BIOSTATISTIQUE 1

Chapitre 1 : Généralités sur les séries statistiques

1. Notions de base et terminologie

Population / Echantillon / Individu

- La population correspond à l'ensemble des individus sur lequel porte l'étude ou la prévision, (il est généralement difficile de l'étudier dans sa totalité), et l'échantillon représente la fraction de cette population qui est réellement observée ou étudiée.

- La notion d'individu est très large : les éléments d'un échantillon ou d'une population sont appelés généralement des individus, cependant cette notion peut être remplacé par plusieurs dénominations: unité statistique, sujet, objet, élément, observation, mesure, doses,…toutefois, dès que la dénomination est choisie aucune ambiguïté ne doit persister.

Caractère / Modalité / Variable

Le caractère

les éléments d'un ensemble sont décrits par un caractère. Cela revient à établir une correspondance entre chaque élément i de l'ensemble E et l'ensemble X des modalités ou des valeurs du caractère : chaque élément de E a une modalité (caractère qualitatif) ou une valeur (caractère quantitatif) et une seule dans X. Ainsi le caractère peut être défini comme une des caractéristiques ou des attributs d'un individu,

Modalité

la modalité (respectivement la mesure) est l'une des formes particulière d'un caractère. Les différentes situations où les éléments de E peuvent se trouver à l'égard d'un caractère qualitatif considéré, sont les différentes modalités du caractère qualitatif X. Dans le cas ou le caractère X est quantitatif, les différentes situations où les éléments de E peuvent se trouver sont des mesures. Ces modalités ou ces mesures doivent être à la fois incompatibles (un élément de E ne peut prendre qu'une seule modalité) et exhaustive (à chaque élément de E doit pouvoir correspondre une modalité de X) de sorte que chaque élément de E ait une modalité et une seule dans X.

Variable

Dans chaque étude statistique il est très important de considérer la nature des données (observations, caractères, attributs) que l'on va tester. D'elle dépend la nature des opérations possibles et donc des statistiques utilisables dans chaque situation. Il est donc primordial de

préciser la nature de chaque variable, ou caractère. Il existe deux types de variables (ou observations), celles-ci peuvent être soit quantitatives soit qualitatives. Ces variables peuvent être mesurées d'où l'importance du choix des échelles de mesures, c'est-à-dire, des règles permettant d'affecter une valeur à chaque individu de la population ou de l'échantillon.

Variable quantitative

c'est un caractère auquel on peut associer un nombre c'est-à dire, pour simplifier, que l'on peut "mesurer" (grandeur mesurable). Les différentes situations où peuvent se trouver les éléments sont des mesures; elles sont ordonnables et la moyenne a une signification. On distingue alors deux types de caractère quantitatif :

Variable discrète ou discontinue

c'est un caractère quantitatif, un tel caractère ne prend qu'un nombre fini de valeurs (valeur entière dénombrable et sans aucune valeur intermédiaire). Les différentes situations où peuvent se trouver les éléments (observations, mesures, valeurs,…) sont des nombres isolés dont la liste peut être établie a priori. Exemple: (nombre d'enfants, nombre de pétales d'une fleur, nombre de dents,..) : (1 ; 2 ; 3 ; 4 ; 5 ;….10 ; 11 ;…)

Variable continue

C'est un caractère quantitatif, un tel caractère peut, théoriquement, prendre toutes les valeurs d'un intervalle de l'ensemble des nombres réels. Toutes les valeurs ne sont pas dénombrables et ne peuvent pas être établi a priori. Ses valeurs sont alors regroupées en classes (taille, temps, poids, vitesse, glycémie, altitude, surfaces,….) (1,60 m ; 1,61 m ; 1,62m ;…..)

Caractère qualitatif

Dans ce type de variable les modalités ne sont pas quantifiables (pas mesurables) (couleur des yeux, douleur, …). Ce sont des noms ou ce qui revient au même des sigles ou des codes. Les différentes modalités ne sont pas ordonnables. Attention, même si les modalités sont des codes numériques, les opérations sur les modalités n'ont aucun sens.

Exemple : type de relief avec trois modalités (plaine, montagne, plateau), ou encore taille d'une niche écologique avec quatre modalités (petite, moyenne, grande, très grande). Les données qualitatives peuvent être assimilées au cas des variables discontinues, en supposant que les différentes variantes du caractère qualitatif sont rangées dans un ordre correspondant par exemple à la suite des nombres entiers positifs (différentes couleurs, différents degrés d'infection...).

