SERIE STATISTIQUE A UNE VARIABLE

I - RAPPELS

– L'effectif d'une classe statistique est le nombre d'éléments de la population observés dans cette classe.

– La fréquence d'une classe statistique est le rapport de l'effectif de cette classe à l'effectif total de la population. (la fréquence peut être exprimée en pourcentage)

$$f_i = \text{ fréquence de } x_i = \frac{\text{effectif de } x_i}{\text{effectif total}} = \frac{n_i}{N}$$

où x_i est une valeur donnée de la variable et n_i l'effectif correspondant.

EXEMPLE 1:
Dans un service de maintenance, on a répertorié le nombre d'interventions par jour sur un mois.
On a obtenu la distribution suivante:

Nombre d'interventions x_i	3	5	6	7	8	9
Nombre de jours n_i	2	4	9	6	3	1
Fréquences f_i (%)						

REPRESENTATIONS GRAPHIQUES

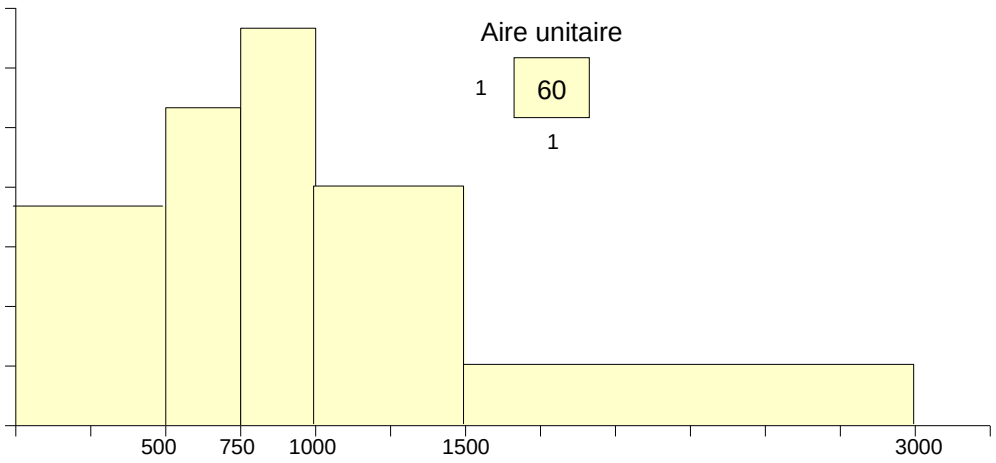
CAS DE DISTRIBUTIONS QUANTITATIVES

Les graphiques correspondant à des distributions quantitatives sont normalement réalisés en portant en abscisse la variable observée, et en ordonnée l'effectif ou la fréquence.

– Dans le cas d'une variable continue, on utilise un histogramme : **L'AIRE DE CHAQUE RECTANGLE EST PROPORTIONNELLE A L'EFFECTIF .**

Exemple 2:
Dans une succursale de banque, on a noté le montant des 2000 versements effectués au guichet pendant la journée.

Montant (en €)	]0 ; 500[	[500 ; 750[	[750 ; 1000[	[1000 ; 1500[	[1500 ; 3000[
effectif	440	320	400	480	360



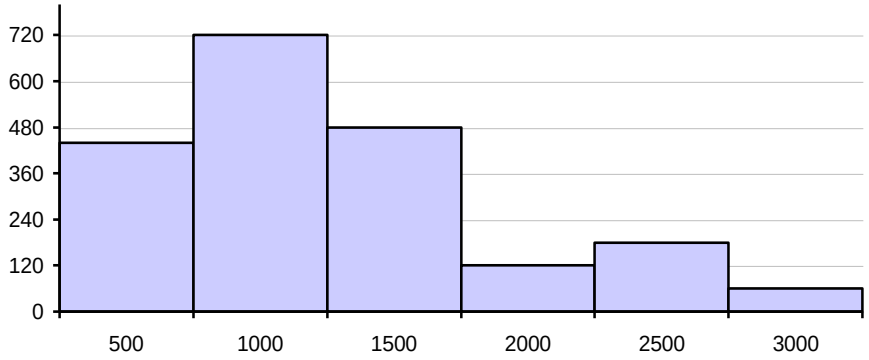
L'axe des abscisses a été gradué en prenant pour unité 250 €.
Chaque rectangle a une base égale à l'amplitude de la classe $[a_i ; a_{i+1}[$
La hauteur h de chaque rectangle est telle que $h \times \text{base} = \text{effectif} \times k$ où k est l'aire unitaire (aire du rectangle représentant un effectif égal à 1).

