Chapitre III : Statistique descriptive à deux ou plusieurs caractères

Introduction

La plupart des phénomènes qui se manifestent dans la nature sont le résultat de l'action ou de l'effet conjugué de plusieurs facteurs. Il est nécessaire alors d'étudier plusieurs caractères à la fois pour connaitre le déroulement de ces phénomènes.

1- Analyse statistique descriptive à deux caractères (variables)

Cette analyse peut être appliquée à 2 caractères qualitatifs ou quantitatifs continus.

a) Cas de 2 variables quantitatives

Considérant 2 caractères quantitatifs sur un ensemble d'individus (taille - poids, âge - taille, ….etc.). On mesure ces caractères sur chacun des individus de l'ensemble. Les résultats obtenus se représentent sous la forme d'une série de couple de valeurs soient : (x_1, y_1), (x_2, y_2), (x_3, y_3),……. (x_n, y_n).

Pour rendre explicite cette série de couple, on fait la répartition de cette série en classes selon les 2 variables respectivement. On obtient un tableau croisé selon X et Y où les colonnes contiendront les classes de X et les lignes les classes de Y.

X / Y	1ère classe C_{1x}	2ème classe C_{2x}	3ème classe C_{3x}	……	……	pème classe C_{px}	**Total $n_{.j}$**
C_{1y}	n_{11}	n_{21}	n_{31}	……	……	n_{p1}	$n_{.1}$
C_{2y}	n_{12}	n_{22}	n_{32}	……	……	n_{p2}	$n_{.2}$
C_{3y}	n_{13}	n_{23}	n_{33}	……	……	n_{p3}	$n_{.3}$
……	……	……	……	……	……	……	……
……	……	……	……	……	……	……	……
C_{qy}	n_{1q}	n_{2q}	n_{3q}	……	……	n_{pq}	$n_{.q}$
Total $n_{i.}$	$n_{1.}$	$n_{2.}$	$n_{3.}$	……	……	$n_{p.}$	N

- Si X est réparti en p classes et Y est réparti en q classes, alors la distribution double ou conjointe obtenue contient p.q classes exclusives ou disjointes ; un individu n'appartient qu'aune seule classe et une seule. Ces classes peuvent être identifiées par le symbole C_{ij} où :

i : indice de classe pour X : $1 \leq i \leq p$ / j : indice de classe pour Y : $1 \leq j \leq q$

- Si l'ensemble étudié comprend N individus et après affectation de ces individus dans les différentes classes croisées, on obtient un tableau des effectifs où chaque classe contient le nombre d'individus qui appartient à cette classe. $N = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$

n_{ij} : effectif ou nombre d'individus de la classe C_{ij}

- En faisant la somme des effectifs sur les lignes du tableau, on obtient les effectifs des différentes classes de Y, on obtient la distribution simple de la variable Y seule, ces effectifs sont appelés effectifs marginaux de Y. On les symboliser par $\mathbf{n_{.j}}$; $n_{.j} = \sum_{i=1}^{p} n_{ij}$

Ex : $n_{.1} = \sum_{i=1}^{p} n_{i1}$; $n_{.2} = \sum_{i=1}^{p} n_{i2}$

$n_{.1}, n_{.2}, \ldots\ldots\ldots\ldots n_{.q}$: sont les effectifs marginaux de Y.

L'effectif global N est égal : $N = \sum_{j=1}^{q} n_{.j}$

- De même, la somme des effectifs du tableau sur les colonnes donne les effectifs par classe de la variable X prise seule, on obtient les effectifs marginaux de X qui sont symbolisés $\mathbf{n_{i.}}$

$n_{i.} = \sum_{j=1}^{q} n_{ij}$

$n_{1.}, n_{2.}, \ldots\ldots\ldots\ldots n_{p.}$: sont les effectifs marginaux de X, et l'effectif global est égal :

$N = \sum_{i=1}^{p} n_{i.}$

Remarque : le tableau de distribution croisée entre les variables X et Y (tableau à double entrée) est appelé aussi tableau de contingence.

- En divisant les effectifs par le nombre total des individus N on obtient les fréquences relatives simples de classes croisées du tableau. Elles sont notées f_{ij}.

$f_{ij} = \frac{n_{ij}}{N}$; $\sum_{i=1}^{p} \sum_{j=1}^{q} f_{ij} = 1 = 100\ \%$.

Elles représentent le poids de la classe croisée C_{ij} par rapport à l'ensemble des classes de la distribution. On peut également déterminer les fréquences relatives simples marginales de X et Y.

$f_{.j} = \frac{n_{.j}}{N} = \sum_{i=1}^{p} f_{ij}$: Fréquence relative marginale de la classe j de Y.

$f_{i.} = \frac{n_{i.}}{N} = \sum_{j=1}^{q} f_{ij}$: Fréquence relative marginale de la classe i de X.

$\sum_{j=1}^{q} f_{.j} = 1$ et $\sum_{i=1}^{p} f_{i.} = 1$

- Grace aux effectifs marginaux ou grâce aux fréquences relatives marginales, on peut déterminer tous les paramètres et les indices sur les variables X et Y.

La moyenne de X : $\bar{x} = \frac{1}{N} \sum_{i=1}^{p} n_{i.}\, x_i$

x_i : Centre de la classe i de X / $n_{i.}$: Effectif marginal de la classe i de X.

