CPES 3 : Examen Statistique 2020-2021

Documents et calculatrice interdits. Le barème indiqué n'est pas définitif.

Exercice 1 : questions de cours (/5)

On répondra aux questions suivantes en une phrase ou une formule (sans justification).

1. Soit $X_1, \ldots, X_n$ une suite i.i.d. de variables aléatoires suivant une loi de Poisson de paramètre λ c'est-à-dire que pour tout entier $k \geq 1$, $\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$. . Écrire la fonction de vraisemblance associée à ces observations.

   $L(\lambda; x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1) \times \ldots \times \mathbb{P}(X_n = x_n) = e^{-\lambda}\frac{\lambda^{x_1}}{x_1!} \times \ldots \times e^{-\lambda}\frac{\lambda^{x_n}}{x_n!} = e^{-\lambda n}\frac{\lambda^{x_1+\ldots+x_n}}{x_1! \ldots x_n!}$,

   donc, la fonction de vraisemblance est

   $$L(\lambda; X_1, \ldots, X_n) = e^{-\lambda n}\frac{\lambda^{X_1+\ldots+X_n}}{X_1! \ldots X_n!}.$$

2. Soient $X_1, \ldots, X_n$ des observations d'une quantité positive. Comment s'appelle la courbe passant par les points $(\widehat{F}(x), \widehat{FQ}(x))$ où $\widehat{F}(x)$ est un estimateur de la fonction de répartition et

   $$\widehat{FQ}(x) = \frac{\sum_{i=1}^n X_i \mathbf{1}_{\{X_i \leq x\}}}{\sum_{i=1}^n X_i} \quad ?$$

   Il s'agit de la courbe de Lorenz.

3. Supposons que les observations de la question précédente sont les salaires des n salariés d'une entreprise. Que représente la quantité $\widehat{FQ}(0.5)$?

   $\widehat{FQ}(0.5)$ représente la part de masse salariale portée par les 50% de salariés les moins payés.

4. Soit F la fonction de répartition d'une variable aléatoire X. Donner la définition générale du quantile d'ordre α.

   $q_\alpha = \inf\{x; F(x) \geq \alpha\}$

5. Soit $X_1, \ldots, X_n$ une suite i.i.d. suivant la loi de Bernoulli de paramètre θ. Écrire la forme de la zone de rejet du test $\mathcal{H}_0 : \theta = 0.3$ contre $\mathcal{H}_1 : \theta \neq 0.3$ en fonction de l'estimateur usuel $\widehat{\theta}$ de θ et d'un seuil t_0. Nous ne demandons pas ici la valeur de t_0.

   $\mathcal{R} = \{|\widehat{\theta} - 0.3| \geq t_0\}$ où $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$.

Exercice 2 : comparaison de moyennes d'échantillons gaussiens (/10)

Soient $X_1, \ldots, X_n \sim_{i.i.d.} \mathcal{N}(\mu_1, \sigma^2)$ et $Y_1, \ldots, Y_n \sim_{i.i.d.} \mathcal{N}(\mu_2, \sigma^2)$ deux échantillons indépendants. L'objectif de cet exercice est de comparer deux tests statistiques des hypothèses :

$$\mathcal{H}_0 : \mu_1 = \mu_2 \text{ contre } \mathcal{H}_1 : \mu_1 \neq \mu_2.$$

1. (0.5 point) Écrire $\widehat{\mu}_1$ et $\widehat{\mu}_2$ les estimateurs usuels de μ_1 et μ_2.

   $\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n X_i$, $\widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n Y_i$.

2. (2 points : 0.5 loi, 0,5 moyenne, 1 variance) Quelle est la loi de $\widehat{\mu}_1 - \widehat{\mu}_2$? On précisera (en justifiant) la moyenne et la variance en fonction de μ_1, μ_2, n et σ^2.

   Comme $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ sont indépendants et de loi normale, $\widehat{\mu}_1 - \widehat{\mu}_2$, qui est combinaison linéaire de $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ suit une loi normale. On a

   $$\mathbb{E}[\widehat{\mu}_1 - \widehat{\mu}_2] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i] = \mu_1 - \mu_2.$$

   et, comme $X_1 - Y_1, \ldots, X_n - Y_n$ est une suite de vecteurs indépendants

   $$\mathrm{Var}(\widehat{\mu}_1 - \widehat{\mu}_2) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n Y_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n (X_i - Y_i)\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i - Y_i).$$

