

Chapitre 16 : Statistique.

*Il existe trois types de mensonges :
les mensonges simples, les sacrés mensonges et les statistiques*
Mark Twain

I Vocabulaire, représentation graphique

L'objectif de la statistique descriptive est d'être capable de donner une information synthétique sur la répartition d'un caractère au sein d'une population. Il s'agit donc de donner des grandeurs caractéristiques pertinentes pour avoir un bon aperçu de la répartition dudit caractère.

On se retrouvera, en statistique, face à deux types de situations différentes mais menant à des traitements mathématiques similaires :

- La répétition d'une expérience dont le résultat, du fait d'un aléa dans l'expérience ou des incertitudes de mesure, n'est pas toujours le même. C'est la situation qui se présente le plus souvent dans les sciences expérimentales (SVT, Physique-Chimie)
- La considération d'un échantillon dans une population. On recueille des informations sur des personnes tirées au hasard dans le but d'obtenir le sentiment général (e.g. sondage), cette approche est plus proche de celle pratiquée par les économistes.

Définition 1

Une population est un ensemble fini dont les éléments sont appelés des individus. Le nombre d'individus d'une population est appelé sa taille.

Très souvent on ne mène pas notre étude sur la population entière mais sur une sous-partie que l'on espère représentative

Définition 2

Un sous-ensemble d'une population est appelé un échantillon de cette population

Définition 3

Un caractère x de la population est une donnée qualitative ou quantitative attachée à chaque individu de la population. On notera x_i la valeur du caractère x pour un individu i . La donnée des valeurs d'un caractère pour les individus d'un échantillon de taille n est une série statistique de taille n . Une série statistique associée au caractère x sera notée (x_1, \dots, x_n) .

- Un caractère est dit quantitatif s'il prend des valeurs quantifiables, souvent des réels mais éventuellement des p -uplets ou des matrices.
- Un caractère est dit qualitatif s'il correspond à une propriété qui ne se quantifie pas (e.g. couleur des cheveux, opinion politique, etc.)

Définition 4

On appelle modalités d'un caractère les valeurs possibles qu'il peut prendre

Exemple 1. Le tableau ci-dessous regroupe les diamètres en cm de 48 pièces prélevées dans la production d'une machine.

1,19	1,26	1,23	1,20	1,22	1,24	1,20	1,24
1,22	1,20	1,21	1,19	1,21	1,22	1,19	1,20
1,21	1,21	1,22	1,21	1,23	1,22	1,21	1,24
1,25	1,23	1,22	1,19	1,20	1,26	1,24	1,25
1,23	1,26	1,25	1,25	1,21	1,22	1,25	1,24
1,23	1,22	1,24	1,24	1,25	1,23	1,25	1,22

Remarque 1.
Population ne signifie pas que l'on considère des personnes. Si vous réalisez plusieurs titrages pour votre T.I.P.E. vous aurez une population de concentrations

Il s'agit d'un échantillon statistique de taille 48 dans la population des pièces fabriquées par la machine. Le caractère étudié est le diamètre de la pièce en centimètres. Les modalités sont 1.19, 1.20, 1.21, 1.22, 1.23, 1.24, 1.25 et 1.26.

A Effectifs, regroupements en classes

Une série statistique peut être donnée sous plusieurs formes.

On peut simplement donner la valeur du caractère pour chaque individu

Par exemple, une série de masse des personnes

Individu	1	2	...	n
Masse	62	80	...	74

Si la taille de l'échantillon est trop grande on préférera donner les nombres d'individus associés à chaque modalité. On appelle cela l'effectif associé à ladite modalité

Par exemple,

Nombre d'enfants	0	1	...	5
Effectif	20	31	...	3

Parfois le nombre de modalités est trop grand pour ce faire, voire infini pour des modalités dites continues (e.g. à valeurs réelles).

Il est alors pertinent de regrouper les modalités en classes disjointes, souvent on regroupe les modalités en des intervalles qui ne sont pas forcément de tailles égales.

Inversement les modalités non regroupées en classe sont dites ponctuelles. Il peut également arriver que nos modalités comportent des classes et des modalités ponctuelles

Par exemple, lorsque l'on étudie la démographie urbaine française on va regrouper les communes en classes selon leur nombre d'habitants :

- Hameaux de 1 à 99 habitants,
- Village de 100 à 1999 habitants,
- Ville de 200 à 99999 habitants,
- Agglomération à partir de 100000 habitants

Les salaires dans une entreprise peuvent aussi donner lieu à un regroupement en classes, par exemple

Salaire mensuel en euros	1173]1173, 1300]]1300, 1500]	...]4000, +∞[
Effectifs					

Définition 5

Considérons une série statistique de taille n admettant une nombre fini de modalités (ce qui est toujours possible quitte à faire des regroupements de classes) notées a_1, \dots, a_n .

- Pour $j \in \llbracket 1, p \rrbracket$, on définit l'effectif n_j associé à la valeur a_j comme étant le nombre d'individus i pour lesquels $x_i = a_j$.
- Pour $j \in \llbracket 1, p \rrbracket$, on définit l'effectif n_j associé à la valeur a_j comme étant la proportion d'individus i pour lesquels $x_i = a_j$, c'est-à-dire $f_j = \frac{n_j}{n}$

Proposition 1

On a

$$\sum_{j=1}^p n_j = n \quad \text{et} \quad \sum_{j=1}^p f_j = 1$$

Dans le cas où les caractères étudiés sont des réels (ils peuvent donc être ordonnés) on va introduire les effectifs cumulés croissants et les fréquences cumulées croissantes

Définition 6

On suppose ici les modalités a_1, \dots, a_p rangées dans l'ordre croissant ($a_1 < a_2 < \dots < a_p$). Pour des intervalles $]a, b]$ et $]c, d]$ cela correspond à $a < b \leq c < d$.

Soit $j \in \llbracket 1, p \rrbracket$.

