P7

STATISTIQUES UNIVARIÉES

Introduction : qu'est-ce que la statistique ?

La **statistique** est la branche des mathématiques qui consiste à la collecte, au classement, à l'analyse et à l'interprétation de données afin d'en tirer des conclusions et de faire des prévisions.

Le mot statistique désigne à la fois un ensemble de données d'observations et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation.

Exemples d'applications : études quantitatives de marchés, prévisions économétriques, analyse de la consommation des ménages, taxation des primes d'assurances et de franchises, gestion de portefeuille, évaluation d'actifs financiers, ...

Objectif d'une étude statistique

En général, les données collectées sont entâchées d'**incertitudes** et présentent des **variations** pour plusieurs raisons :

- le déroulement des phénomènes observés n'est pas prévisible à l'avance avec certitude ;
- toute mesure est entâchée d'erreur ;
- seuls quelques individus sont observés ;
- ...

L'objectif est alors maîtriser au mieux cette incertitude pour extraire des informations utiles des données.

Deux classes de méthodes statistiques

Statistique descriptive : elle a pour but de résumer l'information contenue dans les données de façon synthétique et efficace par :

- Représentations graphiques ;
- Indicateurs de position et de dispersion.

Elle permet de dégager les caractéristiques essentielles du phénoméne étudié et de suggérer des hypothéses pour une étude ultérieure plus poussée. Les probabilités n'ont ici qu'un rôle mineur.

Statistique inférentielle : elle a pour but de faire des prévisions et de prendre des décisions au vu des observations par :

- Estimation paramétrique ;
- Intervalles de confiance, tests d'hypothèses.

Cela nécessite de définir des modéles probabilistes du phénoméne aléatoire et demande donc de bonnes connaissances en probabilités. Cette partie sera étudiée en deuxième année.

Statistique VS probabilités

La statistique repose sur l'observation de phénoménes concrets et utilise les probabilités comme outils d'analyse et de généralisation.

La théorie des probabilités permet de modéliser efficacement certains phénoménes aléatoires et d'en faire l'étude théorique.

I Vocabulaire général

I.1 Population, individu, échantillon

Une **population** est l'ensemble des éléments auxquels se rapportent les données étudiées. En statistique, le terme "population" s'applique à des ensembles de toute nature : étudiants d'une académie, production d'une usine, poissons d'une rivière, entreprises d'un secteur donné ...
Une population doit être bien définie. Sa définition est importante car elle conditionne l'homogénéité des unités observées et aussi la fiabilité des résultats.

Dans une population donnée, chaque élément est appelé **individu**.

Lorsqu'on veut étudier les données relatives aux caractéristiques d'un ensemble d'individus ou d'objets il est difficile d'observer toutes les données lorsque leurs nombres sont élevés. Au lieu d'examiner l'ensemble qu'on appelle population on examine un nombre restreint qu'on appelle **échantillon**.

Les observations obtenues sur une population ou sur un échantillon constituent un ensemble de données auxquelles s'appliquent les méthodes de la statistique descriptive dont le but est de décrire le plus complétement et le plus simplement l'ensemble des observations qu'elles soient relatives à toute la population ou seulement à un sous-ensemble.

Exemple 1. On souhaite étudier les étudiants de ECG.

- la population est l'ensemble des étudiants en ECG en France ;
- un échantillon est, par exemple, la classe ECG1 du lycée Dupuy de Lôme de Lorient ;
- un individu est un étudiant.

I.2 Variable statistique

Pour étudier une population, le statisticien ne retient que les caractères qui l'intéressent. Un caractère étant une **variable statistique** qui caractérise les individus de cette population. Les valeurs possibles d'un caractère sont appelées ses modalités.

Exemple 2. Dans le cas des étudiants en ECG, les variables étudiées peuvent être : le sexe, la taille, le nombre de cafés consommés, etc.

