Examen de statistiques – 29 juin 2016
L1 de psychologie

Réponses

Veuillez rendre ce sujet avec votre copie. **Numéro d'anonymat :**

Le formulaire et la calculatrice sont autorisés. Le soin de la rédaction entrera en compte dans la correction mais dans les questions où des détails ne sont pas explicitement demandés, un résultat correct, donné sans détails de calcul sera accepté.

Vous êtes priés de rendre à la fois votres copie (avec noms et prénoms) et cet énoncé (avec numéro d'anonymat reporté sur votre copie). Vous pouvez répondre soit directement sur l'énoncé (dans les cadres prévus à cet effet), soit sur votre copie (si vous manquez de place sur l'énoncé).

Exercice 1 : Stress des habitants d'un même quartier

On évalue niveau le niveau de stress (X) d'un échantillon d'habitants d'un même quartier d'habitation péri-urbain. On obtient les niveaux de stress suivants :

Niveau de stress (X)	[30 ; 60[	[60 ; 75[	[75 ; 90[	[90 ; 105[	[105 ; 120[	[120 ; 135[	[135 ; 150[
Effectif	7	12	34	33	22	13	4
Fréquences f_i	0,056	0,096	0,272	0,264	0,176	0,104	0,032
Fréq, cum, F_i	0,056	0,152	0,424	0,688	0,864	0,968	1
hauteur $f_i/(a_{i+1}-a_i)$	0,0019	0,0064	0,0181	0,0176	0,0117	0,0069	0,0021

1. Au sein de l'échantillon, combien d'habitants ont un niveau de stress supérieur ou égal à 120 ?

Dans cet échantillon, il y a $13 + 4 = 17$ habitants dont le niveau de stress est supérieur ou égal à 120.

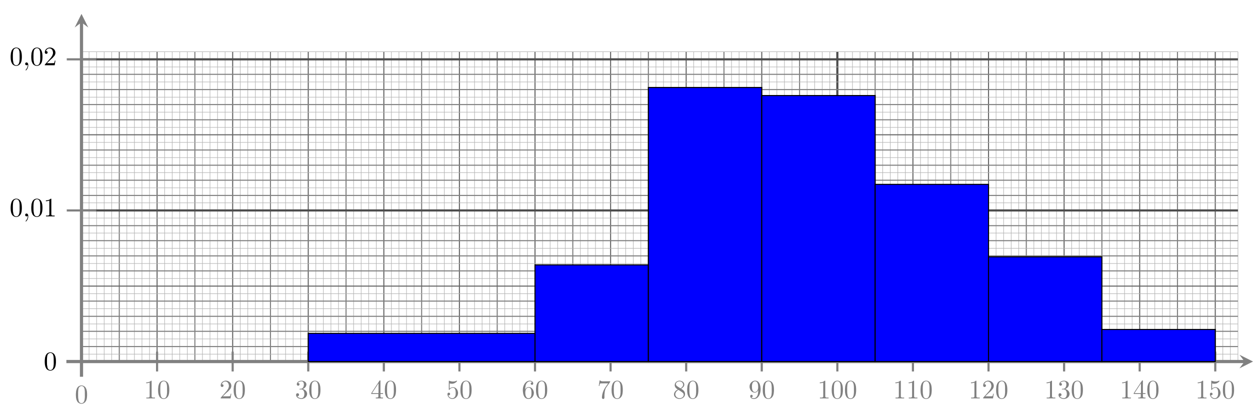
2. Quelle est au sein de cet échantillon la proportion de sujets dont le niveau de stress est inférieur à 120 ?

Dans l'échantillon, $7 + 12 + 34 + 33 + 22 = 108$ habitants ont un niveau de stress inférieur à 120. Comme la taille de l'échantillon est $7 + 12 + 34 + 33 + 22 + 13 + 4 = 125$, on calcule donc la proportion :

$\mathbb{P}_r[X < 120] = \frac{108}{125} = 0{,}864$

3. Déterminer les fréquences et les fréquences cumulées. *Les ajouter dans le tableau au début de l'exercice.*

4. Représenter ces données par un **histogramme** (sur le quadrillage ci-dessous).



5. Déterminer la médiane en détaillant les calculs. En donner l'interprétation.

Classe de la médiane : [90 ; 105[

$M_e = a_i + \frac{a_{i+1}-a_i}{F(a_{i+1})-F(a_i)}(0{,}5 - F(a_i)) \simeq 90 + \frac{105-90}{0{,}688-0{,}424}(0{,}5 - 0{,}424) \simeq 94{,}32$