- On définit l'effectif cumulé croissant associé à la modalité a_j comme le nombre d'individus i pour lesquels $x_i \leq a_j$, c'est-à-dire

$$n_j^c = \text{Card}(\{i \in \llbracket 1, n \rrbracket, x_i \leq a_j\})$$

- On définit la fréquence cumulée (croissante) associée à la modalité a_j comme la proportion d'individus i pour lesquels $x_i \leq a_j$, c'est-à-dire

$$f_j^c = \frac{n_j^c}{n}$$

Proposition 2

On a alors

$$\forall j \in \llbracket 1, p \rrbracket, \quad n_j^c = \sum_{k=1}^j n_k, \quad \text{et} \quad f_j^c = \sum_{k=1}^j f_k$$

et

$$\begin{aligned} n_1^c &\leq n_2^c \leq \dots \leq n_p^c = n \\ f_1^c &\leq f_2^c \leq \dots \leq f_p^c = 1 \end{aligned}$$

B Représentation graphique

Comme un bon dessin vaut parfois mieux qu'un long discours on peut vouloir donner une représentation graphique d'une série statistique. Il y a plusieurs manières usuelles de ce faire que l'on adaptera à la situation à illustrer

- Diagramme en bâtons

Pour établir un diagramme en bâtons on va tracer pour chaque modalité une bâton (un rectangle long et fin) centré en a_j et de hauteur f_j ou n_j (le choix de tracer en fonction des effectifs ou des fréquences se fera pour des questions de renormalisation)

Ce type de graphique est adapté aux données ponctuelles et aux données qualitatives.

- Histogramme

Ce type de graphique n'est utilisé que pour des données quantitatives regroupées en classes. Pour chaque classe on va alors tracer un rectangle dont la largeur vaut l'amplitude de l'intervalle et dont l'aire est proportionnelle à la fréquence ou à l'effectif de la classe.

- Polygone des fréquences cumulées croissantes

La procédure est différente pour des modalités ponctuelles ou regroupées par classes.

- Pour des modalités ponctuelles, on place les points A_i de coordonnées (a_i, f_i^c) . On va alors relier les points $A_i(a_i, f_i^c)$ à $B_i(a_{i+1}, f_i^c)$ puis B_i à $A_{i+1}(a_{i+1}, f_{i+1}^c)$. Schématiquement on trace un trait horizontal puis un trait vertical.
- Pour des modalités regroupées en classe $[\alpha_i, \alpha_{i+1}]$ on place les points A_i de coordonnées (α_{i+1}, f_i^c) . (à la droite de l'intervalle donc). On relie ensuite simplement les points A_i par une ligne brisée.

Pour $c \in \mathbb{R}$ on va alors pouvoir lire simplement la proportion d'individus i pour lesquels $x_i \leq c$ en lisant l'ordonnée du point de la courbe d'abscisse c .

Exemple 2. — On considère la série :

x_i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
n_i	3	5	6	5	6	7	7	10	13	20	25	21	23	12	10	5	7	5	3	2	1	0

Et la même regroupée en classe :

Remarque 2. C'est l'aire du rectangle qui est importante, pas sa hauteur

Remarque 3. Cette représentation est pertinente si les individus sont répartis uniformément au sein de la classe. Si vous avez des raisons de penser que ce n'est pas le cas, alors il faut séparer votre classe en plusieurs classes

x_i	$[0, 5[$	$[5, 10[$	$[10, 15[$	$[15, 20]$
n_i	25	57	91	23

On obtient les diagrammes en bâtons et histogramme.

Figure 12.1 – Diagramme en bâtons

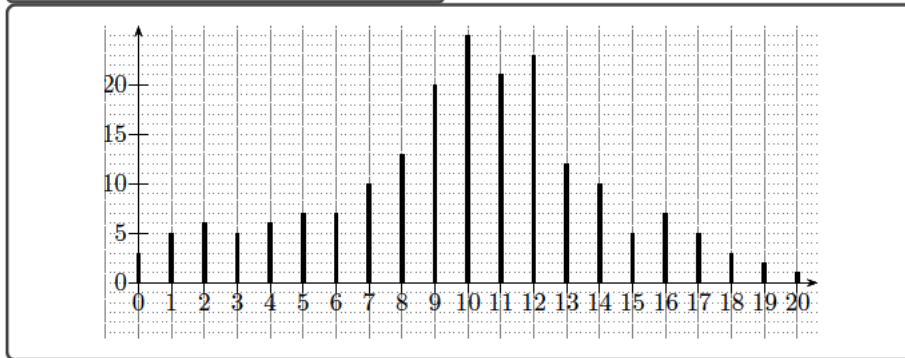
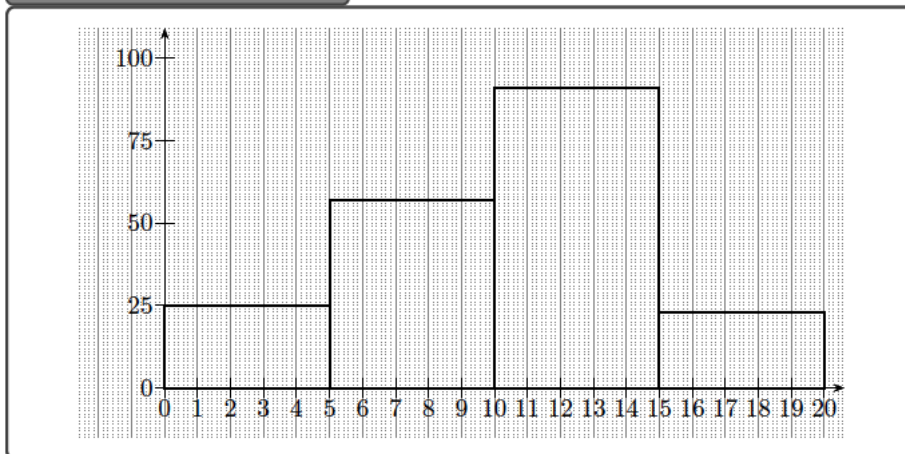
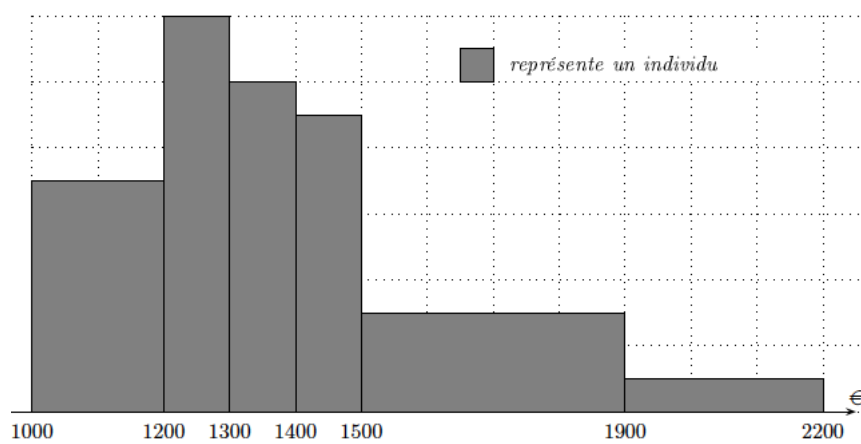