Il y a plusieurs types de variables :

variable qualitative : s'exprimant par l'appartenance à une modalité.
Ex : homme/femme, une couleur, etc.

variable quantitative : s'exprimant par des nombres réels.
Ex : la taille des individus ou les résultats d'un examen.
Dans les variables quantitatives, on peut encore faire deux catégories.

variables quantitatives discrètes : lorsque les valeurs possibles de la variable sont *discrètes*, c'est-à-dire isolées les unes des autres. Pour nous, cela voudra dire, que les valeurs sont entières ou alors en nombre fini.
Ex : nombre de personne dans une famille.

variables quantitatives continues : lorsque toutes les valeurs d'un intervalle de $\mathbb{R}$ sont possibles.
Ex : la taille.

Nous n'étudierons que des variables quantitatives discrètes.

I.3 Série statistique

Lorsque l'étude statistique d'une population concerne un seul caractère (variable) on parle d'une **série statistique univariée**.

Dans la suite, on souhaite étudier une variable statistique x prenant ses valeurs dans Ω sur une population $\mathcal{P}$.
Un échantillon de la population sera un sous ensemble $\mathcal{E}$ de $\mathcal{P}$, c'est-à-dire un ensemble de n individus.
On suppose que Ω est un ensemble fini :

$$\Omega = \{x_1, x_2, \ldots, x_p\}.$$

Ici, $x_1 < x_2 < \cdots < x_p$ et p est le nombre de valeurs possibles pour le caractère étudié x.

Définition

Notons, pour tout $i \in [\![1, p]\!]$, n_i le nombre d'individus de l'échantillon $\mathcal{E}$ pour lesquels x prend la valeur x_i.
L'ensemble des couples $(x_i, n_i)_{i\in[\![1,p]\!]}$ est appelé **une série statistique univariée** associée à l'échantillon $\mathcal{E}$.

Remarque. L'**effectif total** est le nombre d'individus constituant $\mathcal{E}$. Il vaut

$$N = n_1 + n_2 + \cdots + n_p = \sum_{i=1}^{p} n_i.$$

Exemple 3. Lors d'une enquête on a interrogé 50 employés d'une entreprise afin de connaître le nombre de personnes vivant avec eux dans leur foyer (variable x).
Les données brutes sont :

0 3 1 4 3 0 4 1 3 1 5 2 4 2 3 3 2 5 5 2 4 2 2 2 4 1
1 2 3 5 1 0 3 3 4 5 1 2 1 2 3 2 2 2 4 0 3 0 2 2

Les données ordonnées sont :

0 0 0 0 0 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 5 5 5 5 5

On a ici :

$$\Omega = \{0, 1, 2, 3, 4, 5\}$$

La série statistique associée est

$$\{(0, 5)\ ,\ (1, 8)\ ,\ (2, 15)\ ,\ (3, 10)\ ,\ (4, 7)\ ,\ (5, 5)\}\,.$$

Il plus parlant de représenter ceci dans un tableau.

Valeur x_i	0	1	2	3	4	5
Effectif n_i	5	8	15	10	7	5

L'effectif total de l'échantillon est $N = 50$.

II Étude d'une variable quantitative discrète

II.1 Fréquences, fréquences cumulées

Définition

Soit x une variable statistique prenant les valeurs $x_1 < x_2 < \cdots < x_p$ et $\mathcal{E}$ un échantillon de N individus.

- L'**effectif** de x_i est le nombre n_i de fois où la valeur x_i est prise dans l'échantillon.
- La **fréquence** de x_i est
$$f_i = \frac{n_i}{N}.$$
- L'**effectif cumulé** jusqu'à x_i est
$$n_1 + n_2 + \cdots + n_i = \sum_{j=1}^{i} n_j.$$
- La **fréquence cumulée** jusqu'à x_i est
$$F_i = f_1 + f_2 + \cdots + f_i = \sum_{j=1}^{i} f_j = \frac{1}{N}\sum_{j=1}^{i} n_j.$$

Exemple 4. Compléter le tableau de la série statistique précédente pour y faire apparaître les fréquences.

