Université de Tunis — Année Universitaire 2020-2021
Institut Supérieur de Gestion — 1ère LNSE

Statistique Descriptive et Calcul des Probabilités
TD 2 : Les distributions statistiques à deux dimensions

Exercice 1 :

Nous considérons 10 salariés qui sont observés à l'aide de deux variables "âge" et "salaire". Les informations brutes sont données dans le tableau suivant :

Salaire	600	740	750	820	827	890	910	990	995	1075
Age	15	26	20	43	47	37	52	34	50	44

1. Déterminer le tableau de contingence (X : âge, Y : salaire). Pour l'âge et pour le salaire, former respectivement des classes d'amplitudes respectivement de 10 ans et de 100 DT.

Nombre de classes $= \frac{e}{taille\ de\ l'intervalle} = \frac{x_{max}-x_{min}}{taille\ de\ l'intervalle}$

Nombre de classes pour la variable âge : $\frac{52-15}{10} = 3.7 \approx 4$ **classes pour l'âge.**

Nombre de classes pour la variable salaire : $\frac{1075-600}{100} = 4.75 \approx 5$ **classes pour le salaire.**

On regroupe cette série statistique dans le tableau suivant :

En utilisant les hypothèses, nous considérons les classes suivantes: [15, 25[, [25, 35[, [35, 45[, [45, 55[, pour l'âge et [6, 7[, [7, 8[, [8, 9[, [9, 10[, [10, 11[, pour le salaire (*100).

Age/Salaire	[6,7[	[7,8[	[8,9[	[9,10[	[10,11[	$n_{i.}$	$f_{i.}$
[15,25[	1	1	0	0	0	2	0.2
[25,35[	0	1	0	1	0	2	0.2
[35,45[	0	0	2	0	1	3	0.3
[45,55[	0	0	1	2	0	3	0.3
$n_{.j}$	1	2	3	3	1	10	
$f_{.j}$	0.1	0.2	0.3	0.3	0.1		1

2. Calculer et interpréter : f_{12}, n_{45} et f_{33}.
   ***f_{12}=0.1 : 10%* des salariés sont âgés entre 15 et 25 ans et gagnent entre 700 et 800 dinars**
   n_{45}=0 : il n'y a pas de salariées qui sont âgés entre 45 et 55 ans et qui sont payés entre 1000 et 1100 dinars
   ***f_{33}=0.2 : 20%* des salariés sont âgés entre 35 et 45 ans et gagnent entre 800 et 900 dinars**

3. Déterminer les distributions marginales de X et de Y ainsi que les moyennes marginales.

Age (X)	$n_{i.}$	$f_{i.}$	c_i	$f_{i.}c_i$
[15,25[	2	0,2	20	4
[25,35[	2	0,2	30	6
[35,45[	3	0,3	40	12
[45,55[	3	0,3	50	15
Total	10	1		37

Salaire (Y)	$n_{.j}$	$f_{.j}$	c_j	$f_{.j}c_j$
[6,7[	1	0,1	6,5	0,65
[7,8[	2	0,2	7,5	1,5
[8,9[	3	0,3	8,5	2,55
[9,10[	3	0,3	9,5	2,85
[10,11[	1	0,1	10,5	1,05
Total	10	1		8,6

$$\bar{x} = \sum f_{i\bullet} C_i = 37 \text{ ans (Moyenne d'âge)}$$

$$\bar{y} = \sum f_{\bullet j} C_j = 8.6 \times 100 = 860 \text{ dinars (Salaire moyen)}$$

Exercice 2 :

Le tableau suivant donne la répartition de 40 ouvriers d'une entreprise selon le salaire mensuel X (en DT) et l'ancienneté Y (en Années).