Au sein de l'échantillon, 50% des sujets ont un niveau de stress inférieur à 94,32

6. Calculer le premier quartile. En donner l'interprétation.

> Classe du premier quartile : $[75\,;\,90[$
> $Q_1 = a_i + \frac{a_{i+1}-a_i}{F(a_{i+1})-F(a_i)}(0{,}25 - F(a_i)) \simeq 75 + \frac{90-75}{0{,}424-0{,}152}(0{,}25 - 0{,}152) \simeq 80{,}4$
> Au sein de l'échantillon, 25% des sujets ont un niveau de stress inférieur à 80,4

7. En détaillant les calculs, déterminer une valeur approchée de la proportion $\mathbb{P}_r[X \leqslant 86]$. En donner l'interprétation.

> Ce qui l'on demande de calculer est $F(86)$. Or $86 \in [75\,;\,90[$ donc
> $\mathbb{P}_r[X \leqslant 86] \simeq \quad F(86) = F(a_i) + \frac{F(a_{i+1})-F(a_i)}{a_{i+1}-a_i}(86 - a_i) = 0{,}152 + \frac{0{,}424-0{,}152}{90-75}(86 - 75) \simeq 0{,}351$
> Au sein de l'échantillon, environ 35,1% des sujets ont un niveau de stress inférieur ou égal à 86.

Exercice 2 : Dons à des associations

Selon un sondage réalisé par téléphone, 45% de la population adulte française effectue régulièrement des dons à des associations sollicitant la générosité du public. Ce chiffre étant aussi corroboré par les données du ministère de l'économie et des finances, on supposera dans cet exercice qu'il est correct.

Au total, il y a environ 52 000 000 français adultes, et 34 000 000 adultes polonais.

Dans cet exercice, si un calcul de probabilité utilise l'approximation d'une loi binomiale par une loi normale, on demande d'effectuer une correction de continuité.

1. Étant donné un échantillon aléatoire de 14 adultes français, choisis avec remise, on note S_{14} le nombre de donateurs réguliers au sein de l'échantillon.
   (a) Donner la loi de la variable S_{14}.

   > $S_{14} \sim \mathcal{B}(14\ ;\ 0{,}45)$

   (b) Peut-on approximer cette loi par une autre loi ?
   Justifiez pourquoi et précisez – le cas échéant – par quelle loi.

   > La seule approximation dont on dispose pour les lois binômiales est leur approximation par une loi normale.
   > Or cette approximation nécessite $n > 30$ (qui n'est pas le cas), donc on ne peut pas faire d'approximation.

   (c) Calculer la probabilité $\mathbb{P}[S_{14} < 3]$.

   > $$\begin{aligned}\mathbb{P}[S_{14} < 3] &= \mathbb{P}[S_{14} = 0] + \mathbb{P}[S_{14} = 1] + \mathbb{P}[S_{14} = 2] \\ &= \binom{14}{0}(1-0{,}45)^{14} + \binom{14}{1}(0{,}45)(1-0{,}45)^{14-1} + \binom{14}{2}(0{,}45)^2(1-0{,}45)^{14-2}\end{aligned}$$
   > $\mathbb{P}[S_{14} < 3] \simeq 0{,}0002 + 0{,}0027 + 0{,}0141 \simeq 0{,}017$

2. Étant donné un échantillon aléatoire de 1427 adultes français, choisis avec remise, on note S_{1427} le nombre de donateurs réguliers au sein de l'échantillon.
   (a) Donner la loi de la variable S_{1427}.

   > $S_{1427} \sim \mathcal{B}(1427\ ;\ 0{,}45)$

   (b) Peut-on approximer cette loi par une autre loi ?
   Justifiez pourquoi et précisez – le cas échéant – par quelle loi.

   > On a $n = 1427 > 30$, et $n\,p \simeq 642 > 5$, et $n(1-p) \simeq 785 > 5$,
   > donc on peut approximer S_{14} par $\mathcal{N}(642; 18{,}8)$ où $18{,}8 \simeq \sqrt{n\,p(1-p)}$

   (c) En détaillant vos calculs, déterminer la probabilité $\mathbb{P}[630 \leqslant S_{1427} \leqslant 670]$.

