

STATISTIQUES DESCRIPTIVES

Source : Wannacott et Wannacott, Statistique : Economie, Gestion, Sciences , Médecine, Economica, 3° éd., 1984

Les statistiques descriptives sont des calculs opérés sur des observations que l'on notera x_1, \dots, x_p .

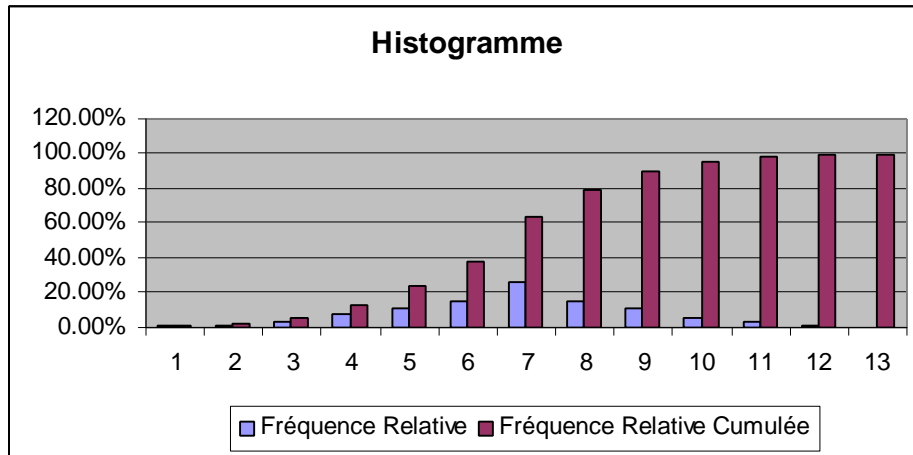
Les fréquences

Les **effectifs (ou fréquences)** représentent le comptage d'un nombre d'individus par valeur (ou par plages de valeurs, si la variable observée est une variable continue). Les **fréquences relatives** représentent les fréquences divisées par le nombre total d'observations effectuées.

Exemple : soit un échantillon de 200 hommes américains, et X leur taille mesurée en inches. On dit de X que c'est une variable continue, puisque les valeurs qu'elle peut prendre varient de façon continue. On ne peut parler de fréquence d'une valeur spécifique (ce qui pourrait être le cas pour une variable discrète (par exemple, le nombre d'enfants d'une famille) puisqu'on n'observera jamais deux fois une taille de exactement 64.328... inches. Par contre, on peut parler de fréquence des tailles appartenant à une classe, comme dans le tableau ci-dessous.

Classes		Centre classe	Comptage	Fréquence	Fréquence Relative	Fréquence Relative Cumulée
Inférieur	Supérieur					
58	62	60	III	3	1.37%	1.37%
62	66	64	II	2	0.91%	2.28%
66	70	68	IIIIII	6	2.74%	5.02%
70	74	72		17	7.76%	12.79%
74	78	76		24	10.96%	23.74%
78	82	80		32	14.61%	38.36%
82	86	84		56	25.57%	63.93%
86	90	88		34	15.53%	79.45%
90	94	92		23	10.50%	89.95%
94	98	96		12	5.48%	95.43%
98	102	100		6	2.74%	98.17%
102	106	104		3	1.37%	99.54%
106	110	108		1	0.46%	100.00%
Total				219		

Le graphique qui reprend l'évolution des fréquences par classes est appelé **histogramme**.



Les **percentiles** représentent la place qu'occupe un individu dans les fréquences relatives cumulées. Par exemple, si un individu mesure 92 inches, on pourra affirmer, suivant la tableau ci-dessous, qu'il est 90ème percentiles. Les **quartiles** sont les percentiles qui partagent les données en quatre. Le 50ème percentile est appelé **médiane** puisqu'il correspond à la valeur centrale qui partage les données en deux parties égales.

Le centre d'une distribution (ou tendance centrale)

Le **mode** est la valeur la plus fréquente. Il correspond au sommet de la distribution. Une distribution peut être bimodale (avoir deux sommets).

La **médiane** est la valeur qui partage la distribution en deux moitiés égales (le 50ème percentile).