Figure 12.2 – Histogramme



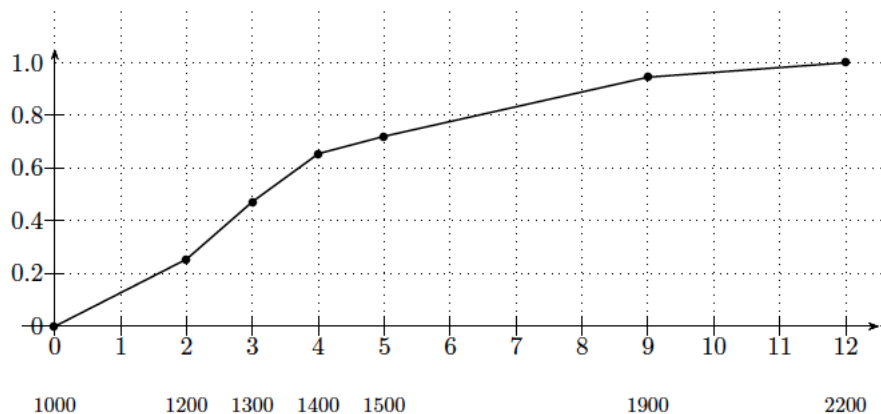
- Une étude portant sur les salaires mensuels des employés en CDI à temps complet d'une entreprise a permis d'établir l'histogramme ci-dessous.



On peut alors retrouver la série statistique associée

x_i	$[1000, 1200[$	$[1200, 1300[$	$[1300, 1400[$	$[1400, 1500[$	$[1500, 1900[$	$[1900, 2200]$
n_i	28	24	20	18	24	6
n_i^c	28	52	72	80	104	110

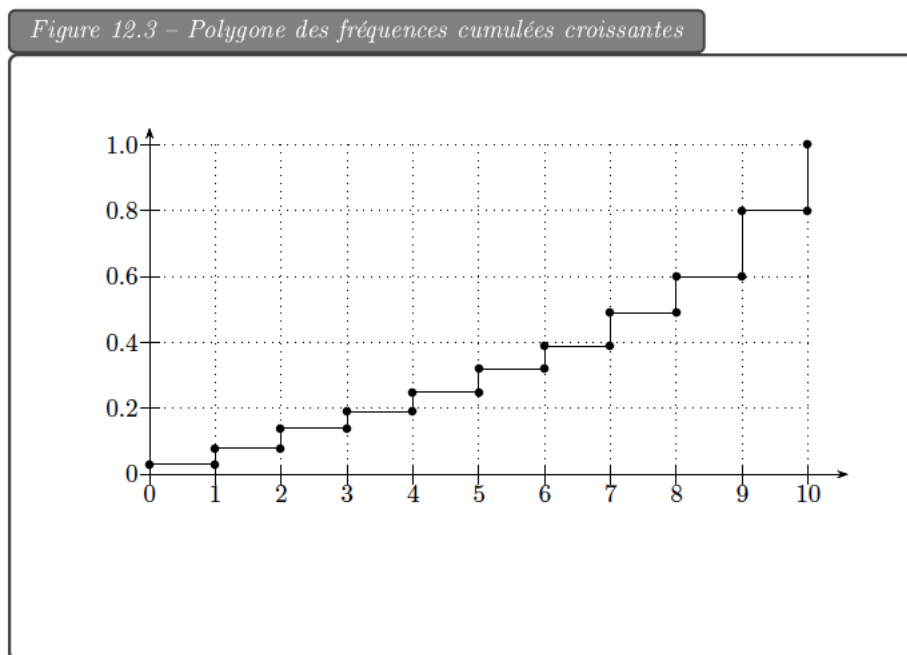
D'où le polygone des fréquences cumulées croissantes.



— On considère la série :

Caractère	0	1	2	3	4	5	6	7	8	9	10
Effectif	3	5	6	5	6	7	7	10	11	20	20
Fréquence	0.015	0.025	0.03	0.025	0.03	0.035	0.035	0.05	0.065	0.1	0.125
Effectif cumulé	3	8	14	19	25	32	39	49	60	80	100
Fréquence cumulée	0.03	0.08	0.14	0.19	0.25	0.32	0.39	0.49	0.6	0.8	1

On obtient la courbe de fréquences cumulées



II Caractéristiques de position

L'idée des caractéristiques de position est de donner une grandeur dont on espère qu'elle résume bien la série statistique.

A Mode(s)

Définition 7

On appelle mode de la série statistique x toute modalité de x dont l'effectif est maximal parmi les effectifs de toutes les modalités.

Lorsque les modes correspondent à des classes on appelle alors classe modale la classe dont l'effectif est maximal.

Remarque 5. — Il est possible qu'une série statistique admette plusieurs modes ou classes modales.

— Il est très facile de déterminer le mode mais son utilité est pratiquement nulle.

