Chapitre 10 : Statistiques descriptives

Population : l'ensemble des individus (unités) sur lesquels porte l'étude statistique.

Échantillon : un ensemble d'individus extraits d'une population étudié de manière à ce que cet ensemble soit représentatif de la population. La taille de l'échantillon est le nombre d'individus formant l'échantillon. Pour chaque individu, on peut mesurer un ou plusieurs caractères (variables).

Les statistiques descriptives sont des nombres ou des graphiques visant à décrire les principales caractéristiques d'un échantillon donné. Pour la suite, on suppose qu'un échantillon de taille *n* est disponible pour lequel une ou plusieurs variables ont été mesurées ou obtenues.

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

Médiane : valeur milieu dans l'échantillon avec les valeurs ordonnées.

si *n* pair : moyenne des deux valeurs au milieu dans la suite ordonnée
ex : 1, 4, 6, 10, 20, 27 => médiane : (6+10)/2 = 8

si *n* impair : valeur milieu dans la suite ordonnée
ex : 1, 4, 6, 10, 20 => médiane : 6

Variance : $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\left[\sum_{i=1}^{n} x_i^2\right] - n\bar{x}^2\right)$

ex. : 1, 4, 6, 10, 20 => $s^2 = 54.2$

Écart-type : $s = \sqrt{s^2}$

Coefficient d'asymétrie

$$\beta = \frac{n}{(n-1)(n-2)\,s^3}\sum_{i=1}^{n}(x_i - \bar{x})^3 \qquad (n>2)$$

ex. x : 1, 4, 6, 10, 20 => $\beta = 1.337$

Covariance : on a deux variables x et y. On calcule :

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) = \frac{1}{n-1}\left(\left[\sum_{i=1}^{n} x_i y_i\right] - n\bar{x}\bar{y}\right)$$

ex. x : 1, 4, 6, 10, 20
y : 5, 3, 7, 9, 14 => $s_{xy} = 29.35$

Corrélation (Pearson) : $r_{xy} = \frac{s_{xy}}{s_x s_y}$

ex. x : 1, 4, 6, 10, 20
y : 5, 3, 7, 9, 14 => $r_{xy} = 0.94493$

Corrélation (Spearman)

La corrélation de Spearman est la corrélation de Pearson calculée sur les rangs correspondant aux valeurs observées plutôt que sur les valeurs observées elles-mêmes. La corrélation de Spearman est utile pour décrire les relations non-linéaires entre les variables.

Valeurs	Rangs correspondants
x : 1, 4, 6, 10, 20	x : 1, 2, 3, 4, 5
y : 5, 3, 7, 9, 14	y : 2, 1, 3, 4, 5

La corrélation entre les rangs vaut ici : 0.9

ex. Soit x=1, 2, 3, ...10 et y=exp(x). $r_{pearson,xy} = 0.717$ et $r_{spearman,xy} = 1$. La corrélation de Spearman reconnaît le lien déterministe unissant les deux variables.

En cas d'égalité de valeurs, on assigne le rang moyen à chacune des valeurs égales.

Valeurs	Rangs correspondants
x : 1, 4, 4, 4, 6, 10, 20	x : 1, 3, 3, 3, 5, 6, 7
y : 5, 3, 7, 9, 9, 14, 18	y : 2, 1, 3, 4.5, 4.5, 6, 7

$r_{spearman} = 0.86$

Quantile (ou centile ou percentile)

Il existe plusieurs méthodes pour définir les quantiles selon la façon dont on définit les p_i associés aux différentes valeurs ordonnées.
Ainsi, on peut avoir pour une observation placée en $i^{ème}$ position (i=1...n):

A- $p_i = i/(n+1)$
B- $p_i = (i-0.5)/n$ (ex. fonction quantile dans Matlab)
C- $p_i = (i-1)/(n-1)$ (ex. fonction centile dans Excel)

A- $p_i = i/(n+1)$ où i est le rang des observations ordonnées (i=1...n)
On prend la valeur q_p correspondant au rang *p (n+1)*

Concrètement, on interpole linéairement entre les valeurs de l'échantillon. On calcule d'abord *p (n+1).* La partie entière identifie la borne inférieure x et la partie fractionnaire l'interpolation sur l'intervalle.

ex. 1, 2, 4, 10, 20 => n=5, n+1=6

- le quantile $q_{0.75}$ => 0.75*6 = 4.5. On part de la 4^e observation et l'on interpole à mi-chemin avec la 5^e, donc $q_{0.75} = 10 + 0.5\ (20-10) = 15$
- le quantile $q_{0.25}$ => 0.25*6 = 1.5. On part de la $1^{ère}$ observation et l'on interpole à mi-chemin avec la seconde. Donc $q_{0.25} = 1.5$.

B- $p_i = (i-0.5)/n$ où i est le rang des observations ordonnées (i=1...n)
On calcule *p n* et l'on interpole linéairement.

ex. 1, 2, 4, 10, 20

- le quantile $q_{0.75}$ => 0.75*5 = 3.75. On part de la 4e observation à laquelle correspond le rang 3.5 et on interpole au ¼ avec la 5e observation à laquelle correspond le rang 4.5, donc $q_{0.75}$ = 10+0.25(20-10) = 12.5
- le quantile $q_{0.25}$ => 0.25*5 = 1.25. On part de la 1ère observation à laquelle correspond le rang 0.5 et on interpole aux ¾ avec la 2e observation à laquelle correspond le rang 1.5, soit $q_{0.25}$ = 1+0.75 (2-1) =1.75.

