TD Statistique en Python.

Proposition 1

On a alors

$$\forall j \in [\![1,p]\!], \qquad n_j^c = \sum_{k=1}^{j} n_k, \qquad \text{et} \qquad f_j^c = \sum_{k=1}^{c} f_k$$

et

$$n_1^c \leqslant n_2^c \leqslant \cdots \leqslant n_p^c = n$$
$$f_1^c \leqslant f_2^c \leqslant \cdots \leqslant f_p^c = 1$$

Exemple 1. — 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

Et la même regroupée en classe :

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

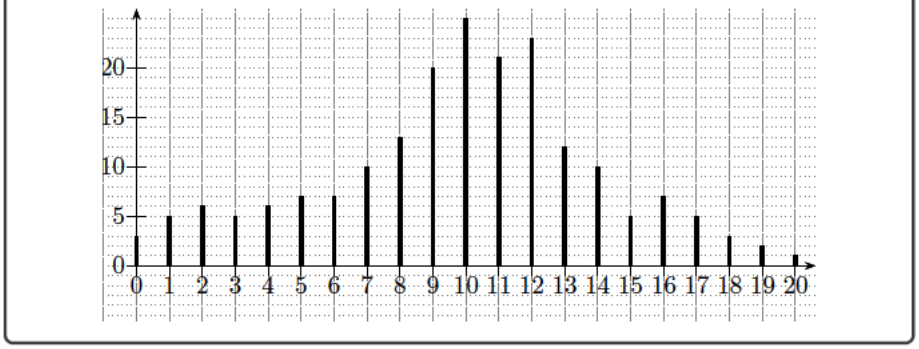
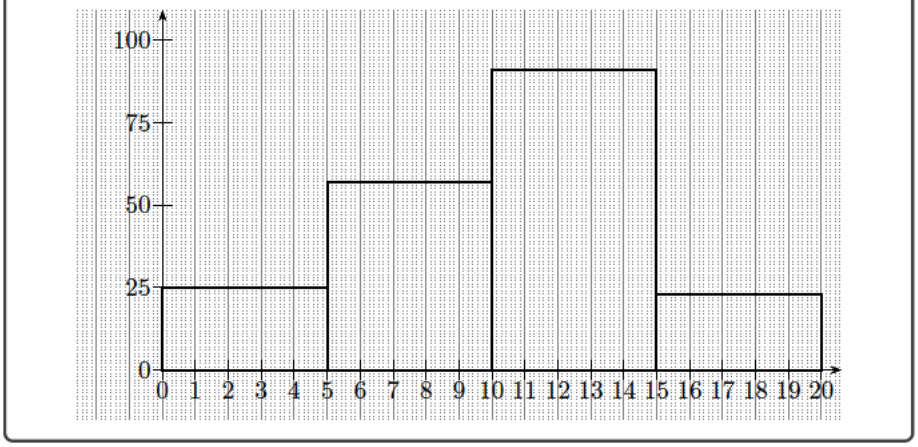
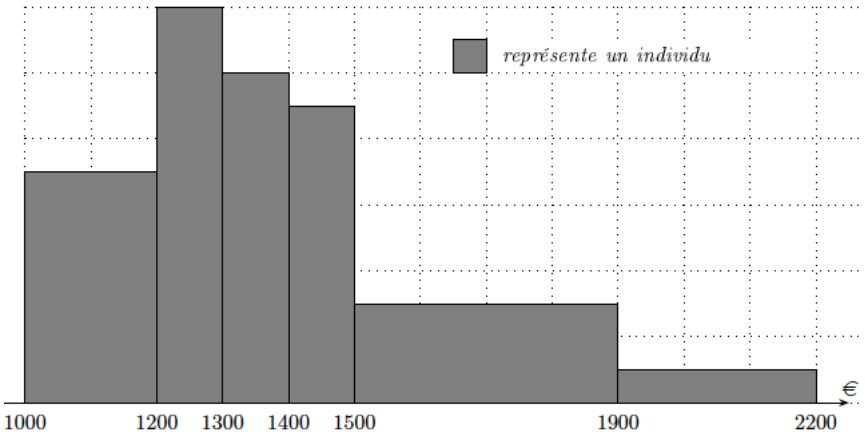


