ENS
$2^{ème}$ année **2015-2016**
Dr. Aicha. Lazraq Khlass

Mathématiques
Statistiques (IV)
Statistiques à 2 variables

IV. 1. Introduction.

Les études statistiques permettent d'analyser et de prévoir une tendance. Le but de ce cours est de déterminer s'il existe un lien de dépendance entre deux caractères qu'on étudie simultanément, ou d'un caractère qu'on étudie à différentes dates. On commence par définir des séries statistiques à deux variables. Puis, on étudiera la possibilité de faire un ajustement affine.

On observe que, dans certains cas, il semble exister un lien entre deux caractères statistiques quantitatifs sur une population ; par exemple, entre le poids et la taille d'un nouveau-né, entre le chiffre d'affaires et le montant des charges d'une société, entre la consommation et la vitesse d'une voiture. Il est alors intéressant d'étudier simultanément ces deux caractères. Nous pouvons alors présenter les résultats sous forme de tableaux ou de graphiques.

.

IV. 2. Définition.

Une série statistique à deux caractères (X, Y) est une série double dont les valeurs sont données par les couples $(x_i; y_i)_{\substack{1\leq i\leq n \\ 1\leq j\leq m}}$.

Deux caractères (X, Y) pouvant être de natures différentes : qualitatif, quantitatif discret ou continu.

.

IV. 3. Tableau de contingence.

Appelé aussi « tableau croisé »

i) Définition.

Le tableau de contingence met en relation deux variables X et Y observées sur une même population. La case à l'intersection de la ligne i et de la colonne j contient le nombre d'individus ayant choisi la modalité i de la variable X et la modalité j de la variable Y.

.

ii) Exemples.

1) Salaire net et âge des livreurs du Pizza Hut

Salaire en euros / Ages	$[170; 200[$	$[200; 230[$	$[230; 260[$	Total
$[20; 22[$	3	1	0	4
$[22; 24[$	2	3	0	5
$[24; 26[$	1	5	1	7
	6	9	1	16

3 individus ont un salaire entre $[170; 200[$ et un âge entre $[20; 22[$.

7 individus ont un âge entre $[24; 26[$.

9 individus ont un salaire entre $[200; 230[$.

.

2) Lien entre la filière du bac général et le sexe

Dans un lycée, on compte le nombre de garçons et de filles de terminale dans chaque filière du bac général. On obtient le tableau croisé suivant :

	L	ES	S	Total
Filles	45	61	76	182
Garçons	11	35	91	137
Total	56	96	167	319

Il y a 319 élèves en terminale dans ce lycée, 182 filles et 137 garçons.

Sur les 319 élèves en terminale S, il y a 91 garçons et 76 filles.

.

iii) Notations mathématiques.

	y_1	y_2	...	y_j	...	y_s	Total
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2s}	$n_{2.}$
...							
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{2.}$
...							
x_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.s}$	n

* X et Y deux variables ayant respectivement r et s modalités.

* n l'effectif total

* $n_{i.}$ la distribution marginale de X : c'est le nombre d'individu qui présente la modalité x_i de X indépendament des valeurs de Y.

* $n_{.j}$ la distribution marginale de Y : c'est le nombre d'individu qui présente la modalité y_i de Y indépendament des valeurs de X.

* nij l'effectif de la case $(i; j)$, effectif partiel : c'est le nombre d'individu présentant à la fois la modalité x_i de X et la modalité y_i de Y.

On a :

$n_{i.} = \sum_{1 \leq j \leq s} n_{ij}$; $n_{j.} = \sum_{1 \leq i \leq r} n_{ij}$; $n = \sum_{1 \leq j \leq s} n_{.j} = \sum_{1 \leq i \leq r} n_{i.}$

.

IV. 4. Lois marginales.

A partir du tableau de contingence, on peut récupérer d'une part les lois de X et d'autres part les lois de Y. Ces deux lois s'appellent les lois marginales. On peut alors appliquer tous les résultats des chapitres précédents : représentation graphique, calcul des caractéristiques de position de dispersion.

i) Moyennes marginales :

$\overline{\overline{X}} = \frac{1}{n} \sum_{1 \leq i \leq r} n_{i.} x_i$

$\overline{\overline{Y}} = \frac{1}{n} \sum_{1 \leq j \leq s} n_{.j} y_j$

ii) Variances marginales :

$V(X) = \frac{1}{n} \sum_{1 \leq i \leq r} n_{i.} x_i^2 - \left(\overline{\overline{X}}\right)^2$

$V(Y) = \frac{1}{n} \sum_{1 \leq j \leq s} n_{.j} y_j^2 - \left(\overline{\overline{Y}}\right)^2$

.

