Rappels de statistiques

F. Wlazinski

Licence d'économie

1 Moyenne arithmétique et variance

1.1 Cas discret non pondéré

Définition 1.1

La *moyenne arithmétique*, notée $\overline{x}$, d'une variable quantitative discrète x qui prend n valeurs x_1, x_2,..., x_n est $\overline{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.

Définition 1.2

La *variance* d'une série est la moyenne arithmétique des carrés des écarts à la moyenne arithmétique.
Autrement dit, avec les notations de la définition 1.1, on a :

$$V(x) = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_p - \overline{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{p} (x_i - \overline{x})^2$$

L'*écart type* σ de la série est la racine carrée de la variance c'est-à-dire $\sigma_x = \sigma(x) = \sqrt{V(x)}$.

Remarque 1.3

Avec les notations de la définition 1.1, on a :

$$\overline{x^2} = \frac{x_1^2 + x_2^2 + \ldots + x_p^2}{n} = \frac{1}{n}\sum_{i=1}^{p} x_i^2 \text{ et } V(x) = \overline{x^2} - \overline{x}^2.$$

Exemple 1.4

On cherche à calculer la moyenne et l'écart-type des notes (non pondérées) d'un étudiant qui sont 05, 08, 12 et 15.

1.2 Cas discret pondéré

Définition 1.5

Soit une variable quantitative x qui prend les valeurs $\underbrace{x_1, \ldots x_1}_{n_1\, fois}$, $\underbrace{x_2, \ldots x_2}_{n_2\, fois}$, ..., $\underbrace{x_p, \ldots x_p}_{n_p\, fois}$

avec $n_1 + n_1 + \ldots n_p = n$

Alors $\overline{x} = \dfrac{n_1 \times x_1 + n_2 \times x_2 + \ldots + n_p \times x_p}{n} = \dfrac{1}{n}\sum_{i=1}^{p} n_i x_i$.

Il s'agit d'une moyenne arithmétique dite *pondérée*.

Propriété 1.6

Avec les notations de la définition 1.5, on a :

$$V(x) = \frac{n_1 (x_1 - \overline{x})^2 + n_2 (x_2 - \overline{x})^2 + \cdots + n_p (x_p - \overline{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \overline{x})^2$$

Remarques 1.7

On conserve les notations de la définition 1.5.

- Si l'on connaît la fréquence $f_i = \dfrac{n_i}{n}$ de chacune des n valeurs x_i alors la moyenne arithmétique est : $\overline{x} = x_1 f_1 + x_2 f_2 + \ldots + x_n f_n = \sum_{i=1}^{n} x_i f_i$.
- De plus, $\overline{x^2} = \dfrac{n_1 \times x_1^2 + n_2 \times x_2^2 + \ldots + n_p \times x_p^2}{n} = \dfrac{1}{n} \sum_{i=1}^{p} n_i x_i^2$.

Exemple 1.8

On cherche à calculer la moyenne et l'écart-type des notes d'un étudiant qui sont 05 coefficient 3, 08 coefficient 2, 12 coefficient 2 et 15 coefficient 1.

Remarque 1.9

Dans le cadre de test et lorsque l'on travaille avec un petit échantillon de la population, on corrige ces valeurs par *le biais de petit échantillon.*

La *variance corrigée* est alors $S_x^2 = V(x) \times \dfrac{n}{n-1}$.

Il s'en suit que l'*ecart-type corrigée* est $S_x = \sqrt{S_x^2} = \sigma_x \times \sqrt{\dfrac{n}{n-1}}$.

Remarque 1.10

Définition 1.11

Le *coefficient de variation* noté CV_x d'une série x est une mesure de dispersion relative et sa valeur est $CV_x = \dfrac{\sigma_x}{\overline{x}}$.

Remarque 1.12

Propriété 1.13

Soient x_1 et x_2 deux séries quantitatives d'effectifs respectifs n_1 et n_2, de moyennes $\overline{x_1}$ et $\overline{x_2}$.

La moyenne de la série x conjointe de x_1 et x_2 est $\overline{x} = \dfrac{n_1 \times \overline{x_1} + n_2 \times \overline{x_2}}{n_1 + n_2}$.

Exemple 1.14

Les moyennes de deux groupes d'étudiants à un examen sont respectivement 8 et 12. Sachant qu'il y a 30 étudiants dans le premier groupe et 50 étudiants dans le deuxième groupe, on cherche la moyenne des deux groupes.

