Statistique descriptive à deux variables

1. Les distribution statistique à deux dimensions

Nous avons vu précédemment que les statistiques à une variable s'intéressaient, pour une population donnée, à un caractère donné. Lorsque l'on s'intéresse à l'étude simultanée de deux caractères d'une même population, on fait ce que l'on appelle des statistiques à deux variables, en étudiant des séries statistiques doubles.

Définition 1.1

On considère une population Ω d'effectif n, si on étudie deux caractères X et Y de cette population, on dit que l'on étudie une série statistique double (bivariante). La Statistique bivariée (à 2 dimension) est une application définie comme suit

$$\begin{aligned}(X,Y):\Omega &\longrightarrow \mathbb{R}^2\\ \omega &\longmapsto (X(\omega),Y(\omega))\end{aligned}$$

Exemple 1.
- Une entreprise veut mener une étude sur la liaison entre les dépenses mensuelles en publicité (en milliers de DA) X et le volume des ventes (en milliers de DA) Y qu'elle réalise.
- On a effectué une enquête sur 100 foyers en observant " les dépenses mensuelles" X et "le revenu mensuel" Y (en milliers de DA).

1.1. Distribution marginale et conditionnelles

Considérons un échantillon comprenant n individus, pour chacun des quels nous avons,

$$X = \{x_1, x_2, \ldots, x_k\}$$
$$Y = \{y_1, y_2, \ldots, y_l\}$$

deux variables qualitatives ayant respectivement k et l modalités.

Tableau de contingences

A chacun des sous ensembles correspond une case du tableau statistique à double entrées, où figurent en lignes les modalités de X et en colonnes les modalités de Y. Le nombre d'éléments du sous ensemble est l'effectif n_{ij} des individus présentant à la fois les modalités x_i et y_j.

X \ Y	y_1	.....	y_j	.....	y_l	Totaux
x_1	n_{11}	.....	n_{1j}	.....	n_{1l}	$n_{1\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	n_{i1}	.....	n_{ij}	.....	n_{il}	$n_{i\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_k	n_{k1}	.....	n_{kj}	.....	n_{kl}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$	.....	$n_{\bullet j}$	.....	$n_{\bullet l}$	$n_{\bullet\bullet}$

$n_{i\bullet}$ est le total des effectifs de la ligne i :

$$n_{i\bullet} = n_{i1} + n_{i2} + \ldots + n_{ij} + \ldots + n_{il} = \sum_{j=1}^{l} n_{ij}$$

$n_{\bullet j}$ est le total des effectifs de la colonne j :

$$n_{\bullet j} = n_{1j} + n_{2j} + \ldots + n_{ij} + \ldots + n_{kj} = \sum_{i=1}^{k} n_{ij}$$

tel que,

$$n = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j} = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}$$

Définition 1.2

On appelle fréquence de l'évènement (x_i, y_j) la proportion des observations qui présente simultanément les modalités x_i et y_j notée par

$$f_{ij} = \frac{n_{ij}}{n}.$$

Définition 1.3

On appelle **distributions marginales** les totaux des fréquences relatives aux différentes lignes et colonnes, c'est-à-dire :

$$f_{i\bullet} = \sum_{j=1}^{l} f_{ij} = \sum_{j=1}^{l} \frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n}$$

$f_{i\bullet}$ est le totale des fréquences de la colonne j.
$f_{\bullet j} = \frac{n_{\bullet j}}{n}$ est le totale des fréquences de la ligne i.

