_1,\dots,\mathscr{P}_{m^*}$ *sont deux à deux disjoints). Alors,*

$$\mu_{\mathscr{P}}(\mathrm{X}) = \frac{1}{m} \sum_{k=1}^{m^*} \mathrm{card}(\mathscr{P}_k)\, \mu_{\mathscr{P}_k}(\mathrm{X}).$$

Exemple. Soit $\mathscr{P}$ (resp. S) la population (resp. la variable statistique) de l'exemple précédent. D'une part, la moyenne du score S dans la population $\mathscr{P}$ est de 3 :

$$\mu_{\mathscr{P}}(\mathrm{S}) = 0.088 \cdot 1 + 0.154 \cdot 2 + \cdots + 0.209 \cdot 5 = 3.333.$$

D'autre part, lors de l'étude statistique, deux groupes de patients ont été distingués :

— La sous-population $\mathscr{P}_1$ de $\mathscr{P}$ formée de 43 patients pratiquant les colonnes de Beck depuis plus de six mois.
— La sous-population $\mathscr{P}_2$ de $\mathscr{P}$ formée de 48 patients pratiquant les colonnes de Beck depuis moins de six mois.

S	1	2	3	4	5
Effectifs $\mathscr{P}_1$	7	9	17	6	4
Proportions $\mathscr{P}_1$	0.163	0.209	0.395	0.139	0.093
Effectifs $\mathscr{P}_2$	1	5	11	16	15
Proportions $\mathscr{P}_2$	0.021	0.104	0.229	0.333	0.312

Alors,

$$\mu_{\mathscr{P}_1}(\mathrm{S}) = 0.163 \cdot 1 + 0.209 \cdot 2 + \cdots + 0.093 \cdot 5 = 2.787$$

et

$$\mu_{\mathscr{P}_2}(\mathrm{S}) = 0.021 \cdot 1 + 0.104 \cdot 2 + \cdots + 0.312 \cdot 5 = 3.808.$$

Ainsi, les patients de la population $\mathscr{P}$ pratiquant les colonnes de Beck depuis plus de six mois attribuent en moyenne un score plus faible à l'émotion après la recherche de pensées alternatives (3) que les autres patients (4). Une prise de recul par rapport à l'exercice semble en accroître l'effet décentrant.

Enfin,

$$\frac{1}{\mathrm{card}(\mathscr{P})}[\mathrm{card}(\mathscr{P}_1)\mu_{\mathscr{P}_1}(\mathrm{S}) + \mathrm{card}(\mathscr{P}_2)\mu_{\mathscr{P}_2}(\mathrm{S})] = 3.325 \approx \mu_{\mathscr{P}}(\mathrm{S}).$$

L'énoncé de la Proposition 1.1.7 est retrouvé dans ce cas particulier. L'erreur de 0.008 est due à l'arrondi des proportions.

Remarque. La moyenne seule est insuffisante pour décrire convenablement une série statistique car elle n'évalue pas la dispersion de cette dernière.

Définition 1.1.8 *La variance de la variable* X *dans la population* $\mathscr{P}$ *est*

$$\mathrm{var}_{\mathscr{P}}(\mathrm{X}) := \sum_{i=1}^{m} p(\mathrm{X} = x_i)[x_i - \mu_{\mathscr{P}}(\mathrm{X})]^2.$$

L'écart-type de la variable X *dans la population* $\mathscr{P}$ *est* $\sigma_{\mathscr{P}}(\mathrm{X}) := \sqrt{\mathrm{var}_{\mathscr{P}}(\mathrm{X})}$.

L'écart-type est un indicateur de dispersion de la série statistique. Il évalue la distance (euclidienne) moyenne entre chaque modalité de la variable X et sa moyenne $\mu_{\mathscr{P}}(\mathrm{X})$ dans la population $\mathscr{P}$.

Proposition 1.1.9 *La variance de la variable* X *dans la population* $\mathscr{P}$ *satisfait :*

$$\mathrm{var}_{\mathscr{P}}(\mathrm{X}) = -\mu_{\mathscr{P}}(\mathrm{X})^2 + \sum_{i=1}^{m} p(\mathrm{X} = x_i) x_i^2.$$

L'expression de la variance énoncée à la Proposition 1.1.9 est plus simple à manipuler lors des calculs que celle de la Définition 1.1.8.

