Statistique

Objectif : Etudier les divers ***Caractères*** d'une ***Population*** P de ***taille*** (***effectif***) p.

Caractère : Un caractère peut être ***quantitatif*** (taille, poids, salaire, etc…) on peut comparer par des mesures. Ou ***qualitatif*** (couleur, saveur, etc,…) on ne peut pas comparer par mesure, on fait alors recours à un codage.

Echantillon : On peut travailler sur toute la population P de taille p, ou sur un ***échantillon*** E de taille N, de cette population.

Un échantillon peut être :

Biaisé : choisi selon les convenances du trieur.

Non biaisé : collecte au hasard de cet échantillon parmi la population P

Biaisé ou non biaisé, ceci donne les ***fluctuations de l'échantillonnage***.
Ceci signifie que d'un échantillon biaisé à un autre, les résultats varient considérablement.

Série statistique : Soit le caractère désigné par X de la population P, étudié sur un échantillon E (supposé non biaisé) de taille N la série des mesures correspondantes est :

$$x_1, x_2, \ldots\ldots\ldots, x_k, \ldots\ldots\ldots, x_{N-1}, x_N,$$

Cette série est notée par $\boldsymbol{x}$, dite ***série statistique***.

Paramètres de position :La série statistique $\boldsymbol{x}$:

$$x_1, x_2, \ldots\ldots\ldots, x_k, \ldots\ldots\ldots, x_{N-1}, x_N,$$

N est l'***effectif*** de la série, $\max_{\boldsymbol{x}}$ son ***maximum*** et $\min_{\boldsymbol{x}}$ son ***minimum*** et

$\bar{x} = (x_1 + x_2 + \ldots\ldots\ldots + x_k + \ldots\ldots\ldots + x_{N-1} + x_N)/N$ sa ***moyenne***.

$\max_{\boldsymbol{x}} - \min_{\boldsymbol{x}}$ est l'***étendue*** de la série.

Série centrée :

$y_1 = x_1 - \sigma$, $y_2 = x_2 - \sigma, \ldots, y_k = x_k - \sigma$, …., $y_{N-1} = x_{N-1} - \sigma$, $y_N = x_N - \sigma$, la série ***centré*** correspondante.

Paramètres de dispersion :

La variance : $\boldsymbol{var(x)} = ((y_1)^2 + (y_2)^2 + \ldots\ldots\ldots + (y_k)^2 + \ldots\ldots + (y_{N-1})^2 + (y_N)^2)/N$ la variance de la série $\boldsymbol{x}$.

L'écart Type (sd = Standard Deviation) :

$\sigma(\boldsymbol{x}) = \sqrt{var(x)}$ est l'***écart type*** de la série $\boldsymbol{x}$.

Le mode ***Mod*** : c'est la valeur qui est la plus répétée dans la série. (s'il y a deux valeurs qui sont les plus et également répétées, on dit qu'il n'y a pas de mode)

La médiane Med : Si on ordonne la série $x_1, x_2, \ldots\ldots\ldots, x_k, \ldots\ldots\ldots, x_{N-1}, x_N,$

On obtient la série formée par les mêmes éléments de x mais dans un ordre différent :

$z_1, z_2, \ldots\ldots, z_k, \ldots\ldots, z_{N-1}, z_N,$

$$z_1 \leq z_2 \leq \ldots\ldots \leq z_k \leq \ldots\ldots \leq z_{N-1} \leq z_N,$$

et on découpe en 2 groupes :

50%	50%

Calcul de la médiane : si N est impair N = 2p+1, ***Med*** = x_{p+1}, c'est le terme central de la série ordonnée.

si N est pair N = 2p, on n'a pas un terme central, une méthode de calcul est celle de la (**norme AFNOR NF 06-003**) qui consiste à (pour la série ordonnée):

$$\boldsymbol{Med} = (x_p + x_{p+1})/2.$$

On applique la même méthode pour le calcul de q_1 et q_3.