Par exemple la hauteur h du rectangle représentant la classe]0 ; 500[est telle que $h \times 2 = 440 \times \frac{1}{60}$

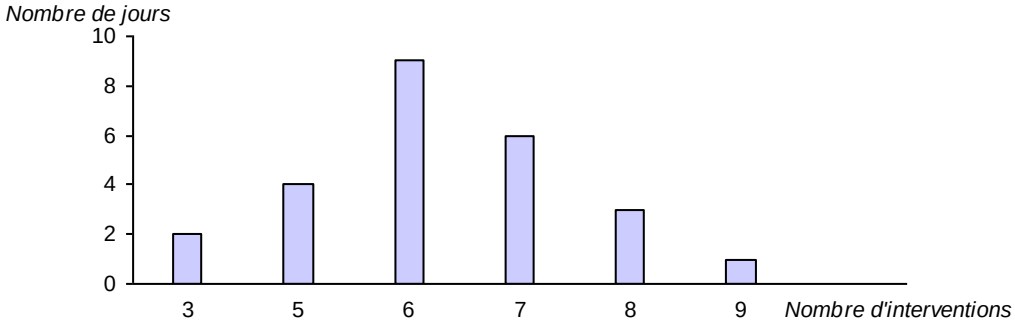
soit en cm : $h = \frac{440}{120} \times 0{,}8 \approx 2{,}9$

Dans le cas où la répartition est faite dans des classes de même amplitude, les hauteurs des rectangles sont alors proportionnelles aux effectifs.

Montant (en €)	]0 ; 500[	[500 ; 1000[	[1000 ; 1500[	[1500 ; 2000[	[2000 ; 2500[	[2500 ; 3000[
effectif	440	720	480	120	180	60



– Dans le cas d'une variable discrète, le graphique représentant la répartition est un diagramme à bâtons : **LA HAUTEUR EST PROPORTIONNELLE A L'EFFECTIF**



II - PARAMETRES DE TENDANCE CENTRALE

Trois paramètres de tendance centrale de la distribution sont utilisés : le mode, la médiane et la moyenne :

LE MODE

Le mode ou valeur modale est la valeur que la variable statistique prend le plus souvent. C'est à dire la valeur du caractère ou de la classe qui a le plus grand effectif.

Dans l'exemple 1 le mode est de 6 interventions.

Attention : Si on fait des regroupements en classes la classe modale dépend du découpage retenu.

Dans l'exemple 2 la classe modale est [1000 ; 1500[par contre si on avait effectué le regroupement par tranche de 500€ la classe modale serait [500 ; 1000[

LA MEDIANE

La médiane d'une série statistique est une valeur de la variable telle qu'il y ait autant d'observations ayant une valeur supérieure à la médiane que d'observations ayant une valeur inférieure à la médiane.

1. Lorsque les observations sont toutes données, pour calculer la médiane de la série statistique on distingue deux cas suivant que l'effectif de la population est pair ou impair :

Dans une série de données :
- si l'effectif total est 2 n + 1 où n est un entier, la médiane est la valeur classée au rang n + 1.
- si l'effectif total est 2 n où n est entier, la médiane est la demi somme des valeurs de rang n et n + 1.

