Chapitre 2 - Statistiques descriptives univariées

Synthèse

Sommaire :

Les statistiques descriptives ont pour objet de rassembler, regrouper, représenter les données en construisant des distributions, des tableaux, des graphes. L'analyse descriptive cherche à résumer les principales caractéristiques de l'ensemble des données en calculant des paramètres et des coefficients. Tous les résultats sont strictement limités à la population étudiée (ou aux objets mesurés).

Pour la partie statistique, on va se baser sur l'exemple suivant :

Note à l'Examen N-1	Effectifs
0	0
1	0
2	1
3	0
4	1
5	0
6	1
7	0
8	1
9	1
10	1
11	2
12	2
13	1
14	1
15	1
16	1
17	0
18	1
19	0
20	0
Total	**15**

1 Définitions des principales notions

Variable : caractère statistique. Dans le cas présent, les notes à l'examen N-1.

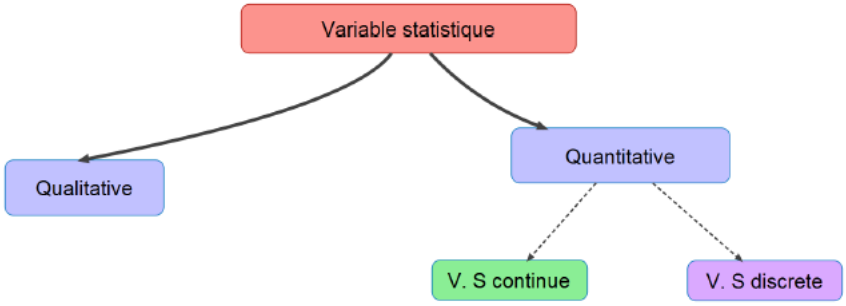
Modalités (xi) : caractéristique ou valeur que peut prendre la variable (les notes).

Effectifs (ni) : nombre de quantités observées par caractéristique (les étudiants/effectifs).

Fréquence : proportion des effectifs observés pour une modalité de la variable / total des effectifs observés pour toutes les modalités.

Exemple :

Note à l'Examen (Xi)	Effectifs (Ni)	Fréquences
0	0	0,00
1	0	0,00
2	1	0,07
…	…	…
18	1	0,07
19	0	0,00
20	0	0,00
Total	**15**	**1**



Variable qualitative : les différentes valeurs que peut prendre la variable sont des modalités qualitatives (sexe : M, F…, profession : cadre, non-cadre…). Ces valeurs peuvent être numériques, mais chaque nombre indique une qualité, une caractéristique.
La variable qualitative est **non mesurable**. Elle est qualifiée par **des modalités**.
On considère qu'il existe 2 types de modalités :

- **Ordinale :** des modalités qu'on peut classer (des rangs), exemple : primaire, secondaire, supérieur…
- **Nominale** : des modalités où le classement est indifférent (ex : yeux bleus, verts…).

Variable quantitative (numérique ou cardinale) : les différentes valeurs que peut prendre la variable appartiennent à une échelle numérique (les opérations arithmétiques ont un sens). Dans le cas présent, les notes à l'examen.

Variable quantitative discrète : est une variable ne prenant que des valeurs **isolées entières** (plus rarement décimales). Le nombre de valeurs distinctes d'une telle variable est habituellement assez faible. Ce sont des **valeurs précises**. Dans le cas présent, les notes à l'examen.