Fréquences absolues, relatives et cumulées

Désignée par « n », « f » ou « F » la notion de fréquence peut être exprimée de plusieurs manières :

* Fréquence absolue (effectif)

* Fréquence relative (ou fréquence)

* Fréquences cumulées

Fréquences absolues = Effectifs

Le terme de fréquence absolue désigne les effectifs : a chaque modalité x_i du caractère X, peut correspondre un ou plusieurs individus dans l'échantillon de taille N. On appelle effectif (ou fréquence absolue) de la modalité x_i, le nombre n_i où n_i est le nombre d'individus de chacune des modalités

Fréquences relatives = Fréquences

On appelle fréquence de la modalité x_i, le nombre f_i tel que : $f_i = n_i / N$.

Remarques :

Rq1 : Le pourcentage est une fréquence exprimée en pour cent. Il est égal à $100f_i$.

Rq2 : L'emploi des fréquences ou fréquences relatives s'avère utile pour comparer deux distributions de fréquences établies à partir d'échantillons de tailles différentes.

Effectifs et Fréquences cumulés

On appelle effectifs cumulés (resp. fréquences cumulées) en x_i, le nombre $\mathbf{N}_i = \sum_{p=1}^{i} \mathbf{n}_p$ (resp. $\mathbf{F}_i = \sum_{p=1}^{i} \mathbf{f}_p$).

Remarques : k étant le nombre de modalités du caractère X.
Rq3 : la taille de l'échantillon est $N = \sum_{i=1}^{k} n_i$.
Rq4 : $\sum_{i=1}^{k} f_i = 1$.

2. Représentation des données

Il existe plusieurs niveaux de description statistique : la présentation brute des données, des présentations par tableaux numériques, des représentations graphiques et des résumés numériques fournis par un petit nombre de paramètres caractéristiques.

Tableaux statistiques

En général une série statistique à caractère discret se présente sous la forme :

Valeurs	X_1	X_2	…	X_k
Effectifs	n_1	n_2	…	n_k
Fréquences	f_1	f_2	…	f_k

Plutôt que réécrire ce tableau on écrira souvent : la série (x_i,n_i), $1 \leq i \leq k$. (On n'indique pas le nombre de valeurs lorsqu'il n'y a pas d'ambigüité). Souvent on notera N l'effectif total de cette série donc $N=n_1+n_2+...+n_p$.

Représentations graphiques et statistiques descriptives

Les représentations graphiques sont très importantes en statistique descriptive. Elles ont l'avantage de renseigner immédiatement sur l'allure générale de la distribution. Elles facilitent l'interprétation des données recueillies. La représentation graphique des données montre la forme générale de la distribution et donne une image de la grandeur des nombres qui constituent les données. D'autres statistiques simples sont utilisées pour représenter le centre de la distribution et les mesures liées à la dispersion des observations autour de cette tendance centrale.

Chapitre 2 : Variables statistiques quantitatives

L'étude descriptive d'une variable statistique quantitative (aussi bien discrète que continue) se fait en 3 étapes :

- Description préliminaire

- Caractéristiques de position centrale

- Caractéristiques de dispersion

Une variable quantitative peut être :

Discrète : si la variable ne prend qu'un nombre fini de valeurs (ces valeurs sont appelées modalités et notées x_i) .

Continue : si la variable prend ses valeurs dans un intervalle (classe)

1. Représentation graphique des séries statistiques

Tableau statistique

Quand la variable statistique X est discrète, on compte, pour chaque valeur de X, le nombre d'individus prenant cette valeur ; c'est l'effectif de la valeur. On aboutit à un tableau du type :

Valeurs de la variable x_i	Effectifs n_i
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_p	n_p

Individus	Variable 1	Variable 2
1	X_1	Y_1
2	X_2	Y_2
$\vdots$	$\vdots$	
N	X_N	Y_N

avec $X_1 < X_2 < \ldots < X_p$.

Quand la variable statistique X est continue, on regroupe les valeurs en classes. Les classes sont des intervalles semi-ouverts $[e_i,e_{i+1}[$. Leur amplitude est le nombre : $(e_{i+1}-e_i)$ et leur centre, le nombre

$((e_{i+1}+e_i)/2)$.

Pour chaque classe, on compte le nombre d'individus qui prennent une valeur supérieure ou égale à e_i et inférieure à e_{i+1} : c'est l'effectif de la classe. On aboutit à un tableau de type :

Classes	Effectifs n_i
$[e_0 ; e_1[$	n_1
$[e_1 ; e_2[$	n_2
$\vdots$	$\vdots$
$[e_{p-1} ; e_p[$	n_p

Remarques

• Quand le nombre de valeurs prises par la variable statistique est trop grand, on traite la variable discrète comme une variable continue.

• Quand on regroupe les valeurs par classes, on essaye d'avoir des classes de même amplitude et pas trop nombreuses. Mais, souvent, les valeurs extrêmes posent problème, c'est pourquoi les premières ou dernières classes sont soit ouvertes, soit d'amplitude différente des autres classes.