Les quartiles :

la série $x_1, x_2, \ldots\ldots, x_k, \ldots\ldots, x_{N-1}, x_N$, étant ordonnée, en :

$z_1, z_2, \ldots\ldots, z_k, \ldots\ldots, z_{N-1}, z_N,$

On découpe en 4 groupes

25%	25%	25%	25%

$\boldsymbol{q_1}$, le premier quartile divise l'effectif en (25%, 75%)
$\boldsymbol{q_2}$ = ***Med***, le second quartile divise l'effectif en (50%, 50%)
$\boldsymbol{q_3}$, le troisième quartile divise l'effectif en (75%,25%)

Si N = 4s (multiple de 4) :
$\boldsymbol{q_1}$, le dernier du premier groupe, $\boldsymbol{q_3}$ le dernier du troisième groupe
Si N ≠ 4s (non multiple de 4) :
$\boldsymbol{q_1}$, le premier du second groupe, $\boldsymbol{q_3}$ le premier du quatrième groupe

$\min_x$	Q_1	Med	q_3	$\max_x$

Ceci donne lieu à :
$[\boldsymbol{q_1 ; q_3}]$ **:** l'***intervalle interquartile*** et
$\boldsymbol{q_3 - q_1}$ **:** l'***écart interquartile***

Les déciles : $x_1, x_2, \ldots\ldots, x_k, \ldots\ldots, x_{N-1}, x_N$, étant ordonnée en

$z_1, z_2, \ldots\ldots, z_k, \ldots\ldots, z_{N-1}, z_N,$

On découpe en 10 groupes

10%	10%	10%	10%	10%	10%	10%	10%	10%	10%

Selon le même principe :

d_1, d_2, d_3, d_4, d_5 = ***Med***, d_6, d_7, d_8, d_9, sont les ***déciles*** de la série ***x***.

d1	d2	d3	d4	d5 = Med	d6	d7	d8	d9

Distribution symétrique : Un caractère a une distribution symétrique si :

Mode, médiane, moyenne sont confondus.

Indice de dissymétrie :

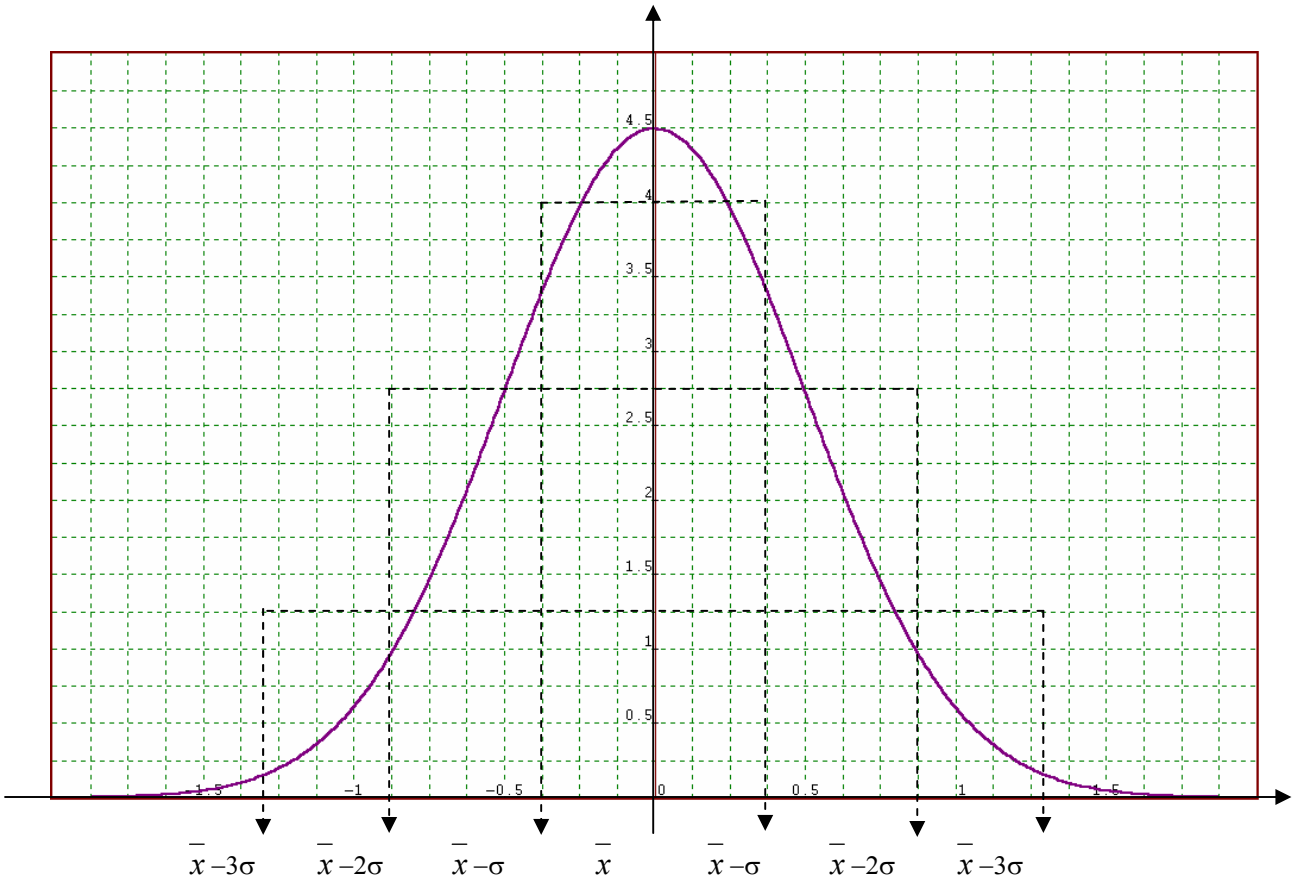
Pour une distribution non symétrique on a un indice de ***dissymétrie*** :

