INSTITUT Mines-Télécom

Statistique pour ingénieur

Thème 0 : Éléments de correction des exercices

F. Delacroix & M. Lecomte, 20 octobre 2016

Exercice 1 : Quelques calculs de statistique descriptive

Dans une entreprise, on a recensé les salariés par tranche d'âge et par sexe. Les résultats sont donnés dans le tableau ci-dessous.

Tranche d'âge	Hommes	Femmes
Moins de 20 ans	32	51
20 — 30	1309	2118
30 — 40	1902	3025
40 — 50	1730	2330
50 — 60	1468	1624
Plus de 60 ans	114	131

1. Quelles sont les caractéristiques étudiées ? Préciser s'il s'agit de caractères discrets ou continus.

Solution:

On étudie l'âge (caractère quantitatif continu discrétisé en classes) et le sexe (caractère qualitatif discret) des individus.

2. Quelle est la proportion de salariés dans les tranches d'âge inférieures ou égales à 40 ans ? Mêmes questions pour les hommes et femmes séparément. Que peut-on en conclure ?

Solution:

Complétons le tableau de l'énoncé avec des données calculées pour les hommes, les femmes, et l'ensemble des employés à partir d'un tableur.

		Population totale					
Représentant x_i	*Classe d'âge* $[a_i,b_i]$	*Effectif* n_i	*Fréquence* f_i	*Effectif cumulé*	*Fréquence cumulée*	$n_i x_i$	$n_i x_i^2$
19	*moins de 20*	*83*	*0,52%*	*83*	*0,52%*	*1577*	*29963*
25	[20,30[	*3427*	*21,64%*	*3510*	*22,17%*	*85675*	*2141875*
35	[30,40[	*4927*	*31,12%*	*8437*	*53,28%*	*172445*	*6035575*
45	[40,50[	*4060*	*25,64%*	*12497*	*78,93%*	*182700*	*8221500*
55	[50,60[	*3092*	*19,53%*	*15589*	*98,45%*	*170060*	*9353300*
62	*plus de 60*	*245*	*1,55%*	*15834*	*100,00%*	*15190*	*941780*
	Somme	*15834*	*100%*			*627647*	*26723993*
		Hommes					
Représentant x_i	*Classe d'âge* $[a_i,b_i]$	*Effectif* n_i	*Fréquence* f_i	*Effectif cumulé*	*Fréquence cumulée*	$n_i x_i$	$n_i x_i^2$
19	*moins de 20*	*32*	*0,49%*	*32*	*0,49%*	*608*	*11552*
25	[20,30[	*1309*	*19,97%*	*1341*	*20,46%*	*32725*	*818125*
35	[30,40[	*1902*	*29,02%*	*3243*	*49,47%*	*66570*	*2329950*
45	[40,50[	*1730*	*26,39%*	*4973*	*75,87%*	*77850*	*3503250*
55	[50,60[	*1468*	*22,40%*	*6441*	*98,26%*	*80740*	*4440700*
62	*plus de 60*	*114*	*1,74%*	*6555*	*100,00%*	*7068*	*438216*
	Somme	*6555*	*100%*			*265561*	*11541793*

		Femmes					
Représentant x_i	*Classe d'âge* $[a_i,b_i]$	*Effectif* n_i	*Fréquence* f_i	*Effectif cumulé*	*Fréquence cumulée*	$n_i x_i$	$n_i x_i^2$
19	*moins de 20*	*51*	*0,55%*	*51*	*0,55%*	*969*	*18411*
25	[20,30[	*2118*	*22,83%*	*2169*	*23,38%*	*52950*	*1323750*
35	[30,40[	*3025*	*32,60%*	*5194*	*55,98%*	*105875*	*3705625*
45	[40,50[	*2330*	*25,11%*	*7524*	*81,09%*	*104850*	*4718250*
55	[50,60[	*1624*	*17,50%*	*9148*	*98,59%*	*89320*	*4912600*
62	*plus de 60*	*131*	*1,41%*	*9279*	*100,00%*	*8122*	*503564*
	Somme	*9279*	*100%*			*362086*	*15182200*

Les représentants x_i des différentes classes d'âges ont été choisis au centre des différentes classes, à l'exception des deux classes extrêmes pour lesquelles les choix de 19 ans et 62 ans comme représentants comportent une part d'arbitraire. Les calculs possibles à ce niveau ne sont qu'approximatifs étant donnée la perte de renseignement due à l'agglomération des données sous forme de classes d'âge.

