Statistique Descriptive à une Seule Variable.

0. Introduction :

Définition 1: On appelle ***statistique*** l'ensemble des outils scientifiques, à partir des quels on peut recueillir, ordonner, analyser et tirer des conclusions d'un certain nombre de données.

D'après cette définition, on en déduit que l'étude des problèmes statistiques se fait en suivant les étapes suivantes :

- Recueillir les données.
- Organiser et ordonner ces données.
- Représenter graphiquement les données.
- Analyser les données.
- Expliquer les résultats obtenus.
- Tirer les conclusions.

1. Vocabulaires statistiques :

- Population (univers statistique) est l'ensemble sur le quel porte l'étude statistique.

- Unité statistique (individu) est un élément de la population.

- Echantillon est une partie (sous ensemble) de la population.

- Caractère est une caractéristique prise par les individus d'une population.

- Modalités sont les différents cas susceptibles d'être pris par un caractère.

Exemples

1) Etude des différentes spécialités choisies par les étudiants de la $2^{ème}$ année S. T.
   - Population : Les étudiants de la $2^{ième}$ année S. T.
   - Individu : Un étudiant.
   - Caractère étudié : Les spécialités choisies.
   - Modalités : électronique, mécaniques, . . . etc.
2) Etude du poids des nouveaux née.
   - Population : Les nouveaux née.
   - Individu : Un nouveau née.
   - Caractère étudié : Le poids.
   - Modalités 3.400 kg, 2.900 kg, . . . etc.

3) Etude du nombre des travailleurs dans un certain nombre de petites entreprises.

- Population : Les petites entreprises.
- Individu : Une petite entreprise
- Caractère étudié : Le nombre des travailleurs.
- Modalités : 10, 25, 5, . . . etc.

On remarque, d'après ces exemples, qu'il existe deux types de caractères :

- Caractère qualitatif dont les modalités ne sont pas mesurables.

- Caractère quantitatif dont les modalités sont mesurables. On lui donne souvent le nom de **variable statistique.**

D'après ces modalités, une variable statistique peut être **discrète** si elle prend des valeurs isolées (dans *IN*), ou **continue** si elle prend des valeurs dans un intervalle

Définition 2 Une **série statistique** est la correspondance entre les individus d'une population et les modalités du caractère étudié.

Notations

- Les caractères seront notés par : *X, Y,* . . ., etc.
- Les modalités sont notées par les lettres minuscules: $x_1, x_2, \ldots, x_k$ et $y_1, y_2, \ldots, y_k$.

Dans la suite de ce chapitre nous allons restreindre l'étude aux variables statistiques.

2. Etude des séries statistiques.

2.1. Représentation des séries statistiques.

Pour étudier un caractère ou une variable statistique, la première opération consiste à recueillir toutes les informations voulues. Elles doivent être ordonnées dans des **tableaux statistiques** ou bien on donne des visualisations **graphiques,** qui donnent un résumé plus clair et facilite l'interprétation des données.

2.1.1. Tableaux statistiques.

(a) Cas d'une variable discrète.

Considérons une population à N individus, décrite suivant une variable statistique discrète X ayant les valeurs $(x_1, x_2, \ldots, x_k)$. On s'intéresse donc à connaître, pour chaque valeur x_i , le nombre d'individus prenant cette valeur, ce nombre est noté par $\boldsymbol{n_i}$, $i = 1,\ldots,k$. Nous obtenons donc le ***tableau statistique*** suivant :

Les valeurs x_i	n_i
x_1	n_1
x_2	n_2
.	.
.	.
.	.
x_k	n_k
Total	N

Définitions 3:

- Le nombre d'individus n_i de la population, pour les quels la variable X prend la valeur x_i, est appelé **effectif** ou **fréquence absolue** de la valeur x_i.
- La **fréquence relative** f_i de la valeur x_i d'effectif n_i est donnée par la formule $f_i = \frac{n_i}{N}$, ou N est l'**effectif total** de la population.
- Le **pourcentage** p_i de la valeur x_i d'effectif n_i est donné par la formule $p_i = f_i \times 100 = \frac{n_i}{N} \times 100$.

Remarques : - $\sum_{i=1}^{k} n_i = N \; et \; 0 \le n_i \le N$, ou k est le nombre des valeurs différentes.

