Contrôle Terminal – 2ème session – 28/06/2019 – 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 : Chaleur estivale et millésimes viticoles

On cherche dans cet exercice à déterminer si la qualité d'un millésime est liée aux températures estivales mesurées cette année là. On relève donc les notes attribuées aux vins rouges bourguignons des années 2005 à 2014 sur le site d'une enseigne de vente de vin. De plus, on considère la température quotidienne (mesurée à Dijon au moment le plus chaud de la journée) dont on fait la moyenne sur les mois de juillet et août de ces mêmes années. On obtient les données suivantes :

Année	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Température	25,8 °C	26,8 °C	23,8 °C	24,8 °C	26,5 °C	25,1 °C	24,1 °C	25,6 °C	26,8 °C	23,9 °C
Note attribuée à ce millésime	9	6	5	7	9	7	7	7	6	8

1. Quel est le coefficient de corrélation linéaire entre cette note et la température ?

moyenne : $m(X) = \frac{\sum x_i}{n} = \frac{25{,}8+26{,}8+\cdots+23{,}9}{10} = \frac{253{,}2}{10} = 25{,}32$
$m(X^2) = \frac{\sum x_i^2}{n} = \frac{25{,}8^2+26{,}8^2+\cdots+23{,}9^2}{10} = \frac{6423{,}24}{10}$
$Var(X) = m(X^2) - m(X)^2 = \frac{6423{,}24}{10} - \left(\frac{253{,}2}{10}\right)^2 \simeq 1{,}222$
Écart-type : $s(X) = \sqrt{Var(X)} \simeq 1{,}11$

moyenne : $m(Y) = \frac{\sum x_i}{n} = \frac{9+6+\cdots+8}{10} = \frac{71}{10} = 7{,}1$
$m(Y^2) = \frac{\sum x_i^2}{n} = \frac{9^2+6^2+\cdots+8^2}{10} = \frac{519}{10}$
$Var(Y) = m(Y^2) - m(Y)^2 = \frac{519}{10} - \left(\frac{71}{10}\right)^2 = 1{,}49$
Écart-type : $s(Y) = \sqrt{Var(Y)} \simeq 1{,}22$

$m(XY) = \frac{\sum x_i y_i}{n} = \frac{25{,}8\times 9+26{,}8\times 6+\ldots+23{,}9\times 8}{10} = \frac{1799{,}7}{10} = 179{,}97$
$Cov(X,Y) = m(XY) - m(X)\,m(Y) = \frac{1799{,}7}{10} - \frac{253{,}2}{10}\frac{71}{10} \simeq 0{,}198$
$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{0{,}198}{\sqrt{1{,}222\times 1{,}49}} \simeq 0{,}147$

2. Une fois que l'on connaîtra les températures moyennes de l'été 2019, serait-il pertinent d'utiliser une droite de régression pour en déduire une prédiction de la note qu'obtiendrait le millésime 2019 ? Si oui, quelle droite considérer ?

Non : Le coefficient de corrélation est trop faible pour que l'utilisation d'une droite de régression soit pertinente.

Exercice 2 : Fautes d'orthographe et vitesse de lecture

Des chercheurs souhaitent savoir s'il est plus facile de lire un texte bien orthographié ou au contraire un texte présentant des fautes d'orthographes. Pour cela, ils considèrent un texte assez court, dont ils ont constaté que les enfants de 15 ans mettent en moyenne 108 secondes à le lire. Ils produisent alors un texte modifié, obtenu en ajoutant délibérément des fautes d'orthographe au texte original, et ils mesurent le temps mis par des enfants pour lire ce texte modifié. Sur un échantillon de 200 enfants de 15 ans, ils obtiennent les données suivantes :

Temps de lecture	[75 ; 91[	[91 ; 107[	[107 ; 123[	[123 ; 139[	[139 ; 155[	[155 ; 171[
Effectif	14	35	61	49	33	8
Fréquence	0,070	0,175	0,305	0,245	0,165	0,040
Fréquence cumulée	0,070	0,245	0,550	0,795	0,960	1,000

1. (a) Calculer les fréquences et les fréquences cumulées.

Vous compléterez les lignes correspondantes du tableau ci-dessus.

(b) Quels sont, au sein de cet échantillon, la moyenne et l'écart type du temps de lecture du texte modifié ?

Dans cette question, on demande de présenter les calculs effectués.

moyenne :

$$m(X) = \frac{\sum_i c_i\, n_i}{n} = \frac{83\times14+99\times35+115\times61+131\times49+147\times33+163\times8}{200} = \frac{24216}{200} = 121{,}08\,\text{s}$$

$$m(X^2) = \frac{\sum_i c_i^2\, n_i}{n} = \frac{83^2\times14+99^2\times35+115^2\times61+131^2\times49+147^2\times33+163^2\times8}{200} = \frac{3012744}{200}$$

$$Var(X) = m(X^2) - m(X)^2 = \frac{3012744}{200} - \left(\frac{24216}{200}\right)^2 \simeq 403{,}35$$

Écart-type : $s(X) = \sqrt{Var(X)} \simeq 20{,}08\,\text{s}$

(c) Quelle en est environ la médiane ?