Exemple 2.
On a effectué une enquête sur 100 foyers en observant " les dépenses mensuelles" X et "le revenu mensuel" Y (en milliers de DA), les résultats sont donnés dans le tableau suivant :

X \ Y	[4,10[	[10,20[	[20,40]	$n_{i\bullet}$
[3,5[	20	10	0	$n_{1\bullet} = 30$
[5,15[	10	20	10	$n_{2\bullet} = 40$
[15,35]	0	10	20	$n_{3\bullet} = 30$
$n_{\bullet j}$	$n_{\bullet 1} = 30$	$n_{\bullet 2} = 40$	$n_{\bullet 3} = 30$	$n = 100$

Les distributions marginales

- de X :

X	[3,5[	[5,15[	[15,35]	Total
$n_{i\bullet}$	30	40	30	100
$f_{i\bullet}$	$\frac{30}{100} = 0.3$	$\frac{40}{100} = 0.4$	$\frac{30}{100} = 0.3$	1

- de Y :

Y	[4,10[	[10,20[	[20,40]	Total
$n_{\bullet j}$	30	40	30	100
$f_{\bullet j}$	$\frac{30}{100} = 0.3$	$\frac{40}{100} = 0.4$	$\frac{30}{100} = 0.3$	1

Remarque 1.1

La connaissance des distributions marginales de X et de Y ne suffit pas, en général, pour déterminer la distribution de couple (X, Y). Ce la n'est possible que si X et Y sont indépendantes.

1.2. Indépendance statistique

Définition 1.4

On dit que les variables X et Y sont indépendantes si, et seulement si :

$$f_{ij} = f_{i\bullet} \times f_{\bullet j} \quad \text{ou} \quad n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}, \quad \forall\, i = 1, \ldots, k \text{ et } \forall\, j = 1, \ldots, l$$

Cette définition donne une interprétation intéressante de d'indépendance ; elle signifie que dans ce cas, les effectifs des modalités conjointes peuvent se calculer uniquement à partir des distributions marginales, supposées « identiques » aux distributions de X et Y dans la population ; en d'autres termes, si X et Y sont indépendantes, les observations séparées de X et de Y donnent la même information qu'une observation conjointe.

Exemple 3.
Pour l'indépendance des variables X et Y de l'exemple 2. Si on choisit $i = 2$ et $j = 1$, nous obtenons

$$n_{21} \times n = 10 \times 100 = 1000,$$

et

$$n_{2\bullet} \times n_{\bullet 1} = 40 \times 30 = 1200.$$

qui sont bien évidement non égaux. Par conséquent, il existe i et j tel que

$$n_{ij} \times n \neq n_{i\bullet} \times n_{\bullet j}$$

Donc, X et Y ne sont pas indépendants.

2. La représentation graphique

2.1. Nuage de points

Définition 1.5

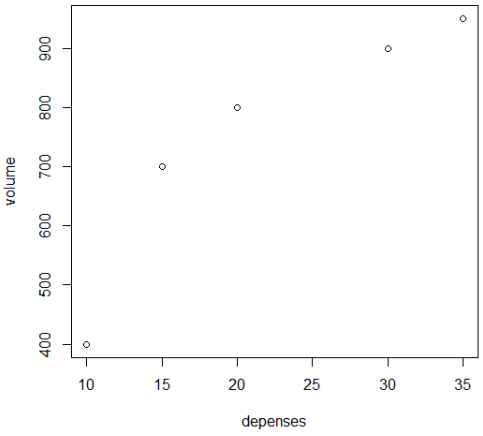
L'ensemble des points M_1, M_2, …, M_n de coordonnées respectives $M_1(x_1; y_1)$, $M_2(x_2; y_2)$, …, $M_n(x_n; y_n)$ dans un repère du plan est appelé nuage de points de la série.

Exemple 4.