Figure 12.2 – Histogramme



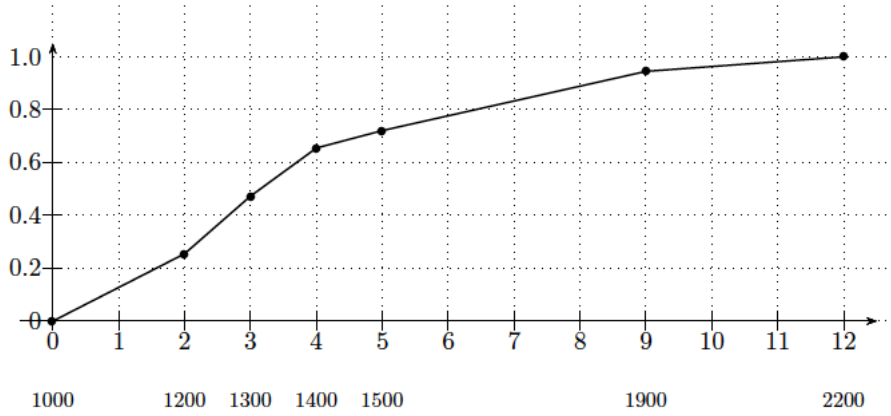
— Une étude portant sur les salaires mensuels des employés en CDI à temps complet d'une entreprise a permis d'établir l'histogramme ci-dessous.



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.

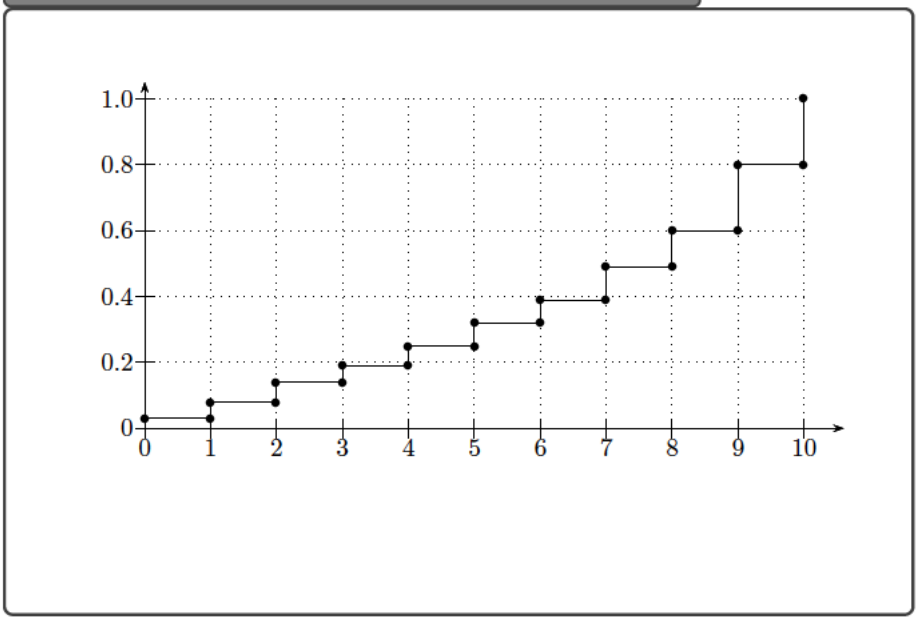


— On considère la série :

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

Figure 12.3 – Polygone des fréquences cumulées croissantes



I 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 1

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 2. — 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.

