Éléments de correction du contrôle continu # 2

Questions de cours (3 points)

1. Énoncer la loi des grands nombres. 1,5 points

   Soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes et identiquement distribuées, telles que $\mathbb{E}[|X_1|] < +\infty$. Alors lorsque n tend vers l'infini, avec probabilité un, on a la convergence

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{n \to +\infty} \mathbb{E}[X_1].$$

2. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Si A, B, et C sont trois évenements, à quelle condition sont-ils mutuellement indépendants ? 1,5 points

   Les événements A, B, et C sont mutuellement indépendants si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C), \quad \mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C),$$

$$\text{et} \quad \mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C).$$

Exercice 1 *La stratégie de l'échec* (9 points)

Un test d'évaluation comporte 100 questions, chaque question ayant 5 réponses possibles A, B, C, D ou E, dont une seule est correcte. Les auteurs du test font en sorte que les bonnes réponses soient réparties uniformément au hasard entre les réponses A, B, C, D et E.

1. Combien y a-t-il de réponses possibles au test ? 1 point

   Pour chaque question, il y a 5 réponses possibles et il y a 100 questions. Le nombre total de réponses distinctes est donc 5^{100}.

2. Un premier candidat répond totalement au hasard. Quelle est alors la loi du nombre de bonnes réponses ? Quelle est la probabilité qu'il ait entre 20 et 22 bonnes réponses (on attend un résultat exact, sous forme littérale) ? 2 points

   Avec cette stratégie, à chaque question, la probabilité d'avoir la bonne réponse est $p = 1/5$. Le nombre total N de bonnes réponses aux $n = 100$ questions suit alors une loi binomiale $B(n, p)$ i.e. $N \sim B(100, 1/5)$ ici. On a alors

$$\mathbb{P}(20 \leq N \leq 22) = \mathbb{P}(N = 20) + \mathbb{P}(N = 21) + \mathbb{P}(N = 22)$$

$$= \binom{100}{20}\left(\frac{1}{5}\right)^{20}\left(\frac{4}{5}\right)^{80} + \binom{100}{21}\left(\frac{1}{5}\right)^{21}\left(\frac{4}{5}\right)^{79} + \binom{100}{22}\left(\frac{1}{5}\right)^{22}\left(\frac{4}{5}\right)^{78}.$$

3. À l'aide d'une approximation de la loi binomiale, donner une estimation numérique de cette probabilité. 2 points

   D'après le cours, comme $n = 100 > 30$, $np = 20 > 5$ et $np(1-p) = 16 > 5$, on peut approcher la loi $B(100, 1/5)$ de la variable N par la loi normale $\mathcal{N}(20, 16)$. Autrement dit $(N - 20)/4 \approx \mathcal{N}(0, 1)$ et

$$\begin{aligned} \mathbb{P}(20 \leq N \leq 22) &= \mathbb{P}\left(0 \leq \frac{N-20}{4} \leq \frac{1}{2}\right) \approx \mathbb{P}\left(0 \leq \mathcal{N}(0,1) \leq \frac{1}{2}\right) \\ &= \mathbb{P}\left(\mathcal{N}(0,1) \leq \frac{1}{2}\right) - \mathbb{P}\left(\mathcal{N}(0,1) < 0\right) = 0.191. \end{aligned}$$

4. Un deuxième candidat suit une autre stratégie : il répond systématiquement la réponse D. Quelle est la loi du nombre de bonnes réponses avec cette stratégie ? Justifier. 2 points
   D'après l'énoncé, les bonnes réponses sont réparties uniformément parmi A, B, C, D, E, aussi à chaque question, le second candidat a une chance sur 5 de donner la bonne réponse. Comme précédemment, le nombre $\widetilde{N}$ du second candidat suit donc une loi binomiale $B(100, 1/5)$.
5. Estimer la probabilité que le second candidat ait entre 20 et 24 bonnes réponses ? 2 points
   On utilise à nouveau l'approximation binomiale :

$$\begin{aligned}\mathbb{P}(20 \leq \widetilde{N} \leq 24) &= \mathbb{P}\left(0 \leq \frac{\widetilde{N}-20}{4} \leq 1\right) \approx \mathbb{P}\left(0 \leq \mathcal{N}(0,1) \leq 1\right) \\ &= \mathbb{P}\left(\mathcal{N}(0,1) \leq 1\right) - \mathbb{P}\left(\mathcal{N}(0,1) < 0\right) = 0.341.\end{aligned}$$