Valeur x_i	0	1	2	3	4	5	Total
Effectif n_i	5	8	15	10	7	5	50
Fréquence f_i	$\frac{5}{50} = 0,1$	0.16	0.3	0.2	0.14	0.1	1
Fréquence cumulée F_i	0.1	0.26	0.56	0.76	0.90	1	–

Diagramme en barre des effectifs

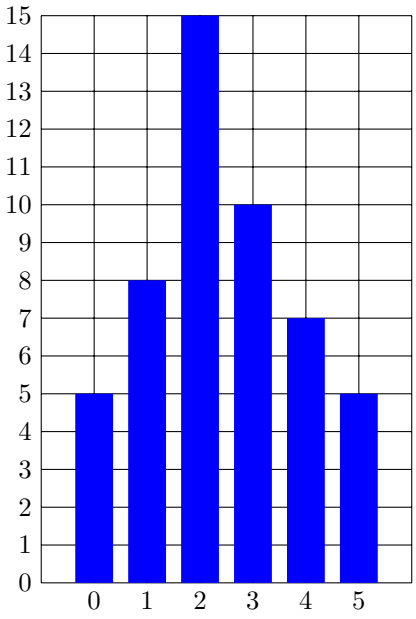


Diagramme en barre des fréquences

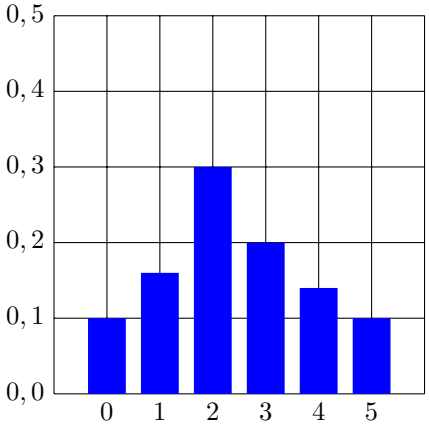
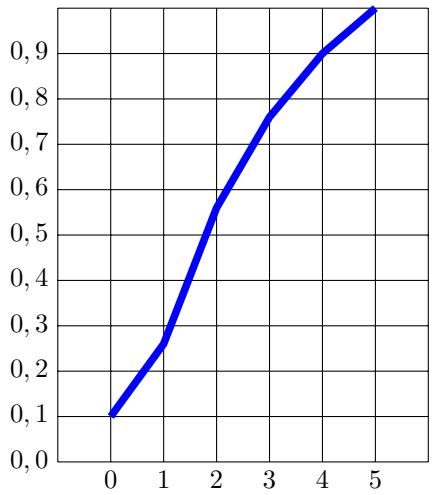


Diagramme des fréquences cumulées



Remarque. Il est aussi possible de représenter les fréquences dans un diagramme en secteurs (camembert). Ce format sera plutôt adapté aux variables statistiques qualitatives.

II.2 Indicateurs de tendance centrale

a) Mode

Définition

Le **mode** d'une série statistique est la valeur ayant le plus grand effectif.

Exemple 5. Le mode de la série précédente est : 2

b) Moyenne

Définition

Moyenne La **moyenne** d'une série statistique x :

Valeur x_i	x_1	x_2	...	x_p	Total
Effectif n_i	n_1	n_2	...	n_p	N

est

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

où $f_i = \dfrac{n_i}{N}$ est la fréquence de la valeur x_i.

Remarque. $\bar{x} = \dfrac{n_1x_1 + n_2x_2 + \ldots n_px_p}{N}$.

Exemple 6. La moyenne de la série

Valeur x_i	0	1	2	3	4	5	Total
Effectif n_i	5	8	15	10	7	5	50

est :

$$\bar{x} = \frac{1}{50}(5 \times 0 + 8 \times 1 + 15 \times 2 + 10 \times 3 + 7 \times 4 + 5 \times 5) = 2,42.$$

Dans l'entreprise, les employés habitent en moyenne avec 2,42 personnes.