X /Y	[0,8[	[8,16[	[16,24[	[24,32[	$n_{i\bullet}$
[200,300[	5	6	1	0	12
[300,400[	2	4	3	3	12
[400,500[	0	2	4	10	16
$n_{\bullet j}$	7	12	8	13	40

1. Calculer et expliquer la signification de $n_{2\bullet}$ et $f_{\bullet 3}$.
   $n_{2\bullet}$ =12 : 12 ouvriers ont un salaire compris entre 300 et 400 Dinars
   $f_{\bullet 3} = \frac{n_{\bullet 3}}{n_{\bullet\bullet}} = \frac{8}{40} = 0,2$: 20% des ouvriers ont de 16 à 24 ans d'ancienneté.
2. Etudier les séries conditionnelles : X/ $Y = y_4$ et Y/ $X = x_2$ et expliquer la signification de $f_{2\backslash Y=y_4}, f_{4\backslash X=x_2}$

X /Y=y4	$n_i/Y=y_4$
[200,300[	0
[300,400[	3
[400,500[	10
Total	13

Y /X=x2	$n_j/X=x_2$
[0,8[	2
[8,16[	4
[16,24[	3
[24,32[	3
Total	12

$f_{2\backslash Y=y_4} = \frac{3}{13} = 0.23 : 23\%$ des ouvriers ayant une ancienneté au travail entre 24 et 32 ans, reçoivent un salaire entre 300 et 400 Dinars

$f_{4\backslash X=x_2} = \frac{3}{12} = 0.25 : 25\%$ des ouvriers ayant des salaires entre 300 et 400 Dinars, ont une ancienneté de 24 à 32 ans.

3. Calculer la moyenne conditionnelle de la distribution : Y/ $X = x_2$. Interpréter.

Y/X=x2	n.j/X=x2	fj/X=x2	cj	fj/x2. cj
[0,8[	2	0,17	4	0,68
[8,16[	4	0,33	12	3,96
[16,24[	3	0,25	20	5,00
[24,32[	3	0,25	28	7,00
Total	12	1		16,64

$\overline{Y}_{/\ X=x_2} = 16.64$ ans : le taux d'ancienneté moyen est de 16.64 ans pour les ouvriers qui ont un salaire entre 300 et 400 dinars.

4. Calculer la covariance entre l'ancienneté et le revenu. Commenter.

$$cov(X,Y) = \sum_i \sum_j f_{ij} C_i C_j - \overline{xy}$$

Calcul de $\overline{x}$ et $\overline{y}$

X	ni.	fi.	ci	fi. Ci
[200,300[	12	0,3	250	75
[300,400[	12	0,3	350	105
[400,500[	16	0,4	450	180
Total	40	1		360

Y	n.j	f.j	cj	f.j cj
[0,8[	7	0,175	4	0,7
[8,16[	12	0,3	12	3,6
[16,24[	8	0,2	20	4
[24,32[	13	0,325	28	9,1
Total	40	1		17,4

Donc : $\overline{x} = 360\ dinars\ \ et\ \ \overline{y}$ =17,4 ans

On utilise la méthode de coin dans le tableau pour calculer $\sum_i \sum_j f_{ij} C_i C_j$:

X\Y	4	12	20	28	Total
250	0.125 125	0.15 450	0.025 125	0 0	700
350	0.05 70	0.1 420	0.075 525	0.075 735	1750
450	0 0	0.05 270	0.1 900	0.25 3150	4320
Total	195	1140	1550	3885	$\sum_i \sum_j f_{ij} C_i C_j$ =6770

$cov(X,Y) = \sum_i \sum_j f_{ij} C_i C_j - \overline{xy} = 6770 - (360 \times 17,4) = 506 > 0$

Corrélation positive entre X et Y. Les 2 variables varient dans le même sens.

5. Déterminer si les variables *X* et *Y* sont indépendants.

Remarque : d'après la question précédente comme cov (X,Y) est différente de 0, alors X et Y ne sont pas indépendantes.

En effet ,

Afin de montrer que *X* et *Y* sont indépendants, il faut que cette condition soit satisfaite :

$$f_{ij}=f_{i.} \times f_{.j}$$

Prenons au hasard : $f_{13}=\frac{1}{40} = 0.025$

D'après les distributions marginales on a : $f_{1.}=\frac{12}{40} = 0.3$ **et** $f_{.3}=\frac{8}{40} = 0.2$

D'où : $f_{1.} \times f_{.3} = 0.3 \times 0.2 = 0.06 \neq f_{13}$

Donc *X* et *Y* ne sont pas indépendants

Exercice 3 :

