Contrôle Terminal, 2ème session – 26/06/2024 – 2h00
Statistiques – L1 de psychologie

Réponses

Veuillez rendre ce sujet et votre copie.

Numéro d'anonymat :

Le formulaire et la calculatrice sont autorisés. Merci d'indiquer dans la case Numéro d'anonymat*, ci-dessus, un numéro que vous reporterez aussi sur votre copie. Vous rendrez l'énoncé et votre copie, et pouvez soit répondre sur l'énoncé, soit détailler certaines questions sur la copie si vous avez besoin de plus de place. Le soin de la rédaction entrera en compte dans la notation mais dans les questions où des détails ne sont pas explicitement demandés, un résultat correct, donné sans détails de calcul sera accepté.*

Exercice 1 : Genre des personnes souffrant de troubles de l'humeur

Il y a en France environ 10 millions de personnes souffrant de troubles de l'humeur. On cherche dans cet exercice à déterminer si ces troubles de l'humeur touchent de manière égale les femmes et les hommes.

1. (a) On considère un échantillon aléatoire de 64 français·es, qui souffrent de troubles de l'humeur. On désigne par X le nombre de femmes au sein de cet échantillon. Quelle serait la loi du nombre X s'il y avait autant d'hommes que de femmes parmi l'ensemble des français·es atteint·e·s de troubles de l'humeur ?

$X \sim \mathcal{B}(64\,;\,0{,}5)$

(b) Justifier que l'on pourrait alors approcher X par la loi $\mathcal{N}(32\,;\,4)$.

$n\,p(1-p) = 64 \times 0{,}5\ (1-0{,}5) = 16 > 10$,
donc on peut approcher $\mathcal{B}(64\ ;\ 0{,}5) \approx \mathcal{N}(64 \times 0{,}5\ ;\ \sqrt{64 \times 0{,}5\ (1-0{,}5)}) = \mathcal{N}(32\ ;\ 4)$, mais il est nécessaire de faire de correction de continuité.

(c) Que vaudrait alors la probabilité $\mathbb{P}[X \geqslant 41]$. *On demande d'effectuer une correction de continuité*

$\mathbb{P}\,[X \geqslant 41] = \mathbb{P}\,[X \geqslant 40{,}5] \simeq 0{,}016\,79$

(d) Si l'on renouvelle les questions 1a, 1b et 1c en supposant désormais que l'échantillon est choisi sans remise, les résultats seraient-ils différents ?

Une brève justification est demandée en plus des résultats

La taille de l'échantillon (64 individus) est beaucoup plus petite que le nombre de personnes parmi lesquelles on pioche (environ 10 million de français·es atteint·e·s de troubles de l'humeur), largement plus de dix fois plus petite. En conséquence, les résultats "avec" et "sans" remise sont extrêmement proches, on obtiendrait les mêmes résultats que ci-avant.

2. En interrogeant plusieurs psychologues sur leur patientèle, on parvient à savoir qu'ils/elles ont justement 64 patient·es souffrants de troubles de l'humeur, à savoir 48 femmes et 16 hommes.

   (a) Au vu des questions précédentes, vous semble-t-il vraisemblable qu'il y ait autant d'hommes que de femmes parmi l'ensemble des français·es atteint·e·s de troubles de l'humeur ?

   S'il y avait autant d'hommes que de femmes parmi l'ensemble des français·es atteint·e·s de troubles de l'humeur, alors il y aurait peu de chances (environ 1,679% d'après la question 1c) d'avoir au moins 41 femmes dans un échantillon de 64 personnes atteintes de troubles alimentaires. Or dans cette patientèle il y a au moins 41 femmes, donc il est peu vraisemblable qu'il y ait autant d'hommes que de femmes parmi l'ensemble des français atteints de troubles alimentaires.

   (b) Estimer, pour la confiance 90%, la proportion de femmes, parmi l'ensemble des français·es atteint·e·s de troubles de l'humeur.

   On déterminera, par un court calcul, un intervalle de confiance correspondant à la confiance 90%.