Exemple. Soit $\mathscr{P}$ (resp. S) la population (resp. la variable statistique) de l'exemple précédent. Il a été établi que $\mu_{\mathscr{P}}(\mathrm{S}) = 3.333$. Donc, d'après la Proposition 1.1.9 :

$$\mathrm{var}_{\mathscr{P}}(\mathrm{S}) = -(3.333)^2 + 0.088 \cdot 12 + \cdots + 0.209 \cdot 52 = 1.464.$$

Enfin,

$$\sigma_{\mathscr{P}}(\mathrm{S}) = \sqrt{\mathrm{var}_{\mathscr{P}}(\mathrm{S})} \approx 1.210.$$

Après la recherche de pensées alternatives, les patients de la population $\mathscr{P}$ attribuent un score à l'émotion compris entre 2 et 4 en moyenne.

1.1.2 Variables quantitatives continues

Soient $\mathscr{P}$ une population formée de $n \in \mathbb{N}^*$ individus, $\mathrm{E} := [a, b]$ avec $a, b \in \mathbb{R}$ tels que $a < b$, et $\mathrm{X} : \mathscr{P} \to \mathrm{E}$ une variable quantitative continue au sens de la Définition 1.1.1.(3).

L'ensemble d'arrivée E de la variable X est subdivisé en $m \in \mathbb{N}^*$ classes

$$\mathrm{E}_1 := [x_0, x_1[, \ldots, \mathrm{E}_{m-1} := [x_{m-2}, x_{m-1}[, \mathrm{E}_m := [x_{m-1}, x_m]$$

avec $x_0, \ldots, x_m \in [a, b]$ tels que :

$$a = x_0 < x_1 < \cdots < x_{m-1} < x_m = b.$$

Définition 1.1.10 *Pour tout* $i \in \{1, \ldots, m\}$*, le nombre* $n(\mathrm{X} \in \mathrm{E}_i)$ *d'individus* $\lambda \in \mathscr{P}$ *tels que* $\mathrm{X}(\lambda) \in \mathrm{E}_i$ *est l'effectif de la classe* E_i *dans la population* $\mathscr{P}$*.*

Définition 1.1.11 *Pour tout $i \in \{1,\dots,m\}$,*

$$p(\mathrm{X} \in \mathrm{E}_i) := \frac{n(\mathrm{X} \in \mathrm{E}_i)}{n}$$

est la proportion d'individus $\lambda \in \mathscr{P}$ tels que

$$\mathrm{X}(\lambda) \in \mathrm{E}_i.$$

Ces proportions, également appelées fréquences, forment la distribution de la variable X *dans la population* $\mathscr{P}$.

Remarque. La somme des proportions des classes $\mathrm{E}_1,\dots,\mathrm{E}_m$ dans la population $\mathscr{P}$ vaut 1 :

$$\sum_{i=1}^{m} p(\mathrm{X} \in \mathrm{E}_i) = \frac{1}{n} \sum_{i=1}^{m} n(\mathrm{X} \in \mathrm{E}_i) = 1.$$

Définition 1.1.12 *Pour tout $i \in \{1,\dots,m\}$,*

$$p(\mathrm{X} \leqslant x_i) := \sum_{j=1}^{i} p(\mathrm{X} \in \mathrm{E}_i)$$

est la proportion d'individus $\lambda \in \mathscr{P}$ tels que $\mathrm{X}(\lambda) \leqslant x_i$.

Par convention, $p(\mathrm{X} \leqslant x_0) := 0$.

En pratique, la série statistique (i.e. la donnée des classes et des effectifs associés dans la population $\mathscr{P}$) est présentée dans un tableau :

X	$[a, x_1[$	$[x_1, x_2[$	…	$[x_{m-1}, b]$
Effectifs	$n(\mathrm{X} \in \mathrm{E}_1)$	$n(\mathrm{X} \in \mathrm{E}_2)$	…	$n(\mathrm{X} \in \mathrm{E}_m)$
Proportions	$p(\mathrm{X} \in \mathrm{E}_1)$	$p(\mathrm{X} \in \mathrm{E}_2)$	…	$p(\mathrm{X} \in \mathrm{E}_m)$

Définition 1.1.13 *Pour tout $i \in \{1,...,m\}$,*

$$f_{\mathrm{X}}(x_i) := \frac{p(\mathrm{X} \in \mathrm{E}_i)}{x_i - x_{i-1}}$$

est la densité de proportion de la variable X*, pour la classe* E_i*, dans la population* $\mathscr{P}$.