Exercice 2 *Hypertension* (8 points)
Dans deux pays A et B aux habitudes alimentaires distinctes, on considère un échantillon de 500 personnes (300 hommes et 200 femmes à chaque fois), dont on mesure la tension artérielle.

1. Dans la population A, on veut tester au niveau de confiance 95%, s'il existe un lien entre le genre d'un individu et sa tension artérielle.

   (a) Parmi les statistiques présentées au verso, laquelle faut-il considérer ici ? Justifier. 1 point
   Il s'agit de faire un test du χ^2 d'indépendance, précisément, au niveau de confiance 95%, on veut tester l'hypothèse H_0 : "tension et genre sont indépendants" contre l'hypothèse H_1 : "tension et genre ne sont pas indépendants". Le tableau des fréquences empiriques est le suivant :

Observé	Hypertension	Tension normale	Hypotension	Total
Homme	72/500	192/500	36/500	300/500
Femme	38/500	118/500	44/500	200/500
Total	110/500	310/500	80/500	500/500

   La statistique empirique du χ^2 associée est la statistique B) du verso, i.e.

$$F_{500}^{obs} = 500 \times \left(\left(\frac{\left(\frac{72}{500} - \frac{110}{500}\frac{300}{500}\right)^2}{\frac{110}{500}\frac{300}{500}}\right) + \ldots + \left(\frac{\left(\frac{44}{500} - \frac{80}{500}\frac{200}{500}\right)^2}{\frac{80}{500}\frac{200}{500}}\right)\right) \approx 9,35.$$

   (b) Expliciter la zone de rejet du test et conclure. 1 point
   Par ailleurs, sous l'hypothèse H_0, la statistique du χ^2 suit une loi du χ^2 à $(3-1)\times(2-1) = 2$ degrés de liberté. D'après les donées numériques de l'énoncé, on a $\mathbb{P}(\chi^2(2) > 5,99) = 95\%$ de sorte que la région de rejet du test d'indépendance est $W = \{F_{500} > 5,99\}$. Ici, on a clairement $F_{500}^{obs} \in W$ de sorte que l'on rejette l'hypothèse H_0 au niveau 95%.

2. On s'intéresse plus particulièrement à la population masculine. On souhaite savoir si les fréquences des hypertendus/"normaux"/hypotendus sont les mêmes dans les deux pays A et B. À cette fin, au niveau de confiance 95%, on met en place un test du χ^2 d'adéquation des fréquences de la population B à la loi cible $\pi = (72/300, 192/300, 36/300)$.

   (a) Parmi les statistiques présentées au verso, laquelle faut-il considérer ici ? Justifier.1 point
   Si p désigne le vecteur des fréquences théoriques dans la population B, il s'agit ici de tester, au niveau 95%, l'hypothèse H_0 : "$p = \pi$" contre H_1 : "$p \neq \pi$". La statistique du χ^2 est associée est la statistique F) du verso i.e.

$$F_{300}^{obs} = 300 \times \left(\left(\frac{\left(\frac{75}{300} - \frac{72}{300}\right)^2}{\frac{72}{300}}\right) + \ldots + \left(\frac{\left(\frac{38}{300} - \frac{36}{300}\right)^2}{\frac{36}{300}}\right)\right) \approx 0,37.$$

(b) Expliciter la zone de rejet du test et conclure. 1 point
Par ailleurs, sous l'hypothèse H_0, la statistique théorique suit une loi du χ^2 à $(3-1) = 2$ degrés de liberté. On a toujours $\mathbb{P}(\chi^2(2) > 5,99) = 95\%$ de sorte que la région du test d'indépendance est $W = \{F_{300} > 5,99\}$. Ici, on a clairement $F_{300}^{obs} \notin W$ de sorte que l'on ne rejette pas l'hypothèse d'indépendance au niveau 95%.