- $\sum_{i=1}^{k} f_i = 1$ et $\sum_{i=1}^{k} p_i = 100$.
- La correspondance entre les valeurs de x_i et leurs effectifs s'appelle **distribution d'effectifs.**

Exemple 1. Un fabriquant de tissue essaie une nouvelle machine, il compte le nombre de défauts sur 75 échantillons de 10 mètres. Il a obtenu les résultats suivants

Nombre de défauts x_i	**Nombre d'échantillons n_i**
0	38
1	15
2	11
3	6
4	3
5	2
Total	75

Effectifs cumulés.

Il peut être intéressant par la lecture du tableau de répondre à des questions de la forme:

- Quel est le nombre d'individus pour les quels la variable X prend au moins x_j ?
- Quel est le nombre d'individus pour les quels la variable X prend au plus x_j ?

La réponse à la 1^{ére} question se fait en additionnant les effectifs à partir de la première valeur n_1 jusqu'à n_j ($1 \leq j \leq k$). Les nombres ainsi obtenus sont appelés **effectifs cumulés croissants** ou **fréquences absolus cumulées croissantes,** notés par $\boldsymbol{n_{ic}\uparrow}$**.**

La réponse à la 2ème question se fait en additionnant les effectifs à partir de n_j ($1 \leq j \leq k$) jusqu'à la dernière valeur n_k. Les nombres ainsi obtenus sont appelés **effectifs cumulés décroissants** ou **fréquences absolus cumulées décroissantes**, notés par $\boldsymbol{n_{ic}\downarrow}$**.**

Par exemple, la 4ème ligne du tableau -1-, se lit aussi :

- 6 échantillons contiennent 3 défauts.
- 70 échantillons contiennent au plus 3 défauts.
- 11 échantillons contiennent au moins 3 défauts.

Le tableau -1- pourra être compléter de la manière suivante

Nombre de défauts x_i	**Nombre d'échantillons n_i**	$\boldsymbol{n_{ic}\uparrow}$	$\boldsymbol{n_{ic}\downarrow}$
0	38	38	75
1	15	53	37
2	11	64	22
3	6	70	11
4	3	73	5
5	2	75	2
Total	75		

Tableau -1-

Remarque : De la même manière on peut définir

- Les **fréquences relatives cumulées** (croissantes et décroissantes).
- Les **pourcentages cumulés** (croissants et décroissants).

b) Cas d'une variable continu.

Dans la cas d'une variable continue, théoriquement les valeurs recueillies sont infinies et très proches l'une de l'autre. Alors, pour simplifier l'étude on construit des **classes** (intervalles) en divisant l'**étendue** de la série statistique en plusieurs intervalles.

Définitions

- L'**étendue** d'une série statistique est la différence entre la plus grande et la plus petite valeur dans la série.
- Les **classes** sont des intervalles de la forme $[a_i\ ,\ a_{i+1}[$, tel que $\overline{\bigcup_{i=1}^{k-1}[a_i, a_{i+1}[} = [a_0, a_k]$; ou a_0 et a_k sont respectivement la plus petite et la plus grande valeur de la série.
- Dans la classe $[a_i\ ,\ a_{i+1}[$, les valeurs a_i et a_{i+1} sont les **bornes** ou les **limites** de cette classe.
- Le nombre $x_i = \frac{a_i + a_{i+1}}{2}$ s'appelle le **centre** de la classe $[a_i\ ,\ a_{i+1}[$.
- Le nombre $\alpha_i = a_{i+1} - a_i$ s'appelle l'**étendue** ou **amplitude** de la classe $[a_i\ ,\ a_{i+1}[$.
- L'effectif n_i la classe $[a_i\ ,\ a_{i+1}[$.correspond au nombre de valeurs appartenant à cette classe

Remarque : Le nombre de classes k ne doit pas être trop petit, perte d'information, ni trop grand, le regroupement en classe est alors inutile. Le nombre de classe qu'on peut construire est donné par la formule $k = \sqrt{N}$.