B Moyenne

Définition 2

Soit $x = (x_1, \cdots, x_n)$ 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 2

Si les modalités $(a_1, \cdots, a_p)$ 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 3

Dans la situation où les modalités $(a_j)_{j\in[\![1,p]\!]}$ 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 3

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

$$\forall i \in [\![1, n]\!], \qquad u_i = ax_i + b$$

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

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

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

C Médiane

On suppose ici que nos modalités sont des réels

Définition 4

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

$$\text{Card}\left(\{i \in [\![1, n]\!]\ ,\ x_i \leqslant m\}\right) \geqslant \frac{n}{2} \quad \text{ et } \text{Card}\left(\{i \in [\![1, n]\!]\ ,\ x_i \geqslant m\}\right) \geqslant \frac{n}{2}$$

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.

Proposition 4

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

— 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 9. — 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 5

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 $\frac{1}{2}$.

II 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 6

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 10. 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 7

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

$$\text{Card}\left(\{i\in[\![1,n]\!]\ ,\ x_i\leqslant Q_1\}\right)\geqslant\frac{n}{4}\quad\text{et Card}\left(\{i\in[\![1,n]\!]\ ,\ x_i\geqslant Q_1\}\right)\geqslant\frac{3n}{4}$$

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

$$\text{Card}\left(\{i\in[\![1,n]\!]\ ,\ x_i\leqslant Q_3\}\right)\geqslant\frac{3n}{4}\quad\text{et Card}\left(\{i\in[\![1,n]\!]\ ,\ x_i\geqslant Q_3\}\right)\geqslant\frac{n}{4}$$

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

On peut définir la notion de décile de manière similaire

Définition 8

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

$$\text{Card}\left(\{i\in[\![1,n]\!]\ ,\ x_i\leqslant d_j\}\right)\geqslant\frac{jn}{10}\quad\text{et Card}\left(\{i\in[\![1,n]\!]\ ,\ x_i\geqslant d_j\}\right)\geqslant\frac{(10-j)n}{10}$$

Ces deux notions se généralisent avec la notion de quantiles

Définition 9

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

$$\text{Card}\left(\{i\in[\![1,n]\!]\ ,\ x_i\leqslant q_t\}\right)\geqslant nt\quad\text{et Card}\left(\{i\in[\![1,n]\!]\ ,\ x_i\geqslant q_t\}\right)\geqslant(1-t)n$$

Remarque 14. — 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 10

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 $d_9 - d_1$.

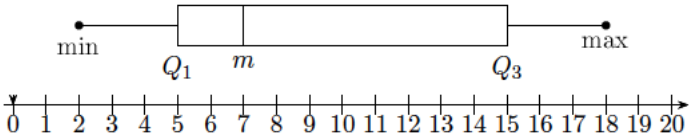
— intervalle interdécile l'intervalle $[d_1, d_9]$

Rem
moit
popu
dans
inter
Rem
moin
popu
dans
inter

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

— au centre une boite 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.



C Variance et écart-type

Définition 11

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 :

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

$\mathbb{V}_x$ e
carré
moye

On a également

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

Remarque 17. — La variance est toujours positive.

— Dans la 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 concentré 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 5

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

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

Rem
donc
carré
de la

Démonstration. On a

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

□

Remarque 19.
— 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 milieu 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 12

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

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

Rem
trouv
nota

Remarque 21. 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 6

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

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

□

III 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 $\mathbb{R}^2$

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

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

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

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

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

Pour $(i, j) \in [\![1, p]\!] \times [\![1, q]\!]$, on pose

$$n_{\blacksquare,1} = \sum_{j=1}^{\ell} ni, j \qquad \text{et } n_{j,\blacksquare} = \sum i = 1^k n_{i,j}$$

On a alors

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

Définition 13

Pour $(i, j) \in [\![1, p]\!] \times [\![1, q]\!]$ 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 $f_{i,\blacksquare} = \frac{n_{i,\blacksquare}}{n}$
- La fréquence marginale de la modalité a_i par $f_{\blacksquare,j} = \frac{n_{\blacksquare,j}}{n}$

On a alors

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

Définition 14

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

B Indicateurs statistiques

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

Définition 15

On définit

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

On a encore

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

et

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

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

Définition 16

On définit

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

et

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

Définition 17

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

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

On a également

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

Remarque 26. — Si $\mathrm{Cov}(x,y) < 0$ alors x et y ont tendance à varier dans des sens opposés (quand l'un augment l'autre diminue), si $\mathrm{Cov}(x,y) > 0$ alors ils ont tendance à varier dans le même sens.