$$s = \frac{(q_3 - Med) - (Med - q_1)}{(q_3 - Med) + (Med - q_1)}$$

si s= 0 la distribution est ***symétrique***

si s<0 la distribution est ***dissymétrique à gauche*** (étalement vers la droite)

si s>0 la distribution est ***dissymétrique à droite*** (étalement vers la gauche)



Plages de normalité : (Intervalles de confidence de la ***donnée***)

On admet que

1) 68% de la donnée est située dans $[\bar{x}-\sigma\ ;\ \bar{x}+\sigma]$

Si au moins 68% de la donnée est située dans $[\bar{x}-\sigma\ ;\ \bar{x}+\sigma]$, on dit que la distribution est normale à 68%.

2) 95% de la donnée est située dans $[\bar{x}-2\sigma\ ;\ \bar{x}+2\sigma]$

Si au moins 95% de la donnée est située dans $[\bar{x}-2\sigma\ ;\ \bar{x}+2\sigma]$, on dit que la distribution est normale à 95%.

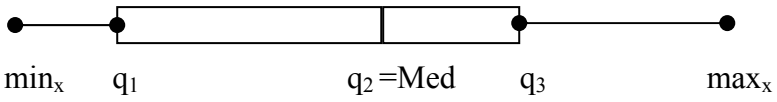
3) 99,7% de la donnée est située dans $[\bar{x}-3\sigma\ ;\ \bar{x}+3\sigma]$

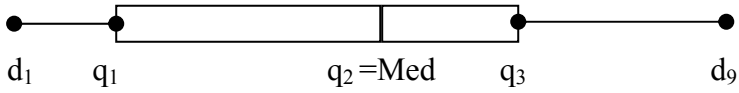
Si au moins 99,7% de la donnée est située dans $[\bar{x}-3\sigma\ ;\ \bar{x}+3\sigma]$, on dit que la distribution est normale à 99,7%.

Intervalles de confidence de la ***Médiane*** :
on admet que à 95% la médiane ***Med*** est situé dans l'intervalle

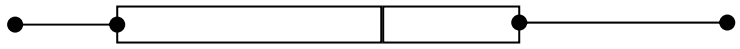
$$[\boldsymbol{Med} - 1{,}5\frac{q_3 - q_1}{\sqrt{n}};\ \boldsymbol{Med} + 1{,}5\frac{q_3 - q_1}{\sqrt{n}}]$$

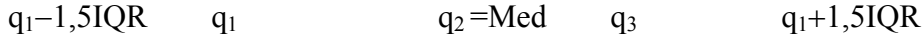
Boite à moustaches :





IQR = $\boldsymbol{q_3 - q_1}$ écart interquartile





John Tukey 1977 (tem and leaf) : ***Tige et feuille*** :

on représente la donnée 11 13 13 14 18 19 21 22 22 27 27 par
1 : 1 3 3 4 8 9
2 : 1 2 2 7 7

Statistique

1°) ***Expérience enquête (notes)*** *:*
Population : les élève de 5V (2000-2001)
Chaque élève de cette classe est un ***individu*** .
Notes de 5V : interrogation 27 septembre 200 :
Dans l'ordre lexicographique des noms :
Nombre d'élèves (***effectif total***) : **N = 30**.
Les notes : la ***donnée*** (**data**) :

6, 18, 11, 6, 9, 14, 10, 7, 15, 15, 12, 18, 10, 4, 8,
11, 4, 2, 2, 10, 8, 8, 3, 18, 10, 4, 13, 4, 3, 8

En général c'est une suite de **N** éléments :