Exemple 2 Une étude concernant le poids de 80 nouveaux née dans une maternité a donné les résultats suivant

Les classes	**Les centres de classes x_i**	**Les effectifs n_i**	$n_{ic}\uparrow$	$n_{ic}\downarrow$
[2.2 , 2.5[	2.35	2	2	80
[2.5 , 2.8[	2.65	5	7	78
[2.8 ,3.1[	2.95	20	27	73
[3.1 , 3.4[	3.25	19	46	53
[3.4 , 3.7[	3.55	20	66	34
[3.7 , 4.0[	3.85	8	74	14
[4.0 , 4.3[	4.15	4	78	6
[4.3 , 4.6[	4.45	2	80	2
Total		80		

Tableau -2-

2.1.2. Représentation graphique.

Aux tableaux statistiques, on associe des représentations graphiques qui permettent une compréhension plus globale des données. Selon le besoin, on représente :

- Les effectifs et les effectifs cumulés.
- Les fréquences relatives et les fréquences relatives cumulées.

Les représentations graphiques sont différentes selon le type de la variable.

(a) Variable discrète.

- Diagramme en bâtons. Cette représentation, se fait en portant sur l'axe des abscisses, les valeurs x_i prisent par la v. s. puis, on trace à partir de chaque point x_i un bâton dont la longueur est proportionnelle à n_i ou f_i.

- Le polygone des fréquences (ou des effectifs) Cette représentation est obtenue en joignant les sommets bâtons.

- Le polygone des effectifs cumulés (ou courbe cumulative des effectifs)

On définie la fonction qui associe à chaque valeur $x \in \mathbb{R}$, la somme des effectifs de tout les $x_i < x$ et qu'on appelle **fonction de distribution des effectifs**. La représentation graphique de cette fonction est appelée **polygone des effectifs cumulés.**

(b) Variable continue.

b1) Cas des classes à étendues égale

(i) Histogramme. Sur l'axe des abscisses, on représente les bornes des différentes classes et on associe à chaque classe un rectangle, dans la base est une partie de l'axe des abscisses comprise entre les bornes de cette classe et dont la longueur est proportionnelle à n_i ou f_i.

(ii) Le polygone des fréquences (ou des effectifs) Cette représentation est obtenue en joignant les points (x_i , n_i) par des segments de droite. On complète par 2 classes extrêmes de même amplitude.

Remarques.

1- L'aire de touts les rectangles est égale à 1 si on représente les fréquences relatives et n si on représente les effectifs.

2- L'aire comprise une le polygone des effectifs et l'axe des abscisses est égale à l'aire de l'histogramme.

b2) Cas des classes à différents étendues.

Reprenons l'exemple 2 et on regroupe les classes 1 et 2 en une seule classe, ainsi que les classes 6, 7 et 8. Donc le tableau statistique devient

Les classes	x_i	n_i	Les effectifs rectifiables $\tilde{n}_i$
[2.2 , 2.8[	2.5	7	(2+5)/2 = 3.5
[2.8 ,3.1[	2.95	20	20
[3.1 , 3.4[	3.25	19	19
[3.4 , 3.7[	3.55	20	20
[3.7 , 4.6[	4.15	14	4.66
Total		80	

Tableau -3-

Dans ce cas, les rectangles de l'histogramme ont une longueur proportionnelle à $\tilde{n}_i$, l'effectif rectifiable de chaque classe.

(ii) Le polygone des effectifs cumulés croissants On obtient le polygone des effectifs cumulés croissants en joignant, par des segments droits, les points ayant pour abscisse les bornes supérieurs des classes et pour ordonnés les effectifs cumulés (ou les fréquences relatives cumulées) croissants correspondant à la classe considérée. Le premier point est $(a_0,0)$.

(iv) Le polygone des effectifs cumulés décroissants On obtient le polygone des effectifs cumulés décroissants en joignant, par des segments droits, les points ayant pour abscisse les bornes inférieurs des classes et pour ordonnés les effectifs cumulés (ou les fréquences relatives cumulées) décroissants correspondant à la classe considérée. Le premier point est $(a_k,0)$.

2.2. Paramètres caractéristiques.

Une distribution statistique peut être résumé par quelques valeurs appelées **paramètres caractéristiques**, classées en 3 catégories.

- Les caractéristiques à tendance centrale (position).
- Les caractéristiques de dispersion.
- Les caractéristiques de forme.

2.2.1. Les caractéristiques à tendance centrale

A1) Le mode *Mo.* Le mode est la valeur de la variable d'effectif maximum.

(i) Si la v. s. est discrète, le mode représente la valeur qui correspond au plus grand effectif.

(ii) Si la v. s. est continue, le mode est le centre de la classe modale, c.-à-d. la classe qui correspond au plus grand effectif.

Exemples - Dans l'exemple 1, le mode Mo = 0.