Exemple 1 (cas discret)

Afin d'étudier la production en fruits d'une certaine variété de fraises, 150 arbrisseaux ont été sélectionnés dans un champ. On a dénombré le nombre de fruits qu'ils portent. Le tableau suivant a été obtenu :

Nombre de fruits	8	10	16	20	24	32	42
Nombre d'arbrisseaux	12	23	41	24	22	16	12

Exemple 2 : (Cas continu)

Dans le but de déterminer le temps de réaction des gens par rapport au son, l'expérience suivante a été entreprise : 50 personnes ont été réunies et pour chacune d'elles, on a enregistré le temps mis pour réagir après avoir entendu un signal sonore. Les résultats de l'expérience ont été reportés dans le tableau suivant :

T. de réaction	$[0,45;0,51[$	$[0,51;0,57[$	$[0,57;0,63[$	$[0,63;0,69[$	$[0,69;0,75[$	$[0,75;0,81[$
Nre d'individus	2	8	18	16	4	2

Comment représenter une série statistique ?

- Pour représenter une variable statistique discrète, on utilise un diagramme en bâtons (chaque bâton a une hauteur proportionnelle à l'effectif et/ou à la fréquence).

- Pour représenter une variable statistique continue, on trace un histogramme. L'histogramme est constitué de rectangles juxtaposés dont la surface est proportionnelle à l'effectif de la classe correspondante.

Si les classes ont des amplitudes égales, la hauteur des rectangles est proportionnelle à l'effectif. Si les classes ont des amplitudes inégales, on représente la classe ayant la plus petite amplitude ; puis on compense une amplitude k fois plus grande par une hauteur k fois plus petite.

2. Paramètres de position centrale

Les paramètres de position centrale sont un ensemble de valeurs caractéristiques qui permettent une représentation de l'information contenue dans la série statistique. Certaines de ces caractéristiques nous renseigneront sur la position (mode, médiane, quartiles). D'autres, en plus de cette information, serviront à résumer les données en notre possession (moyennes).

Les caractéristiques de position centrale sont :

1 - Le mode

2 - Les quantiles

3 - Les moyennes (arithmétique, géométrique, harmonique).

1. **Le mode** : (M_0)

- Dans le cas d'une variable discrète, le mode est la valeur la plus fréquente de la variable statistique, c'est à dire celle qui correspond au plus grand effectif.

Remarque :

Le mode peut ne pas être unique. Certaines distributions peuvent présenter plusieurs valeurs dont les effectifs sont égaux et qui sont les plus grands de la distribution.

Dans le cas de l'exemple 1 le plus grand effectif est 41, le mode est donc : $M_0=16$.

Dans le cas d'une variable continue, nous ne pouvons pas parler de mode car les modalités sont des classes et non des valeurs. La notion de mode correspond à une idée d'intensité plutôt qu'à une idée d'effectifs. Mais pour la variable continue l'amplitude entre en jeu. Les classes peuvent être d'inégale amplitudes et une casse a le plus grand effectif n'est pas nécessairement la classe où le caractère est le plus intense. Nous définirons donc la classe modale comme suit : C'est la classe qui correspond au plus grand rapport $((n_i)/(a_i))$, a_i étant l'amplitude de la classe $[e_{i-1},e_i[$.

Dans le cas de l'exemple 2 la classe modale est [0,57-0,63[.

La moyenne arithmétique

• Quand la série statistique est discrète, de taille N, on appelle moyenne de X le nombre :

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i X_i = \sum_{i=1}^{k} f_i X_i$$

où $X_1,\ldots,X_k$ sont les différentes valeurs et $n_1,\ldots,n_k$ leurs effectifs correspondants satisfaisant $n_1+\ldots+n_k=N$.

Ainsi, la moyenne arithmétique pour l'exemple 1 est 20,04.

• Quand la série statistique est continue, de taille N, pour calculer la moyenne, on utilise la formule précédente en remplaçant x_i par le centre c_i de l'intervalle $[e_{i-1},e_i[$.

La moyenne de X est alors le nombre :

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{p} n_i c_i = \sum_{i=1}^{p} f_i c_i$$

Ainsi, la moyenne arithmétique pour l'exemple 2 est 0,6216.

2. Les quantiles :

Définition : on appelle quantiles les valeurs du caractère qui définissent les bornes d'une partition en classes d'effectifs égaux.

2.1 **La médiane**

• La médiane est le nombre qui sépare la série ordonnée en valeurs croissantes en deux groupes de même effectif. Pour la trouver, on écrit la liste de toutes les valeurs de la série par ordre croissant, chacune d'elles étant répétée autant de fois que son effectif.

On distingue ensuite deux cas :

-- si l'effectif total N est un nombre impair, la médiane est le terme de rang ((N+1)/2).