   On a $n\,p_e(1-p_e) = 64 \times \frac{48}{64}\left(1-\frac{48}{64}\right) = 12 > 10$, donc on peut utiliser la procédure du formulaire pour estimer la proportion p.
   On a $F(1{,}645) \simeq 0{,}95$ d'où $z_\alpha \simeq 1{,}645$
   d'où $a_\alpha = z_\alpha\sqrt{\frac{p_e(1-p_e)}{n}} = 1{,}645 \times \sqrt{\frac{\frac{48}{64}(1-\frac{48}{64})}{64}} \simeq 0{,}089$ et $p_e - a_\alpha \simeq \frac{48}{64} - 0{,}089 = 0{,}661$ et $p_e + a_\alpha \simeq \frac{48}{64} + 0{,}089 = 0{,}839$.
   On estime donc que p est dans l'intervalle $[0{,}661; 0{,}839]$ avec la confiance $c = 0{,}9$.

Exercice 2 : Impact socioprofessionnel de maladies dans l'enfance

Madame Dupuis, qui est médecin généraliste, dispose des dossiers médicaux de ses patient·e·s, et cherche à savoir si leurs revenus à l'âge adulte dépendent du fait qu'ils/elles aient été hospitalisé·e·s au cours de leur enfance. On pourrait en effet s'attendre à ce que les hospitalisations nuisent à la réussite des études, et influent donc directement sur leur revenu à l'âge adulte.

Pour cela, elle divise sa patientèle en deux groupes : le groupe A est constitué d'adultes qui n'ont jamais été hospitalisé·e·s pendant plus d'un mois au cours de leur enfance, et le groupe B est constitué d'adultes qui, au cours de leur enfance, ont été hospitalisé·e·s au moins une fois pendant plus d'un mois.

Afin de calculer des intervalles de confiance, on pourra supposer dans cet exercice que les revenus suivent une loi normale.

1. Cette médecin commence par interroger 26 patient·e·s choisi·e·s au hasard (avec remise) au sein du groupe B. elle obtient les revenus annuels suivants : 5 488 €, 15 186 €, 7 055 €, 25 347 €, 24 183 €, 24 536 €, 10 048 €, 23 111 €, 14 318 €, 9 377 €, 19 302 €, 14 164 €, 14 832 €, 16 236 €, 30 084 €, 13 974 €, 18 623 €, 11 528 €, 18 032 €, 8 661 €, 22 383 €, 18 107 €, 21 228 €, 21 524 €, 25 155 € et 37 926 €.

   (a) Déterminer la moyenne et l'écart type du revenu annuel au sein de cet échantillon.
   On demande dans cette question de justifier la réponse par un court calcul.

   moyenne : $m(X) = \frac{\sum x_i}{n} = \frac{5\,488+15\,186+7\,055+\cdots+37\,926}{26} = \frac{470\,408}{26} \simeq 18\,092{,}62\,€$
   $m(X^2) = \frac{\sum x_i^2}{n} = \frac{5\,488^2+15\,186^2+7\,055^2+\cdots+37\,926^2}{26} = \frac{9\,918\,729\,646}{26}$
   $Var(X) = m(X^2) - m(X)^2 = \frac{9\,918\,729\,646}{26} - \left(\frac{470\,408}{26}\right)^2 = 54\,146\,870$
   Écart-type : $s(X) = \sqrt{Var(X)} \simeq 7\,358{,}46\,€$

(b) Déterminer aussi le revenu médian au sein de cet échantillon.

La médiane est la valeur numéro $\frac{26+1}{2} = 13{,}5$ (en ordonnant par ordre croissant), ou plutôt le milieu entre les valeurs numéro 13 et 14. C'est donc le milieu entre 18032 et 18107 c'est-à-dire $\frac{18\,032+18\,107}{2} = 18\,069{,}5\,€$.

(c) Si l'on utilise cet échantillon pour estimer le revenu moyen au sein du groupe B, montrer que pour la confiance 95%, on obtient l'intervalle [15 062 ; 21 124].

Comme $n = 26 \leqslant 30$, on cherche t_α à partir de la table inverse de Student avec $p = \frac{\alpha}{2} = 0{,}025$ et $n - 1 = 25$ degrés de liberté (ddl)
On lit $t_\alpha \simeq 2{,}0595$ d'où $a_\alpha = t_\alpha \frac{s_e}{\sqrt{n-1}} \simeq 2{,}0595 \times \frac{7\,358{,}46}{\sqrt{26-1}} \simeq 3\,030{,}9$.
On estime donc que μ est dans l'intervalle $[18\,092{,}62 - 3\,030{,}9\,;\, 18\,092{,}62 + 3\,030{,}9] \simeq [15\,062\,;\, 21\,124]$ avec la confiance $c = 0{,}95$

2. Elle interroge ensuite 250 patient·e·s choisi·e·s au hasard (avec remise) au sein du groupe A. Elle constate qu'au sein de cet échantillon de 250 patients, le revenu annuel moyen est de 21 792,84 €, avec un écart type de 13 027,38 €.
Estimer, avec la confiance 95%, le revenu moyen de l'ensemble des patient·e·s du groupe A.
Répondre en calculant un intervalle de confiance.

