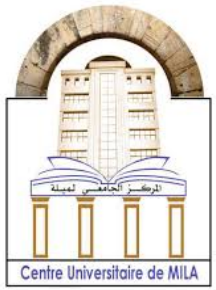
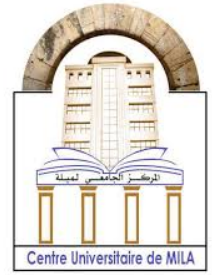
People's Democratic Republic Of Algeria
Ministry Of Higher Education And Scientific Research



Centtre Univercitaire Abdelhafid Boussouf Mila



Rappels sur les statistiques descriptive

L'objectif du cours est d'étudier sur une popolution de N individus, deux variables différentes et de chercher s'il y a un lien ou une corrélation entre ces deux caractères. Chacune des deux variables peut être, soit quantitative, soit qualitative. On étudie deux cas :

- Les deux variables sont qualitatives.
- Les deux variables sont quantitatives.

Exemples de relations possibles entre deux charactères : Taille et poids des individus, Effet et dosage, tabagisme et cancers du poumon, rendement et quantité d'engrais utilisée, ...

Statistique bivariée

- Variables qualitatives .

Statistique bivariée

- Variables qualitatives .
- Données observées-Tableau de contingence .

Statistique bivariée

- Variables qualitatives .
- Données observées-Tableau de contingence .
- Tableau des fréquences .

- Variables qualitatives .
- Données observées-Tableau de contingence .
- Tableau des fréquences .
- Variables quantitatives .

- Variables qualitatives .
- Données observées-Tableau de contingence .
- Tableau des fréquences .
- Variables quantitatives .
- Paramètres de position et de dispersion .

Données observées-Tableau de contingence
On considère une population statistique décrite selon deux caractères $X = (x_1, x_2, ..., x_k)$ et $Y = (y_1, y_2, ..., y_p)$: Leurs valeurs distinctes sont notées respectivement :

Données observées-Tableau de contingence

Les données observées peuvent être regroupées sous la forme d'un tableau de contingence :

	y_1	y_2	$\dots$	y_j	$\dots$	y_p	$n_{i\cdot}$
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	n_2
$\vdots$							
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\cdot}$
$\vdots$							
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	n_k
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot j}$	$\dots$	$n_{\cdot p}$	$N = n_{\cdot\cdot}$

Données observées-Tableau de contingence

$-n_{ij}$ représente le nombre de fois que les modalités x_i et y_j apparaissent ensemble. - N est l'effectif total de la population :

$$N = n.. = \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij}.$$

- Les $n_{i.}$ et $n_{.j}$ sont appelés les effectifs marginaux.
- $n_{i.}$ représente le nombre de fois que la modalité x_i apparaît,

$$\sum_{j=1}^{p} n_{ij} = n_{i1} + n_{i2} + \ldots + n_{ip} = n_{i.}, \text{ pour tou } i = 1, \ldots, k$$

- $n_{.j}$ représente le nombre de fois que la modalité y_j apparaît,

$$\sum_{i=1}^{k} n_{ij} = n_{1j} + n_{2j} + \ldots + n_{kj} = n_{.j} \text{ pour tou } j = 1, \ldots, p$$

$$\sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{p} n_{.j} = N = n \ldots$$

Fréquences partielles :

$$f_{ij} = \frac{n_{ij}}{n_{..}}, \quad i = 1, \dots k, \quad j = 1, \dots, p,$$

Fréquences marginales : La fréquence marginale de la variable X est donnée par :

$$f_{i.} = \frac{n_{i.}}{n_{..}}, \quad i = 1, \dots k,$$

Données observées-Tableau de fréquences :

	y_1	y_2	$\dots$	y_j	$\dots$	y_p	$f_{i.}$
x_1	f_{11}	f_{12}	$\dots$	f_{1j}	$\dots$	f_{1p}	$f_{1.}$
x_2	f_{21}	f_{22}	$\dots$	f_{2j}	$\dots$	f_{2p}	$f_{2.}$
$\vdots$							
x_i	f_{i1}	f_{i2}	$\dots$	f_{ij}	$\dots$	f_{ip}	$f_{i.}$
$\vdots$							
x_k	f_{k1}	f_{k2}	$\dots$	f_{kj}	$\dots$	f_{kp}	$f_{k.}$
$f_{.j}$	$f_{.1}$	$f_{.2}$	$\dots$	$f_{.j}$	$\dots$	$f_{.p}$	1

La fréquence marginale de la variable Y est donnée par :

$$f_{.j} = \frac{n_{.j}}{n_{.}}, \quad j = 1, \dots p,$$

$$\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} = 1, \quad \sum_{i=1}^{k} f_{i} = 1, \quad \sum_{j=1}^{p} f_{.j} = 1.$$

Fréquences conditionnelles :

La fréquence conditionnelle de la variable X par rapport à $Y_j, j = 1, \ldots, p$ est donnée par :

$$f_i^j = \frac{n_{ij}}{n_{.j}}, \quad j = 1, \ldots p,$$

La fréquence conditionnelle de la variable Y par rapport à $X_i, i = 1, \ldots, k$ est donnée par :

$$f_j^i = \frac{n_{ij}}{n_{i.}}, \quad i = 1, \ldots k$$

$$f_{ij} = f_{i.} \times f_j^i = f_{.j} \times f_i^j.$$

Relations entre fréquences marginales et fréquences conditionnelles :