-- si l'effectif total N est un nombre pair, la médiane est le centre de l'intervalle formé par les termes de rang (N/2) et (N/2)+1..

• Quand la série est regroupée par classes, on détermine la médiane graphiquement à partir du polygone des effectifs ou des fréquences cumulés.

On calcule pour chaque classe $[e_{i-1},e_i[$ l'effectif cumulé croissant N_i, c'est-à-dire le nombre d'individus qui prennent une valeur inférieure à e_i. On place ensuite dans un repère les points (e_i,N_i), on obtient ainsi le polygone des effectifs cumulés croissants.

La médiane est l'abscisse du point dont l'ordonnée est (N/2). On peut calculer une valeur plus précise par interpolation linéaire.

2.2 Le premier et troisième quartile

- le premier quartile Q_1 est la valeur de la variable au-dessous de laquelle on trouve le quart de l'effectif. Si la série est discrète, c'est la valeur de la variable dont le rang est égal ou immédiatement supérieur au quart de l'effectif. Si la série est continue, on lit la valeur correspondant à 25 % de l'effectif sur la case du tableau des fréquences ou des effectifs cumulés. On peut calculer une valeur plus précise par interpolation linéaire.

- le troisième quartile Q_3 est la valeur de la variable au-dessous de laquelle on trouve les trois-quarts de l'effectif. Si la variable est discrète, c'est la valeur de la variable dont le rang est égal ou immédiatement supérieur aux trois-quarts de l'effectif. Si la série est continue, on lit la valeur correspondant à 75 % de l'effectif sur la case du tableau des fréquences ou des effectifs cumulés. On peut aussi calculer une valeur plus précise par interpolation linéaire.

2.3 Les déciles et les centiles

- Les déciles partagent la série en 10 parties de même effectif. Ce sont les 9 valeurs de X qui permettent de découper la distribution en dix classes d'effectifs égaux. 0n les note $D_1,D_2,\ldots,D_9$
- Les centiles partagent la série en 100 parties de même effectif. Les centiles sont les 99 valeurs de X qui permettent de découper la distribution en 100 classes d'effectifs égaux. 0n les note $C_1,C_2,\ldots,C_{99}$
- Ainsi, Les déciles correspondent à N/10, 2N/10,…,9N/10. Le 5ième décile est la médiane. Pour les centiles : N/100,2N/100,…,99N/100. Le 50ième centile est la médiane.

3. Interpolation linéaire

Soit f une fonction définie sur $\mathbb{R}$, [a;b] un intervalle de $\mathbb{R}$ et c un nombre réel . Quand il n'est pas possible de calculer l'image de c par f , on utilise une interpolation linéaire, cela consiste à remplacer f(c) par g(c) où g est la fonction affine telle que : g(a)=f(a) et g(b)=f(b).

Cela consiste à remplacer la courbe représentative de f sur [a;b] par la droite (AB) (On dit que l'on a déterminé f(c) par interpolation linéaire). Les détails et explications donnés en cours peuvent

être trouvés dans le fascicule 1 de Ahmed Chibat 'cours de statistique descriptive" disponible en grande quantité à la bibliothèque.

3. Caractéristiques de dispersion

Une fois qu'une caractéristique de position centrale est cernée, il s'agit de voir comment se développe le phénomène autour de ce centre. C'est à dire : comment il se disperse ?

Les caractéristiques de dispersion sont : L'étendue, l'écart interquartile, la variance, l'écart-type

Etendue : (notée E)

- Quand la série statistique est discrète, l'étendue est la différence entre la plus grande valeur et la plus petite valeur de la série.

- Quand la série statistique est continue, l'étendue est la longueur de l'intervalle sur lequel se disperse la variable. Pour l'exemple 1, nous avons E=42-8=34. Pour l'exemple 2, nous avons E=0,81-0,45=0,36.

Ecart interquartile : (noté I_Q)

C'est la différence entre le troisième quartile et le premier quartile. $I_Q=Q_3-Q_1$.

Pour l'exemple 1, nous avons $I_Q=24-16=8$. Pour l'exemple 2, nous avons $I_Q=0,666-0,578=0,088$.

Variance : (notée Var(X))

• Quand la série statistique est discrète, de taille N, on appelle variance de X le nombre :

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i X_i^2 - \bar{X}^2 = \sum_{i=1}^{k} f_i X_i^2 - \bar{X}^2$$

où $X_1,\ldots,X_k$ sont les différentes valeurs et $n_1,\ldots,n_k$ leurs effectifs correspondants satisfaisant $n_1+\ldots+n_k=N$. La variance est le paramètre qui, par sa construction même, tient effectivement compte de la dispersion de tous les individus.