Proposition (Transformation affine)

Soit x une variable statistique. Soient a et b deux réels et soit $y = ax + b$ une autre variable statistique, obtenue à partir de x par transformation affine. Alors $\boxed{\bar{y} = a\bar{x} + b.}$

Proposition (Associativité)

On suppose que l'échantillon $\mathcal{E}$ est séparé en deux échantillons $\mathcal{E}_a$ et $\mathcal{E}_b$ ($\mathcal{E} = \mathcal{E}_a \cup \mathcal{E}_b$).
On note N_a l'effectif de $\mathcal{E}_a$ et N_b l'effectif de $\mathcal{E}_b$.
On note $\bar{x}_a$ la moyenne de x sur l'échantillon $\mathcal{E}_a$ et $\bar{x}_b$ la moyenne de x sur l'échantillon $\mathcal{E}_b$.
Alors, la moyenne globale $\bar{x}$ sur l'échantillon total est :

$$\bar{x} = \frac{N_a\bar{x}_a + N_b\bar{x}_b}{N_a + N_b}.$$

Remarque. Cette formule permet :

- de calculer la moyenne globale sur une population constituée de plusieurs groupes que l'on a déjà étudiés ;
- de recalculer facilement la moyenne en cas d'ajout d'une observation.

Exemple 7. Le salaire moyen des huits 8 employés d'une entreprise est 30 000 euros.
L'entreprise recrute deux nouveaux employés qualifiés dont le revenu moyen est de 100 000 euros.
Calculer le nouveau revenu moyen de l'entreprise.

Échantillon A : 8 employés, $\overline{x_a} = 30000$
Échantillon B : 2 employés, $\overline{x_b} = 100000$
Alors

$$\bar{x} = \frac{8 \times 30000 + 2 \times 100000}{8 + 2} = 44000$$

. Le nouveau revenu moyen est de 44 000 euros.

Remarque. La moyenne est sensible à la présence de valeurs aberrantes.
Exemple :

- série 1 : (1,1,2,2,2,2,3,3) $\rightarrow \bar{x} = 2$
- série 2 : (1,1,2,2,2,2,3,300) $\rightarrow \bar{x} = 39,125$

c) Médiane

Définition (Médiane)

La médiane d'une variable statistique x sur un échantillon de données rangées dans l'ordre croissant est la valeur de x séparant les données de la série en deux sous-ensembles de tailles égales.
Plus précisément, notant N la taille de l'échantillon,

- si N est impair, la médiane est la valeur de rang $\dfrac{N+1}{2}$;
- si N est pair, il y a deux valeurs centrales et la médiane est la moyenne entre ces deux valeurs.

Exemple 8.

- série 1 : (0,0,1,1,2,2,4). Ici $N = 7$, impair. Le rang du milieu est $\dfrac{N+1}{2} = 4$. La médiane est donc la quatrième valeur, soit 1.
- série 2 : (0,0,1,1,2,2,3,4). Ici, $N = 8$, pair. Il y a deux rangs centraux : les rangs 4 et 5. La médiane est la moyenne de la quatrième et de la cinquième valeur, soit $\dfrac{1+2}{2} = 1,5$.

Remarque. Contrairement à la moyenne, la médiane n'est pas sensible à la présence de valeurs aberrantes.
Exemple :

- série 1 : (1,1,2,2,2,2,3,3) $\rightarrow$ médiane = 2
- série 2 : (1,1,2,2,2,2,3,300) $\rightarrow$ médiane = 2

Proposition (Transformation affine)

Soit x une variable statistique. Soient a et b deux réels et soit $y = ax + b$ une autre variable statistique, obtenue à partir de x par transformation affine.
Notons m_x la médiane de x et m_y la médiane de y. Alors

$$m_y = am_x + b.$$

II.3 Indicateurs de dispersion

a) Étendue

Définition

L'**étendue** d'une série statistique est la différence entre la plus grande valeur et la plus petite.