Exemple. On s'intéresse à une éventuelle relation entre les caractères : "être fumeur" (plus de 20 cigarettes par jour, pendant 10 ans) et "avoir un cancer de la gorge", sur une population de 1000 personnes, dont 500 sont atteintes d'un cancer de la gorge. Les résultats observés sont présentés dans le tableau suivant :

Observé	cancer	non cancer
fumeur	342	258
non fumeur	158	242

Relations entre fréquences marginales et fréquences conditionnelles :

On calcule l'effectif total :

Observé	cancer	non cancer	Total
fumeur	342	258	600
non fumeur	158	242	400
Total	500	500	200

En utilisant les relations précédentes, on obtient le tableau des fréquences (fréquences marginales) :

Observé	cancer	non cancer	Total
fumeur	0.342	0.258	0.6
non fumeur	0.158	0.242	0.4
Total	0.5	0.5	1

Quelques exemples :

Fréquences marginales :

$$f_{11} = \frac{n_{11}}{n_{...}} = \frac{342}{1000} = 0.342, f_{1.} = \frac{n_1}{n_{..}} = \frac{600}{1000} = 0.6$$

Relations entre fréquences marginales et fréquences conditionnelles :

Fréquences conditionnelles de X sachant Y :

$$f_1^2 = \frac{n_{12}}{n.2} = \frac{258}{500} = 0.516, f_2^2 = \frac{n_{22}}{n.2} = \frac{242}{500} = 0.484$$

Fréquences conditionnelles de Y sachant X :

$$f_2^1 = \frac{n_{12}}{n_1} = \frac{258}{600} = 0.43, f_2^2 = \frac{n_{22}}{n_2.} = \frac{242}{400} = 0.605$$

Paramètres de position et de dispersion :

Dans le cas où X et Y sont des variables quantitatives, on peut associer à chacune des séries marginales définies par le tableau de contingence des caractéristiques de tendance centrale et de dispersion. On considère un tableau de contingence comme celui défini précédemment. Moyenne marginale de X. La moyenne marginale de X, notée $\overline{\overline{x}}$:

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

Moyenne marginale de Y. La moyenne marginale de Y, notée $\overline{\overline{y}}$:

$$\overline{\overline{y}} = \frac{1}{N}\sum_{j=1}^{p} n_{-j} y_i = \sum_{j=1}^{p} f_{\cdot j} y_j.$$

Paramètres de position et de dispersion :

Variance marginale de X. La variance marginale de X, notée $V(X)$:

$$V(X) = \frac{1}{N}\sum_{i=1}^{k} n_{i\cdot}\left(x_i - \overline{\overline{x}}\right) = \frac{1}{N}\sum_{i=1}^{k} n_i \cdot x_i^2 - \overline{\overline{x}}^2.$$

L'écart type marginale de X : $\sigma_X = \sqrt{V(X)}$. Varaince marginale de Y. La moyenne marginale de Y, notée $V(Y)$:

$$V(Y) = \frac{1}{N}\sum_{j=1}^{p} n_{\cdot j}\left(y_j - \overline{\overline{y}}\right) = \frac{1}{N}\sum_{j=1}^{p} n_{-j} y_j^2 - \overline{\overline{y}}^2.$$

L'écart type marginale de Y : $\sigma_Y = \sqrt{V(Y)}$ Covariance. La covriance est la mesure de la variation simultanée de deux variables X et Y.

$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}\left(x_i - \overline{\overline{x}}\right)\left(y_j - \overline{\overline{y}}\right) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} x_i y_j - \overline{\overline{x}}\,\overline{\overline{y}}$$

Propriétés :

- Une covariance peut être positive, négative ou nulle.
- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X, Y) = \mathrm{acov}(Y, X)$
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{cov}(X, Y)$

Exemple

Le Tableau suivant présente la répartition d'un ensemble de consommateurs selon leurs revenus et leurs dépenses :

X/Y	25	30	35	40	Total
20	4	2	1	0	7
25	5	1	0	0	6
30	3	2	1	1	7
Total	12	5	2	1	20

X représente les revenus, et Y, les dépenses de consommation. Les revenus moyens :

$$\bar{x} = \frac{7 \times 20 + 6 \times 25 + 7 \times 30}{20} = 25.$$

La variance des revenus :

$$V(X) = \frac{1}{20}\left(7 \times 20^2 + 6 \times 25^2 + 7 \times 30^2\right) - 25^2 = 17.5.$$

Les dépenses de consommation qu'effectuent en moyenne chacun des consommateurs :

$$\bar{y} = \frac{12 \times 25 + 5 \times 30 + 2 \times 35 + 1 \times 40}{20} = 28.$$

Exemple

Les dépenses de consommation qu'effectuent en moyenne chacun des consommateurs :

$$\bar{y} = \frac{12 \times 25 + 5 \times 30 + 2 \times 35 + 1 \times 40}{20} = 28.$$

La variance des dépenses de consommation :

$$V(Y) = \frac{1}{20}\left(12 \times 25^2 + 5 \times 30^2 + 2 \times 35^2 + 1 \times 40^2\right) - 28^2 = 18.5.$$

La covariance de X et Y.

X/Y	25	30	35	40	Total	$\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}x_i y_j$
20	4	2	1	0	7	3900
25	5	1	0	0	6	3875
30	3	2	1	1	7	6300
Total	12	5	2	1	20	14075

$$\mathrm{Cov}(X, Y) = \frac{14075}{20} - 25 \times 28 = 3.75$$

Fin du partie 1