$x_1, x_2, \ldots\ldots, x_N$

La ***variable*** statistique correspondante est ***x*** qui est une valeur générique dans la donnée.
Les ***modalités*** (caractères) : Soit la suite des données
S = (6,18,11,6,9,14,10,7,15,15,12,18,10,4,8,11,4,2,2,10,8,8,3,18,10,4,13,4,3,8)
L'ensemble E correspondant
E = {6,18,11,6,9,14,10,7,15,15,12,18,10,4,8,11,4,2,2,10,8,8,3,18,10,4,13,4,3,8 }
Cet ensemble contient des éléments répétés, si on élimine les répétitions des éléments, (pour ce faire, il vaut mieux ordonner d'abord cette suite)
S' = (2,2,3,3,4,4,4,4,6,6,7,8,8,8,8,9,10,10,10,10,11,11,12,13,14,15,15,18,18,18)

E = {6,18,11,6,9,14,10,7,15,15,12,18,10,4,8,11,4,2,2,10,8,8,3,18,10,4,13,4,3,8}
={**2**,2,**3**,3,**4**,4,4,4,**6**,6,**7**,**8**,8,8,8,9,**10**,10,10,10,**11**,11,**12**,**13**,**14**,**15**,15,**18**,18,18}
L'ensemble E s'écrit aussi : E ={2,3,4,6,7,8,9,10,11,12,13,14,15,18}
Chaque élément de cet ensemble est une ***modalité*** (caractère) :

Les ***fréquences*** : on appelle fréquence d'une modalité, le nombre de fois que celle-ci apparaît dans la donné : ceci donne lieu à un ***tableau de fréquence*** :

modalité	2	3	4	6	7	8	9	10	11	12	13	14	15	18
effectifs	2	2	4	2	1	4	1	4	2	1	1	1	2	3

En général les ***modalités*** forment une suite $y_1, y_2, \ldots, y_p$. Dont les effectifs partiels correspondants : $n_1, n_2, \ldots\ldots, n_p$.
Effectifs cumulés croissants : c'est la suite :
$v_1 = n_1$.
$v_2 = n_1 + n_2 = v_1 + n_2$.
$v_3 = n_1 + n_2 + n_3 = v_2 + n_3$.
…….
$v_j = v_{j-1} + n_j$.
…….
$v_p = n_1 + n_2 + n_3 + \ldots\ldots + n_p = v_{p-1} + n_p = N$

On a : pour l'Effectif ***total*** :

$$\boxed{N = n_1 + n_2 + \ldots\ldots + n_p}$$

L'***effectif total*** vaut la somme des ***effectifs partiels***.
Les ***fréquences respectives*** :

$$\boxed{f_1 = \frac{n_1}{N} , f_2 = \frac{n_2}{N}, \ldots\ldots , f_p = \frac{n_p}{N}}$$

(Ce sont des proportions, on peut les écrire en pourcentages)
Les ***fréquences cumulées croissantes*** :
$c_1 = f_1$.
$c_2 = f_1 + f_2 = c_1 + f_2$.
$c_3 = f_1 + f_2 + f_3 = c_2 + f_3$.
…….
$c_j = c_{j-1} + f_j$.
…….
$d_p = f_1 + f_2 + f_3 + \ldots\ldots + f_p = c_{p-1} + f_p = 1$
Les ***fréquences cumulées décroissantes*** :
$d_1 = f_1 + f_2 + f_3 + \ldots.. + f_p = 1$
$d_2 = f_2 + f_3 + \ldots.. + f_p = d_1 - f_1$.
$d_3 = f_3 + \ldots.. + f_p = d_2 - f_2$.
…….
$d_j = d_{j-1} + f_{j-1}$.
…….
$d_p = f_p$.
Les ***classes*** : Si on regroupe la donnée en des paquets de valeurs prises par intervalles, chaque paquet est appelé une ***classe***.

Exemple : Grouper la donnée en quatre classes : [0 ;5[, [5 ;10[, [10 ;15[, [15 ;20] :

[0;5[	2+2+4 = 8
[5;10[	2+1+4+1 = 8
[10;15[	4+2+1+1+1= 9
[15;20]	2+3 = 5

Ce tableau contient les classes et la ''fréquence'' de chacune.

L'***étendue*** : une donnée a une étendue c'est la différence entre la plus grande valeur et la plus petite :

$$\text{Etendue}(\mathbf{x}) = \text{Max}(\mathbf{x}) - \text{Min}(\mathbf{x})$$

Dans l'exemple : Max(x) = 18 ; min(x) = 2

Etendue(**x**) = Max(**x**) – Min(**x**) = 18 – 2 = 16

N.B. L'étendue est une caractéristique de ***dispertion*** : plus l'étendue est petite plus que la donnée est concentrée, plus que l'étendue est grande, plus que la donnée est dispersée.