Une entreprise veut mener une étude sur la liaison entre les dépenses mensuelles en publicité (en milliers de DA) et le volume des ventes (en milliers de DA) qu'elle réalise. Nous avons obtenu au cours des cinq derniers mois les données suivantes :

X	10	15	20	30	35
Y	400	700	800	900	950

Cette distribution peut être représentée par le nuage de points



3. Description numérique

1. **caractéristique des séries marginales**
   Dans le cas d'une variable statistique à deux dimensions X et Y, les moyennes sont données respectivement par

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i = \sum_{i=1}^{k} f_{i\bullet}x_i \quad \text{et} \quad \bar{y} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}y_j = \sum_{j=1}^{l} f_{\bullet j}y_j$$

   Dans le cas continu,x_i et y_j représentent respectivement le centre des classes de X et Y.
   Nous définissions maintenant la variance de X et la variance de Y comme suit

$$V(X) = \overline{x^2} - (\bar{x})^2 \quad \text{avec} \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i^2 = \sum_{i=1}^{k} f_{i\bullet}x_i^2$$

   et

$$V(Y) = \overline{y^2} - (\bar{y})^2 \quad \text{avec} \quad \overline{y^2} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}y_j^2 = \sum_{j=1}^{l} f_{\bullet j}y_j^2$$

   Les écarts-type de X et de Y sont donnés, respectivement, par

$$\sigma_X := \sqrt{V(X)} \quad \text{et} \quad \sigma_Y := \sqrt{V(Y)}$$

Exemple 5.
Nous calculons $\bar{x}$ et $\bar{y}$ pour l'exemple traité précédemment (l'exercice 5 et 6 TD2). Pour

X	10	15	20	30	35
Y	400	700	800	900	950

Nous avons,

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i \\ &= \frac{1}{5}(10+15+20+30+35) = 22\end{aligned}$$

$$\begin{aligned}V(X) &= \overline{x^2} - (\bar{x})^2 \\ &= \frac{1}{5}(10^2+15^2+20^2+30^2+35^2) - 22^2 = 86\end{aligned}$$

$$\sigma_X = \sqrt{V(X)} = 9.27$$

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}y_j \\ &= \frac{1}{5}(400+700+800+900+950) = 750\end{aligned}$$

$$\begin{aligned}V(Y) &= \overline{y^2} - (\bar{y})^2 \\ &= \frac{1}{5}(400^2+700^2+800^2+900^2+950^2) - 750^2 \\ &= 38000\end{aligned}$$

$$\sigma_Y = \sqrt{V(Y)} = 194.93$$

Exemple 6.
Pour

X \ Y	[4,10[	[10,20[	[20,40]	$n_{i\bullet}$
[3,5[	20	10	0	$n_{1\bullet} = 30$
[5,15[	10	20	10	$n_{2\bullet} = 40$
[15,35]	0	10	20	$n_{3\bullet} = 30$
$n_{\bullet j}$	$n_{\bullet 1} = 30$	$n_{\bullet 2} = 40$	$n_{\bullet 3} = 30$	$n = 100$

Nous avons,

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i \\ &= \frac{1}{100}[(30\times 4)+(40\times 10)+(30\times 25)] = 12.7\end{aligned}$$

$$\begin{aligned}V(X) &= \overline{x^2} - (\bar{x})^2 \\ &= \frac{1}{100}(30\times 4^2+40\times 10^2+30\times 25^2) - 12.7^2 = 71.01\end{aligned}$$

$$\sigma_X = \sqrt{V(X)} = 8.42$$

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}y_j \\ &= \frac{1}{100}(30\times 7+40\times 15+30\times 30) = 17.1\end{aligned}$$

$$\begin{aligned}V(Y) &= \overline{y^2} - (\bar{y})^2 \\ &= \frac{1}{100}(30\times 7^2+40\times 15^2+30\times 30^2) - 17.1^2 = 82.29\end{aligned}$$

$$\sigma_Y = \sqrt{V(Y)} = 9.07$$

2. **Notion de covariance**
   Nous notons par $Cov(X,Y)$ la covariance entre les variables X et Y. La covariance est un paramètre qui donne la variabilité de X par rapport à Y.