Remarque 4. Si vos classes sont de tailles différentes alors la classe modale n'est pas forcément la classe qui a le « plus haut » rectangle dans l'histogramme

B Moyenne

Définition 8

Soit $x = (x_1, \dots, x_n)$ une série statistique quantitative. La moyenne de la série, notée \bar{x} , est définie par

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Proposition 3

Si les modalités (a_1, \dots, a_p) sont ponctuelles alors

$$\bar{x} = \frac{1}{n} \sum_{j=1}^p n_j a_j = \sum_{j=1}^p f_j a_j$$

Quand on travaille avec des données regroupées par classes cette définition n'est pas utilisable. Dans ce situation on va alors considérer que les valeurs sont uniformément réparties dans les intervalles et prendre pour moyenne de la série la moyenne des milieux des intervalles pondérés par les effectifs

Définition 9

Dans la situation où les modalités $(a_j)_{j \in \llbracket 1, p \rrbracket}$ correspondent à des intervalles $[b_j, c_j[$ on définit alors la moyenne par

$$\bar{x} = \frac{1}{n} \sum_{j=1}^p n_j \left(\frac{b_j + c_j}{2} \right) = \sum_{j=1}^p f_j \left(\frac{b_j + c_j}{2} \right)$$

Proposition 4

Soit x et y deux séries statistiques quantitatives dont les modalités sont à valeurs dans le même ensemble et d'effectifs respectifs n et m .

— Soit $(a, b) \in \mathbb{R}^2$ et soit u la série statistique définie par

$$\forall i \in \llbracket 1, n \rrbracket, \quad u_i = ax_i + b$$

Alors $\bar{u} = a\bar{x} + b$

— Soit z la série statistique obtenue en concaténant les séries x et y (on donc $z = (x_1, \dots, x_n, y_1, \dots, y_m)$). Alors

$$\bar{z} = \frac{n\bar{x} + m\bar{y}}{n + m}$$

C Médiane

On suppose ici que nos modalités sont des réels

Définition 10

On appelle médiane d'une série statistique de taille n tout réel m tel que

$$\text{Card}(\{i \in \llbracket 1, n \rrbracket, x_i \leq m\}) \geq \frac{n}{2} \quad \text{et} \quad \text{Card}(\{i \in \llbracket 1, n \rrbracket, x_i \geq m\}) \geq \frac{n}{2}$$

En pratique on prend souvent comme médiane la valeur d'une modalité. Dans ce cas, un individu dont le caractère correspond à la médiane est dit être un individu médian.

Remarque 6. On peut interpréter la moyenne comme le barycentre des modalités affectés de poids proportionnels aux effectifs

Remarque 7. Il s'agit d'une approximation pour la situation où on ne dispose que des effectifs par classe, en général on n'a pas $\frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{j=1}^p n_j \left(\frac{b_j + c_j}{2} \right)$

Remarque 8. Si on voit les moyennes comme des barycentres alors ce résultat est simplement l'associativité du barycentre

Remarque 9. Il y a en général plusieurs médianes

Remarque 10. m partage la série en deux parties comportant chacune au moins la moitié des individus

Proposition 5

Soit x une série statistique de taille n dont les modalités sont données dans l'ordre croissant $a_1 < a_2 < \dots < a_n$.

- Si n est impair alors $a_{\frac{n+1}{2}}$ est une médiane.
- Si n est pair alors tout nombre de l'intervalle $[a_{\frac{n}{2}}, a_{\frac{n}{2}+1}]$ est une médiane.

Remarque 12. — La médiane a , sur la moyenne, l'avantage d'être peu influencée par les valeurs extrêmes. Elle est alors plus représentative que la moyenne lorsque la série comporte des valeurs très grandes ou très petites.

Par exemple, en France en 2014 le salaire moyen mensuel était de 1934 euros pour les femmes et 2389 euros pour les hommes tandis que le salaire médian mensuel était de 1619 euros pour les femmes et 1882 euros pour les hommes. La différence s'explique par le fait que les très hauts salaires, même s'ils sont peu nombreux, tirent la moyenne vers la haut.

- Sur le polygone des fréquences cumulées croissantes on lit une médiane assez simplement. Il suffit de chercher l'abscisse du point de la courbe d'ordonnée $\frac{1}{2}$.

Définition 11

Soit x une série statistique dont les modalités sont regroupées en classes. On définit la médiane de x comme l'abscisse du point de la courbe des fréquences cumulées croissantes d'ordonnée $\frac{1}{2}$.

Remarque 11. En particulier $\frac{a_{\frac{n}{2}} + a_{\frac{n}{2}+1}}{2}$ est une médiane mais ce n'est pas la seule.

III Caractéristique de dispersion

L'idée des caractéristiques de dispersion est de donner une idée de la répartition de la série autour de sa moyenne ou de sa médiane. Les valeurs sont elles relativement proches de la moyenne ou existe-t-il des valeurs très grandes et très petites ?

A Valeurs extrêmes et étendue

Définition 12

Soit x une série statistique de taille n à valeurs réelles. On appelle valeurs extrêmes de la série x les nombres $\min_{i \in [1, n]} x_i$ et $\max_{i \in [1, n]} x_i$. Il s'agit donc des modalités maximales et minimales.

Si on travaille avec des données regroupées en classes on prendra comme valeurs extrémales la borne supérieure de la classe maximale et la borne inférieure de la classe minimale.

On appelle étendue de la série statistique la différence entre la valeur maximale et la valeur minimale.

Remarque 13. L'étendue est facile à déterminer mais ne délivre que très peu d'informations car elle est très fortement affectée par les valeurs extrêmes. Par exemple en France le revenu annuel se situe entre 0 euros et environ 7 millions, ce qui ne nous donne pas vraiment une idée de la répartition des salaires dans la population.

B Quantiles

Définition 13

Soit x une série statistique de taille n à valeurs réelles. On appelle premier quartile de la série x tout réel Q_1 tel que

$$\text{Card}(\{i \in \llbracket 1, n \rrbracket, x_i \leq Q_1\}) \geq \frac{n}{4} \quad \text{et} \quad \text{Card}(\{i \in \llbracket 1, n \rrbracket, x_i \geq Q_1\}) \geq \frac{3n}{4}$$

De même on appelle troisième quartile de la série x tout réel Q_3 tel que

$$\text{Card}(\{i \in \llbracket 1, n \rrbracket, x_i \leq Q_3\}) \geq \frac{3n}{4} \quad \text{et} \quad \text{Card}(\{i \in \llbracket 1, n \rrbracket, x_i \geq Q_3\}) \geq \frac{n}{4}$$

La médiane de la série x est aussi appelée deuxième quartile.