Dans cette question, on demande à nouveau de présenter les calculs effectués.

Classe de la médiane : [107 ; 123[

$$\text{Méd} \simeq a_i + \frac{a_{i+1}-a_i}{F_X(a_{i+1})-F_X(a_i)}(0{,}5 - F_X(a_i)) \simeq 107 + \frac{123-107}{0{,}55-0{,}245}(0{,}5-0{,}245) \simeq 120{,}38\,\text{s}$$

(d) Quel est enfin le troisième quartile ? Remplir la phrase de conclusion ci-dessous pour exprimer ce que signifie ce troisième quartile.

Classe du troisième quartile : [123 ; 139[

$$Q_3 \simeq a_i + \frac{a_{i+1}-a_i}{F_X(a_{i+1})-F_X(a_i)}(0{,}75 - F_X(a_i)) \simeq 123 + \frac{139-123}{0{,}795-0{,}55}(0{,}75-0{,}55) \simeq 136{,}06\,\text{s}$$

Dans cet échantillon, 75 % des enfants ont mis moins que 136,06 secondes à lire le texte modifié.

2. (a) Estimer le temps de lecture moyen du texte modifié pour l'ensemble des enfants de 15 ans.

Vous calculerez un intervalle de confiance, associé au risque d'erreur $\alpha = 0{,}01$.

Comme n= 200 >30, on cherche z_α tel que $F(z_\alpha) = \frac{0{,}99+1}{2} = 0{,}995$
On lit sur la table inverse que pour la confiance 0,99 on a $z_\alpha \simeq 2{,}576$ d'où $a_\alpha = z_\alpha \frac{s_e}{\sqrt{n-1}} \simeq 3{,}667$
On estime donc que μ est dans l'intervalle $[121{,}08 - 3{,}667\,;\, 121{,}08 + 3{,}667] \simeq [117{,}41\,;\, 124{,}75]$, avec la confiance $c = 0{,}99$.

(b) Conclure : peut-on affirmer, avec la confiance 99%, que les enfants de 15 ans mettent en moyenne plus longtemps à lire le texte modifié que l'original ?

La valeur 108 se trouve à l'extérieur de l'intervalle, on peut bien affirmer que le temps moyen de lecture du texte modifié est supérieur à 108 s.

Exercice 3 : Anxiété en milieu rural ou urbain.

En France, 20% de la population vit en milieu rural. On souhaite savoir si ces habitants de milieux ruraux sont plus sujets à l'anxiété que le reste de la population française (sachant que chaque année en France, 15% de population environ a un trouble anxieux).

On considère pour cela un échantillon aléatoire de personnes ayant eu un trouble anxieux au cours de l'année 2019. On note par X le nombre d'individus de l'échantillon qui habitent en milieu rural.

1. On considère un échantillon aléatoire de 62 français·e·s ayant eu un trouble anxieux en 2019. Montrer que si l'on suppose que les habitants de milieux ruraux et urbain ont la même probabilité de développer un trouble anxieux, on peut décrire la loi de X par une loi binomiale (vous préciserez laquelle).

Il y a environ $65000000 \times \frac{15}{100} = 9750000$ français·e·s qui ont eu un trouble anxieux en 2019. La taille de l'échantillon (62) est plus de 10 fois plus petite que cela, donc on peut utiliser une loi binomiale.
Donc $X \sim \mathcal{B}(62\ ;\ 0{,}2)$.

2. Justifiez que cette loi peut s'approcher par une loi normale, dont vous préciserez la moyenne et l'écart type.

On a $n = 62 > 30$, et $np = 12{,}4 > 5$, et $n(1-p) = 49{,}6 > 5$ donc on peut approximer X par $\mathcal{N}(12{,}4; 3{,}15)$ où $3{,}15 \simeq \sqrt{np(1-p)}$

3. Quelle serait alors la probabilité $\mathbb{P}[1 \leqslant X \leqslant 23]$?
On utilisera la loi normale de la question précédente et une correction de continuité

$$\begin{aligned}\mathbb{P}[1 \leqslant X \leqslant 23] &= \mathbb{P}[0{,}5 \leqslant X \leqslant 23{,}5] = \mathbb{P}[\tfrac{0{,}5-12{,}4}{3{,}15} < \tfrac{X-12{,}4}{3{,}15} < \tfrac{23{,}5-12{,}4}{3{,}15}] \\ &\simeq \mathbb{P}[-3{,}78 < Z < 3{,}52] \simeq F(3{,}52) - F(-3{,}78) \\ &\simeq 0{,}9998 - (1 - 0{,}999\,9) \simeq 0{,}999\,8 - 0{,}000\,1 \\ &\simeq 1\end{aligned}$$

4. Sur un échantillon aléatoire de 62 français·e·s ayant eu un trouble anxieux en 2019, si l'on constatait qu'il y ait 29 habitants de milieux ruraux, cela serait-il compatible avec l'hypothèse selon laquelle les habitants de milieux ruraux et urbain ont la même probabilité de développer un trouble anxieux ?