   > $$\begin{aligned}\mathbb{P}[630 \leqslant S_{1427} \leqslant 670] &= \mathbb{P}[629{,}5 \leqslant S_{1427} \leqslant 670{,}5] = \mathbb{P}[\tfrac{629{,}5-642}{18{,}8} < \tfrac{S_{1427}-642}{18{,}8} < \tfrac{670{,}5-642}{18{,}8}] \\ &\simeq \mathbb{P}[-0{,}66 < Z < 1{,}52] \simeq F(1{,}52) - F(-0{,}66) \\ &\simeq 0{,}9357 - (1 - 0{,}7454) \simeq 0{,}9357 - 0{,}2546\end{aligned}$$
   > $\mathbb{P}[630 \leqslant S_{1427} \leqslant 670] \simeq 0{,}68$

3. On cherche désormais à déterminer s'il y a une plus grande proportion de donateurs réguliers parmi les polonais que parmi les français. On interroge donc un échantillon aléatoire de 1041 polonais choisis sans remise, parmi lesquels 722 affirment donner régulièrement à des associations.
   (a) Estimer la proportion de donateurs au sein de la population adulte polonaise.
   Vous déterminerez un intervalle de confiance avec la confiance 90%.

   On a $n = 1041 > 30$, et $n\,p_e = 722 > 5$, et $n(1 - p_e) = 319 > 5$ donc on peut utiliser la pocédure du formulaire pour estimer la proportion p.
   On a $F(1{,}645) \simeq 0{,}95$ d'où $z_\alpha \simeq 1{,}645$
   d'où $a_\alpha = z_\alpha\sqrt{\frac{p_e(1-p_e)}{n}} \simeq 0{,}0235$.
   On estime donc que p est dans l'intervalle $[0{,}6701; 0{,}7171]$ avec la confiance $c = 0{,}9$.

   (b) Pouvez-vous conclure, avec la confiance 90%, que la proportion de donateurs est plus importante en Pologne qu'en France ?

   Oui : cet intervalle permet d'affirmer, avec la confiance 90%, que la proportion de donateurs en Pologne est supérieure à 45%,

Exercice 3 : Apprentissage de l'anglais
Un lycée comporte 5 classes de seconde. Parmi tous les élèves de seconde il y en a 25 qui ont effectué au moins un séjour linguistique dans un pays anglophone au cours de leur scolarité. On souhaite comparer, au moyen d'un test sur cent points, le niveau en anglais des élèves, selon qu'ils aient fait, ou non, un tel séjour.

1. **Ensemble des élèves de seconde**
   Dans un premier temps, on fait passer le test à l'ensemble des élèves de seconde de ce lycée. Une fois regroupées en classes, les notes obtenues sont les suivantes :

Notes	[10 ; 25[	[25 ; 40[	[40 ; 55[	[55 ; 70[	[70 ; 85[	[85 ; 100[
Effectifs	1	15	46	59	29	8

   (a) Combien y a t'il d'élèves de seconde dans ce lycée ?

   Il y a $1 + 15 + 46 + 59 + 29 + 8 = 158$ élèves de seconde dans ce lycée.

   (b) Calculer le moyenne et l'écart-type des notes de ces élèves.

   $m(X) \simeq 59{,}27$
   $s(X) \simeq 15{,}5$

   (c) On suppose pour simplifier que ces élèves forment un échantillon représentatif de l'ensemble des 540 920 élèves de seconde en France. Si l'on faisait passer le même test à tous les élèves de seconde de France, quelle note moyenne estimez-vous que l'on obtiendrait ?
   Vous déterminerez un intervalle de confiance, avec la confiance 90%.

   Comme n= 158 >30, on cherche z_α tel que $F(z_\alpha) = \frac{c+1}{2}$
   On a $F(1{,}645) \simeq 0{,}95$ d'où $z_\alpha \simeq 1{,}645$ d'où $a_\alpha = z_\alpha \frac{s_e}{\sqrt{n-1}} \simeq 2{,}035$
   On estime donc que μ est dans l'intervalle $[59{,}27 - 2{,}035\,;\,59{,}27 + 2{,}035] \simeq [57{,}23\,;\,61{,}31]$, avec la confiance $c = 0{,}9$.

2. **Élèves ayant fait un séjour linguistique dans un pays anglophone**
   Dans un second temps, on regroupe les copies des 25 élèves, parmi les secondes de ce lycée, qui ont déjà effectué un séjour linguistique dans un pays anglophone. On constate que leur moyenne est de 71,96, avec un écart type de 11,05.
   Si l'on faisait passer le même test à tous les élèves français de seconde ayant fait un séjour linguistique dans un pays anglophone, quelle note moyenne estimez-vous que l'on obtiendrait ?
   Vous déterminerez un intervalle de confiance, avec la confiance 90%.