L'entreprise « CONFORT » est spécialisée dans la fabrication et la commercialisation de plusieurs modèles de chaussures, les données qui suivent indiquent le nombre de modèle de chaussures fabriqués par « CONFORT », son bénéfice au cours des cinq derniers semestres :

Semestres	Bénéfices (en 1000 dinars) Y	Le nombre de modèles de chaussures X	x_i^2	y_i^2	$x_i y_i$
1	50	10	100	2500	500
2	60	12	144	3600	720
3	70	18	324	4900	1260
4	90	24	576	8100	2160
5	130	36	1296	16900	4680
Total	400	100	2440	36000	9320

1. Calculer les moyennes et variances. Commenter.

$$\overline{X} = \frac{\sum x_i}{n} = \frac{10+12+18+24+36}{5} = \frac{100}{5} = 20$$

$$\overline{Y} = \frac{\sum y_i}{n} = \frac{50+60+70+90+130}{5} = \frac{400}{5} = 80$$

$$V(X) = \frac{1}{n}\sum x_i^2 - \overline{X}^2 = \frac{2440}{5} - (20)^2 = 488 - 400 = 88$$

$$V(Y) = \frac{1}{n}\sum y_i^2 - \overline{Y}^2 = \frac{36000}{5} - (80)^2 = 7200 - 6400 = 800$$

$\overline{X} = 20$: Le nombre moyen de modèle de chaussures fabriqués est égal à 20 modèles.

$\overline{Y} = 80$: L'entreprise CONFORT réalise un bénéfice moyen de 80000 dinars.

$$CV(x) = \frac{\sqrt{V(x)}}{\overline{x}} = \frac{\sqrt{88}}{20} = 0.47$$

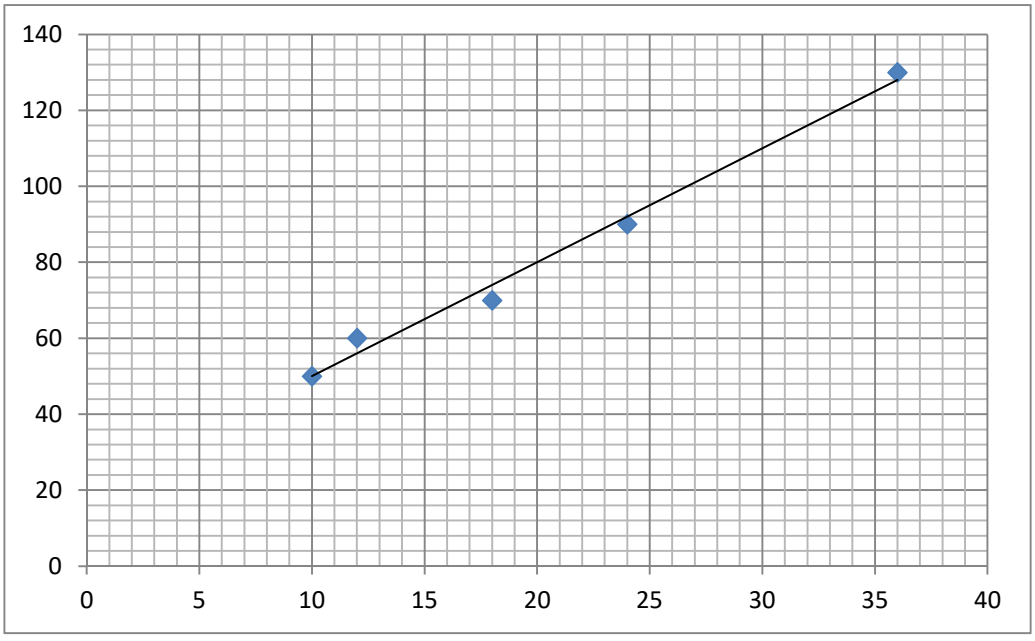
$$CV(y) = \frac{\sqrt{V(y)}}{\overline{y}} = \frac{\sqrt{800}}{80} = 0.35$$

Les 2 distributions sont fortement dispersées autour de leurs valeurs moyennes

2. Discuter la validité du modèle d'ajustement linéaire : $y = a\,X + b$

1ère méthode :

Pour discuter la validité du modèle d'ajustement linéaire, on doit représenter le nuage de point pour ces deux variables.