Si les habitants de milieux ruraux et urbain avaient la même probabilité de développer un trouble anxieux, alors on aura du avoir $1 \leqslant X \leqslant 23$ (avec une probabilité très proche de 100%). Si on observe au contraire $X = 29$, cela indique qu'en fait les urbains et ruraux n'ont pas la même probabilité de subir des troubles anxieux.

Exercice 4 : Genre et goût pour les films d'action

Pendant une durée d'une semaine, la gérante d'un cinéma décide de collecter des statistiques sur le genre (Femme / Homme) de ses client·e·s. Elle constate alors que parmi les 500 billets vendus au cours de cette semaine, 51% ont été vendus à des femmes et 49% à des hommes.

De plus, elle constate que parmi ces 500 billets vendus, il y en a 160 qui ont été vendus pour des films d'action. Parmi ces billets pour des films d'actions, 45% ont été vendus à des femmes, et 55% à des hommes.

Pour simplifier les questions posées, on suppose que les clients étaient 500 personnes distinctes (c'est à dire qu'aucune personne n'a acheté plusieurs billets cette semaine là).

1. (a) Parmi les femmes qui (au cours de cette semaine) ont acheté un billet dans ce cinéma, combien ont opté pour un film d'action ?

$160 \times \frac{45}{100} = 72$

(b) Parmi les hommes (qui ont acheté un billet dans ce cinéma au cours de cette semaine), quelle proportion a opté pour un film d'action ?

Il s'agit de $160 \times \frac{55}{100} = 88$ personnes, parmi un total de $500 \times \frac{49}{100} = 245$ hommes.
Ça correspond donc à la proportion $\frac{88}{245} \simeq 0{,}359$ (ie 35,9%).

2. On souhaite désormais déduire, à partir des observations sur cet échantillon de clients, des proportions au sein de l'ensemble des clients de cinéma en France.

   (a) Estimer, parmi l'ensemble des billets de cinéma vendus en France à des hommes, la proportion qui correspondent à des films d'action.

   Vous calculerez un intervalle de confiance, associé à la confiance $c = 90\%$.

On a $n = 245 > 30$, et $p_e = \frac{88}{245} \simeq 0{,}359$, d'où $n\,p_e = 88 > 5$, et $n(1-p_e) = 157 > 5$ donc on peut utiliser la procédure du formulaire pour estimer la proportion p.
On lit sur la table inverse que pour la confiance 0,9 on a $z_\alpha \simeq 1{,}645$
d'où $a_\alpha = z_\alpha\sqrt{\frac{p_e(1-p_e)}{n}} \simeq 0{,}0504$.
On estime donc que p est dans l'intervalle $[0{,}3088; 0{,}4096]$ avec la confiance $c = 0{,}9$.

   (b) Estimer, parmi l'ensemble des billets de cinéma vendus en France à des femmes, la proportion qui correspondent à des films d'action.

   Vous calculerez un intervalle de confiance, associé à la confiance $c = 90\%$.

Dans l'échantillon, il y a $500 \times \frac{51}{100} = 255$ femmes dont 72 ont été à un film d'action.
On a $n = 255 > 30$, et $p_e = \frac{72}{255} \simeq 0{,}282$, d'où $n\,p_e = 72 > 5$, et $n(1-p_e) = 183 > 5$ donc on peut utiliser la procédure du formulaire pour estimer la proportion p.
On lit sur la table inverse que pour la confiance 0,9 on a $z_\alpha \simeq 1{,}645$
d'où $a_\alpha = z_\alpha\sqrt{\frac{p_e(1-p_e)}{n}} \simeq 0{,}0464$.
On estime donc que p est dans l'intervalle $[0{,}236; 0{,}3288]$ avec la confiance $c = 0{,}9$.

   (c) À partir de ces estimations, peut-on conclure (avec la confiance 90%) que lorsqu'elles vont au cinéma, les femmes ont moins tendance que les hommes à aller voir un film d'action ?

Les intervalles se chevauchent, et ne permettent pas de déterminer (avec la confiance 90%) si cette proportion est plus élevée chez les hommes que chez les femmes.