IV. 5. Caractéristiques conditionnelles.

i) Moyenne conditionnelle de Y sachant $X = x_i$:

$\overline{Y_i} = \frac{1}{n_{i.}} \sum_{1 \leq j \leq s} n_{ij} y_j$

ii) Variance conditionnelle de Y sachant $X = x_i$:

$V_i(Y) = \frac{1}{n_{i.}} \sum_{1 \leq j \leq s} n_{ij} y_j^2 - \left(\overline{Y_i}\right)^2$

iii) Moyenne conditionnelle de X sachant $Y = y_i$:

$\overline{X_j} = \frac{1}{n_{.j}} \sum_{1 \leq i \leq r} n_{ij} x_i$

iv) Variance conditionnelle de Y sachant $X = x_i$:

$V_j(X) = \frac{1}{n_j} \sum_{1 \leq i \leq r} n_{ij} x_i^2 - \left(\overline{X_j}\right)^2$

.

IV. 6. Covariance.

i) Définition.

On appelle covariance de la série statistique double X et Y le nombre réel, noté $cov(X, Y)$, défini par :

$cov(X, Y) = \frac{1}{n} \sum_{1 \leq i \leq r} \sum_{1 \leq j \leq s} n_{ij} \left(x_i - \overline{\overline{X}}\right) \left(y_j - \overline{\overline{Y}}\right)$

.

ii) Propriété.

* Formule de Huyghens-Konig :

$cov(X, Y) = \frac{1}{n} \sum_{1 \leq i \leq r} \sum_{1 \leq j \leq s} n_{ij} x_i y_j - \overline{\overline{XY}}$

* $cov(X, X) = cov(X)$

* Changement de variables affines :

$\forall\, (a, \lambda) \in \mathbf{R}^+,\ \forall\, (b, \mu) \in \mathbf{R}$, on a :

$cov(aX + b, \lambda Y + \mu) = a\lambda cov\,(X, Y)$

.

iii) Remarques.

* Si $cov(X, Y) > 0$ alors X et Y varient dans le même sens.

* Si $cov(X, Y) < 0$ alors X et Y varient dans le sens inverse.

.

IV. 7. Coefficient de corrélation linéaire.

i) Définition.

On appelle coefficient de corélation entre X et Y, le nombre :

$\mathcal{P}\,(X, Y) = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$

.

ii) Remarque.

On a toujours $-1 \leq \mathcal{P}\,(X, Y) \leq 1$.

.

iii) Définition.

On dit que X et Y sont indépendantes si $cov(X, Y) = 0$ donc si $cov(X, Y) \neq 0$, alors X et Y sont dépendantes

IV. 8. Exemple.

Le concours d'accès à un établissement porte sur deux épreuves : Technique de communication et informaique. Les condidats qui se sont présentés à ce concours se répartissent en fonction des notes obtenus à ces deux épreuves de la manière suivante :

	$[1,5[$	$[5,9[$	$[9,11[$	$[11,13[$	$[13,17[$	Total
$[6,8[$	0	3	9	7	11	30
$[8,10[$	10	13	18	16	13	70
$[10,12[$	9	11	14	17	14	65
$[12,16[$	12	9	7	5	2	35
Total	31	36	48	45	40	200

X : Note sur 20 obtenue en technique de communication

Y : Note sur 20 obtenue en informatique.

1) Préciser la nature des caractères X et Y.

2) Calculer les moyennes et les variances marginales.

3) Tracer l'histogramme de Y.

4) Calculer les médianes marginales de X et Y.

5) Calculer $cov(X,Y)$ ainsi que le coefficient de corrélation.

6) Comment peut on interprêter le signe de la covariance.

7) Les variables X et Y sont elle indépendantes?

8) Calculer la moyenne et la variance conditionnelle de Y sachant que X appartient à $[8,10[$.