On peut définir la notion de décile de manière similaire

Définition 14

Soit x une série statistique de taille n à valeurs réelles. Pour $j \in \llbracket 1, 9 \rrbracket$, on appelle j -ème décile de la série x tout réel d_j tel que

$$\text{Card}(\{i \in \llbracket 1, n \rrbracket, x_i \leq d_j\}) \geq \frac{jn}{10} \quad \text{et} \quad \text{Card}(\{i \in \llbracket 1, n \rrbracket, x_i \geq d_j\}) \geq \frac{(10-j)n}{10}$$

Ces deux notions se généralisent avec la notion de quantiles

Définition 15

Soit x une série statistique de taille n à valeurs réelles. Soit $t \in [0, 1]$, on appelle t -quantile de la série x tout réel q_t tel que

$$\text{Card}(\{i \in \llbracket 1, n \rrbracket, x_i \leq q_t\}) \geq nt \quad \text{et} \quad \text{Card}(\{i \in \llbracket 1, n \rrbracket, x_i \geq q_t\}) \geq (1-t)n$$

Remarque 17. — Le premier quartile est alors un $\frac{1}{4}$ quantile, etc. Si la série statistique est donnée via un regroupements en classes alors on déterminera un t -quantile en prenant l'abscisse d'un point du polygone des fréquences cumulées croissantes d'ordonnée t .

Définition 16

Soit x une série statistique de taille n à valeurs réelles. On appelle

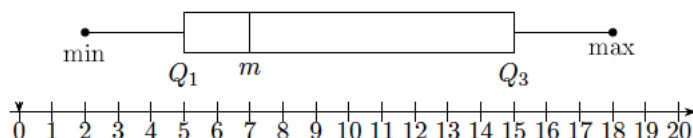
- écart interquartile la différence $Q_3 - Q_1$.
- intervalle interquartile l'intervalle $[Q_1, Q_3]$

De même on appelle

- écart interdécile la différence $d_9 - d_1$.
- intervalle interdécile l'intervalle $[d_1, d_9]$

On peut représenter de manière graphique l'étendue, les quartiles et la médiane en dessinant un diagramme dit « boîte à moustaches » conçu de la manière suivante :

- au centre une boîte allant du premier au troisième quartile, séparée en deux par la médiane ;
- de chaque côté une moustache allant du minimum au premier quartile pour l'une, et du troisième quartile au maximum pour l'autre.



Remarque 14. Q_1 partage la série en deux parties comportant, d'un côté au moins un quart de la population, et de l'autre côté au moins trois quarts de la population

Remarque 15. d_j partage la série en deux parties comportant, d'un côté au moins j dixièmes de la population, et de l'autre côté au moins $10-j$ dixièmes de la population

Remarque 16. Comme pour les quartiles et les déciles il n'y pas unicité du t -quantile

Remarque 18. La moitié au moins de la population se trouve dans l'intervalle interquartile

Remarque 19. Au moins 80% de la population se trouve dans l'intervalle interdécile

C Variance et écart-type

Définition 17

La variance d'une série statistique quantitative à valeurs réelles $x = (x_1, x_2, \dots, x_n)$ est le nombre \mathbb{V}_x défini par :

$$\mathbb{V}_x = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

On a également

$$\mathbb{V}_x = \frac{1}{n} \sum_{j=1}^p n_j (a_j - \bar{x})^2 = \sum_{j=1}^p f_j (a_j - \bar{x})^2$$

\mathbb{V}_x est la moyenne des carrés des écarts à la moyenne

Remarque 20. — La variance est toujours positive.

- Dans le cas d'une série regroupée en classes on prendra pour valeurs a_j les centres des classes.
- Plus la variance est grande et plus la série est « étalée ». Inversement, plus la variance est proche de zéro et plus la série est concentrée autour de sa moyenne.
- La variance ne donne pas d'informations sur une éventuelle asymétrie de la série (par exemple dans le cas de la répartition des salaires en France).
- On trouve parfois la notation s_x^2 pour la variance.

Proposition 6

Soit x une série statistique quantitative à valeurs réelles. On a

$$\mathbb{V}_x = \bar{x^2} - \bar{x}^2$$

Remarque 21. \mathbb{V}_x est donc la moyenne des carrés moins le carré de la moyenne

Démonstration. On a

$$\begin{aligned} \mathbb{V}_x &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (x_i^2 + \bar{x}^2 - 2x_i\bar{x}) \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 + \frac{1}{n} \sum_{i=1}^n \bar{x}^2 - 2\frac{\bar{x}}{n} \sum_{i=1}^n x_i \\ &= \bar{x^2} + \bar{x}^2 - 2\bar{x}^2 \\ &= \bar{x^2} - \bar{x}^2 \end{aligned}$$

□

Remarque 22. — Cette formule n'est pas la définition de la variance mais c'est celle que l'on utilisera la plupart du temps pour calculer la variance

- Certains français (essentiellement dans les milieux des CPGE) appellent cette formule la formule de Koenig-Huygens. Les allemands l'appellent plutôt théorème de translation de Steiner. On trouve aussi l'appellation formule de Leibniz. Pour l'essentiel des mathématiciens il s'agit d'une formule triviale qui ne mérite pas de nom.

Définition 18

On définit l'écart-type d'une série statistique quantitative réelle, noté σ_x comme la racine carrée de la variance

$$\sigma_x = \sqrt{\mathbb{V}_x}$$

Remarque 23. On trouve parfois la notation s_x

Remarque 24. L'intérêt de l'écart-type par rapport à la variance est que l'écart-type s'exprime dans les mêmes unités que les modalités de la série. On pourra alors faire des calculs faisant intervenir modalités, moyenne et écart-type (par exemple dans des situations d'estimation de paramètres ou de test statistique d'hypothèses).