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

Proposition 7

On a

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

Démonstration.

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

□

Proposition 8

On a

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

Démonstration.

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

□

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

Proposition 9

On a

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

où encore

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

Démonstration. On définit l'application

$$\begin{array}{rccc} P & : & \mathbb{R} & \to & \mathbb{R} \\ & & t & \mapsto & \mathbb{V}_{x+ty} \end{array}$$

On a alors

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

P est donc une application polynomiale de degré 2.
De plus, d'après les propriétés de la variance on a

$$\forall t \in \mathbb{R}, \qquad P(t) \geqslant 0$$

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

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

D'où

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

Par croissance de la fonction racine carrée on a alors

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

□

Définition 18

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

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

Rem
résul
nous
$\rho_{x,y}$

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

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

Théorème 10

Soit (x, y) une série statistique double constituée d'une suite de couples $((x_k, y_k))_{1 \leqslant k \leqslant n}$.
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 :

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

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

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

On procède à l'étude de la fonction f.

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

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 30.

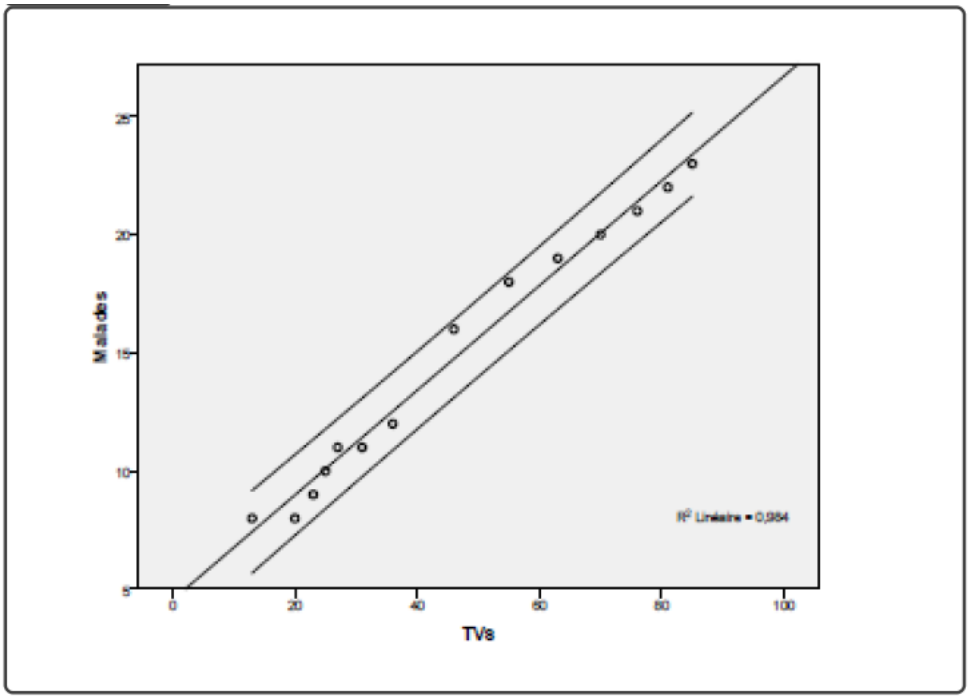
— Cette droite passe par le point moyen du nuage de coordonnées $(\bar{x}, \bar{y})$

— 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 19

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

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

Théorème 11

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

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

Démonstration.

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

□

Proposition 12

On a

$$0 \leqslant r^2 \leqslant 1$$

$r^2 = 1$ 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.