1.3 Cas continu

Définition 1.15

Soient $a < b$ deux réels. Le centre de tout intervalle de a à b est la moyenne $\dfrac{a+b}{2}$.

Définition 1.16

Soit x une variable quantitative continue dont les valeurs sont dans les intervalles $[x_i; x_{i+1}[$ pour i compris entre 1 et $p-1$. On appelle *centres des classes* de x les centres des intervalles $[x_i; x_{i+1}[$. Ils sont généralement noté c_i pour i de 1 à $p-1$.

Exemple 1.17

Les notes d'un groupe d'étudiants sont réparties de la façon suivante :

Intervalles	$[0;8[$	$[8;12[$	$[12;18[$	$[18;20]$
n_i	21	15	10	2

Principe

Définition 1.18

La moyenne arithmétique, notée $\overline{x}$ ou $E(x)$, d'une variable quantitative continue x continue dont les valeurs sont dans les les intervalles $[x_i; x_{i+1}[$ de centre c_i pour i de 1 à $p-1$ et dont les effectifs respectifs sont n_1, n_2, …, n_p est $\overline{x} = \dfrac{n_1 \times c_1 + n_2 \times c_2 + \ldots + n_p \times c_p}{n_1 + n_2 + \ldots + n_p} = \dfrac{1}{n}\sum_{i=1}^{p} n_i c_i$.

Exemple 1.19

Définition 1.20

Avec les notations de la définition 1.18, la variance de x et l'écart type sont :

$V(x) = \dfrac{n_1 (c_1 - \overline{x})^2 + n_2 (c_2 - \overline{x})^2 + \cdots + n_p (c_p - \overline{x})^2}{n} = \dfrac{1}{n}\sum_{i=1}^{p} n_i (c_i - \overline{x})^2$

$\sigma_x = \sigma(x) = \sqrt{V(x)}$.

Remarque 1.21

Exemple 1.22

2 Statistique à deux variables

2.1 Fréquences

Définition 2.1

Soient X et Y deux variables statistiques d'une population d'effectif total n.
Si la variable X possède p modalités et si la variable Y possède q modalités, on appelle *tableau de contingence* de X et Y le tableau où sont indiquées dans la première colonne les p modalités $(x_i)_{i=1,p}$ de la variable X et dans la première ligne les q modalités $(y_j)_{j=1,p}$ de la variable Y. A l'intersection de la ligne x_i et de la colonne y_i, on trouve l'effectif *partiel* $n_{i,j}$ correspondant à la conjonction des dites modalités pour la population.

Définition 2.2

Les totaux par ligne et par colonne sont appelées les *distributions marginales.*

Remarque 2.3

L'effectif total n s'obtient en considérant les sommes des distributions marginales.
Si X prend les valeurs $(x_1, x_2, \ldots, x_p)$ et si Y prend les valeurs $(y_1, y_2, \ldots, y_q)$, la somme de la ligne x_i est $n_{i,1} + n_{i,2} + \ldots + n_{i,q}$ c'est-à-dire $\sum_{j=1}^{q} n_{i,j}$ et la somme de la ligne y_j est $n_{1,j} + n_{2,j} + \ldots + n_{p,j}$ c'est-à-dire $\sum_{i=1}^{p} n_{i,j}$. L'effectif total n est $\sum_{i=1}^{p}\left(\sum_{j=1}^{q} n_{i,j}\right)$ ou $\sum_{j=1}^{q}\left(\sum_{i=1}^{p} n_{i,j}\right)$ suivant si l'on a fait les sommes par ligne ou par colonne en premier.

Définition 2.4

Si X prend les valeurs $(x_i)_{i=1,p}$, si Y prend les valeurs $(y_j)_{j=1,q}$ et si $n_{i,j}$ est l'effectif à la conjonction des modalités x_i et y_j alors la *fréquence empirique* est $f_{i,j} = \dfrac{n_{i,j}}{n}$.
La *fréquence marginale* de la modalité x_i du caractère X est le quotient de l'effectif marginal de cette valeur de caractère par l'effectif total de la population c'est-à-dire $\dfrac{\sum_{j=1}^{q} n_{i,j}}{n}$.
La *fréquence marginale* de la modalité y_j du caractère Y est le quotient de l'effectif marginal de cette valeur de caractère par l'effectif total de la population c'est-à-dire $\dfrac{\sum_{i=1}^{p} n_{i,j}}{n}$.