Exemple 9.
- série 1 : (1,1,2,2,2,2,3,3) → étendue $= 3 - 1 = 2$
- série 2 : (1,1,2,2,2,2,3,300) → étendue $= 300 - 1 = 299$

L'étendue est donc sensible à la présence de valeurs aberrantes.

b) Quantiles

La notion de quantile généralise celle de médiane.

Définition (Quantiles)

Le **quantile d'ordre p** ($p \in [0,1]$) d'une variable statistique x est la valeur de x qui permet de scinder la population étudiée en deux sous-populations dont les effectifs respectifs sont égaux à p et $1-p$ de l'effectif total.

Exemple 10. Les **quartiles** sont les quantiles d'ordre $\frac{1}{4}$, $\frac{1}{2}$ et $\frac{3}{4}$.
Le premier quartile Q_1 est la plus petite valeur de la série telle que (au moins) $\frac{1}{4}$ des valeurs prises par x sont inférieures ou égales à Q_1.
Le second quartile est simplement la médiane.
Le troisième quartile Q_3 est la plus petite valeur de la série telle que (au moins) $\frac{3}{4}$ des valeurs prises par x sont inférieures ou égale à Q_3.

Voyons cela sur un exemple :

Valeur	0	1	2	5	6	7	10	Total
Effectif	4	2	3	1	5	4	1	20
Fréquence cumulée	0.2	0.3	0.45	0.5	0.75	0.95	1	

$Q_1 = 1$ car au moins $1/4$ $(= 0.25)$ des données sont inférieures ou égales à 1, ce qui n'est pas le cas de 0. On regarde ici le moment où on passe 0.25.

Médiane $= \dfrac{5+6}{2} = 5.5$ (moyenne de la dixième et onzième valeur).

$Q_3 = 6$ (première valeur où la fréquence cumulée est supérieure ou égale à 0.75)

Définition (Écart interquartile)

L'**écart interquartile** est $Q_3 - Q - 1$.

Remarque. Puisqu'il ne tient compte que de la moitié centrale de la série, l'écart interquartile n'est pas sensible aux valeurs aberrantes.

Remarque. Les **déciles** sont les quantiles d'ordre 0.1, 0.2, ..., 0.9. Ils partagent la population en 10 parts.

c) Variance et écart-type

Définition (Variance)

La **variance** d'une série statistique x :

Valeur x_i	x_1	x_2	...	x_p	Total
Effectif n_i	n_1	n_2	...	n_p	N

est

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2 = \sum_{i=1}^{p} f_i(x_i - \bar{x})^2$$

où $\bar{x}$ est la moyenne de x et $f_i = \dfrac{n_i}{N}$ est la fréquence de la vakeur x_i.

Remarque. La variance est la moyenne des écarts à la moyenne au carré.

Proposition

1. La variance est un réel positif ou nul.
2. **Formule de Koenig**
$$s_x^2 = \left(\sum_{i=1}^{p} f_i(x_i)^2\right) - (\bar{x})^2.$$

Définition (Écart-type)

L'**écart-type** de x est :

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

Remarque. Le choix de notation est bien sûr cohérent.

Exemple 11. Considérons la série statistique suivante :

Valeur x_i	0	1	2	3	4	5	Total
Effectif n_i	5	8	15	10	7	5	50

Sa moyenne est $\bar{x} = 2,42$ (déjà calculée)
Variance : avec la formule de Huygens

$$s_x^2 = \left(\frac{5}{50}\times 0^2 + \frac{8}{50}\times 1^2 + \frac{15}{50}\times 2^2 + \frac{10}{50}\times 3^2 + \frac{7}{50}\times 4^2 + \frac{5}{50}\times 5^2\right) - (2,42)^2 \approx 2,04.$$

Écart-type : $s_x = \sqrt{s_x^2} \approx 1,43$.