- Dans l'exemple 2, on a 2 classes modales qui sont [2.8 , 3.1[et [3.4 ,3.7[, donc il y 2 mode $Mo_1 = 2.95$ et $Mo_2 = 3.55$.

A2) La moyenne arithmétique. La **moyenne** d'une série statistique $(x_1 , x_2 , \ldots , x_k)$ d'effectifs $(n_1 , n_2 , \ldots , n_k)$ est la valeur réelle notée par $\bar{x}$ et donnée par la formule

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i \ .$$

Pour trouver la valeur de la moyenne, on rajoute une colonne dans le tableau statistique dont laquelle on calcule le produit $n_i x_i$, pour tout i. Si la v. s. est continue, les valeurs $x_1 , x_2 , \ldots , x_k$ représentent les centres de classes.

Exemples - Pour l'exemple 1 on a, $\bar{x} = \frac{1}{75}\sum_{i=1}^{6} n_i x_i = \frac{77}{75} = \underline{1.0267 = \bar{x}}.$

Nombre de défauts x_i	**Nombre d'échantillons n_i**	$n_i x_i$
0	38	0
1	15	15
2	11	22
3	6	18
4	3	12
5	2	10
Total	75	77

- Pour l'exemple 2 on a $\bar{x} = \frac{1}{80}\sum_{i=1}^{8} n_i x_i = \frac{266}{80} = \underline{3.325 = \bar{x}}.$

• **Les classes**	**Les centres de classes x_i**	**Les effectifs n_i**	$n_i x_i$
[2.2 , 2.5[	2.35	2	4.7
[2.5 , 2.8[	2.65	5	13.25
[2.8 ,3.1[	2.95	20	59
[3.1 , 3.4[	3.25	19	61.75
[3.4 , 3.7[	3.55	20	71
[3.7 , 4.0[	3.85	8	30.8
[4.0 , 4.3[	4.15	4	16.6
[4.3 , 4.6[	4.45	2	8.9
Total		80	266

Les propriétés de la moyenne

- La moyenne ne change pas si on remplace un nombre déterminé de valeurs par leur moyenne multipliée par la somme de leurs effectifs.
- La moyenne conserve les changements de l'axe et l'origine

$$X(x_i, n_i) \to Y(y_i = ax_i + b, n_i)$$
$$\bar{x} \mapsto \bar{y} = a\bar{x} + b$$

A3) La médiane *Me* est la valeur qui partage la population en deux parties d'effectifs égaux. Si la série statistique est ordonnée, il y a autant d'observations rangées avant que d'observations rangées après la valeur de la médiane.

(i) Variable discrète. Soit x_1 , x_2 , . . ., x_N une série ordonnée de valeurs de la v. s.

- Si N est impaire i. e. $N = 2p + 1$, alors la médiane est donnée par. $Me = x_{p+1}$.

- Si N est impaire i. e. $N = 2p$, alors la médiane est donnée par $Me = \frac{x_{p+1} + x_p}{2}$.

Exemples :

1) Soient les résultats obtenus par un étudiant dans le module de statistiques

10 9 12 10 13 14 18 13 15

La série ordonnée est : **9 10 10 12** **13** **13 14 15 15**

4 valeurs **4 valeurs**

$n = 9 = 2 \times 4 + 1$ alors la médiane $Me = x_{k+1} = x_5 =$ **13 = *Me***.

2) On garde la même série et on ajoute la valeur **11**, donc la série ordonnée devient :

9 10 10 11 12 **13 13 14 15 15**

5 valeurs **5 valeurs**

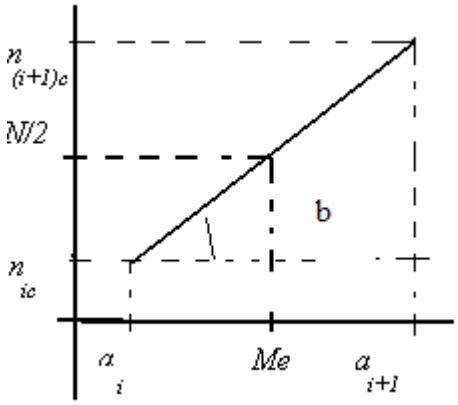
$n = 10 = 2 \times 5$, alors la médiane $Me = (x_k + x_{k+1})/2 = (x_5 + x_6)/2 =$ **12.5 = *Me***.