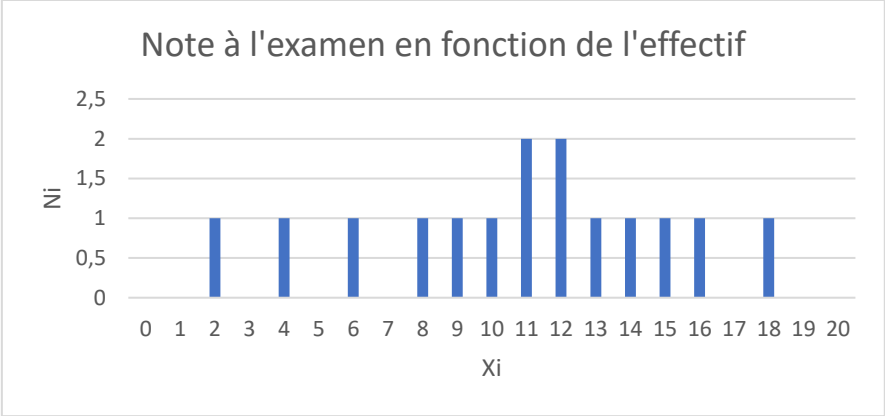
Exemple :

La variable statistique "couleur de maisons d'un quartier" est-elle ?
La variable statistique "revenu brut" est-elle ?
La variable statistique "nombre de maisons vendues par ville" est-elle ?

Solution : Pour le premier cas, la variable statistique est qualitative. Pour le deuxième cas, la variable statistique est quantitative continue. Pour le troisième cas, la variable statistique est quantitative discrète.

De façon générale, pour représenter une variable quantitative discrète, on peut utiliser un diagramme en bâtons :

Exemple :



Néanmoins cette forme se prête difficilement à l'interprétation. Pour y remédier, il faut créer des classes de notes (nombre d'individus ayant obtenu des notes comprises entre 0 et 4, entre 4 et 8 . . .) ; cette approche nous permet d'obtenir une variable dite classée ou continue.

Variable quantitative continue : Une variable quantitative est dite continue lorsque les observations qui lui sont associées ne sont pas des valeurs **précises, mais des intervalles**. C'est le cas lorsque nous avons un grand nombre d'observations distinctes.
Dans le cas d'un caractère quantitatif continu, l'établissement du tableau de fréquences implique d'effectuer au préalable une **répartition en classes** des données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou **intervalle de classe.**

En règle générale, on choisit des classes de même **amplitude**. Pour que la distribution en fréquence est un sens, il faut que chaque classe comprenne un nombre suffisant de valeurs (n_i). Diverses formules empiriques permettent d'établir le **nombre de classes** pour un échantillon de taille n.

- La règle de **STURGE** : Nombre de classes = 1+ (3,3 log n)
- La règle de **YULE** : Nombre de classes = $2.5\sqrt[4]{N}$

L'**intervalle** entre chaque classe est obtenu ensuite de la manière suivante :

- Intervalle de classe = (X max - X min) / Nombre de classes,

avec X max et X min, respectivement la plus grande et la plus petite valeur de X dans la série statistique.

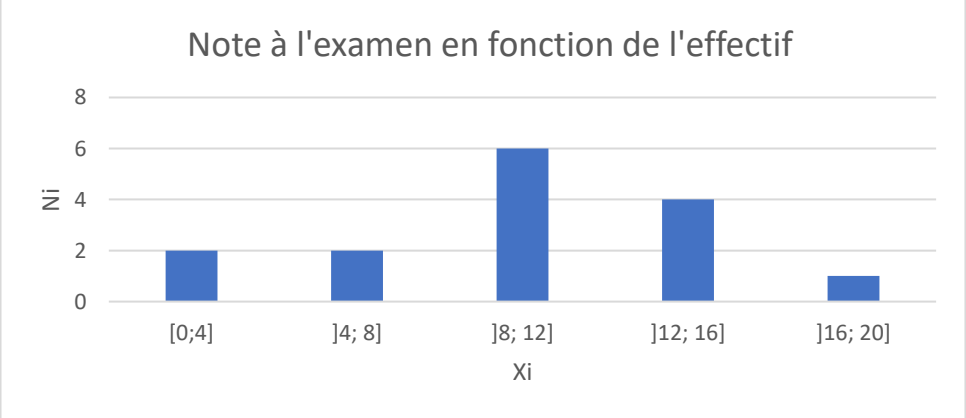
Exemple :

Sturge : k = 1 + 3.3 log(N) = 1 + 3.3 log(15) = 4,88 => 5 classes (arrondir au supérieur)