Proposition 7

Soit x une série statistique quantitative réelle, $(a, b) \in \mathbb{R}$ et y la série statistique $y = ax + b$.
On a alors

$$\mathbb{V}_y = a^2 \mathbb{V}_x$$

et donc

$$\sigma_y = |a| \sigma_x$$

Démonstration. On a

$$\begin{aligned} \mathbb{V}_y &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (ax_i + b - (a\bar{x} + b))^2 \\ &= \frac{1}{n} \sum_{i=1}^n a^2 (x_i - \bar{x})^2 \\ &= a^2 \mathbb{V}_x \end{aligned}$$

□

IV Statistique descriptive bivariée

On a vu dans le chapitre précédent diverses manières d'extraire de l'information d'un échantillon statistique. Lorsque l'on ne dispose plus d'un seul mais de plusieurs échantillons statistiques, on peut, au delà de la simple étude des échantillons, étudier les éventuels liens entre eux, c'est l'objet de ce chapitre. Pour des raisons de simplicité on se limitera à deux échantillons.

A Contexte

On va s'intéresser ici à deux caractères quantitatifs d'une même population. On notera n la taille de l'échantillon étudié et (x, y) les deux caractères étudiés. L'observation des valeurs des caractères se traduit par un échantillon d'éléments de \mathbb{R}^2

$$((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n))$$

Notons (a_1, \dots, a_p) les modalités du caractère x (éventuellement des classes) et (b_1, \dots, b_q) les modalités du caractère y (idem).

Le plus souvent on regroupe les individus par modalités, on obtient alors le tableau d'effectifs suivant :

	b_1	\dots	b_j	\dots	b_q	Totaux
a_1	$n_{1,1}$	\dots	$n_{1,j}$	\dots	$n_{1,q}$	$n_{1\blacksquare}$
\vdots	\vdots		\vdots		\vdots	\vdots
a_i	$n_{i,1}$	\dots	$n_{i,j}$	\dots	$n_{i,q}$	$n_{i\blacksquare}$
\vdots	\vdots		\vdots		\vdots	\vdots
a_p	$n_{p,1}$	\dots	$n_{p,j}$	\dots	$n_{p,q}$	$n_{p\blacksquare}$
Totaux	$n_{\blacksquare 1}$	\dots	$n_{\blacksquare j}$	\dots	$n_{\blacksquare q}$	n

$n_{i,j}$ est le cardinal de l'ensemble des individus présentant à la fois les modalités a_i et b_j .

Pour $(i, j) \in \llbracket 1, p \rrbracket \times \llbracket 1, q \rrbracket$, on pose

$$n_{\blacksquare, 1} = \sum_{j=1}^q n_{i,j} \quad \text{et} \quad n_{j, \blacksquare} = \sum_{i=1}^p n_{i,j}$$

On a alors

$$n = \sum_{i=1}^p \sum_{j=1}^q n_{i,j} = \sum_{i=1}^p n_{i,\bullet} = \sum_{j=1}^q n_{\bullet,j}$$

Définition 19

Pour $(i, j) \in \llbracket 1, p \rrbracket \times \llbracket 1, q \rrbracket$ on définit

- Les fréquences des couples de modalités (a_i, b_j) par $f_{i,j} = \frac{n_{i,j}}{n}$.
- La fréquence marginale de la modalité a_i par $f_{i,\bullet} = \frac{n_{i,\bullet}}{n}$
- La fréquence marginale de la modalité a_i par $f_{\bullet,j} = \frac{n_{\bullet,j}}{n}$

On a alors

$$\sum_{i=1}^p \sum_{j=1}^q f_{i,j} = \sum_{i=1}^p f_{i,\bullet} = \sum_{j=1}^q f_{\bullet,j} = 1$$

Définition 20

On appelle nuage point associé à l'échantillon (x, y) le tracé de tous les points de coordonnées $M(x_k, y_k)$ pour $k \in \llbracket 1, n \rrbracket$.

B Indicateurs statistiques

On définit les moyennes et variances de manière similaire au cas univarié.

Définition 21

On définit

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$$

On a encore

$$\bar{x} = \frac{1}{n} \sum_{i=1}^p n_{i,\bullet} a_i = \sum_{i=1}^p f_{i,\bullet} a_i$$

et

$$\bar{y} = \frac{1}{n} \sum_{k=1}^n y_k = \frac{1}{n} \sum_{j=1}^q n_{\bullet,j} b_j = \sum_{j=1}^q f_{\bullet,j} b_j$$

Le point de coordonnées (\bar{x}, \bar{y}) est appelé point moyen du nuage.

Définition 22

On définit

$$\mathbb{V}_x = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^p n_{i,\bullet} (a_i - \bar{x})^2 = \sum_{i=1}^p f_{i,\bullet} (a_i - \bar{x})^2$$

et

$$\mathbb{V}_y = \frac{1}{n} \sum_{k=1}^n (y_k - \bar{y})^2 = \frac{1}{n} \sum_{j=1}^q n_{\bullet,j} (b_j - \bar{y})^2 = \sum_{j=1}^q f_{\bullet,j} (b_j - \bar{y})^2$$

Remarque 25. Si les données sont regroupées en classes, on prend les centres des classes pour a_i .

On n'a par contre pas $\frac{1}{n} \sum_{k=1}^n x_k = \frac{1}{n} \sum_{i=1}^p n_{i,\bullet} a_i$

Remarque 26. Là aussi, si les données sont regroupées en classes on prend pour a_i et b_j les centres des classes

Remarque 27. Là encore on a $\mathbb{V}_x = \overline{x^2} - \bar{x}^2$ et $\mathbb{V}_y = \overline{y^2} - \bar{y}^2$

Définition 23

On définit la covariance de x et de y noté $\text{Cov}(x, y)$ ou $\sigma_{x,y}$ par

$$\text{Cov}(x, y) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})$$

On a également

$$\text{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^q n_{i,j} (a_i - \bar{x})(b_j - \bar{y})$$

Remarque 29. — Si $\text{Cov}(x, y) < 0$ alors x et y ont tendance à varier dans des sens opposés (quand l'un augmente l'autre diminue), si $\text{Cov}(x, y) > 0$ alors ils ont tendance à varier dans le même sens.