Comme n= 250 >30, on cherche z_α tel que $F(z_\alpha) = \frac{0{,}95+1}{2} = 0{,}975$
On lit sur la table inverse que pour la confiance 0,95 on a $z_\alpha \simeq 1{,}96$
d'où $a_\alpha = z_\alpha \frac{s_e}{\sqrt{n-1}} \simeq 1{,}96 \times \frac{13\,027{,}38}{\sqrt{250-1}} \simeq 1618$
On estime donc que μ est dans l'intervalle $[21\,792{,}84 - 1\,618\,;\, 21\,792{,}84 + 1\,618] \simeq [20\,175\,;\, 23\,411]$, avec la confiance $c = 0{,}95$.

3. Peut-on conclure que, comme s'y attendait le docteur Dupuis, les patient·e·s du groupe B aient en moyenne un revenu plus faible que celles et ceux du groupe A ? *(on répondra avec la confiance 95%)*

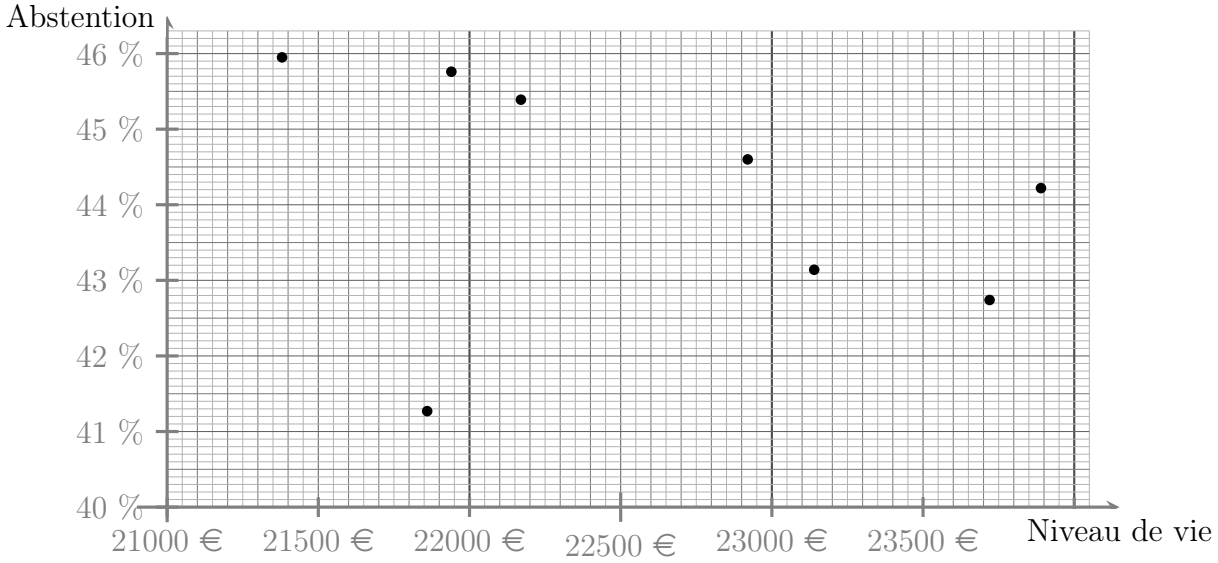
Les intervalles calculés se chevauchent, ils ne permettent donc pas d'affirmer, avec la confiance 95%, quel groupe de patient·e·s à le revenu moyen le plus élevé. Ceci ne suffit donc pas à confirmer l'hypothèse du docteur Dupuis.