Intervalle : (20-0)/4,88 = 4,10. On arrondit à l 'inférieur, **4.**

Note à l'Examen (Xi)	Effectifs (Ni)	Fréquences
[0;4]	2	0,13
]4; 8]	2	0,13
]8; 12]	6	0,40
]12; 16]	4	0,27
]16; 20]	1	0,07
Total	15	1

De façon générale, pour représenter le tableau ci-dessus, on peut utiliser un histogramme :



Synthèse :

Variable quantitative	Distribution	Répartition
Discrète	Graphique en bâtons	Graphique en escalier
Continue	Histogramme	Courbe cumulative

2 Synthétiser des données grâce aux tableaux croisés dynamiques

Dans de nombreuses situations, il est nécessaire de synthétiser les données avant de pouvoir les analyser. Les tableaux croisés dynamiques d'Excel sont l'outil idéal pour cette tâche. Les tableaux croisés dynamiques vont permettre de générer un tableau de contingence.

Le tableau de contingence synthétise les données collectées. Il permet de réaliser une analyse croisée pour améliorer la performance marketing.
Une **table de contingence** est un moyen pour présenter simultanément et de manière **croisée** deux données statistiques. Elle permet d'estimer la dépendance et la relation entre deux caractères observés sur une même population.

Exemple : Nous partons sur la totalité des notes des 2 évaluations des étudiants de BUT en L1. Objectif : réaliser un tableau synthétique afin de connaître le nombre d'étudiants qui ont obtenu une notre entre 0 - 1, 1- 2…

Nom	Prénom	1ère éval.	2ème éval.	Moyenne du semestre
ALAMEH	ALINE	13,75	12	12,875
BARTHOD	ANGELE	4,5	16,5	10,5
BEGOC	GERMAIN	8,75	15	11,875
BERHAULT	CORENTIN	7,75	17	12,375
BILLARD	ELOUEN	14	20	17
CARTAUX	MANON	14,5	16,5	15,5

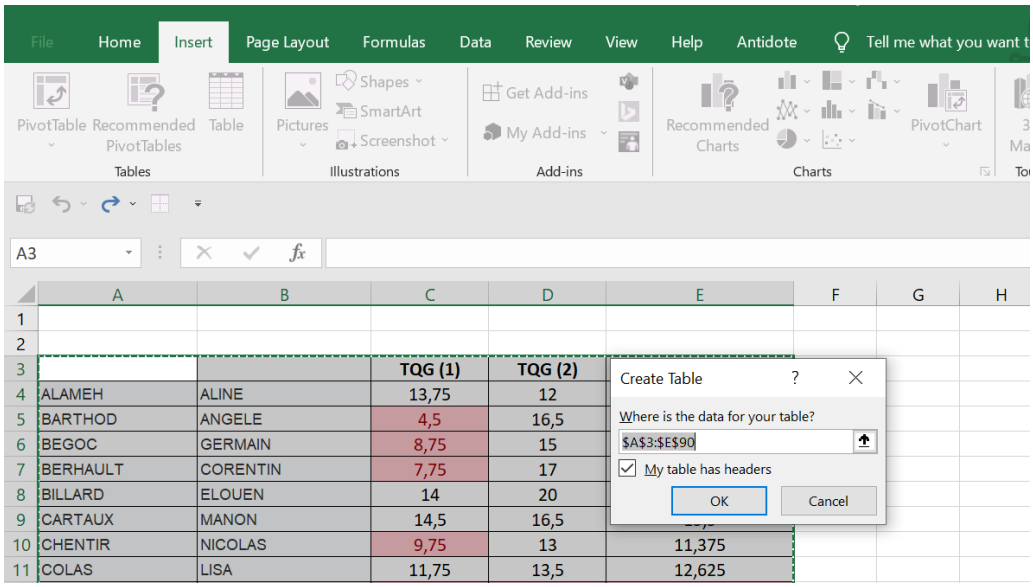
…

PENVERN	THEO	13,5	14,5	14
PILARD	ELOISE	14	19	16,5
POULIQUEN	THOMAS	16,25	16	16,125
RAOULT	ORIANE	18,5	19	18,75

Pour commencer à analyser les notes, vous devez d'abord disposer de données structurées, c'est-à-dire :

- Chacune des colonnes possède un en-tête et représente un champ représentant un type de données homogène.
- Chaque ligne représente un enregistrement.