Le nuage de point indique une relation linéaire positive entre X (le nombre de modèle de chaussures fabriqués) et Y (les bénéfices réalisés).

On peut donc calculer le coefficient de corrélation linéaire :

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

$$Cov(X,Y) = \frac{1}{n}\sum x_i y_i - \overline{X}\,\overline{Y} = \frac{9320}{5} - (20 \times 80) = 1864 - 1600 = 264$$

$$\rho_{X,Y} = r(X,Y) = \frac{Cov\,(X,Y)}{\sigma_X\ \sigma_Y} = \frac{264}{\sqrt{(88\times800)}} = \frac{264}{\sqrt{70400}} = \frac{264}{265{,}33} = 0{,}995$$

2ème méthode:

Il faut calculer le coefficient de corrélation linéaire :

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

$$Cov(X,Y) = \frac{1}{n}\sum x_i y_i - \overline{X}\,\overline{Y} = \frac{9320}{5} - (20 \times 80) = 1864 - 1600 = 264$$

$$\rho_{X,Y} = r(X,Y) = \frac{Cov\,(X,Y)}{\sigma_X\ \sigma_Y} = \frac{264}{\sqrt{(88\times800)}} = \frac{264}{\sqrt{70400}} = \frac{264}{265{,}33} = 0{,}995$$

L'interprétation: le coefficient de corrélation linéaire tend vers 1 donc il y a une relation linéaire positive entre X et Y.

3. Utiliser la méthode des moindres carrés pour estimer les paramètres a et b du modèle précédent.

$$y_i = a\,x_i + b$$

$$\hat{a} = \frac{Cov\,(X,Y)}{V(X)}$$

On sait que $Cov(X,Y) = 264$

$$\hat{a} = \frac{Cov\,(X,Y)}{V(X)} = \frac{264}{88} = 3$$

$$\hat{b} = \overline{Y} - \hat{a}\,\overline{X} = 80 - (3 \times 20) = 80 - 60 = 20$$

$$y_i = \hat{a}\,x_i + \hat{b}$$

La droite de régression est alors : $y_i = 3\,x_i + 20$

4. Evaluer numériquement la qualité de l'ajustement linéaire précédent.

On calcule le coefficient de détermination :

$$R^2 = \frac{SCE}{SCT} = \frac{\frac{1}{N}\sum_{i=1}^{N} \hat{y}_i^{\,2} - \bar{y}^2}{V(y)}$$

Sachant que $\hat{y}_i = 3\,x_i + 20$

x_i	y_i	$\hat{y}_i$	$\hat{y}_i^2$
10	50	50	2500
12	60	56	3136
18	70	74	5476
24	90	92	8464
36	130	128	16384
Total			35960

$$R^2 = \frac{\frac{1}{N}\sum_{i=1}^{N} \hat{y}_i^{\,2} - \bar{y}^2}{V(y)} = \frac{\frac{1}{5} \times 35960 - (80)^2}{800} = \frac{792}{800} = 0,99$$

R^2 tend vers 1 donc la qualité d'ajustement linéaire entre le nombre de modèle de chaussures fabriqués et les bénéfices réalisés est bonne voire même parfaite.