La covariance se calcule par l'expression suivante

$$Cov(X,Y)=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}n_{ij}(x_i-\bar{x})(y_j-\bar{y})=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}n_{ij}x_iy_j-\bar{x}\bar{y}$$

Exemple 7.
Pour

X	10	15	20	30	35
Y	400	700	800	900	950

$$\begin{aligned}Cov(X,Y)&=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}n_{ij}x_iy_j-\bar{x}\bar{y}\\&=\frac{1}{5}(10\times400+15\times700+20\times800\\&\quad+30\times900+35\times950)-22\times750\\&=1650.\end{aligned}$$

Exemple 8.
Pour

Y / X	[4,10[	[10,20[	[20,40]	$n_{i\bullet}$
[3,5[	20	10	0	$n_{1\bullet}=30$
[5,15[	10	20	10	$n_{2\bullet}=40$
[15,35]	0	10	20	$n_{3\bullet}=30$
$n_{\bullet j}$	$n_{\bullet1}=30$	$n_{\bullet2}=40$	$n_{\bullet3}=30$	$n=100$

$$\begin{aligned}Cov(X,Y)&=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}n_{ij}x_iy_j-\bar{x}\bar{y}\\&=\frac{1}{100}[(20\times4\times7)+(10\times4\times15)\\&\quad+(0\times4\times30)+(10\times10\times7)\\&\quad+(20\times10\times15)+(10\times10\times30)\\&\quad+(20\times10\times15)+(0\times25\times7)\\&\quad+(10\times25\times15)+(20\times25\times30)]\\&\quad-(12.7\times17.1)\\&=48.93.\end{aligned}$$

Remarque 1.2

La covariance vérifie les propriétés suivante

- $Cov(X,Y)=Cov(Y,X)$,
- $Cov(X,X)=0$,
- $Cov(aX+b,cY+d)=a\,c\,Cov(X,Y)$ où a, b, c et $d\in\mathbb{R}$,
- $|Cov(X,Y)|\leqslant\sigma_X\cdot\sigma_Y$,
- $V(X+Y)=V(X)+V(Y)+2Cov(X,Y$
- Si X et Y sont indépendantes alors (réciproque est fausse) $Cov(X,Y)=0$.

},

3. **Moyennes et variance conditionnelle**

$$\overline{x_j}=\frac{1}{n_{\bullet j}}\sum_{i=1}^{k}n_{ij}x_i\ ,\quad\text{et}\quad V(X_j)=\frac{1}{n_{\bullet j}}\sum_{i=1}^{k}n_{ij}x_i^2-(\overline{x_j})^2,\quad j=1,\ldots,l.$$

$$\overline{y_i}=\frac{1}{n_{i\bullet}}\sum_{j=1}^{l}n_{ij}y_i\ ,\quad\text{et}\quad V(Y_i)=\frac{1}{n_{i\bullet}}\sum_{j=1}^{l}n_{ij}y_j^2-(\overline{y_i})^2,\quad i=1,\ldots,k.$$

Exemple 9.
Distribution conditionnelle de X, pour $y_i \in [10,20[$

X \ [10,20[	n_{i2}	f_{i2}
[3,5[	10	$\frac{10}{40} = 0.25$
[5,15[	20	$\frac{20}{40} = 0.5$
[15,35]	10	$\frac{10}{40} = 0.25$
Total	40	1

$$\begin{aligned}\overline{X_2} &= \frac{1}{n_{\bullet 2}}\sum_{i=1}^{k} n_{i2}x_i\ ,\\ &= 0.25\times 4 + 0.5\times 10 + 0.25\times 25\\ &= 12.25\end{aligned}$$

$$\begin{aligned}V(X_2) &= \frac{1}{n_{\bullet 2}}\sum_{i=1}^{k} n_{i2}x_i^2 - (\overline{X_2})^2,\\ &= 0.25\times 4^2 + 0.5\times 10^2 + 0.25\times 25^2 - 12.25^2\\ &= 60.19\end{aligned}$$

avec 4, 10 et 25 sont les centres respectives des intervalles $[3,5[$, $[5,15[$ et $[15,35]$.