— On a $\text{Cov}(x, x) = \mathbb{V}_x$

Proposition 8

On a

$$\text{Cov}(x, y) = \overline{xy} - \bar{x}\bar{y}$$

Démonstration.

$$\begin{aligned} \text{Cov}(x, y) &= \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \\ &= \frac{1}{n} \sum_{k=1}^n (x_k y_k - \bar{x} y_k - \bar{y} x_k + \bar{x} \bar{y}) \\ &= \frac{1}{n} \sum_{k=1}^n x_k y_k - \frac{1}{n} \sum_{k=1}^n \bar{x} y_k - \frac{1}{n} \sum_{k=1}^n \bar{y} x_k + \frac{1}{n} \sum_{k=1}^n \bar{x} \bar{y} \\ &= \bar{x} \bar{y} - \bar{x} \bar{y} - \bar{x} \bar{y} + \bar{x} \bar{y} \\ &= \bar{x} \bar{y} - \bar{x} \bar{y} \end{aligned}$$

□

Proposition 9

On a

$$\mathbb{V}_{x+y} = \mathbb{V}_x + 2\text{Cov}(x, y) + \mathbb{V}_y$$

Démonstration.

$$\begin{aligned} \mathbb{V}_{x+y} &= (x + y)^2 - x + y^2 \\ &= x^2 + 2\bar{x}y + y^2 - (\bar{x} + \bar{y})^2 \\ &= \bar{x}^2 + 2\bar{x}\bar{y} + \bar{y}^2 - \bar{x}^2 - \bar{y}^2 - 2\bar{x}\bar{y} \\ &= \bar{x}^2 - \bar{x}^2 + \bar{y}^2 - \bar{y}^2 + 2(\bar{x}\bar{y} - \bar{x}\bar{y}) \\ &= \mathbb{V}_x + 2\text{Cov}(x, y) + \mathbb{V}_y \end{aligned}$$

□

On a une « inégalité de Cauchy-Schwarz » pour la covariance

Proposition 10

On a

$$|\text{Cov}(x, y)| \leq \sqrt{\mathbb{V}_x} \sqrt{\mathbb{V}_y}$$

où encore

$$|\sigma_{x,y}| \leq \sigma_x \sigma_y$$

Remarque 28. La covariance des x et de y exprime comment x et y varient relativement l'un à l'autre, comment ils « covarient »

Remarque 30. C'est la moyenne des produits moins le produit des moyennes

Démonstration. On définit l'application

$$P : \mathbb{R} \rightarrow \mathbb{R} \\ t \mapsto \mathbb{V}_{x+ty}$$

On a alors

$$\forall t \in \mathbb{R}, \quad P(t) = \sigma_x^2 + 2t\sigma_{x,y} + t^2\sigma_y^2$$

P est donc une application polynomiale de degré 2.

De plus, d'après les propriétés de la variance on a

$$\forall t \in \mathbb{R}, \quad P(t) \geq 0$$

Comme P est de signe constant il ne peut pas admettre deux racines réelles distinctes, son discriminant est négatif ou nul, c'est-à-dire

$$\Delta = 4\sigma_{x,y}^2 - 4\sigma_x^2\sigma_y^2 \leq 0$$

D'où

$$\sigma_{x,y}^2 \leq \sigma_x^2\sigma_y^2$$

Par croissance de la fonction racine carrée on a alors

$$|\sigma_{x,y}| \leq \sigma_x\sigma_y$$

□

Définition 24

Lorsque $\mathbb{V}_x \neq 0$ et $\mathbb{V}_y \neq 0$, on définit le coefficient de covariance noté $\rho_{x,y}$ ou $r_{x,y}$ par

$$\rho_{x,y} = \frac{\text{Cov}(x,y)}{\sqrt{\mathbb{V}_x\mathbb{V}_y}} = \frac{\sigma_{x,y}}{\sigma_x\sigma_y}$$

Remarque 31. Le résultat précédent nous dit que $\rho_{x,y} \in [-1, 1]$

C Ajustement affine

L'idée de l'ajustement affine est la suivante : On dispose de séries de données (souvent expérimentales) x et y et on soupçonne qu'il existe une relation les liant de la forme $y = ax + b$.

On veut alors chercher la droite d'équation $y = ax + b$ qui passe « le mieux » par notre nuage de points.

Le problème est : Comment définir « le mieux » ?

On retient le critère suivant : la somme des carrés des écarts verticaux entre les valeurs y_i observés et celles prédites $ax_i + b$ doit être minimale : c'est la méthode des moindres carrés.

Ainsi, on veut $(a, b) \in \mathbb{R}^2$ rendant minimale la somme

$$S(a, b) = \sum_{k=1}^n (ax_k + b - y_k)^2$$

Remarque 32. Parfois on sait que la relation existe et on veut déterminer a et b

Théorème 11

Soit (x, y) une série statistique double constituée d'une suite de couples $((x_k, y_k))_{1 \leq k \leq n}$. La droite de régression par la méthode des moindres carrés de y en x a pour équation :

$$y = \frac{\sigma_{x,y}}{\sigma_x^2}(x - \bar{x}) + \bar{y}$$

Démonstration. Avec les hypothèses ci-dessus :

$$\frac{\partial S}{\partial b}(a, b) = -2 \sum_{k=1}^n (ax_k + b - y_k) = -2 \left[a \sum_{k=1}^n x_k + b \underbrace{\sum_{k=1}^n 1}_n - \sum_{k=1}^n y_k \right] = 0 \Leftrightarrow b = \frac{\sum_{k=1}^n y_k}{n} - a \frac{\sum_{k=1}^n x_k}{n} = \bar{y} - a\bar{x}$$

On remplace b par $\bar{y} - a\bar{x}$. On peut remarquer que cela signifie que la droite de régression passe par le point moyen de coordonnées (\bar{x}, \bar{y}) .

$$f(a) = S(a, \bar{y} - a\bar{x}) = \sum_{k=1}^n [a(x_k - \bar{x}) - (y_k - \bar{y})]^2 = a^2 \sum_{k=1}^n (x_k - \bar{x})^2 - 2a \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) + \sum_{k=1}^n (y_k - \bar{y})^2$$