3) Dans l'exemple 1, on a $n = 75 = 2 \times 37 + 1$, alors $Me = (x_k + x_{k+1})/2$.

D'après le tableau -1- on a : $Me = (x_{34} + x_{34+1})/2 = Me = (0 + 0)/2 =$ **0 = *Me***.

(ii) Variable continue.

La lecture du tableau nous permet de déterminer la classe à partir de laquelle la moitié des effectifs est atteinte. La médiane appartient donc à cette classe. Cette dernière est appelé **classe médiane**. La valeur exacte de la médiane est calculée en utilisant l'**interpolation linéaire**. Si $[a_i , a_{i+1}[$ est la classe médiane, alors on a,

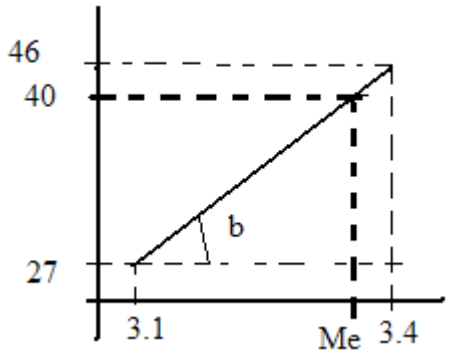


$$tgb = \frac{\sum_{j=1}^{i} n_j - \sum_{j=1}^{i-1} n_j}{a_{i+1} - a_i} = \frac{\frac{N}{2} - \sum_{j=1}^{i-1} n_j}{Me - a_i}, \text{ où } n_{ic} = \sum_{j=1}^{i-1} n_j \text{ et } n_{(i+1)} = \sum_{j=1}^{i} n_j$$

Donc $$Me = a_i + \frac{\frac{N}{2} - \sum_{j=1}^{i-1} n_j}{\sum_{j=1}^{i} n_j - \sum_{j=1}^{i-1} n_j}(a_{i+1} - a_i).$$

Exemple :

Reprenons l'exemple 2, comme $N/2 = 80/2 = 40$, on en déduit d'après le tableau -2- que $Me \in [3.1, 3.4[$, donc on utilisant l'interpolation linéaire on obtient



$$tg\alpha = \frac{46-27}{3.4-3.1} = \frac{40-27}{Me-3.1} \Leftrightarrow Me = 3.1 + \frac{46-27}{3.4-3.1}(3.4-3.1) = 3.3053kg = Me$$

2.2.2 Les caractéristiques de dispersion.

Deux séries statistiques peuvent avoir la même moyenne, la même médiane, le même mode et être cependant très différentes au sens ou les observations faites peuvent être plus au moins « ***dispersées*** »par rapport à une même valeur centrale.

Ce phénomène est mis en évidence par le calcul des nombres appelés **caractéristiques de dispersion.**

(i) L'étendue d'une distribution est la valeur donnée par $E = x_{\max} - x_{\min}$.

Dans le cas d'une variable continue les valeurs $x_{\max}$ et $x_{\min}$ sont respectivement les centres de la dernière et de la première classe.

(ii) L'écart type – La variance.

La **variance** d'une série statistique (x_1 , x_2 , . . ., x_k) d'effectifs (n_1 , n_2 , . . ., n_k) est la valeur réelle notée par V_X ou *var* (*X*) est donné par

$$V_X = Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 .$$

Lorsque la v. s. est continue, les x_i représente les centres de classes.

On appelle **écart type** de la distribution, la racine carrée positive de la variance, soit

$$\sigma_X = \sqrt{\frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2} \quad .$$

Plus l'écart type est petit, plus la distribution est rassemblée autour de la moyenne.

Remarque. Souvent, en utilise la variance sous la forme

$$V_X = \frac{1}{N}\left(\sum_{i=1}^{k} n_i {x_i}^2\right) - \bar{x}^2$$

Pour trouver la valeur de la moyenne, on rajoute une colonne dans le tableau statistique dont laquelle on calcule le produit $n_i {x_i}^2$, pour tout *i*.

Exemple Reprenons l'exemple 1, on a

x_i	n_i	$n_i x_i^2$
0	38	0
1	15	15
2	11	44
3	6	54
4	3	48
5	2	50
Total	75	211

Donc $V_X = \frac{1}{75}\left(\sum_{i=1}^{6} n_i {x_i}^2\right) - \bar{x}^2 = \frac{211}{75} - \bar{x}^2 = 1.7593 = V_x$ et dans ce cas σ_X = 1.3264.