Exemple 10.
Distribution conditionnelle de Y, pour $x_i \in [15,35[$

Y \ [15,35]	[4,10[	[10,20[	[20,40]	Total
$n_{3æ}$	0	10	20	30
f_{3j}	$\frac{0}{30} = 0$	$\frac{10}{30} = 0.33$	$\frac{20}{30} = 0.67$	1

$$\begin{aligned}\overline{Y_3} &= \frac{1}{n_{3\bullet}}\sum_{j=1}^{l} n_{3j}y_i\ ,\\ &= 0\times 7 + 0.33\times 15 + 0.67\times 30\\ &= 25.05\end{aligned}$$

$$\begin{aligned}V(Y_3) &= \frac{1}{n_{3\bullet}}\sum_{j=1}^{l} n_{3j}y_j^2 - (\overline{y_i})^2\\ &= (0\times 7^2 + 0.33\times 15^2 + 0.67\times 30^2)\\ &= 50.20\end{aligned}$$

avec 7,15 et 30 sont les centres respectives des intervalles $[4,10[$, $[10,20[$ et $[20,40]$.

4. Régression et corrélation

Dans le cas où on peut mettre en évidence l'existence d'une relation linéaire significative entre deux caractères quantitatifs continus X et Y (la silhouette du nuage de points est étirée dans une direction), on peut chercher à formaliser la relation moyenne qui unit ces deux variables à l'aide d'une équation de droite qui résume cette relation. Nous appelons cette démarche l'ajustement linéaire.

Définition 1.6

La quantité

$$\rho_{XY} := \frac{Cov(X,Y)}{\sigma_X\,\sigma_Y}$$

s'appelle le coefficient de corrélation.

Proposition 1.1

Le coefficient ρ_{XY} est compris entre $[-1,1]$, ou encore

$$|\rho_{XY}| \leqslant 1$$

Remarque 1.3

- Si ρ_{XY} est proche de -1 ou de $+1$ ceci indique une très forte corrélation entre les deux variables.
- Si $\rho_{XY} = \mp 1$, alors il y a une corrélation maximale entre X et Y.
- $\rho_{XY} = 0$, alors il y a absence de corrélation entre X et Y.

Droite de régression L'idée est de transformer un nuage de point en une droite. Celle-ci doit être la plus proche possible de chacun des points. On cherchera donc à minimiser les écarts entre les points et la droite. Pour cela, on utilise la méthode des moindres carrées. Cette méthode vise à expliquer un nuage de points par une droite qui lie Y à X, c'est à dire,

$$Y = aX + b$$

telle que la distance entre le nuage de points et droite soit minimale. Donc, la droite de régression, qui rend la distance entre elle et les points minimale, est donnée par

$$D(Y/X) : Y = aX + b$$

avec

$$a = \frac{Cov(X,Y)}{V(X)} \quad \text{et} \quad b = \bar{y} - a\bar{x}$$

Ou bien

$$D(X/Y) : X = a'Y + b'$$

avec

$$a' = \frac{Cov(X,Y)}{V(Y)} \quad \text{et} \quad b' = \bar{x} - a\bar{y}$$

Exemple 11.
Pour

X (Dépenses publicitaires en milliers de DA)	10	15	20	30	35
Y (Volume des des ventes en milliers de DA)	400	700	800	900	950

Donc $a = \frac{Cov(X,Y)}{V(X)} = \frac{1650}{86} \approx 19.19$ et $b = \bar{y} - a\bar{x} = 750 - 19.19 \times 22 = 327.82$
La droite ajustement de,

$$Y = 19.19X + 327.82$$

Le coefficient de corrélation

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{1650}{9.27 \times 194.93} = 0.91$$

$\rho_{XY} = 0.91 \approx 1$, donc il y a une corrélation linéaire positive entre X et Y. Le volume des ventes pour un mois où les dépenses publicitaires sont de 40 000 DA($x = 40$) est

$$y = 19.19 \times 40 + 327.82 = 1095.42$$

c'est-à-dire, 1095420 DA.