On peut lire dans ces tableaux (colonne fréquence cumulée) que 53,28% des employés ont au plus 40 ans. Chez les hommes, 49,47% ont au plus 40 ans et chez les femmes elles sont 55,98% à figurer dans ces classes d'âge.

Ce qu'on peut en conclure est que la distribution observée des âges n'est pas la même selon le genre des employés. Notamment, plus de la moitié des femmes employées ont moins de 40 ans.

3. Déterminer l'âge moyen, l'âge médian, les quartiles et l'écart-type pour les hommes. Mêmes questions pour les femmes.

Solution:

En utilisant les x_i comme représentants des classes, la moyenne s'écrit

$$\overline{x} = \frac{1}{n}\sum_i n_i x_i \quad \text{où} \quad n = \sum_i n_i.$$

De même, l'écart-type est donné par

$$s = \sqrt{\frac{1}{n}\sum_i n_i x_i^2 - \overline{x}^2}.$$

La détermination de fractiles tels que la médiane ou les quartiles est plus délicate. On va procéder par interpolation linéaire, en deux étapes :

1. *identification de la classe d'âge $[a_i,b_i[$ à laquelle le fractile (noté q) cherché appartient, par observation des fréquences cumulées,*
2. *détermination du fractile q par la formule d'égalité des coefficients directeurs :*

$$\frac{f(q)-f(a_i)}{q-a_i} = \frac{f(b_i)-f(a_i)}{b_i-a_i} \quad \text{soit} \quad q = a_i + (b_i-a_i)\frac{f(q)-f(a_i)}{f(b_i)-f(a_i)}$$

où $f()$ désigne la fréquence cumulée.

Par exemple pour la détermination de la médiane Me, on a $f(Me) = 50\%$. Pour les hommes, la classe médiane est [40,50[. Alors

$$Me = 40 + 10 \times \frac{50 - 49{,}47}{75{,}87 - 49{,}47} \simeq 40{,}20.$$

Les résultats sont résumés dans le tableau suivant.

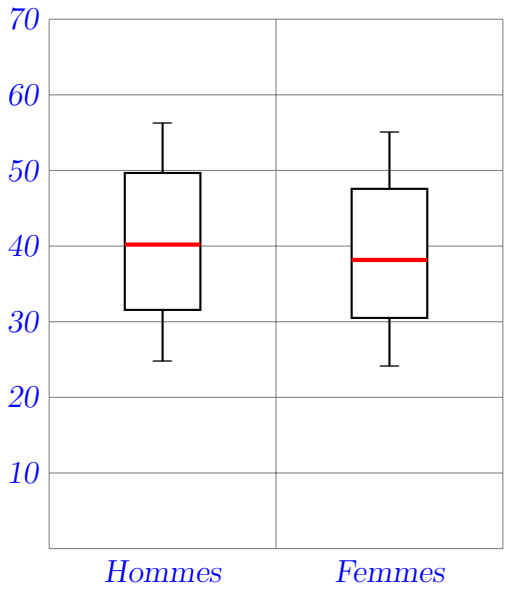
	Hommes		*Femmes*	
	Classe	*Valeur*	*Classe*	*Valeur*
Q1	[30,40[	*31,56*	[30,40[	*30,50*
Me	[40,50[	*40,20*	[30,40[	*38,17*
Q3	[40,50[	*49,67*	[40,50[	*47,57*
$\overline{x}$	*40,51*		*39,02*	
s	*10,93*		*10,65*	

4. Comparer les deux sous-populations (hommes et femmes) à l'aide de boîtes à moustaches.

Solution:

Afin de «couper les moustaches», on détermine comme précédemment les 1er et 9ème déciles des hommes et des femmes :

	Hommes	*Femmes*
D1	*24,76*	*24,14*
D9	*56,31*	*55,09*



Exercice 2 : Étude d'une corrélation

On a relevé la taille (X, exprimée en cm) et le poids (Y, exprimé en kg) d'une population humaine donnée. Les résultats sont regroupés en classes et les effectifs conjoints notés dans le tableau de contingence suivant.