Dans l'exemple 1 le nombre de journées d'intervention est impair la médiane est le nombre d'interventions de la treizième journée c'est à dire 6 interventions. En effet il y a 12 jours avec un nombre d'interventions inférieur ou égal à 6 et 12 jours avec un nombre d'interventions supérieur ou égal à 6.

2. Dans le cas d'un regroupement par classe des données on détermine la classe médiane puis on calcule la médiane par interpolation linéaire.

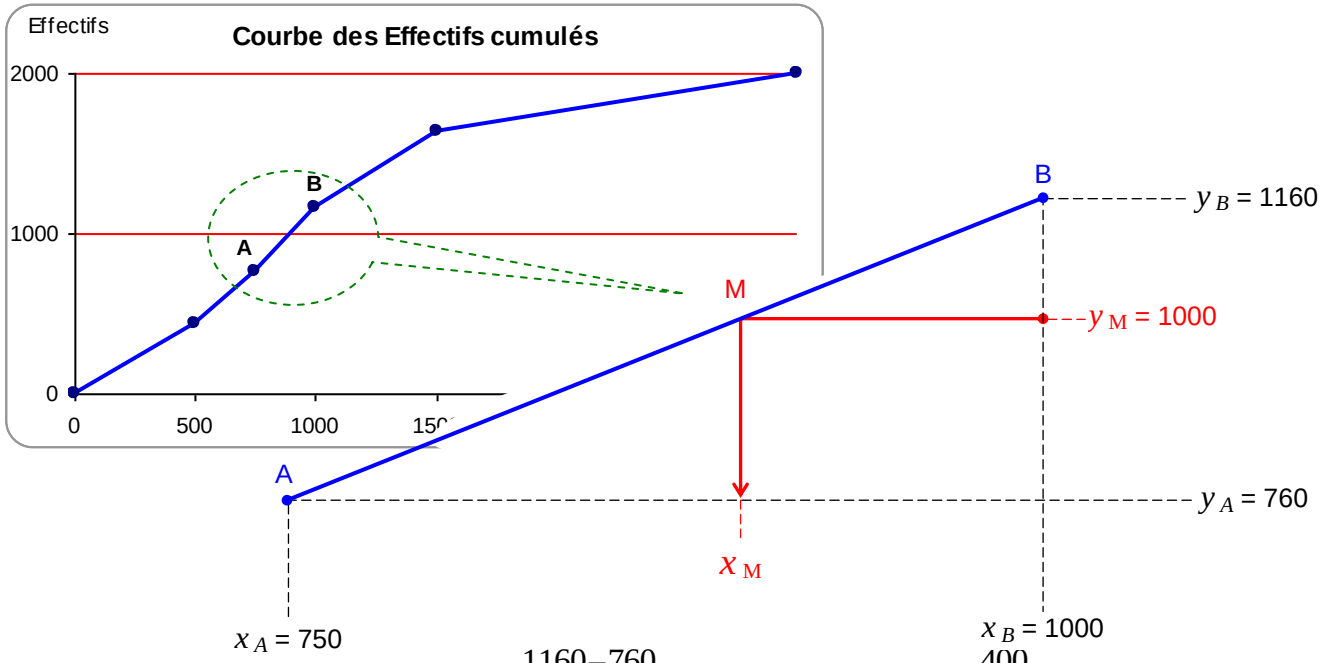
$[x_A ; x_B [$ est l'intervalle médian y_A et y_B sont les effectifs cumulés (*ou les fréquences cumulées*) correspondants respectivement aux valeurs x_A et x_B.
On note A et B les points de la courbe des effectifs cumulés (*ou des fréquences cumulées*) d'abscisses respectives x_A et x_B .

L'équation de la droite (AB) est $y - y_A = \dfrac{y_B - y_A}{x_B - x_A}\ x - x_A$.

La médiane est l'abscisse x_M du point M de la droite (AB) dont l'ordonnée y_M est la moitié de l'effectif total (*ou* 0,5 *dans le cas des fréquences cumulées*).

$$\text{Médiane } x_M = \frac{x_B - x_A}{y_B - y_A}\ y_M - y_A\ + x_A$$

Dans l'exemple 2 la classe médiane est [750 ; 1000[. La médiane est calculée par interpolation linéaire.