5. Calculer le bénéfice prévisionnel de l'entreprise si elle décide de fabriquer 40 modèles de chaussures.

$$y_i = 3\,x_i + 20 = (3 \times 40) + 20 = 120 + 20 = 140$$

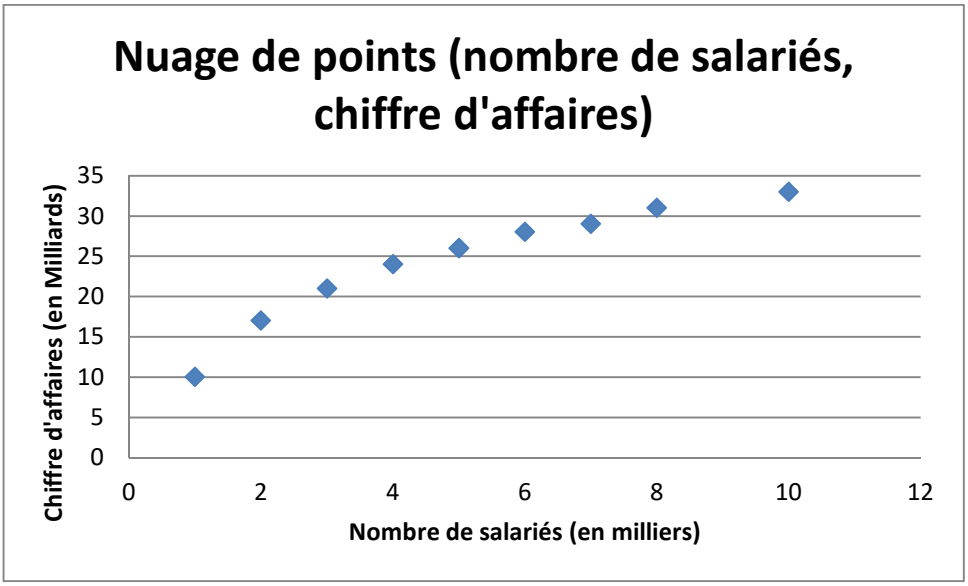
Si l'entreprise décide de fabriquer 40 modèles de chaussures, le bénéfice prévisionnel sera donc égal à 140000 dinars.

Exercice 4

On dispose pour un secteur industriel donné et sur une période de 10 années de la série du nombre de salariés x (en milliers) et du chiffre d'affaires y (en dizaine de milliards)

Années	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
X	8	6	5	7	2	5	3	4	10	1
Y	31	28	26	29	17	26	21	24	33	10

1. Représenter le nuage de points (x_i, y_i).



2. Selon ce graphe, quel type de relation peut-on suggérer entre X et Y.
   D'après l'allure du nuage, on peut conclure à une relation logarithmique entre X et Y.
3. Estimez cette relation

$$y = a\ln(x) + b$$

On pose: $Z = \ln(x)$
On obtient: $y = aZ + b$: Equation d'une droite.
Le nuage (Z, y) est donc aligné avec:

$$\begin{cases} \hat{a} = \dfrac{cov(Z,y)}{V(Z)} \\ \hat{b} = \bar{y} - \hat{a}\bar{Z} \end{cases}$$

Observation	X	Y	log X=Z	Z²	Y²	ZY
1	8	31	2,08	4,32	961	64,46
2	6	28	1,79	3,21	784	50,17
3	5	26	1,61	2,59	676	41,85
4	7	29	1,95	3,79	841	56,43
5	2	17	0,69	0,48	289	11,78
6	5	26	1,61	2,59	676	41,85
7	3	21	1,10	1,21	441	23,07
8	4	24	1,39	1,92	576	33,27
9	10	33	2,30	5,30	1089	75,99
10	1	10	0,00	0,00	100	0,00
Total		245	14,52	25,41	6433	398,86
Moyennes		24,5	1,45	2,54	643,3	39,89
Variances		43,05	0,43			

$$Cov(Z,Y) = \frac{1}{n}\sum z_i y_i - \overline{Z}\,\overline{Y} = 39.89 - (1.45 \times 24.5) = 4.365$$

$$V(Z) = \frac{1}{n}\sum {z_i}^2 - \overline{Z}^2 = 2.54 - (1.45)^2 = 0.43$$

$$\begin{cases} \hat{a} = \dfrac{cov(Z,y)}{V(Z)} = \dfrac{4.365}{0.43} = 10.15 \\ \hat{b} = \bar{y} - \hat{a}\bar{Z} = 24.5 - (10.15 \times 1.45) = 9.78 \end{cases}$$

$$y = 10.15\ln(x) + 9.78$$

4. En 2020, l'industriel prévoit de comptabiliser un nombre total de salariés égal à 12000. Quel sera le chiffre d'affaire prévisionnel

On utilise l'équation d'ajustement :

$$y = 10.15\ln(x) + 9.78$$

Avec x=12

$$y = 10.15\ln(12) + 9.78 = 35\ \textit{Milliars de Dinars}$$