X \ Y	]50,60]	]60,70]	]70,80]	]80,90]
]150,155]	24	11	2	0
]155,160]	22	27	10	1
]160,165]	13	30	14	3
]165,170]	3	6	15	7
]170,180]	0	2	3	7

1. Déterminer les lois marginales.

Solution:

En calculant pour chaque cellule le rapport entre l'effectif de la cellule et l'effectif total (qui vaut 200), on détermine la loi conjointe de X et Y. En sommant ensuite les cellules d'une ligne ou d'une colonne on obtient les lois marginales. On obtient les résultats suivants.

X \ Y	]50,60]	]60,70]	]70,80]	]80,90]	*Loi marginale en X*
]150,155]	12%	5,5%	1%	0%	18,5%
]155,160]	11%	13,5%	5%	0,5%	30%
]160,165]	6,5%	15%	7%	1,5%	30%
]165,170]	1,5%	3%	7,5%	3,5%	15,5%
]170,180]	0%	1%	1,5%	3,5%	6%
Loi marginale en Y	31%	38%	22%	9%	

2. En choisissant les centres des classes comme représentants, calculer :
- la taille moyenne de cette population,
- son poids moyen,
- les écart-types correspondants,
- la covariance de X et de Y,
- le coefficient de corrélation linéaire.

Solution:

À l'aide des lois marginales et en choisissant le centre des classes comme représentant, on peut directement calculer la taille moyenne :

$$\begin{aligned}\overline{x} &= 152{,}5 \times 0{,}185 + 157{,}5 \times 0{,}30 + 162{,}5 \times 0{,}30 + 167{,}5 \times 0{,}155 + 175 \times 0{,}06 \\ &= 160{,}675\, cm.\end{aligned}$$

De même, le poids moyen est :

$$\begin{aligned}\overline{y} &= 55 \times 0{,}31 + 65 \times 0{,}38 + 75 \times 0{,}22 + 85 \times 0{,}09 \\ &= 65{,}9\, kg.\end{aligned}$$

Pour calculer les variances de X et Y, on utilise la formule suivante :

$$\mathbb{V}(X) = \sum_i f_i . x_i^2 - \overline{x}^2$$

où les x_i sont les représentants des classes et f_i. la fréquence marginale en X :

$$\begin{aligned}\mathbb{V}(X) &= 152{,}5^2 \times 0{,}185 + 157{,}5^2 \times 0{,}30 + 162{,}5^2 \times 0{,}30 + 167{,}5^2 \times 0{,}155 + 175^2 \times 0{,}06 \\ &\quad - 160{,}675^2 \\ &= 35{,}919375 \quad \textit{donc} \quad \sigma(X) \simeq 5{,}99\, cm.\end{aligned}$$

De même,

$$\begin{aligned}\mathbb{V}(Y) &= 55^2 \times 0{,}31 + 65^2 \times 0{,}38 + 75^2 \times 0{,}22 + 85^2 \times 0{,}09 - 65{,}9^2 \\ &= 88{,}19 \quad \textit{donc} \quad \sigma(Y) \simeq 9{,}39\, kg.\end{aligned}$$

Pour le calcul de la covariance, on utilise la formule «moyenne des produits moins produit des moyennes» :

$$\begin{aligned}\mathbb{C}\text{ov}\,(X,Y) &= \sum_{i,j} f_{ij} x_i x_j - \overline{x}\,\overline{y}.\\ &= 55 \times 152{,}5 \times 0{,}12 + 55 \times 157{,}5 \times 0{,}11 + \cdots - 160{,}675 \times 65{,}9\\ &= 32{,}2675 \text{ (unité : } cm.kg)\end{aligned}$$