III Comparer deux séries statistiques

III.1 Boîte à moustaches : médiane – écart interquartile – étendue

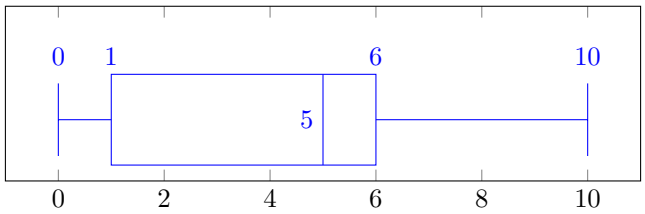
Une **boîte à moustaches** (box plot en anglais) permet de représenter le minimum, Q_1, la médiane, Q_3 et le maximum.

Exemple 12.

Série 1

Valeur	0	1	2	5	6	7	10	Total
Effectif	4	2	3	2	4	4	1	

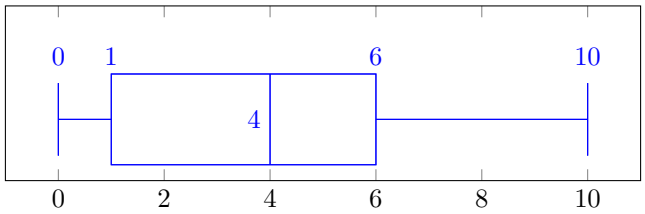
$min = 0$ $Q_1 = 1$ médiane $= 5$ $Q_3 = 6$ $max = 10$



Série 2

Valeur	0	1	3	5	6	8	10	Total
Effectif	2	3	5	3	4	1	2	

$min = 0$ $Q_1 = 1$ médiane $= 4$ $Q_3 = 6$ $max = 10$

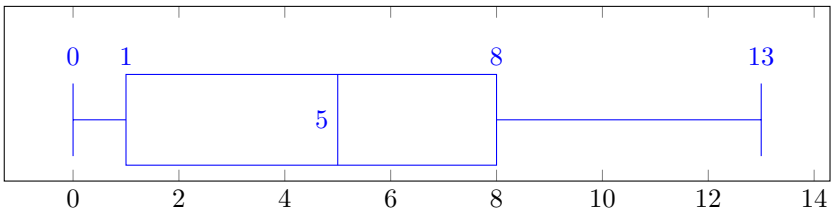


Seul changement avec la série 1 : la médiane. Ici les valeurs dans l'écart interquartile sont mieux distribuées.

Série 3

Valeur	0	1	2	5	7	8	10	13	Total
Effectif	2	4	3	2	3	2	1	3	

$min = 0$ $Q_1 = 1$ médiane $= 5$ $Q_3 = 8$ $max = 13$



Ici, la série est plus étendue.

III.2 Moyenne – écart-type

Une autre façon de comparer les séries statistiques est de comparer leurs (moyenne, écart-type).

Exemple 13.

Série 1

Valeur	0	1	2	5	6	7	10	Total
Effectif	4	2	3	2	4	4	1	

moyenne $= 4$
écart-type $\approx 3,03$

Série 2

Valeur	0	1	3	5	6	8	10	Total
Effectif	2	3	5	3	4	1	2	

moyenne $= 4{,}25$
écart-type $\approx 2,91$

Série 3

Valeur	0	1	2	5	7	8	10	13	Total
Effectif	2	4	3	2	3	2	1	3	

moyenne $= 5{,}3$
écart-type $\approx 4,39$

En termes de position, on compare les moyennes et on en déduit que la série 3 comportent les valeurs les plus grandes, suivie de la série 2.

Concernant la dispersion, on va ici comparer les écarts-types. Les valeurs de la série 2 sont les plus resserrées et les valeurs de la série 3 sont les plus dispersées.

Conclusion :
Les statistiques univariées permettent de décrire des séries statistiques et de comparer la position et la dispersion. Cela reste très basique. Les statistiques bivariées (au programme de ECG2) permettrons elles de croiser les données, de déterminer des corrélations, etc. en ECG2, vous aurez également un aperçu de la statistique iinférentielle permettant de faire des estimations, des prédictions à partir de séries statistiques.