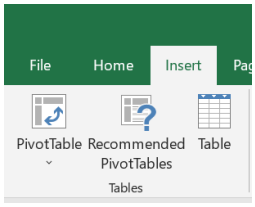
Sélectionner les données et créer un tableau :



	A	B	C	D	E
3			TQG (1)	TQG (2)	
4	ALAMEH	ALINE	13,75	12	
5	BARTHOD	ANGELE	4,5	16,5	
6	BEGOC	GERMAIN	8,75	15	
7	BERHAULT	CORENTIN	7,75	17	
8	BILLARD	ELOUEN	14	20	
9	CARTAUX	MANON	14,5	16,5	
10	CHENTIR	NICOLAS	9,75	13	11,375
11	COLAS	LISA	11,75	13,5	12,625

Dans l'onglet **données**, activez une cellule contenue dans la table, par exemple la cellule **A8**.

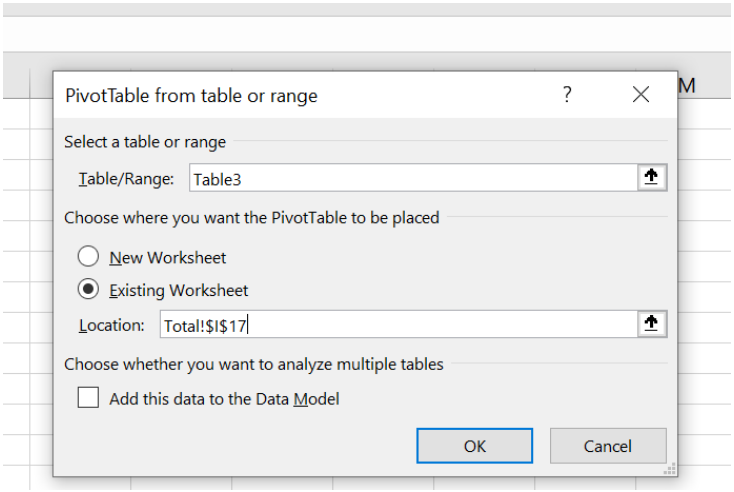
Dans le ruban, onglet **Insertion** - groupe **Tableaux**, cliquez sur **Tableau croisé dynamique**.



La boîte de dialogue **Créer un tableau croisé dynamique** apparaît à l'écran.

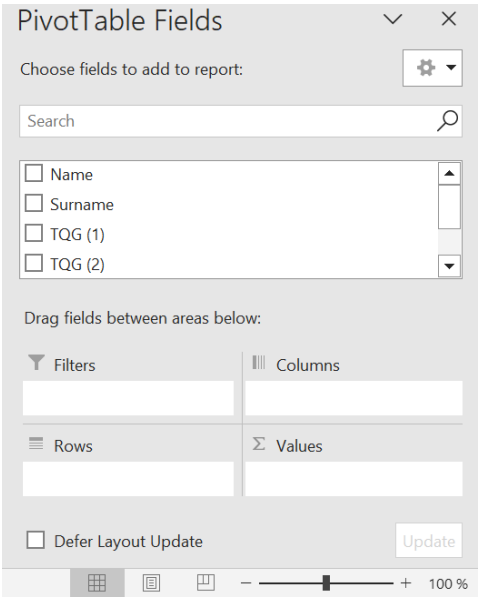
Vérifiez que la zone **Sélectionner un tableau ou une plage** contient les données à analyser, c'est-à-dire le **Tableau1**.

Dans la zone **Emplacement**, activez la cellule **F2**.