• Quand la série statistique est continue, de taille N, pour calculer la variance, on utilise la formule précédente en remplaçant x_i par le centre c_i de l'intervalle $[e_{i-1},e_i[$.

La variance de X est alors le nombre :

$$Var(X) = \frac{1}{N}\sum_{i=1}^{p} n_i c_i^2 - \bar{X}^2 = \sum_{i=1}^{p} f_i c_i^2 - \bar{X}^2.$$

Pour l'exemple 1 (voir le tableau récapitulatif de l'exemple 1) Var(X)=489,253-(20,04)²=87,651.

Pour l'exemple 2 (voir le tableau récapitulatif de l'exemple 2) Var(X)=0,3906-(0,621)²=0,0042.

Ecart-type : (noté σ(X))

Lorsque nous avons établi la formule pour le calcul de la variance les distances ont été prises au carré. De cette façon l'unité de la variable statistique figurera au carré dans la quantité représentant la variance. Pour revenir à l'unité initiale nous allons introduire un nouveau paramètre appelé écart-type. C'est la racine carrée de la variance.

$$\sigma_X = \sqrt{Var(X)}.$$

Pour l'exemple 1 $\sigma_X = \sqrt{87,651} = 9,36$
Pour l'exemple 2 $\sigma_X = \sqrt{0,0042} = 0,065$

Tableau récapitulatif de l'exemple 1 :

X_i	n_i	f_i	n_iX_i	f_iX_i	$n_iX_i^2$	$f_iX_i^2$
8	12	0,080	96	0,64	768	5,12
10	23	0,153	230	1,53	2300	15,3
16	41	0,273	656	4,368	10496	69,88
20	24	0,160	480	3,2	9600	64
24	22	0,147	528	3,528	12672	84,67
32	16	0,107	512	3,424	1638	109,56
42	12	0,080	504	3,36	21168	141,12
Total	150	1	3006	**20,05**	73388	**489,66**
Total/N			**20,04**		**489,253**	

Tableau récapitulatif de l'exemple 2 :

classes	n_i	f_i	c_i	$n_i.c_i$	$f_i.c_i$	$n_i.c_i^2$	$f_i.c_i^2$
$[0,45;0,51[$	2	0,04	0,48	0,96	0,0192	0,4608	0,00921
$[0,51;0,57[$	8	0,16	0,54	4,32	0,0864	2,3328	0,04665
$[0,57;0,63[$	18	0,36	0,60	10,8	0,216	6,48	0,1296
$[0,63;0,69[$	16	0,32	0,66	10,56	0,2112	6,9696	0,13939
$[0,69;0,75[$	4	0,08	0,72	2,88	0,0576	2,0736	0,04147
$[0,75;0,81[$	2	0,04	0,78	1,56	0,0312	1,2168	0,02433
Total	50	1		31,08	0,6216	19,5336	0,3906

Paramètre de dispersion relative

La comparaison des paramètres de dispersion absolue de deux caractères n'a de sens que si les deux caractères sont de même nature et de même ordre de grandeur. Dans le cas contraire, la comparaison n'est possible qu'en ayant recours à des mesures de dispersion relative, c'est à dire en effectuant le rapport entre un paramètre de dispersion absolue et la valeur centrale qui lui tient de référence. Ainsi,

Dispersion relative = Paramètre de dispersion absolue/Valeur centrale

Les plus courants sont :

- **le coefficient de variation** (C.V.) = écart-type/moyenne
- **le coefficient interquartile relatif** (CIR) = $(Q3-Q1)/Q_2$

Le coefficient de variation est particulièrement intéressant car il permet une interprétation assez précise des résultats. En effet, plus le coefficient de variation est faible, plus la dispersion est faible.

Le coefficient de variation est un indicateur de l'homogénéité de la population. On considère qu'un coefficient de variation **inférieur à 15% indique que la population est homogène, tandis** qu'un coefficient **supérieur à 15% indique que les valeurs sont** relativement dispersées. Le coefficient de variation est une mesure sans unité et indépendante de l'ordre de grandeur. On peut donc l'utiliser pour comparer la dispersion de variables statistiques avec des ordres de grandeur et des unités différentes.