Propriétés de la variance.

1) La variance est toujours positive ou nulle.

2) Changement d'échelle et d'origine

$$X(x_i, n_i) \to Y(y_i = ax_i + b, n_i)$$
$$V_X \mapsto V_Y = a^2 V_X$$

(iii) Coefficient de variation.

Le coefficient de variation C_{vX} est le rapport de l'écart type à la moyenne, c – à – d :

$$Cv_X = 100 \times \frac{\sigma_X}{\bar{x}}.$$

Il s'exprime sans unité et est donné en pourcentage

Le coefficient permet de comparer les dispersions des séries statistiques qui ne sont pas exprimée dans les mêmes unités de mesure ou des séries ayant des moyennes très différentes.

(iv) Ecart interquartile.

Les **quartiles** sont les valeurs qui partagent la série statistique ordonnée en 4 parties de même effectifs.

- Le premier quartile est le nombre $\boldsymbol{Q_1}$ tel que 25% des valeurs sont inférieur ou égale à Q_1.
- Le troisième quartile est le nombre $\boldsymbol{Q_3}$ tel que 75% des valeurs sont inférieur ou égale à Q_3.
- Le deuxième quartile $\boldsymbol{Q_2}$ est la médiane.

Remarques :

- Le premier quartile Q_1 est la médiane de la première moitié de la série statistique.
- Le troisième quartile Q_3 est la médiane de la deuxième moitié de la série statistique.
- La méthode de calcul des quartiles est donc identique à celle du calcul de la médiane.

L'**écart interquartile** est le nombre *IQR* tel que $IQR = Q_3 - Q_1$. Il donne l'étendue de la moitié centrale des observations.

Exemples :

1) Soient les résultats obtenus par un étudiant dans le module de statistiques

10 9 12 10 13 14 18 13 15

La série ordonnée est : **9 10 10 12** **13** **13 14 15 15**

4 valeurs **4 valeurs**

La 1er moitié de la série content 4 (= 2 × 2 = 2 × k) valeurs, donc la médiane de cette partie est

$$Q_1 = (x_k + x_{k+1})/2 = (x_2 + x_3)/2 = \underline{\mathbf{10 = Q_1}} \ .$$

La 2ème moitié de la série contient aussi 4 valeurs, donc donc la médiane de cette moitié est

$$Q_3 = (x_{4+1+k} + x_{4+1+k+1})/2 = (x_7 + x_8)/2 = (14 + 15)/2 = \underline{\mathbf{14.5 = Q_1}}$$

2) On garde la même série et on ajoute la valeur **11**, donc la série ordonnée devient :

9 10 10 11 12 13 13 14 15 15 .

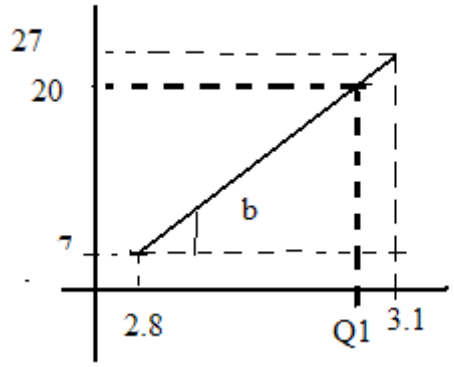
La 1er moitié de la série content 5 (= 2 × 2 +1 = 2 × k +1) valeurs, donc la médiane de cette partie est

$$Q_1 = x_{k+1} = x_3 = \underline{\mathbf{10 = Q_1}} \ .$$

La 2ème moitié de la série contient aussi 5 valeurs, donc la médiane de cette moitié est

$$Q_3 = x_{5+k+1} = (x_7 + x_8)/2 = 14 \ = \underline{\mathbf{14 = Q_1}}$$

Reprenons l'exemple 2 , $N/4$ = 20, alors d'après le tableau -2- Q_1 est [2.8 , 3.1 [. Donc en utilisant l'interpolation linéaire on obtient



$$tgb = \frac{27-7}{3.1-2.8} = \frac{20-7}{Q_1-2.8} \Leftrightarrow Q_1 = 2.8 + \frac{27-7}{3.1-2.8}(3.1-2.8) = \underline{2.995kg = Q_1} \ .$$