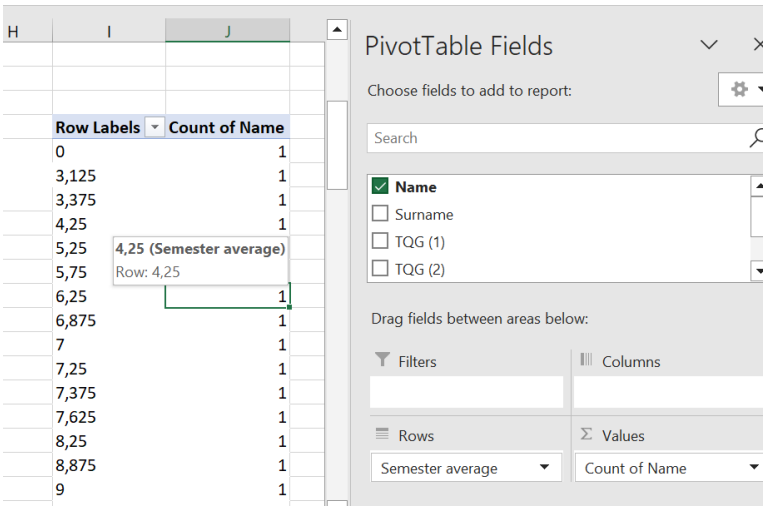


Cliquez sur le bouton **OK**.

Sur la droite de l'écran, le volet **Champs de tableau croisé dynamique** contient la liste des champs du tableau de données.

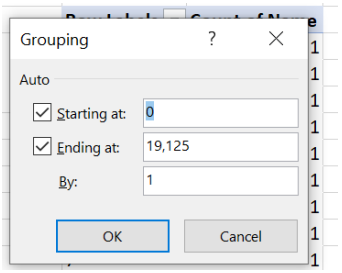


Glisser dans la case colonne, « moyenne du semestre » et la case valeur « nom ».



Pour créer un intervalle d'amplitude de 1 :

- sélectionner une cellule de la 1ère colonne,
- clique droit, « grouper » :



Notes	Nombre d'étudiants
0-1	1
3-4	2
4-5	1
5-6	2
6-7	2
7-8	4
8-9	2
9-10	9
10-11	9
11-12	5
12-13	8
13-14	6
14-15	7
15-16	6
16-17	10
17-18	8
18-19	3
19-20	1
Grand Total	**86**

3 Analyse des données avec une seule variable (analyse univariée)

Pour la partie statistique, on va reprendre l'exemple suivant :

Note à l'Examen	Effectifs
0	0
1	0
2	1
3	0
4	1
5	0
6	1
7	0
8	1
9	1
10	1
11	2
12	2
13	1
14	1
15	1
16	1
17	0
18	1
19	0
20	0
Total	**15**

La première étape de l'analyse descriptive consiste à analyser les variables une à une pour avoir une première compréhension des données disponibles. On appelle cette analyse l'analyse univariée en opposition avec les analyses multivariées qui traitent de plusieurs variables à la fois. L'analyse descriptive univariée comprend des mesures souvent connues, mais parfois mal interprétées, nous allons donc passer en revue les indices statistiques présentant les tendances centrales (la moyenne, la médiane, le mode), la dispersion (la variance, l'écart-type) et la forme des données (asymétrie et aplatissement). Enfin, nous aborderons en complément l'intervalle de confiance.

Les tendances centrales

- *Moyenne arithmétique*

Une moyenne est une technique de réduction de l'information. Elle permet d'appréhender la tendance centrale des données.

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

m_x = somme des x_i / N avec N = somme des effectifs observés

m_x = (2+4+6+8+9 +…+16+18)/15 => m_x = 10,73

Ou bien si la distribution est regroupée en classe **:** mx = somme des nixi / N

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{K} n_i x_i = \sum_{i=1}^{K} f_i x_i,$$

Fonction Excel : moyenne ; avec pondération : sommeprod(… ;…) / total des effectifs.