Chapitre 3. Les distributions statistiques à deux caractères

L'objectif de cette étude statistique est d'étudier sur une même population de N individus, deux caractères différents (ou modalités différentes) et de rechercher s'il existe un lien ou corrélation entre ces deux variables. Exemple de relations possibles entre les variables suivantes : taille et âge ; diabète et poids ; taux de cholestérol et régime alimentaire ; niche écologique et population ; ensoleillement et croissance végétale ; toxine et réaction métabolique ; survie et pollution ; effets et doses; organe 1 et 2 ; organe et fonction biologique ; …

1. Vocabulaire, tableaux, graphiques

Les caractères étudiés peuvent être aussi bien qualitatifs que quantitatifs. Les résultats sont généralement représentés sous forme d'un tableau à double entrée, appelé tableau à deux dimensions, ou tableau croisé ou tableau de contingence, ou parfois tableau de corrélation. Les tableaux croisés permettent :

- De synthétiser l 'information
- De faire le lien entre deux variables

Désignons par (X,Y) le couple de caractères étudiés. A chaque observation conjointe (X_i,Y_i) est associée le nombre d'individus ayant simultanément la valeur X_i pour le caractère X et la valeur Y_i pour le caractère Y. Ce nombre est noté n_{ij} et appelé l'effectif associé à l'observation (X_i,Y_j).

Y_j / X_i	Y_1	Y_2	…	Y_j	…	Y_p	Total	
X_1	n_{11}	n_{12}		n_{1j}		n_{1p}	$\sum_{j=1}^{p} n_{1j} = n_{1.}$	Effe-
X_2	n_{21}	n_{22}		n_{2j}		n_{2p}	$\sum_{j=1}^{p} n_{2j} = n_{2.}$	ctifs
⋮								marg-
X_i	n_{i1}	n_{i2}		n_{ij}		n_{ip}	$\sum_{j=1}^{p} n_{ij} = n_{i.}$	inaux
⋮								de
X_k	n_{k1}	n_{k2}		n_{kj}		n_{kp}		X
Total	$\sum_{i=1}^{k} n_{i1} = n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.p}$	N	
	Effectifs			Marginaux			de Y	

La ligne et la colonne 'Total' correspond aux marges du tableau.

Calcul des fréquences d'une statistique à deux variables

Fréquences relatives

La fréquence de l'observation (X_i,Y_j) s'exprime par l'expression f_{ij}. Elle correspond à la proportion d'individus qui possèdent simultanément les valeurs X_i et Y_j. Elle est obtenue par la formule suivante :

$\mathbf{f}_{ij}=\frac{n_{ij}}{N}$ il est à remarquer que $\sum_{i=1}^{k}\sum_{j=1}^{p}\mathbf{f}_{ij}=\mathbf{1}$.

Fréquences relatives marginales $\mathbf{f_{i.}}$ et $\mathbf{f_{.j}}$

Il s'agit des fréquences relatives des distributions marginales.

$$f_{i.}=\frac{n_{i.}}{N} \text{ et } f_{.j}=\frac{n_{.j}}{N}$$

Calcul des moyennes marginales d'une statistique à deux variables

Dans certaines distributions statistiques bidimensionnelles il est possible de calculer les moyennes, les variances et les écart-types marginaux. Nous expliciterons ces calculs à travers un exemple.

Exemple : Afin d'étudier la relation existante entre le nombre de feuilles et le nombre de fruits d'une certaine variété de fraises, 150 arbrisseaux ont été sélectionnés dans un champ. On a dénombré les feuilles et les fruits de chaque arbrisseau et le tableau suivant a été obtenu :

X/Y	6	11	14	16	18	Total
8	8	2	1	1	0	**12**
10	4	16	2	0	1	**23**
16	3	10	15	8	5	**41**
20	2	4	14	3	1	**24**
24	4	5	6	5	2	**22**
32	3	1	2	8	2	**16**
42	0	0	6	4	2	**12**
Total	**24**	**38**	**46**	**29**	**13**	**150**

X représente le nombre de fruits, Y représente le nombre de feuilles

Calculer respectivement : 1- Les moyennes marginales de X puis de Y

2- Les variances et l'écart-type marginaux de X puis de Y

Pour les moyennes et les variances :

$$\bar{X}_M=\frac{1}{N}\sum_{i=1}^{k}n_{i.}X_i,\quad \bar{Y}_M=\frac{1}{N}\sum_{j=1}^{p}n_{.j}Y_j.$$

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{k} n_{i.}X_i^2 - \bar{X}^2,\ Var_M(Y) = \frac{1}{N}\sum_{j=1}^{p} n_{.j}Y_j^2 - \bar{Y}^2$$

Tableau de distribution marginale de X et de Y

X	$n_{i.}$	$n_{i.}X_i$	$n_{i.}X_i^2$
8	12	96	768
10	23	230	2300
16	41	656	10496
20	24	80	9600
24	22	528	12672
32	16	512	16384
42	12	504	21168
Total	150	3006	73388

et

Y	$n_{.j}$	$n_{.j}Y_j$	$n_{.j}Y_j^2$
6	24	144	864
11	38	418	4598
14	46	644	9016
16	29	464	7412
18	13	234	4212
Total	150	1904	26114

Applications numériques :
a) $\bar{X}_M = 20,04$.
b) $\bar{Y}_M = 12,69.3$
c) $Var_M(X) = 87,651$.
d) $Var_M(Y) = 12,981$.