On procède à l'étude de la fonction f .

$$f'(a) = 2a \sum_{k=1}^n (x_k - \bar{x})^2 - 2 \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \Leftrightarrow a = \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^n (x_k - \bar{x})^2} = \frac{\sigma_{x,y}}{\sigma_x^2}$$

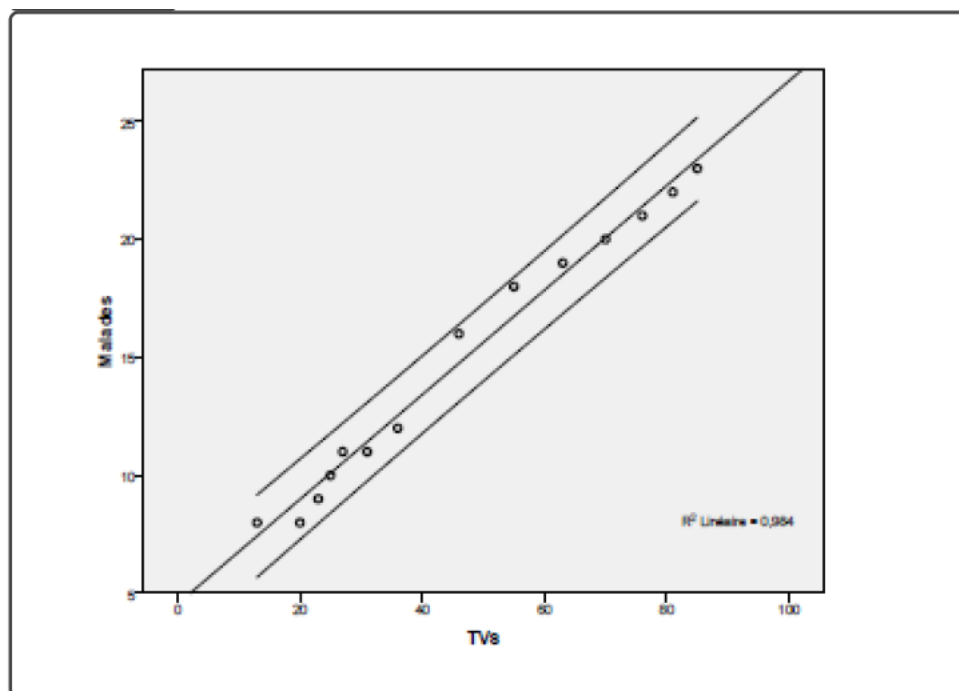
On a donc déterminer le coefficient directeur de la droite de régression qui passe par le point moyen donc

$$y = \frac{\sigma_{x,y}}{\sigma_x^2}(x - \bar{x}) + \bar{y}$$

□

Remarque 33. — Cette droite passe par le point moyen du nuage de coordonnées (\bar{x}, \bar{y})

- Selon la forme du nuage, nos connaissances et notre intuition on considérera parfois les échantillons $\ln(x)$, x^2 , etc
- Pourquoi parle-t-on de « régression linéaire » ? La réponse est une erreur de traduction. Le mathématicien anglais Sir Galton étudiait les tailles des fils (y_j en fonction de la taille de leur père (x_j et a noté un « retour à la moyenne » : Les grands individus ont en moyenne des enfants plus petits qu'eux et les petits individus ont des enfants plus grand qu'eux.
En anglais le terme pour « retour à la moyenne » est « regression to the mean », ce terme a ensuite été mal transposé au français.
- Cette méthode nous permet d'établir un lien de corrélation entre x et y . C'est une erreur fondamentale de logique que de confondre lien de corrélation et lien de causalité. Par exemple la régression linéaire suivante entre taux d'équipement en téléviseurs de la population (en %) et taux de malades mentaux (nombre pour mille habitants) sur des données de Grande Bretagne ou encore l'article du Monde en fin de chapitre.



Comment évaluer la « justesse » d'un ajustement ? La réponse n'est pas simple. Pour y répondre on définit un nouvel indicateur statistique : le coefficient de détermination.

Définition 25

On définit le coefficient de détermination r^2 par

$$r^2 = \frac{\sum_{k=1}^n (ax_k + b - \bar{y})^2}{\sum_{k=1}^n (y_k - \bar{y})^2} = 1 - \frac{S(a, b)}{\sum_{k=1}^n (y_k - \bar{y})^2}$$

où

$$a = \frac{\sigma_{x,y}}{\sigma_x^2} \quad b = \bar{y} - \frac{\sigma_{x,y}}{\sigma_x^2} \bar{x}$$

Théorème 12

Le coefficient de détermination est le carré du coefficient de corrélation, c'est-à-dire

$$r^2 = \rho_{x,y}^2 = \frac{\text{Cov}(x, y)^2}{\mathbb{V}_x \mathbb{V}_y}$$

Démonstration.

$$\begin{aligned} r^2 &= \frac{\sum_{k=1}^n (ax_k + b - \bar{y})^2}{\sum_{k=1}^n (y_k - \bar{y})^2} \\ &= \frac{\sum_{k=1}^n \left(\frac{\sigma_{x,y}}{\sigma_x^2} x_k + \bar{y} - \frac{\sigma_{x,y}}{\sigma_x^2} \bar{x} - \bar{y} \right)^2}{n\sigma_y^2} \\ &= \frac{1}{n} \frac{\left(\frac{\sigma_{x,y}}{\sigma_x^2} \right)^2 \sum_{k=1}^n (x_k - \bar{x})^2}{\sigma_y^2} \\ &= \frac{\sigma_{x,y}^2 \sigma_x^2}{\sigma_x^4 \sigma_y^2} \\ &= \frac{\sigma_{x,y}^2}{\sigma_x^2 \sigma_y^2} \\ &= \rho_{x,y}^2 \end{aligned}$$

□

Proposition 13

On a

$$0 \leq r^2 \leq 1$$

$r^2 = 1$ correspond à une adéquation parfaite tandis que r^2 proche de 0 indique une faible liaison linéaire ce qui peut signifier qu'il n'y a pas de lien entre x et y ou bien que x et y sont liés par une relation non-affine.