   Comme, de plus, pour tout i, X_i et Y_i sont indépendants,

   $$\mathrm{Var}(X_i - Y_i) = \mathrm{Var}(X_i) + \mathrm{Var}(Y_i) = \sigma^2 + \sigma^2 = 2\sigma^2.$$

   D'où

   $$\widehat{\mu}_1 - \widehat{\mu}_2 \sim \mathcal{N}(\mu_1 - \mu_2, 2\frac{\sigma^2}{n}).$$

3. (1 point : 0.5 pour la loi de T_1 sous H_0, 0.25 pour la loi de T_1 sous H_1, 0.25 pour la réponse à la dernière question) En déduire que, sous $\mathcal{H}_0$, la quantité

   $$T_1 = \frac{\sqrt{n}(\widehat{\mu}_1 - \widehat{\mu}_2)}{\sqrt{2}\sigma} \sim \mathcal{N}(0, 1).$$

   Quelle est sa loi sous $\mathcal{H}_1$. Peut-on utiliser T_1 comme statistique de test ?

D'après 2.

$$T_1 \sim \mathcal{N}\left(\frac{\sqrt{n}(\mu_1-\mu_2)}{\sqrt{2}\sigma}, 1\right)).$$

Sous $\mathcal{H}_0$, $\mu_1 - \mu_2 = 0$ donc on obtient bien la loi donnée. Sous $\mathcal{H}_1$, la moyenne de T_2 est différente de 0 et tend vers $+\infty$ lorsque $n \to +\infty$. T_1 dépend de σ inconnu donc ne peut pas être utilisé comme statistique de test.

4. Soit $\alpha \in]0,1[$, sans justifier les différentes étapes du raisonnement mais en définissant bien toutes les notations, donner des intervalles de confiances asymptotiques au niveau α pour

   (a) (1 point) μ_1

   $$IC_{\mu_1} = \left[\widehat{\mu}_1 \pm \frac{\widehat{\sigma}_1}{\sqrt{n}}\phi^{-1}(1-\alpha/2)\right],$$

   où $\widehat{\sigma}_1^2$ est, par exemple, défini dans l'énoncé de la question 5. et Φ est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

   (b) (1 point) $\mu_1 - \mu_2$.

   $$IC_{\mu_1-\mu_2} = \left[\widehat{\mu}_1 - \widehat{\mu}_2 \pm \frac{\widehat{\sigma}_{1,2}}{\sqrt{n}}\phi^{-1}(1-\alpha/2)\right],$$

   où $\widehat{\sigma}_{1,2}^2 = \widehat{\sigma}_1^2 + \widehat{\sigma}_2^2$ avec $\widehat{\sigma}_2$ défini dans l'énoncé de la question 5.

5. Nous étudions maintenant la quantité :

   $$T_2 = \frac{\sqrt{n}(\widehat{\mu}_1 - \widehat{\mu}_2)}{\sqrt{\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2}},$$

   avec

   $$\widehat{\sigma}_1^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu}_1)^2 \text{ et } \widehat{\sigma}_2^2 := \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_2)^2.$$

   Nous admettons que, sous $\mathcal{H}_0$, T_2 converge en loi vers la loi normale centrée réduite $\mathcal{N}(0,1)$.

   (a) (1 point) Pour quelle(s) valeur(s) de t_0 a-t'on

   $$\lim_{n\to+\infty} \mathbb{P}_{\mathcal{H}_0}(|T_2| \geq t_0) = \alpha?$$

   Comme, sous $\mathcal{H}_0$, T_2 tend en loi vers une loi normale centrée réduite, nous avons :

   $$\lim_{n\to+\infty} \mathbb{P}_{\mathcal{H}_0}(|T_2| \geq t_0) = \mathbb{P}(|Z| \geq t_0),$$

avec $Z \sim \mathcal{N}(0,1)$. On cherche les valeurs de t_0 pour lesquelles $\mathbb{P}(|Z| \geq t_0) = \alpha$. On a, en utilisant la symétrie de la loi normale,

$$\begin{aligned}\mathbb{P}(|Z| \geq t_0) &= \mathbb{P}(\{Z \geq t_0\} \cup \{Z \leq -t_0\}) = \mathbb{P}(Z \geq t_0) + \mathbb{P}(Z \leq -t_0) \\ &= \mathbb{P}(Z \geq t_0) + \mathbb{P}(-Z \leq -t_0) = 2\mathbb{P}(Z \geq t_0) = 2(1 - \mathbb{P}(Z < t_0)) \\ &= 2(1 - \phi(t_0)).\end{aligned}$$