Exercice 3 : Niveau de vie et taux d'abstention

Dans cet exercice, on cherche à mettre en évidence un lien entre le revenu des électeurs et le fait de voter ou de s'abstenir. Pour cela, on compare les taux d'abstention dans les différents départements de Bourgogne-Franche-Comté (chiffres du ministère de l'intérieur, pour les élections européennes du 9 juin 2024) avec leur niveau de vie médian (chiffres de l'INSEE pour l'année 2021) :

Département	Côte d'Or	Doubs	Jura	Nièvre	Haute-Saône	Saône-et-Loire	Yonne	Territoire de Belfort
Niveau de vie	23 720 €	23 890 €	23 140 €	21 380 €	21 860 €	22 170 €	21 940 €	22 920 €
Taux d'abstention	42,74 %	44,22 %	43,14 %	45,95 %	41,27 %	45,39 %	45,76 %	44,6 %
x'_i	7	8	6	1	2	4	3	5
y'_i	2	4	3	8	1	6	7	5
$(x'_i - y'_i)^2$	25	16	9	49	1	4	16	0

1. Représenter ces données sous la forme d'un nuage de points.



2. Déterminer le coefficient de corrélation linéaire entre le niveau de vie et le taux d'abstention de ces départements. Que peut-on en conclure ? *Dans cette question, on demande d'indiquer les calculs effectués.*

moyenne : $m(X) = \frac{\sum x_i}{n} = \frac{23\,720+23\,890+23\,140+\cdots+22\,920}{8} = \frac{181\,020}{8} = 22\,627{,}5$
$m(X^2) = \frac{\sum x_i^2}{n} = \frac{23\,720^2+23\,890^2+23\,140^2+\cdots+22\,920^2}{8} = \frac{4\,101\,993\,000}{8}$
$Var(X) = m(X^2) - m(X)^2 = \frac{4\,101\,993\,000}{8} - \left(\frac{181\,020}{8}\right)^2 = 745\,368{,}75$
Écart-type : $s(X) = \sqrt{Var(X)} \simeq 863{,}35$

moyenne : $m(Y) = \frac{\sum x_i}{n} = \frac{42{,}74+44{,}22+43{,}14+\cdots+44{,}6}{8} = \frac{353{,}07}{8} \simeq 44{,}13$
$m(Y^2) = \frac{\sum x_i^2}{n} = \frac{42{,}74^2+44{,}22^2+43{,}14^2+\cdots+44{,}6^2}{8} = \frac{15\,601{,}180\,7}{8}$
$Var(Y) = m(Y^2) - m(Y)^2 = \frac{15\,601{,}180\,7}{8} - \left(\frac{353{,}07}{8}\right)^2 \simeq 2{,}36$
Écart-type : $s(Y) = \sqrt{Var(Y)} \simeq 1{,}54$

$m(X\,Y) = \frac{\sum x_i\, y_i}{n} = \frac{23\,720\times 42{,}74+23\,890\times 44{,}22+\ldots+22\,920\times 44{,}6}{8} = \frac{7\,985\,544}{8} = 998\,193$
$Cov(X,Y) = m(X\,Y) - m(X)\,m(Y) = \frac{7\,985\,544}{8} - \frac{181\,020}{8}\,\frac{353{,}07}{8} \simeq -443{,}416$
$r(X,Y) = \frac{Cov(X,Y)}{s(X)\,s(Y)} = \frac{-443{,}416}{863{,}35\times 1{,}54} \simeq -0{,}334.$
Cela peut suggérer un lien entre revenu et taux d'abstention (l'abstention serait plus faible quand le revenu est plus élevé), mais dans ce cas, le lien n'est pas très fort.

3. Déterminer aussi le coefficient de corrélation des rangs de Spearman. Que peut-on en conclure ?

On calcule tout d'abord les rangs x'_i et y'_i, entrés dans la table au début de l'exercice.
le coefficient de corrélation des rangs de Spearman est donc
$1 - \left(6 \times \frac{(7-2)^2+(8-4)^2+\cdots+(5-5)^2}{8\,(8^2-1)}\right) \simeq -0{,}429$
À nouveau, cela suggère un lien entre niveau de vie et abstention (l'abstention serait plus faible quand le revenu est plus élevé), mais ce lien n'est pas très fort.

4. Si l'on voulait estimer le taux d'abstention dans un département où le niveau de vie médian est 23 750, quelle droite pourrait-on utiliser ? Calculer l'équation de cette droite, et commenter la pertinence de son utilisation.

On utiliserait la droite $D_{Y|X}$, mais la pertinence ne serait pas si claire car les variables sont peu corrélées.
Pour son équation, on pose $a = \frac{Cov(X,Y)}{Var(X)} \simeq \frac{-443{,}416}{745\,368{,}75} \simeq -0{,}00059$ et $b = m(Y) - a\,m(X) \simeq 44{,}134 - (-0{,}001) \times 22\,627{,}5 \simeq 66{,}7615$
D'où l'équation de la droite $D_{Y|X} : Y = -0{,}000\,59\,X + 66{,}8$