L'équation de la droite (AB) est $y - 1160 = \dfrac{1160 - 760}{1000 - 750}\ x - 750$ soit $y - 1160 = \dfrac{400}{250}\ x - 750$

La médiane est obtenue pour un effectif de 1000: $x_M = \left(\dfrac{250}{400}\right) \times\ 1000 - 760\ + 750$ soit $M_e = 900$

LA MOYENNE

La moyenne d'une série statistique est le quotient de la somme de toutes les valeurs de cette série par l'effectif total.

L'effectif total est $N = n_1 + \ldots + n_p$ on le note $N = \sum_{i=1}^{n} n_i$.

Les fréquences sont notées f_i

– La moyenne est donnée par la relation : $\bar{x} = \frac{1}{N}\left(n_1 \times x_1 + n_2 \times x_2 + \cdots + n_p \times x_p\right)$

$$\bar{x} = \frac{\sum_{i=1}^{p} n_i\ x_i}{N} = \frac{somme\ des\ produits\ "effectif \times valeur"}{effectif\ total} \quad \text{ou} \quad \bar{x} = \sum_{i=1}^{p} f_i\ x_i$$

Dans l'exemple 1 le nombre moyen d'interventions par jour est 6,2

$$\bar{x} = \frac{2\times 3 + 4\times 5 + 9\times 6 + 6\times 7 + 3\times 8 + 1\times 9}{25} = 6,2$$

Dans l'exemple 2 le calcul du montant moyen s'effectue en utilisant les centres des classes comme valeurs de la variable x_i

$$\bar{x} = \frac{440\times 250 + 320\times 625 + 400\times 875 + 480\times 1250 + 360\times 2250}{2000} = 1035$$

PROPRIETES DE LA MOYENNE

1. Linéarité de la moyenne

Si on multiplie chaque valeur de la série par un réel a ($a \neq 0$), alors la moyenne est multipliée par a.

Preuve :

On note $N = n_1 + \ldots + n_p$ l'effectif total, m est la moyenne de la série de valeurs ax_i

$$m = \frac{1}{N}\ n_1 \times a \times x_1 + \cdots + n_p \times a \times x_p\ = \frac{a}{N}\ n_1 \times x_1 \cdots + n_p \times x_p\ = a\bar{x}$$

Si on ajoute à chaque valeur de la série le réel b, alors la moyenne augmente de b.

Preuve :

On note $N = n_1 + \ldots + n_p$ l'effectif total, m est la moyenne de la série de valeurs $x_i + b$

$$m = \frac{1}{N}\left[n_1 \times x_1 + b + \cdots + n_p \times x_p + b\right] = \frac{1}{N}\left[n_1 \times x_1 + \cdots + n_p \times x_p + n_1 \times b + \cdots + n_p \times b\right]$$

$$m = \frac{1}{N}\ n_1 \times x_1 + \cdots + n_p \times x_p\ + \frac{1}{N}\ n_1 \times b + \cdots + n_p \times b$$

$$m = \frac{1}{N}\ n_1 \times x_1 \cdots + n_p \times x_p\ + b \times \frac{1}{N}\ n_1 + n_2 + \cdots + n_p\ = \bar{x} + b$$

On regroupe ces deux propriétés dans l'énoncé suivant :

Si une série de valeurs x_i a pour moyenne $\bar{x}$, la série de valeurs $ax_i + b$ a pour moyenne $a\bar{x} + b$. On parle de linéarité de la moyenne.

2. Ecarts à la moyenne

« La moyenne des écarts à la moyenne » est nulle.

Preuve : Il suffit d'appliquer la propriété précédente en prenant $b = \bar{x}$

3. Moyennes partielles

Si une série est partagée en deux séries d'effectifs N et P, et de moyennes $\bar{x}$ et $\bar{y}$ alors la moyenne de la série totale est $\bar{z} = \dfrac{N \times \bar{x} + P \times \bar{y}}{N + P}$.