Le coefficient de corrélation linéaire est donc

$$r(X,Y) = \frac{\mathbb{C}\text{ov}\,(X,Y)}{\sigma(X)\,\sigma(Y)} \simeq \frac{32{,}2675}{5{,}99 \times 9{,}39} \simeq 0{,}57.$$

3. Déterminer la loi conditionnelle de Y sachant $\{150 < X \leqslant 155\}$. Calculer la moyenne conditionnelle de Y sachant sachant $\{150 < X \leqslant 155\}$.

4. Mêmes questions avec les autres classes de la variable X.

Solution:

À l'aide d'un tableur, on calcule les fréquences conditionnelles $f_{j/i} = \frac{f_{ij}}{f_{i\cdot}}$, qui représentent pour chaque classe C_i (de X de représentant x_i) la fréquence de chaque classe de Y (représentée par y_j) sachant que $X = x_i$.

X \ Y	]50,60]	]60,70]	]70,80]	]80,90]	$\overline{y\|X \in C_i}$
]150,155]	65,86%	29,73%	5,41%	0%	59,05 kg
]155,160]	37,67%	45%	16,67%	1,67%	63,33 kg
]160,165]	21,67%	50%	23,33%	5%	66,17 kg
]165,170]	9,68%	19,35%	48,39%	22,58%	73,39 kg
]170,180]	0%	16,67%	25%	58,33%	79,17 kg

La moyenne conditionnelle de y sachant $X \in C_i$ s'écrit alors

$$\overline{y|X \in C_i} = \sum_j y_j f_{j/i}$$

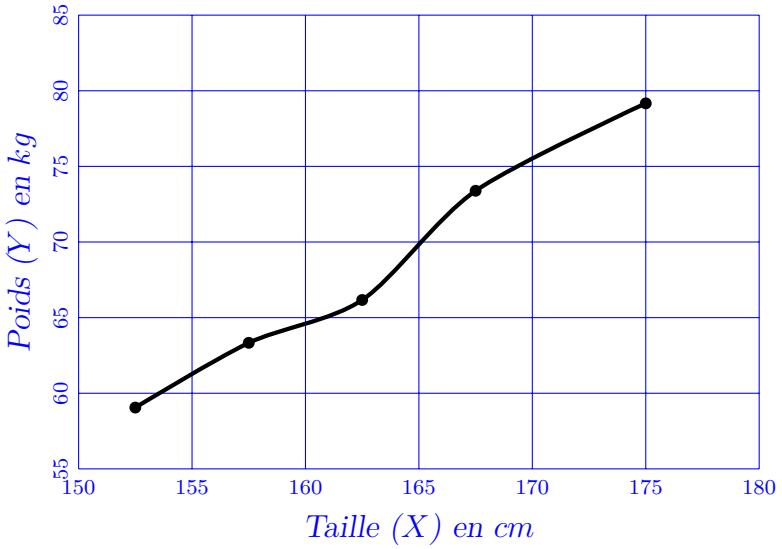
Les résultats ont été introduits dans la dernière colonne du tableau précédent.

5. Représenter graphiquement les points de coordonnées $\left(x_i, \overline{y_i|X \in C_i}\right)$ où :

- C_i désigne l'une des classes de taille,
- x_i est le centre de la classe C_i,
- $\overline{y_i|X \in C_i}$ est la moyenne conditionnelle de Y sachant $\{X \in C_i\}$.

Construire une courbe de régression de Y en X, c'est-à-dire une courbe passant par les points précédemment représentés. Conclure.

Solution:



Il semble, d'après cette représentation graphique, qu'il existe bien une relation linéaire entre poids et taille. Cela mérite toutefois d'être confirmé par une analyse approndie, ce que nous étudierons notamment lors du thème 4 de ce MOOC, consacré à la régression linéaire.