C- $p_i = (i-1)/(n-1)$ où i est le rang des observations ordonnées (i=1...n)

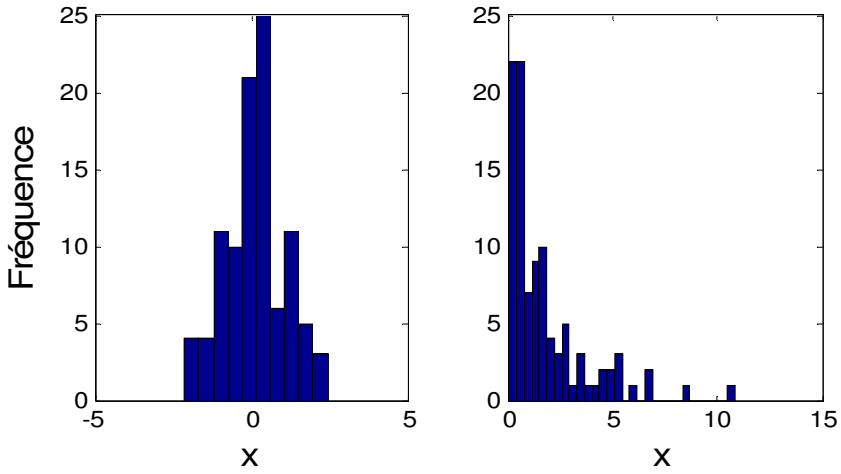
On calcule p(n-1) + 1 et l'on interpole linéairement.
ex. 1, 2, 4 10, 20

- $q_{0.75}$ => 0.75*4 +1 = 4 => $q_{0.75}$ = 10
- $q_{0.25}$ => 0.25*4+1 = 2 => $q_{0.25}$ = 2

Histogramme

On peut voir l'histogramme comme l'analogue expérimental de la fonction de densité. On divise le support de la v.a. en classes (habituellement d'égale largeur) et on compte combien d'observations de l'échantillon tombent dans chaque classe. L'ordonnée est la fréquence observée.

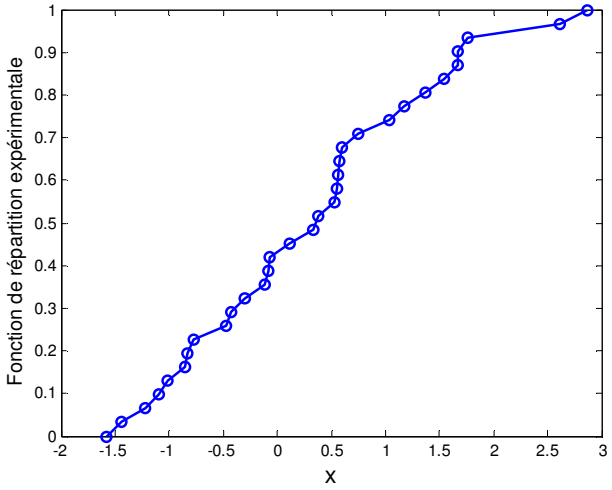
ex :



Fonction de répartition expérimentale

Cette fonction est l'analogue expérimental de la fonction de répartition. On ordonne les valeurs *x*. Soit x_i les valeurs ordonnées avec les rangs de 1 à n. À chaque x_i, on associe la valeur $f(x_i) = i/(n+1)$ où *i* est le rang de l'observation. On complète la fonction en extrapolant avec $f(x_{min}) = 0$ et $f(x_{max}) = 1$, où x_{min} et x_{max} sont les valeurs minimales et maximales que l'on puisse observer pour la variable x. On relie ces points par des segments de droite ou par une courbe lissée non-décroissante.

ex.

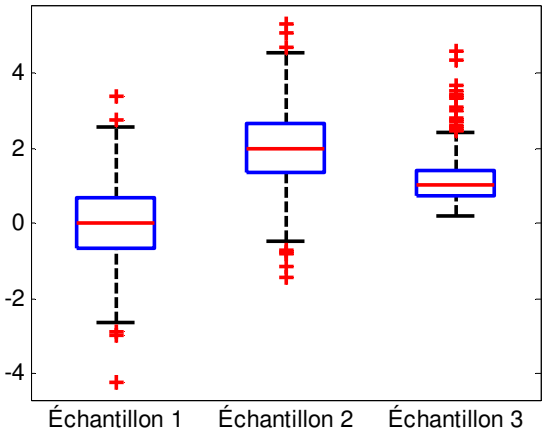


Note : On peut aussi représenter cette fonction comme une fonction en escalier, surtout dans le cas où la variable étudiée est discrète.

Diagramme en boîte à moustaches (diagramme de Tukey, « Box-plot »)

Ce diagramme résume la distribution échantillonnale d'une variable par différentes mesures caractéristiques comme les quartiles (centile 0.25 et 0.75), la médiane, la moyenne, etc. Très utile pour comparer rapidement les distributions de différents échantillons.

Ex. (fonction boxplot de Matlab)



Les deux premiers échantillons montrent des distributions symétriques, le 3^{e} échantillon a une asymétrie positive. Le 2^{e} échantillon présente des valeurs supérieures au premier. Les extrémités des moustaches sont positionnées à

$$q_{0.25} - 1.5\,(q_{0.75}-q_{0.25}) \text{ et } q_{0.75} + 1.5\,(q_{0.75}-q_{0.25})$$

Les valeurs hors intervalle sont considérées « extrêmes ». Le facteur 1.5 peut être changé au besoin.