- *Mode ou classe modale*

La modalité ou la valeur la plus fréquente. Le nombre présentant la plus grande occurrence dans un groupe de nombres. **Exemple** : 2,3,5,5,5,6 : le mode est 5.

Fonction Excel : mode.simple ou mode.multiple

- *Médiane*

Représente le nombre intermédiaire (50%) d'un groupe de nombre. En d'autres termes, la moitié des nombres ont des valeurs supérieures à la médiane et l'autre moitié des valeurs inférieures.
Dans le cas d'une variable quantitative discrète :

- nombre impair : après avoir rangé les valeurs dans l'ordre croissant, la médiane, est la valeur de rang **(N+1)/2**. N étant l'effectif.
- nombre pair : Il y a un nombre pair de valeurs, on a donc 2 valeurs centrales. La médiane est alors la **moyenne** de ces deux valeurs.

Exemple : Nombre impair

Note à l'Examen de TQG	Effectifs
2	1
4	1
6	1
8	1
9	1
10	1
11	1
11	1
12	1
12	1
13	1
14	1
15	1

50 % partie basse

Médiane vaut 11

50 % partie haute

11

16	1
18	1

(15+1) / 2 = 8ème rang => 11.

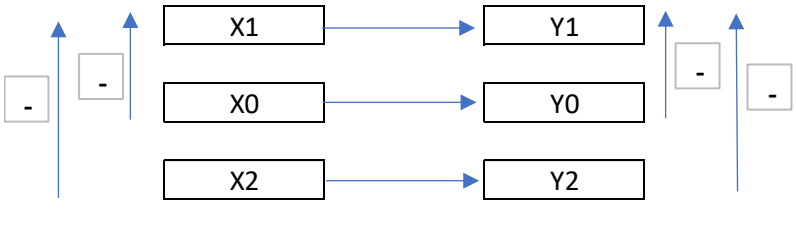
Dans le cas d'un nombre pair : On supprime la note 2. Il reste 14 valeurs. Les 2 valeurs centrales sont 11 et 12. Médiane, (11+12)/2 = 11,5

Note à l'Examen de TQG	Effectifs
4	1
6	1
8	1
9	1
10	1
11	1
11	1

12	1
12	1
13	1
14	1
15	1
16	1
18	1
Total	14

Dans le cas d'une **variable quantitative continue**, résolution par interpolation linéaire :

Note à l'Examen de TQG	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées
[0;4]	2	2,00	0,13	0,13
]4; 8]	2	**4,00**	0,13	**0,27**
]8; 12]	6	**10,00**	0,40	**0,67**
]12; 16]	4	14,00	0,27	0,93
]16; 20]	1	15,00	0,07	1,00



$$\frac{x0 - x1}{x2 - x\,1} = \frac{y0 - y1}{y2 - y1}$$

Recherche du rang : (N+1)/2, (15+1)/2 = 8ème rang (8ème effectif), soit la classe]8; 12]. C'est à ce rang qu'on dépasse les 50%. La classe médiane est]8; 12]. La médiane se situe à l'intérieur de cette classe médiane.

8 => 4

M => 8

12 => 10

$\frac{M-8}{12-8} = \frac{8-4}{10-4}$ = M- 8 = (12-8) x 4/6 = M = 10,67.

Il est possible d'utiliser les fréquences cumulées :

$$\frac{M-8}{12-8} = \frac{53-27}{67-27}$$

53 = 8/15. Par approximation, il est possible d'utiliser 50%, correspondant à la définition de la médiane.

Fonction Excel : médiane

Remarque : moyenne/médiane et l'effet Bill Gates.

Si vos données comprennent des valeurs aberrantes (c'est-à-dire largement distantes des autres données observées), alors la médiane est une valeur plus représentative que la moyenne.

Les valeurs aberrantes ont tendance à tirer la moyenne vers le haut ou le bas.

Exemple : calcul du salaire moyen et médian des habitants d'un petit village.