Comme $n = 25 \leqslant 30$, on cherche t_α à partir de la table inverse de Student avec $p = \frac{\alpha}{2} = 0{,}05$ et $n - 1 = 24$ degrés de liberté (ddl)
On lit $t_\alpha \simeq 1{,}7109$ d'où $a_\alpha = t_\alpha \frac{s_e}{\sqrt{n-1}} \simeq 3{,}8591$,
On estime donc que μ est dans l'intervalle $[68{,}1; 75{,}82]$ avec la confiance $c = 0{,}9$

3. **Conclusion**
Peut-on conclure, avec la confiance 90%, que les élèves ayant fait un séjour linguistique dans un pays anglophone auraient en moyenne de meilleurs résultats à ce test d'anglais ?

Oui :
Les intervalles estimées dans les deux questions précédentes ne se chevauchent pas et on peut affirmer, avec la confiance 90%, que les élèves ayant fait un séjour dans un pays anglophone ont en moyenne de meilleures notes

Exercice 4 : Acquisition du vocabulaire
On étudie l'acquisition du vocabulaire chez des enfants de 4 à 6 ans, dont on détermine environ le nombre de mots qu'ils maîtrisent (c'est à dire les mots qu'ils comprennent et qu'ils utilisent). Au sein d'un échantillon de 12 enfants, on obtient les données ci-dessous :

Enzo (56 mois) : 7 900 mots	Lilou (58 mois) : 4 600 mots	Clément (49 mois) : 5 500 mots	Manon (72 mois) : 5 700 mots
Nolan (70 mois) : 10 000 mots	Romane (67 mois) : 8 100 mots	Camille (55 mois) : 5 400 mots	Léna (57 mois) : 6 600 mots
Inès (55 mois) : 6 100 mots	Hugo (52 mois) : 3 600 mots	Raphaël (71 mois) : 9 200 mots	Eva (48 mois) : 3 300 mots

1. Quelle est, au sein de cet échantillon, la proportion d'enfants qui maîtrisent moins de 7 300 mots ?

Au sein de cet échantillon, il y a 8 enfants qui maîtrisent moins de 7 300 mots.
Cela correspond à une proportion $\frac{8}{12} \simeq 0{,}667$.

2. Déterminer le coefficient de corrélation des rangs ("*de Spearman*").

On commence par regrouper les données (en notant X l'age en mois et Y le nombre de mots maîtrisés,) et calculer les rangs :

Sujet	Enzo	Lilou	Clément	Manon	Nolan	Romane	Camille	Léna	Inès	Hugo	Raphaël	Eva
X	56	58	49	72	70	67	55	57	55	52	71	48
Y	7900	4600	5500	5700	10000	8100	5400	6600	6100	3600	9200	3300
X''	6	8	2	12	10	9	4,5	7	4,5	3	11	1
Y''	9	3	5	6	12	10	4	8	7	2	11	1
$(X''-Y'')^2$	9	25	9	36	4	1	0,25	1	6,25	1	0	0

le coefficient de correlation des rangs de spearman est donc $1-\left(6\,\frac{9+25+9+36+4+\cdots+0}{12\,(12^2-1)}\right) \simeq 0{,}677$

3. Déterminer la moyenne et la variance des nombres de mots maîtrisés par les enfants de l'échantillon.

moyenne : $m(Y) = \frac{\sum x_i}{N} = \frac{7900+4600+\cdots+3300}{12} = \frac{76000}{12} \simeq 6333{,}33$
$m(Y^2) = \frac{\sum x_i^2}{N} = \frac{7900^2+4600^2+\cdots+3300^2}{12} = \frac{530340000}{12}$
$Var(Y) = m(Y^2) - m(Y)^2 = \frac{530340000}{12} - \left(\frac{76000}{12}\right)^2 \simeq 4083888{,}89$
remarque (non-demandé dans l'énnoncé) : l'écart-type est $s(Y) = \sqrt{Var(Y)} \simeq 2020{,}86$

4. Déterminer le coefficient de corrélation **linéaire**.

$m(X) \simeq 59{,}17$ $s(X) \simeq 8{,}23$ $Cov(X\ ;\ Y) \simeq 11886$
Donc $r = \frac{Cov(X\ ;\ Y)}{s(X)\,s(Y)} \simeq 0{,}714$.

5. Déterminer l'équation de la droite qui permet d'estimer le nombre de mots en fonction de l'age.

on pose $a = \frac{Cov(X,Y)}{Var(X)} \simeq \frac{11886{,}111}{67{,}81} \simeq 175{,}286$
et $b = m(Y) - a\,m(X) \simeq 6333{,}333 - 175{,}286 \times 59{,}167 \simeq -4037{,}814$
D'où l'équation de la droite $D_{Y|X} : Y = 175\,X - 4038$