La **moyenne** est la somme des observations divisée par le nombre d'observations

effectuées : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Si l'on considère que chaque individu représente la même

masse ponctuelle (chaque individu a le même poids dans la distribution), la moyenne représente le **centre de gravité** de la distribution.

Le choix entre le mode, la moyenne et la médiane est fonction du type de distribution que l'on observe :

- le **mode** fait apparaître le comportement le plus fréquent. Dans le cas d'une distribution fortement asymétrique, cette valeur peut être peu représentative du comportement de l'ensemble des observations effectuées.
- la **moyenne** donne la même importance à chaque observation. Elle est donc sensible aux observations extrêmes.
- la **médiane**, qui partage en deux la distribution, n'est pas sensible à l'importance de l'éloignement des données extrêmes. Elle est en ce sens plus robuste.

La dispersion d'une distribution empirique

L'étendue (range) : il s'agit de la différence entre la plus grande et la plus petite valeur de la variable.

L'étendue interquartile (EIQ) : afin de réduire l'influence des observations extrêmes, l'étendue interquartile se calcule comme la différence entre le troisième et premier quartile.

L'écart absolu moyen (mean absolute deviation) : il s'agit de la moyenne des écarts en valeur absolue entre les observations et leur moyenne, soit $EAM = \frac{1}{n} \sum |x_i - \bar{x}|$.

L'écart quadratique moyen (mean absolute deviation ou variance empirique) : il s'agit de la moyenne des écarts au carré entre les observations et leur moyenne, soit $EQM = \frac{1}{n} \sum [x_i - \bar{x}]^2$.

X	X-E(X)	Abs(X-E(X))	(X-E(X))^2
10	-30	30	900
20	-20	20	400
30	-10	10	100
50	10	10	100
90	50	50	2500
Moyenne	40	EAM	24 EQM

Les relations entre deux variables

De nombreuses approches peuvent être suivies pour analyser les relations qui peuvent exister entre deux variables. Nous en étudierons un certain nombre de manière plus approfondie ultérieurement. Deux indicateurs sont toutefois classiquement utilisés :

La **covariance empirique** mesure la façon dont deux variables x et y varient simultanément. Elle se définit comme suit : $s_{xy} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$. Comme le montre l'exemple ci-dessous, la covariance permet bien de saisir l'interaction entre deux variables.

- Distribution conjointe de x et y

X	Y				Moyenne = 25
	10	20	30	40	
10	20.00%	4.00%	1.00%		25.00%
20	10.00%	36.00%	9.00%		55.00%
30		5.00%	10.00%		15.00%
40				5.00%	5.00%
Moyenne = 25	30.00%	45.00%	20.00%	5.00%	

- Calcul de la covariance

Ecart à la moyenne pour X				
X	Y			
	10	20	30	40
10	(15)	(15)	(15)	(15)
20	(5)	(5)	(5)	(5)
30	5	5	5	5
40	15	15	15	15
Ecart à la moyenne pour Y				
X	Y			
	10	20	30	40
10	(15)	(5)	5	15
20	(15)	(5)	5	15
30	(15)	(5)	5	15
40	(15)	(5)	5	15
Covariance				
X	Y			
	10	20	30	40
10	45	3	(1)	-
20	8	9	(2)	-
30	-	(1)	3	-
40	-	-	-	11
				74

L'utilisation de la covariance soulève toutefois un problème. Si le signe obtenu est aisément interprétable, la valeur est quant à elle dépendante du choix d'unité effectué. Ainsi, selon que l'une des variables est mesurée en mètre ou en kilomètre, le résultat varie.

Le **coefficient de corrélation empirique** résout ce problème. Il se définit comme

$$r = \frac{s_{xy}}{s_x s_y}$$

Il est borné entre -1 et 1, est insensible à l'unité de mesure choisie et

présente trois valeurs remarquables :

- -1 : les deux variables sont parfaitement corrélées négativement. A toute variation de x correspond une variation de y strictement proportionnelle en sens opposé.
- 1 : les deux variables sont parfaitement corrélées positivement. A toute variation de x correspond une variation de y strictement proportionnelle de même sens.
- 0 : les deux variables sont non corrélées.

On notera que la covariance et le coefficient de corrélation mesurent le degré d'association **linéaire** entre variables.