Sans Bill Gates

Habitant	Salaire mensuel
1	2 017
2	1 423
3	1 796
4	1 736
5	2 030
6	2 915
7	2 377
8	2 173
9	2 852
10	1 251
11	2 849
12	1 964
13	1 952
14	1 631
15	1 447

Moyenne **2 028**

Médiane **1 964**

Avec Bill Gates

Habitant	Salaire mensuel
1	2 017
2	1 423
3	1 796
4	1 736
5	2 030
6	2 915
7	2 377
8	2 173
9	2 852
10	1 251
11	2 849
12	1 964
13	1 952
14	1 631
15	1 447
16	**477 239**

Moyenne **31 728**

Médiane **1 991**

La dispersion

Les indices de dispersion sont un complément indispensable aux indices de tendance centrale, car ils permettent d'évaluer si l'échantillon est homogène ou hétérogène. Ainsi, si nous analysons un portefeuille clients, les indices de tendance centrale résument les informations sur les clients (comportements, attitudes, etc.) et les indices de dispersion nous indiquent s'il existe ou non de grandes variations autour de ce résumé.

- *Étendue*

Correspond à la différence entre la plus grande valeur et la plus petite valeur de la variable.

$$e_X = \max_{1 \le i \le N} x_i - \min_{1 \le i \le N} x_i.$$

Cas présent : 18 – 2 = 16.

- *Variance*

Correspond à la moyenne arithmétique des carrés des écarts à la moyenne arithmétique. Pour une distribution non regroupée, on a :

Formule 1 :

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2\,;$$

= somme $(x_i – m_x)^2$ / N

Formule 2 :

$$\mathrm{Var(X)} = \left(\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{x_i^2}\right) - (\overline{\mathrm{x}})^2$$

= [somme $(xi)^2$ / N] - $(mx)^2$

- *L'écart-type*

Correspond à la racine carrée de la variance de cette variable : $\sigma_X = \sqrt{\mathrm{Var(X)}}.$

L'écart-type sert à mesurer la dispersion, ou l'étalement, d'un ensemble de valeurs autour de leur moyenne. Plus l'écart-type est faible, plus l'échantillon est homogène.

Exemple : cas d'une variable discrète :

Note à l'Examen xi	Effectifs ni	$(Xi - Mx)^2$
2	1	76,27
4	1	45,34
6	1	22,40
8	1	7,47
9	1	3,00

10	1	0,54
11	1	0,07
11	1	0,07
12	1	1,60
12	1	1,60
13	1	5,14
14	1	10,67
15	1	18,20
16	1	27,74
18	1	52,80
Total	15	272,93

Moyenne	10,73
Variance	18,20
Écart Type	4,27

Variance = 272,93 / 15

Écart Type = √(18,20).

Avec les xi pondérés : = somme **ni**$(x_i - m_x)^2$ / N ou [somme **ni**$(xi)^2$ / N] - $(mx)^2$

Note à l'Examen de TQG xi	Effectifs ni	$(Xi - Mx)^2$	$Ni(Xi - Mx)^2$
2	1	76,27	76,27
4	1	45,34	45,34
6	1	22,40	22,40
8	1	7,47	7,47
9	1	3,00	3,00
10	1	0,54	0,54
11	2	0,07	0,14
12	2	1,60	3,21
13	1	5,14	5,14
14	1	10,67	10,67
15	1	18,20	18,20
16	1	27,74	27,74
18	1	52,80	52,80
Total	15	271,26	272,93

Moyenne	10,73
Variance	18,20
Écart Type	4,27

Exemple : cas d'une variable continue :

Note à l'Examen de TQG xi	Valeur centrale xi	Effectifs ni	(Xi - Mx)²	Ni(Xi - Mx)²
[0;4]	2	2	64,00	128,00
]4; 8]	6	2	16,00	32,00
]8; 12]	10	6	0,00	0,00
]12; 16]	14	4	16,00	64,00
]16; 20]	18	1	64,00	64,00
			Total	288,00

Moyenne	10
Variance	19,2
Écart Type	4,38

Fonction Excel : VAR.P.N et ECARTYPE.PEARSON