Chapitre 4. Outils d'analyse

Jusqu'ici on ne s'est intéressé aux variables que prises isolément. Nous avons vu leurs caractéristiques de position. Nous avons vu comment elles se dispersent. Mais qu'en est-il de leur relation l'une par rapport à l'autre ? Quand peut-on dire qu'elles varient dans le même sens ou dans le sens contraire ? comment peut-on mesurer la force de leur liaison ? Nous présentons dans ce qui suit quelques réponses à ces questions.

1. Covariance

Une première approche pour évaluer la relation éventuelle des valeurs d'une variable X avec les valeurs d'une variable Y est donnée par le calcul de la covariance. La covariance du couple (X,Y), notée Cov(X,Y), correspond à la moyenne de $(X - \bar{X})(Y - \bar{Y})$. La formule est donc la suivante :

$$Cov(X,Y) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}(X_i - \bar{X}_M)(Y_j - \bar{Y}_M)$$

Par analogie aux formules précédentes les formules pratiques de calculs de la covariance peuvent aussi s'écrire :

$$Cov(X,Y) = \left[\frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} X_i Y_j \right] - [\bar{X}_M \bar{Y}_M].$$

Propriétés de la covariance

- $Cov(X,X) = var(X)$
- $|Cov(X,Y)| \leq \sigma_X \sigma_Y$

- Le signe de la Covariance est un indicateur de la tendance de la relation sens positif ou négatif (direction d'étirement du nuage de point). Une covariance positive indique une tendance « croissante » des valeurs de Y en fonction de X, une covariance négative une tendance « décroissante ».

Exemple. En reprenant notre exemple du début du chapitre, on calcule la covariance entre X (le nombre de fruits) et Y (le nombre de feuilles).

Il s'agit de dresser le tableau qui répond à cette demande. Soit donc le tableau suivant :

X/Y	6	11	14	16	18	TOTAL
8	8 **384**	2 **176**	1 **112**	1 **128**	0 **0**	 **800**
10	4 **240**	16 **1760**	2 **280**	0 **0**	1 **180**	 **2460**
16	3 **288**	10 **1760**	15 **3360**	8 **2048**	5 **1440**	 **8896**
20	2 **240**	4 **880**	14 **3920**	3 **960**	1 **360**	 **6360**
24	4 **576**	5 **1320**	6 **2016**	5 **1920**	2 **864**	 **6696**
32	3 **576**	1 **352**	2 **896**	8 **4096**	2 **1152**	 **7072**
42	0 **0**	0 **0**	6 **3528**	4 **2688**	2 **1512**	 **7728**
TOTAL	**2304**	**6248**	**14112**	**11840**	**5508**	**40012**

Dans chaque case intérieure, nous remarquons l'existence de deux nombres l'un en petits caractères : c'est l'effectif n_{ij}.

L'autre en gros caractères : c'est le produit $n_{ij}X_iY_j$. La somme des nombres sur la dernière ligne est égale à la somme des nombres sur la dernière colonne, c'est le total des totaux. C'est :

$\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}X_iY_j$ **et on obtient directement la covariance**. **C'est** $Cov(X,Y) = \frac{40012}{150} - (20,04.12,693) = 12,379$

2. Coefficient de corrélation

La covariance n'est pas un indicateur indépendant de l'ordre de grandeur des variables impliquées (de l'unité employée, par exemple). Le coefficient de corrélation, noté r, permet de résoudre cette difficulté. Ce coefficient pour le couple (X,Y) s'écrit selon la formule suivante :

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

où σ_X et σ_Y désignent respectivement l'écart-type de la série statistique X et celui de la série statistique Y.

Propriétés de r :

- r est toujours compris entre -1 et 1, c'est une covariance « réduite »

- quand (|r|=1), les points représentatifs des couples (X_i,Y_i), sont parfaitement alignés sur le graphique :

- quand (|r| est voisin de 1), il existe une forte corrélation entre X et Y. Néanmoins (attention), ceci ne veut pas dire qu'il existe une relation de cause à effet entre elles.

- pour r=1, la droite de la pente est croissante

- Si 0<r<1, la corrélation est positive, X et Y varient dans le même sens.

- Si -1<r<0, la corrélation est négative, X et Y varient dans le sens contraire.

- pour r=-1, la droite de la pente est décroissante

- quand (r=0), aucune tendance linéaire ne peut être déterminée mais il se peut qu'il y ai une autre corrélation non linéaire.

Exemple de calcul. Reprenons notre exemple du début du chapitre. Nous avons $r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{12,379}{9,36.3,6} = 0.367.$

Nous déduisons de ce résultat que, pour les fraisiers, il n y a pas de relation linéaire entre le nombre de feuilles et le nombre de leurs fruits.