Preuve :

	Série X			Série Y		
Série Z	x_1	…	x_k	y_1	…	y_j
effectifs	n_1	…	n_k	p_1	…	p_j

On note N et P l'effectif total respectif des séries partielles X et Y, la série Z a pour effectif total $N + P$. Les moyennes des séries X et Y sont: $\bar{x} = \dfrac{1}{N}\ n_1x_1 + \cdots + n_kx_k$ et $\bar{y} = \dfrac{1}{P}\ p_1y_1 + \cdots + p_jy_j$.

$$\frac{N\bar{x} + P\bar{y}}{N + P} = \frac{n_1 \times x_1 + \cdots + n_p \times x_p \ + \ p_1 \times y_1 + \cdots + p_j \times y_j}{N + P} = \bar{z}$$

IV - PARAMETRES DE DISPERSION

Les caractéristiques de position (Mode, Médiane, Moyenne) sont insuffisants comme on peut le voir dans l'exemple suivant

Vérifier que la moyenne, la médiane et le mode de ces deux séries de données sont identiques.

Série X	35	75	85,5	99,9	100	104,5	124	138,5	185
effectifs	12	29	48	65	44	50	27	17	8

Série Y	28,25	42,5	62,5	99,9	100	114	139,5	195,5	288,45
effectifs	18	48	52	55	40	32	35	24	10

Plusieurs paramètres de dispersion peuvent être utilisés : l'étendue, écarts interdéciles, écarts interquartiles et écart-type.

1. L'ETENDUE

L'étendue est la différence entre les deux valeurs extrêmes observées.

L'étendue de la série X est : $185 - 35 = 150$, celle de la série Y est : $288{,}45 - 28{,}25 = 260{,}2$.

2. LES QUANTILES

1) Les quartiles au nombre de trois (Q_1, Q_2 et Q_3) partagent l'ensemble étudié de n éléments préalablement classés par valeurs croissantes, en 4 sous ensembles.

2) Les déciles au nombre de neuf (D_1, D_2 … D_9) partagent l'ensemble étudié de n éléments préalablement classés par valeurs croissantes, en 10 sous ensembles.

Les valeurs d'une série d'effectif n sont rangées par ordre croissant : $x_1 \leq x_2 \leq ... \leq x_n$

- Le premier quartile Q_1 de la série est la valeur x_i dont l'indice i est le plus petit entier supérieur ou égal à $\frac{n}{4}$.
- Le deuxième quartile Q_2 de la série est la valeur x_i dont l'indice i est le plus petit entier supérieur ou égal à $\frac{2n}{4} = \frac{n}{2}$.
- Le troisième quartile Q_3 de la série est la valeur x_i dont l'indice i est le plus petit entier supérieur ou égal à $\frac{3n}{4}$.
- Le premier décile D_1 de la série est la valeur x_i dont l'indice i est le plus petit entier supérieur ou égal à $\frac{n}{10}$.
- Le neuvième décile D_9 de la série est la valeur x_i dont l'indice i est le plus petit entier supérieur ou égal à $\frac{9n}{10}$.

L'intervalle interquartile est égal à la différence entre le troisième et le premier quartile. Il contient au moins 50% des observations.

L'intervalle interdécile est égal à la différence entre le neuvième et le premier décile. Il contient au moins 80% des observations.

L'intervalle qui sépare deux quantiles extrêmes améliore la notion d'étendue en éliminant les valeurs extrêmes.

Exemples

Dans la série X l'effectif est de 300, le rang du premier quartile est $\frac{300}{4} = 75$ soit $Q_1 = 85{,}5$ .On calcule de même $Q_2 = 99{,}9$ et $Q_3 = 104{,}5$. L'intervalle interquartile est : $Q_3 - Q_1 = 104{,}5 - 85{,}5 = 19$.