Exemple 2.5

En vue d'un éventuel remplacement, on s'intéresse aux nombres annuels (Y) d'interventions pour réparation sur les machines d'une entreprise en fonction de leur ancienneté (X) en années. Les données sont résumées par le tableau suivant.

Nombres de pannes / Ancienneté	0	1	2	
1	5	6	4	
3	10	2	4	
5	17	6	7	
7	2	6	11	

2.2 Moyenne, variance et covariance

Principe

Comme pour les séries à une variable, on pourra calculer les moyennes variances, et covariance uniquement pour les séries à deux variables quantitatives.

Définition 2.6

On utilise les notations de la définition 2.4.
La moyenne arithmétique (marginale) de X est $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{p}\left(x_i \sum_{j=1}^{q} n_{i,j}\right) = \sum_{i=1}^{p} f_i x_i$.
La moyenne arithmétique (marginale) de Y est $\overline{Y} = \dfrac{1}{n}\sum_{j=1}^{q}\left(x_i \sum_{i=1}^{p} n_{i,j}\right) = \sum_{j=1}^{q} f_j y_j$.

Remarque 2.7

On définit comme pour les séries d'une seule variable :
$$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{p}\left(x_i^2 \sum_{j=1}^{q} n_{i,j}\right) = \sum_{i=1}^{p} f_i x_i^2 \text{ et } \overline{Y^2} = \frac{1}{n}\sum_{j=1}^{q}\left(x_i^2 \sum_{i=1}^{p} n_{i,j}\right) = \sum_{j=1}^{q} f_j y_j^2.$$

Remarque 2.8

Exemple 2.9

On reprend l'exemple 2.5.

Nombres de pannes / Ancienneté	0	1	2	
1	5	6	4	
3	10	2	4	
5	17	6	7	
7	2	6	11	

Définition 2.10

La *covariance* des variables X et Y est $\text{cov}(X;Y) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{p}\sum_{j=1}^{q}\big(n_{i,j} \times (\overline{X} - x_i)(\overline{Y} - y_i)\big)$.

Remarque 2.11

Exemple 2.12

Remarque 2.13

2.3 Ajustement affine

Principe

On s'intéresse à l'étude, sur une population donnée, du ou des liens qui peuvent exister entre deux variables X et Y.

Individu	1	2	3	...	n
Valeurs de X	x_1	x_2	x_3	...	x_n
Valeurs de Y	y_1	y_2	y_3	...	y_n

Exemple 2.14

Le tableau suivant fournit les résultats scolaires de 6 étudiants à un examen ainsi que le temps qu'ils ont consacré à la révision pour cet examen.

Etudiants	1	2	3	4	5	6
X nombres d'heures de révision	6	16	10	3	1	12
Y notes à l'examen	11	20	8	8	1	9

Définition 2.15

Soient $X = (x_i)_{i=1,n}$ et $Y = (y_i)_{i=1,n}$ deux séries statistiques.
On appelle *nuage de points* l'ensemble des points $(x_i, y_i)_{i=1,n}$.

Définition 2.16

Le *point moyen* G du nuage de points est le point de coordonnées $(\overline{X};\overline{Y})$.

Remarque 2.17

Exemple 2.18

On reprend l'exemple 2.14 :

Etudiants		A	B	C	D	E	F		
Nombres d'heures de révision	x_i	6	16	10	3	1	12		
Notes à l'examen	y_i	11	20	8	8	1	9		

Remarque 2.19

Définition 2.20

Soient X et Y deux variables statistiques, de variances non nulles, sur une même population. On appelle *coefficient de corrélation linéaire* entre X et Y et on note $r_p(X,Y)$ ou $\rho(X,Y)$ le réel :

$$r_p = r_p(X,Y) = \frac{\text{cov}(X;Y)}{\sigma_X \sigma_Y} = \frac{\text{cov}(X;Y)}{\sqrt{V(X)V(Y)}}.$$

Remarque 2.21

Exemple 2.22

On reprend l'exemple 2.14 :

Etudiants		A	B	C	D	E	F		
Nombres d'heures	x_i	6	16	10	3	1	12		
Notes	y_i	11	20	8	8	1	9		

Définition 2.23

On appelle *droite de régression*, *droite d'ajustement* ou *droite des moindres carrés* de Y en X la droite d'équation $y = ax + b$ où $a = \dfrac{\text{cov}(X;Y)}{V(X)}$ et $b = E(Y) - a \times E(X)$.

Remarque 2.24

La droite de régression de Y en X passe par le point moyen $G\left(\overline{X}, \overline{Y}\right)$.

Exemple 2.25