3. Sans faire de distinction entre homme et femme, on souhaite enfin déterminer, au niveau de confiance 90%, si les proportions d'hypertendus sont les mêmes dans les pays A et B.

(a) Répondre à question en effectuant un test du χ^2 d'adéquation . 2 points
Dans la population hypertendue, on désigne par p la proportion d'individus issus du pays A, de sorte que la proportion d'individus issus de B est $1-p$. Il s'agit d'effectuer, au niveau de confiance 90%, un test du χ^2 d'adéquation de H_0 : "$(p, 1-p) = (1/2, 1/2)$" contre l'hypothèse H_1 : "$(p, 1-p) \neq (1/2, 1/2)$". Les proportion empiriques sont ici $\widehat{p} = 110/215$ et $1-\widehat{p} = 105/215$ de sorte que la statistique empirique du χ^2 vaut

$$F_{215}^{obs} = 215 \times \left(\left(\frac{\left(\frac{110}{215} - \frac{1}{2}\right)^2}{\frac{1}{2}} \right) + \left(\frac{\left(\frac{105}{215} - \frac{1}{2}\right)^2}{\frac{1}{2}} \right) \right) \approx 0,12.$$

Par ailleurs, sous l'hypothèse H_0, la statistique théorique du χ^2 suit une loi du χ^2 à $(2-1) = 1$ degré de liberté. D'après les données numériques du verso de l'énoncé, on a $\mathbb{P}(\chi^2(1) > 2,71) = 90\%$ de sorte que la région de rejet du test d'indépendance est $W = \{F_{215} > 2,71\}$. Ici, on a clairement $F_{215}^{obs} \notin W$ de sorte que l'on ne rejette pas rejette l'hypothèse H_0 au niveau 90%.

(b) Répondre à question en effectuant un test basé sur une approximation gaussienne de la loi binomiale. 2 points
Si dans la population hypertendue, on désigne encore par p la proportion d'individus issus du pays A, il s'agit de tester l'hypothèse H_0 : "$p = 1/2$" contre l'hypothèse H_1 : "$p \neq 1/2$". Ici la fréquence empirique observée est $\widehat{p}^{obs} = 110/215$. Sous l'hypothèse H_0, le nombre de personnes issues de A dans la population hypertendue suit une loi binomiale $B(215, 1/2) \approx \mathcal{N}((215/2, 215/4)$ et la fréquence théorique suit donc approximativement une loi $\mathcal{N}(1/2, 1/(4 \times 215))$. Autrement dit, sous H_0,

$$2\sqrt{215}\left(\widehat{p} - \frac{1}{2}\right) \sim \mathcal{N}(0,1).$$

D'après les données numériques $\mathbb{P}(|\mathcal{N}(0,1)| > 1,64) = 10\%$ de sorte que sous H_0 :

$$\mathbb{P}\left(\widehat{p} \notin \left[\frac{1}{2} - \frac{1,64}{2\sqrt{215}}, \frac{1}{2} + \frac{1,64}{2\sqrt{215}}\right]\right) = 10\%.$$

Autrement dit l'ensemble

$$W = [0,1] \backslash \left[\frac{1}{2} - \frac{1,64}{2\sqrt{215}}, \frac{1}{2} + \frac{1,64}{2\sqrt{215}}\right] \approx [0,1] \backslash [0.44,\ 0.56] \approx [0,\ 0.44] \cup [0.56,\ 1]$$

constitue une zone de rejet pour le test au niveau 90%. Ici $\widehat{p}^{obs} = 110/215 \approx 0,51$ de sorte que $\widehat{p}^{obs} \notin W$ et comme plus haut, on ne rejette pas l'hypothèse H_0 au niveau de confiance 90%.