Les déciles sont : $D_1 = 75$ … $D_9 = 124$. L'intervalle interdécile est : $D_9 - D_1 = 124 - 75 = 49$.

Dans la série Y l'effectif est de 314, l'indice du premier quartile est **79** car $\left(\frac{314}{4} = 78{,}5\right)$ et 79 est le plus petit entier supérieur ou égal à 78,5 ainsi $Q_1 = 62{,}5$.

$\left(\frac{3 \times 314}{4} = 235{,}5\right)$ d'où l'indice 236 du troisième quartile et $Q_3 = 114$.

L'intervalle interquartile est : $Q_3 - Q_1 = 114 - 62{,}5 = 51{,}5$.

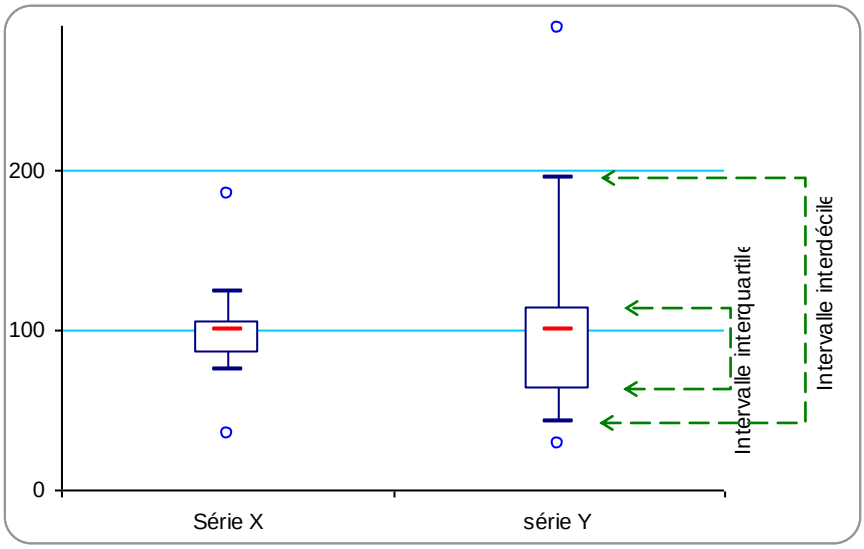
Les indices des premier et neuvième déciles sont respectivement 32 et 283 ainsi : $D_1 = 42{,}5$ et $D_9 = 195{,}5$

L'intervalle interdécile est : $D_9 - D_1 = 195{,}5 - 42{,}5 = 153$.

3. BOITES A MOUSTACHES

La représentation graphique de la dispersion d'une série statistique se fait à l'aide de graphiques appelés « boîte à moustaches » ou « box-plot ».
Pour une catégorie donnée, on construit, en face d'un axe permettant de repérer les quantiles de la variable étudiée, un rectangle dont la longueur est égale à l'intervalle interquartile, la médiane est représentée par un trait. Deux traits repèrent le premier et neuvième décile. Les observations n'appartenant pas à l'intervalle interdécile sont représentées à l'aide de points. (*On se contente parfois des valeurs extrêmes*)

Graphiques boîtes à moustaches des séries *X* et *Y*



4. VARIANCE ET ECART TYPE

La moyenne des écarts à la moyenne étant nulle elle ne peut pas servir d' indicateur de dispersion.

1. Théorème

La moyenne $\bar{x}$ est le nombre qui minimise la somme $S(x)=\sum_{i=1}^{p} n_i \left(x_i - x\right)^2$

Preuve :

$$S(x)=n_1\left(x_1-x\right)^2+\cdots+n_p\left(x_p-x\right)^2=n_1\left(x_1^2-2x_1x+x^2\right)+\cdots+n_p\left(x_p^2-2x_px+x^2\right)$$

$$S(x)=\left(n_1x_1^2+\cdots+n_px_p^2\right)-2x\left(n_1x_1+\cdots+n_px_p\right)+x^2\left(n_1+\cdots+n_p\right)$$

On note $N = n_1 + \ldots + n_p$ l'effectif total d'où $S(x)=N\,x^2-2x\sum_{i=1}^{p} n_ix_i+\sum_{i=1}^{p} n_ix_i^2$.