La variance de X : var (x) = $\frac{1}{N} \sum_{i=1}^{p} n_{i.} (x_i - \bar{x})^2$

La moyenne de Y : $\bar{y} = \frac{1}{N} \sum_{j=1}^{q} n_{.j}\, y_j$

y_j : Centre de la classe j de Y / $n_{.j}$: Effectif marginal de la classe j de Y.

La variance de Y : var (y) = $\frac{1}{N} \sum_{j=1}^{q} n_{.j} \left(y_j - \bar{y}\right)^2$

Moyenne conditionnelle

Dans le tableau de contingence donné par les effectifs ou les fréquences, on peut considérer une classe fixée d'une variable et obtenir la moyenne de l'autre variable pour cette classe. Le résultat obtenu est appelé moyenne conditionnelle.

Définition : on appelle moyenne conditionnelle de X sachant que Y est fixé, la quantité qui est notée : $\bar{x}/_{y=j} = \frac{1}{n.j} \sum_{i=1}^{p} n_{ij}\, x_i$

- De même, on appelle moyenne conditionnelle de Y sachant que X est fixé, la quantité qui est notée : $\bar{y}/_{x=i} = \frac{1}{ni.} \sum_{j=1}^{q} n_{ij}\, y_j$

Ex : Considérant le tableau de contingence qui donne la distribution de 400 individus selon le poids et la taille.

X / Y	130-140	140-150	150-160	160-170	170-180	180-190	Total n.j
40-50			15				42
50-60			17				83
60-70	20	25	33	18	16	13	125
70-80			25				90
80-90			15				60
Total ni.	38	62	105	85	67	43	400

- Le poids moyen des individus ayant une taille comprise entre 150 et 160 cm ($3^{ème}$ classe) :

$\bar{y}/_{x=3} = \frac{1}{105}(15.45 + 17.55 + 33.65 + 25.75 + 15.85) = 65{,}76$ kg.

Les individus dont la taille comprise entre 150 et 160 cm ont un poids moyen de 65,76 kg.

- La taille moyenne des individus ayant un poids compris entre 60 et 70 kg ($3^{ème}$ classe) :

$\bar{x}/_{y=3} = \frac{1}{125}(20.135 + 25.145 + 33.155 + 18.165 + 16.175 + 13.185) = 155.92$ cm

Les individus dont le poids compris entre 60 et 70 kg ont une taille moyenne de 155.92 cm.

Variance conditionnelle

C'est la variance qui mesure la dispersion d'une variable sachant que la deuxième variable est fixée.

Définition : On appelle variance conditionnelle de X sachant que Y est fixée, la quantité :

$$\text{Var}\,(x/_{y=j}) = \frac{1}{n.j} \sum_{i=1}^{p} n_{ij}\,(x_i - \bar{x}/_{y=j})^2$$

- De même, on appelle la variance conditionnelle de Y sachant que X est fixée, la quantité :

$$\text{Var}\,(y/_{x=i}) = \frac{1}{ni.} \sum_{j=1}^{q} n_{ij}\,\left(y_j - \bar{y}/_{x=i}\right)^2$$

- On peut donc calculer les moyennes et les variances conditionnelles de la 1ère variable par rapport à la 2ème, par rapport à chacune des classes de la seconde variable. On peut donc comparer les moyennes de la 1ère variable selon les différentes classes de la seconde variable ainsi que les variances.
- Si les variances conditionnelles d'une variable (X ou Y) sont égales pour chacune des classes de seconde variable, on dit quelles sont homogènes ou égales et on parle d'homoscédasticité.
- Si les variances conditionnelles sont différentes ont dit qu'elles sont hétérogènes ou qu'il y a hétéroscédasticité.
- On peut tracer également les histogrammes ou la fonction de répartition grâce aux fréquences cumulées pour la variable X sachant une classe donnée, fixée de Y et inversement.

- On peut également calculer le lien entre les deux variables X et Y par un coefficient appelé **coefficient de corrélation** qui mesure le degré d'influence qu'exerce chaque variable sur l'autre.

Le coefficient de corrélation entre 2 variables dans une distribution croisée est qu'on le note par $\varphi\,(x,y)$.

$$\varphi\,(x,y) = \frac{\mathrm{Cov}\,(x,y)}{\sigma(x).\sigma(y)}$$

Cov (x,y) : covariance de X et Y ; σ (x) : écart type de X ; σ (y) : écart type de Y.

La covariance est calculée par :

$$\mathrm{Cov}\,(x,y) = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\,(x_i-\bar{x})(y_j-\bar{y}) = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\,x_i\,y_j - \bar{x}\,\bar{y}$$

Φ (x,y) en général est compris entre -1 et 1 : $-1 \leq \Phi\,(x,y) \leq 1$

- Si $\Phi\,(x,y) > 0$: cela indique que les variables X et Y ont une influence mutuelle positive (les variables X et Y variaient dans le même sens).
- Si $\Phi\,(x,y) < 0$: cela indique que les variables X et Y ont une influence mutuelle négative (les variables X et Y variaient en sens contraire).
- Si $\Phi\,(x,y) = 0$: cela indique que les variables X et Y n'ont pas d'influence l'une sur l'autre (X et Y sont indépendants).

Dans le cas pratique, lorsque les variables sont indépendants, il est rarement possible d'obtenir une valeur nulle de ce coefficient, mais seulement on compare la valeur obtenue qui est proche de zéro à une valeur critique tirée de certains tableaux de loi statistique connue.