Valeur milieux : $\text{milieu}(\mathbf{x}) = \dfrac{Max(x) + Min(x)}{2} = 10$

Le mode : c'est la modalité la plus fréquente : ***mode*** = 4 ou 8.

La moyenne (***mean***) : c'est la somme de toute la donnée divisée par l'effectif total N:

$$\bar{x} = \frac{x_1 + x_2 + \cdots\cdots + x_N}{N}$$

Une autre formule qui donne cette moyenne :

$$\bar{x} = \frac{n_1 y_1 + n_2 y_2 + \cdots\cdots + n_p y_p}{n_1 + n_2 + \cdots\cdots + n_p}$$

N.B. La ***moyenne*** est une caractéristique centrale, elle donne le centre de gravité de la donnée.

Variable centré : c'est la nouvelle série statistique **z** obtenue en retranchant la moyenne $\bar{x}$:

$$z_1 = x_1 - \bar{x}\ ,\ z_2 = x_2 - \bar{x}\ ,\ \ldots\ldots\ ,\ z_N = x_N - \bar{x}$$

Variance : c'est la moyenne quadratique des écarts des valeurs de la donnée par rapport à sa moyenne $\bar{x}$ (c'est la moyenne des carrés de la variable centrée) :

$$\text{var}(\mathbf{x}) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots\cdots + (x_N - \bar{x})^2}{N}$$

L'écart Type (standard deviation) : c'est la racine carrée de la variance :

$$\sigma\,(x) = \sqrt{\text{var}(x)}$$

$$\sigma\,(x) = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots\cdots + (x_N - \bar{x})^2}{N}}$$

Les quartiles : ce sont des valeurs q_1, q_2, q_3. qui divisent [Min(x) ; Max(x)] en quatre intervalles :

[***Min(x) ; q_1***] , [q_1; q_2] , [q_2; q_3] , [q_3; ***Max(x)***]

Chacune de ces intervalles, contient le quart de l'effectif total. q_1, q_2, q_3. (Ne figurent pas forcément dans la donnée

q_1 est une valeur telle que :

25% de la donnée sont $\leq q_1$

75% de la donnée sont $\geq q_1$

Il y des fois plusieurs valeurs qui vérifient ceci.

La norme ***afnor*** consiste à prendre la moyenne des deux modalités de la donnée qui sont le plus proche de q_1.

La médiane $Me = q_2$:

Q_2 est une valeur telle que :

50% de la donnée sont $\leq q_2$

50% de la donnée sont $\geq q_2$

La norme ***afnor*** consiste à prendre la moyenne des deux modalités de la donnée qui sont le plus proche de ***Me***.

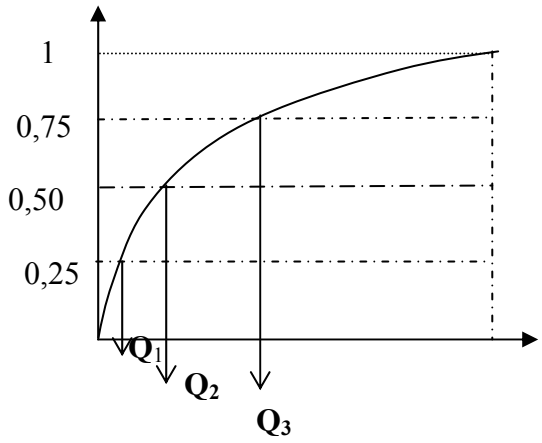
Q_3 est une valeur telle que :

75% de la donnée sont $\leq q_3$

25% de la donnée sont $\geq q_3$

La norme ***afnor*** consiste à prendre la moyenne des deux modalités de la donnée qui sont le plus proche de q_3.

Calcul pratique des quartiles : On dessine la courbe des ***fréquences cumulées croissantes*** par les points : (y_1, c_1),……., (y_p, c_p)



On divise l'intervalle [0 ; 1] de l'axe des ordonnées en 4 parties égales, les points obtenus 0,25 ; 0,50 ; 0,75 correspondent aux abscisses q_1, q_2, q_3.

Les déciles : ce sont neuf valeurs d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, qui divisent [Min(x) ; Max(x)] en dix intervalles

[Min(x) ; d_1] , [d_1; d_2] , [d_2; d_3] , [d_3; d_4] , [d_4; d_5] ,
[d_5; d_6] , [d_6; d_7] , [d_7; d_8] , [d_8; d_9] , [d_9; Max(x)]

Dont chacun contient le dixième de l'effectif total :

N.B. On peut appliquer la même méthode graphique précédente.

Les centiles : On fait une définition semblable en découpant en cent parties égales.