Ainsi la somme S se présente sous la forme d'un polynôme du second degré en x dont le coefficient N de x^2 est positif

La somme S est donc minimale pour $x=\dfrac{2\sum_{i=1}^{p} n_ix_i}{2N}$ soit $x=\bar{x}$.

Pour obtenir un indicateur de dispersion on utilise la somme des carrés des écarts à la moyenne.

2. LA VARIANCE

La variance est la moyenne des carrés des écarts à la moyenne. C'est un nombre positif.

On note $N = n_1 + \ldots + n_p$ l'effectif total et f_i la fréquence

$$V(x) = \frac{n_1 \left(x_1 - \bar{x}\right)^2 + \cdots + n_p \left(x_p - \bar{x}\right)^2}{N} = \frac{\sum_{i=1}^{p} n_i \left(x_i - \bar{x}\right)^2}{N} \quad \text{ou} \quad V(x) = \sum_{i=1}^{p} f_i \left(x_i - \bar{x}\right)^2$$

Pour simplifier les calculs de la variance on préfère utiliser les formules :

$$V(x) = \frac{n_1 \times x_1^2 + \cdots n_p \times x_p^2}{N} - \bar{x}^2 = \frac{\sum_{i=1}^{p} n_i \times x_i^2}{N} - \bar{x}^2 \quad \text{ou} \quad V(x) = \left[\sum_{i=1}^{p} f_i x_i^2\right] - \bar{x}^2$$

Preuve avec les effectifs (*la démonstration avec les fréquences étant similaire*):

$$V(x) = \frac{n_1 \left(x_1 - \bar{x}\right)^2 + \cdots + n_p \left(x_p - \bar{x}\right)^2}{N} = \frac{n_1 x_1^2 - 2x_1\bar{x} + \bar{x}^2 + \cdots + n_p x_p^2 - 2x_p\bar{x} + \bar{x}^2}{N}$$

$$V(x) = \frac{n_1 x_1^2 + \cdots + n_p x_p^2 - 2\bar{x}\left(n_1 x_1 + \cdots + n_p x_p\right) + \bar{x}^2\left(n_1 + \cdots + n_p\right)}{N}$$

$$V(x) = \frac{n_1 x_1^2 + \cdots + n_p x_p^2 - 2\bar{x}\left(\sum_{i=1}^{p} n_i x_i\right) + \bar{x}^2 \sum_{i=1}^{p} n_i}{N} = \frac{n_1 x_1^2 + \cdots + n_p x_p^2 - 2\bar{x}\, N\bar{x} + N\bar{x}^2}{N}$$

$$V(x) = \frac{n_1 x_1^2 + \cdots + n_p x_p^2 - N\bar{x}^2}{N} = \frac{\sum_{i=1}^{p} n_i x_i^2}{N} - \bar{x}^2$$

Exemple

Série X	35	75	85,5	99,9	100	104,5	124	138,5	185
effectifs	12	29	48	65	44	50	27	17	8
$n_i x_i^2$	14700	163125	350892	648700,65	440000	546012,5	415152	326098,25	273800

$V(x) = \frac{3178480,4}{300} - 100^2 = \frac{44620,1}{75}$ et l'arrondi à 10^{-3} près de $V(x)$ est : 594,935

Série Y	28,25	42,5	62,5	99,9	100	114	139,5	195,5	288,45
effectifs	18	48	52	55	40	32	35	24	10
$n_i y_i^2$									

$V(y) =$

3. L'écart type

Pour des raisons de concordance des unités on utilise la racine carrée de la variance.

L'écart type d'une série est égal à la racine carrée de la variance $s_x = \sqrt{V(x)}$

Vérifier à l'aide de la calculatrice que l'écart type de la série X est 24,391 et celui de la série Y: 55, 281.