Chapitre 5. Ajustement linéaire-Regression

L'une des méthodes simple d'étude de la corrélation entre 2 variables consiste à rechercher une courbe d'équation Y=f(X) qui passe au plus proche de tous les points expérimentaux. Une telle courbe permet d'avoir une idée sur la tendance de la relation entre les variables étudiées et de formuler d'éventuelles prévisions.

1. Droite de régression linéaire

Une droite de régression linéaire s'écrit selon l'équation : «y=ax+b ». Cette approche de corrélation repose sur l'hypothèse que la relation entre deux variables est de nature linéaire.

En partant de l'équation y=ax+b , a et b doivent être choisis convenablement de sorte que la droite passe au plus proche (ou par le plus possible) des points expérimentaux. Pour ce faire, on utilise la méthode des moindres carrés : On cherche les coefficients a et b de la droite qui minimise la somme des carrés des distances entre les points expérimentaux et la droite de régression (les points théoriques).

- les coefficients a (pente) et b (ordonnée à l'origine) se déterminent comme suit :

$$\hat{a} = \frac{cov(X,Y)}{var(X)},\ \hat{b} = \bar{Y}_M - \hat{a}\bar{X}_M.$$

Ainsi la droite de régression de Y en X a pour équation :

$$Y = \hat{a}X + \hat{b},$$

Ces équations permettent de définir deux droites différentes de régression à l'intérieur du nuage de point. Néanmoins cette inversion, qui permet d'obtenir l'équation X=a′Y+b′ (régression de X en Y) n'est pas souvent intéressante, car en général, Y est une variable à exprimer et X est une variable potentiellement explicative.

Propriété de ces deux droites de régression :

1) les deux droites de régression se coupent en un point qui a pour coordonnées les moyennes de X et de Y (en remplaçant dans l'equation X par sa moyenne, il est ainsi possible de retrouver Y (qui correspond à la moyenne de Y)).

2) les coefficients a et a′ (qui sont les pentes) sont toujours de même signe (soit -- (corrélation négative) soit + (corrélation positive)), ainsi les deux droites sont orientées dans le même sens que le nuage de point.

3) l'angle maximum des deux droites de régression est de 90° (droites perpendiculaires). Dans ce cas, les points sont dispersés dans tout le plan. La corrélation est nulle. Les droites sont respectivement parallèles à l'axe des x et à l'axe des y.

Exemple. Afin d'étudier la relation qui existe entre le temps de réaction au son et le temps de réaction à la lumière, un groupe de personne a été ensuite divisé en 6 sous-groupes après les résultats préliminaires, des plus lents au pus rapides. L'expérience a été ensuite entreprise et on a enregistré le temps moyen dans chaque sous-groupe relatif au deux réactions. Le temps est mesuré en unités conventionnelles.

Le tableau suivant a été obtenu :

Temps de réaction au son (X)	2	4	6	8	10	12
Temps de réaction à la lumière (Y)	5,5	9,2	11,8	15,2	18,5	22

Pour calculer les coefficients de la droite d'ajustement nous avons donc besoin de calculer les moyennes marginales des deux variables, la variance de X et la covariance entre X et Y.

Pour cela nous sommes conduit à dresser un tableau à quatre colonnes.

	X_i	Y_i	X_i^2	$X_{i.}Y_i$
	2	5,5	4	11
	4	9,2	16	36,8
	6	11,8	36	70,8
	8	15,2	64	121,6
	10	18,5	100	185
	12	22	144	264
Total	42	684	364	689,2
Total/N	$\bar{X}$ =**7**	$\bar{Y}$ =**13,7**	**60,67**	**114,867**

$Var_M(X)=60,67-7^2=11,67$; $Cov(X,Y)=114,867-(7)(13,7)=18,97$.

Ainsi les coefficients de la droite d'ajustement de Y en fonction de X s'obtiennent par :

$$\begin{cases} a = \frac{Cov(X,Y)}{Var(X)} = \frac{18,97}{11,67} = 1,63 \\ b = \bar{X} - a\bar{Y} = 13,7 - 1,67.(7) = 2,32 \end{cases}$$

la droite d'ajustement aura donc pour équation : Y=1,63X+2,32.

Cette approche de corrélation repose sur l'hypothèse que la relation entre deux variables est de nature linéaire. En faite, il est possible de soupçonner une relation différente entre ces variables :

- courbe de puissance - courbe exponentielle - courbe logarithmique, - courbe hyperbolique, etc…

Cependant, il existe de nombreuses méthodes permettant de « linéariser » un grand nombre de ces courbes. Ainsi, on se retrouve souvent dans des situations où il est alors possible de tester l'existence d'une relation linéaire entre les variables auxiliaires.