Chapitre 15 : Résumé Statistique.

I Statistique descriptive univariée

Soit X une série statistique en modalité $\{x_1, x_2, ..., x_p\}$ d'effectif total n. Les effectifs de chaque modalité seront noté n_i et les fréquence $f_i = \dfrac{n_i}{n}$:

x_i	x_1	x_2	...	x_p
n_i	n_1	n_2	...	n_p

Proposition 1 (Effectif total)

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

Proposition 2 (effectifs et fréquences cumulés croissants)

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

$$\text{donc} \qquad n_1^c \leqslant n_2^c \leqslant \cdots \leqslant n_p^c = n \qquad \text{et} \qquad f_1^c \leqslant f_2^c \leqslant \cdots \leqslant f_p^c = 1$$

Définition 1 (mode)

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.

Proposition 3 (moyenne)

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

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

Proposition 4 (Transformation affine d'une série statistique)

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

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

Définition 2 (Médiane)

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

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

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 5 (Médiane)

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.

Définition 3 (Quantiles)

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

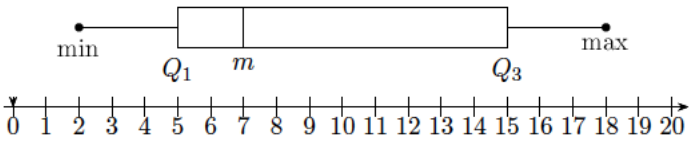
$$\text{Card}\left(\{i \in [\![1,n]\!],\ x_i \leqslant Q_3\}\right) \geqslant \frac{3n}{4} \quad \text{et Card}\left(\{i \in [\![1,n]\!],\ x_i \geqslant Q_3\}\right) \geqslant \frac{n}{4}$$

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

Définition 4 (Écart inter-quartile)

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 $[Q_1, Q_3]$



Définition-Proposition 6 (Variance et écart-type)

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

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

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

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

$$\mathbb{V}_y = a^2\mathbb{V}_x \qquad \text{et} \qquad \sigma_y = |a|\sigma_x$$

II Statistique descriptive bivariée

Soit (X, Y) une série statistique bivariée :

x_i	x_1	x_2	...	x_p
y_i	y_1	y_2	...	y_p

Définition 5

On définit

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \text{et} \qquad \bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$$

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

Définition 6 (Variance)

On définit

$$\mathbb{V}_x = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$

Idem pour Y.

Définition 7

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

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

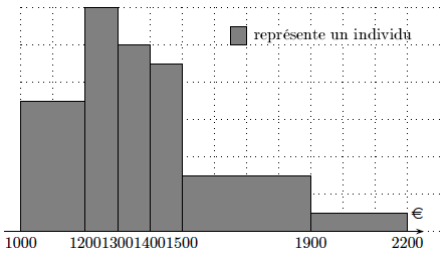
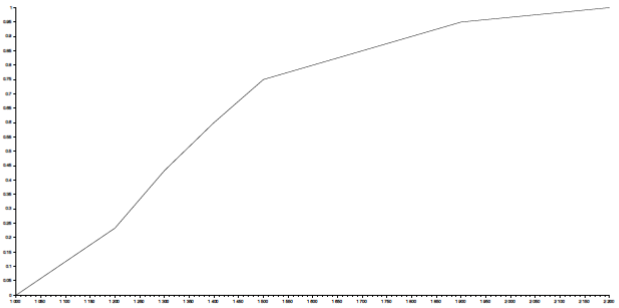
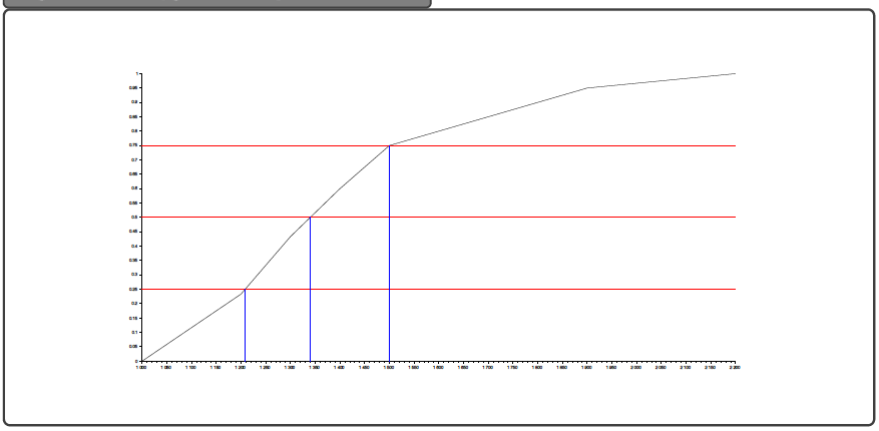


Figure 12.2 – Polygone des fréquences cumulées



Exemple :

Figure 12.3 – Polygone des fréquences cumulées



Définition 8

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

$$\rho_{x,y} = \frac{\text{Cov}(x,y)}{\sqrt{\mathbb{V}_x \mathbb{V}_y}} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} \qquad rho_{x,y} \in [-1,1]$$

A 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 8

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 :

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

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.

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

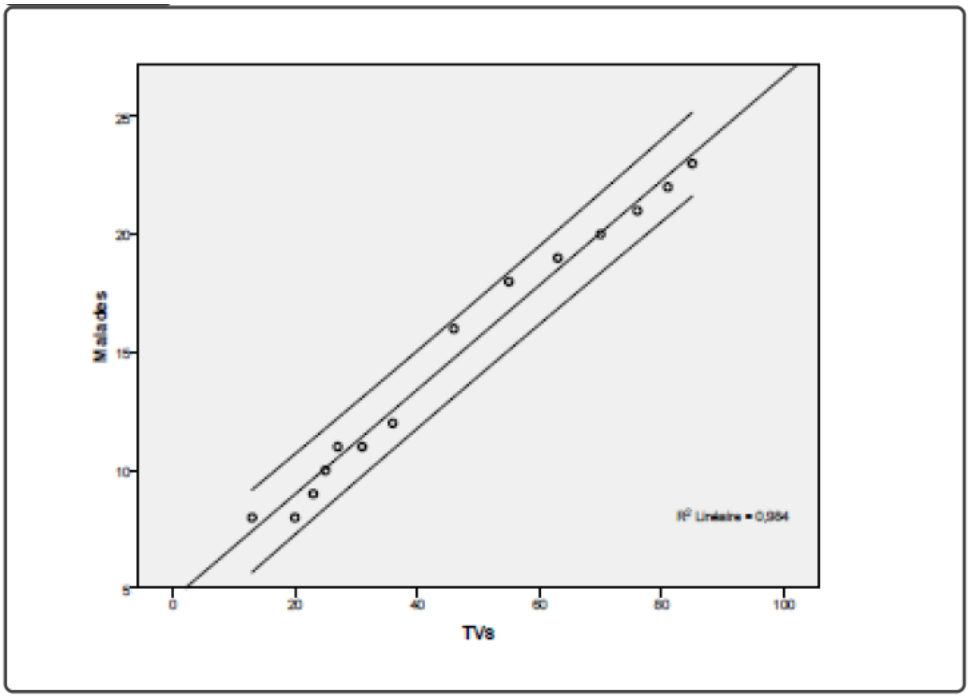
On a donc déterminer le coefficient directeur de la droite de régression qui passe par le point moyen donc

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

□

Remarque 2. — 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 9

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 9

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.

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

□

Proposition 10

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.