Donc

$$\mathbb{P}(|Z| \geq t_0) = \alpha \Leftrightarrow 2(1 - \phi(t_0)) = \alpha \Leftrightarrow \phi(t_0) = 1 - \alpha/2 \Leftrightarrow t_0 = \phi^{-1}(1 - \alpha/2).$$

(b) (0.5 point) En déduire un test asymptotique de niveau α de

$$\mathcal{H}_0 : \mu_1 = \mu_2 \text{ contre } \mathcal{H}_1 : \mu_1 \neq \mu_2.$$

$$\varphi(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = \mathbf{1}_{\{|T_2| \geq \phi^{-1}(1-\alpha/2)\}}.$$

6. Nous admettrons maintenant que

$$(n-1)\frac{\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2}{\sigma^2} \sim \chi^2(2n-2)$$

et que $\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2$ est indépendant de $\widehat{\mu}_1 - \widehat{\mu}_2$.

(a) (1 point) En déduire la loi de T_2.

D'après 1., $T_1 \sim \mathcal{N}(0,1)$ et d'après les résultats admis de l'exercice $U = (n-1)\frac{\widehat{\sigma}_1^2+\widehat{\sigma}_2^2}{\sigma^2} \sim \chi^2(2n-2)$, et U et T_1 sont indépendantes donc

$$\frac{T_1}{\sqrt{U/(2n-2)}} \sim t(2n-2).$$

Or

$$\frac{T_1}{\sqrt{U/(2n-2)}} = \frac{\frac{\sqrt{n}(\widehat{\mu}_1 - \widehat{\mu}_2)}{\sqrt{2}\sigma}}{\sqrt{(n-1)\frac{\widehat{\sigma}_1^2+\widehat{\sigma}_2^2}{\sigma^2}/(2n-2)}} = T_2.$$

Donc T_2 suit une loi de Student à $2n-2$ degrés de liberté.

(b) (1 point) Trouver u_0 en fonction d'un quantile d'une loi à préciser tel que

$$\mathbb{P}_{\mathcal{H}_0}(|T_2| \geq u_0) = \alpha.$$

Comme la loi $\mathcal{N}(0,1)$ est symétrique, la loi de Student est symétrique aussi, donc, par le même raisonnement que dans la question 5.(a)., nous trouvons $u_0 = q^t_{2n-2,1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi de Student à $2n-1$ degrés de liberté.

(c) (0.5 point) En déduire un test au niveau α de

$$\mathcal{H}_0 : \mu_1 = \mu_2 \text{ contre } \mathcal{H}_1 : \mu_1 \neq \mu_2.$$

$$\varphi(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = \mathbf{1}_{\{|T_2| \geq q^t_{2n-2,1-\alpha/2}\}}.$$

7. (0.5 point) Quel est l'avantage du test défini à la question 6.(c) par rapport à celui défini à la question 5.(b).

   Le test défini à la question 6.(c) est de niveau α quel que soit le nombre d'observations n alors que le test de la question 5.(b) n'est de niveau α que quand $n \to +\infty$.

Exercice 3 : régression linéaire (/5)

Nous souhaitons vérifier une méthode de prédiction du rapport capital privé/revenu national d'un pays sur une année à partir d'observations sur les années précédentes. La prédiction est faite sur l'année 2010 et nous vérifions l'adéquation de la prédiction avec les observations.

Nous notons :

- O_i : le rapport capital privé/revenu national observé en 2010 pour le i-ème pays,
- P_i : le rapport capital privé/revenu national prédit pour 2010 pour le i-ème pays à partir d'observations réalisées sur les années antérieures.

	Observations	Prédictions
Etats-Unis	410%	400%
Japon	601%	616%
Allemagne	412%	510%
France	575%	526%
Royaume-Uni	522%	371%
Italie	676%	644%
Canada	416%	438%
Australie	518%	419%

Table 1: Rapports capital privé/revenu national observés et prédits en 2010. Source : *Le capital au 21e siècle*, T. Piketty, Éditions du Seuil, Septembre 2013, annexe technique sur le site de l'auteur piketty.pse.ens.fr/fr/capital21c.

Nous supposons vérifiée la relation suivante :

$$P_i = \beta O_i + \varepsilon_i,$$

avec $\varepsilon_1, \ldots, \varepsilon_n \sim_{i.i.d} \mathcal{N}(0, \sigma^2)$ et β est un paramètre inconnu. Les valeurs observées $O_1, \ldots, O_n$ sont supposées non aléatoires.

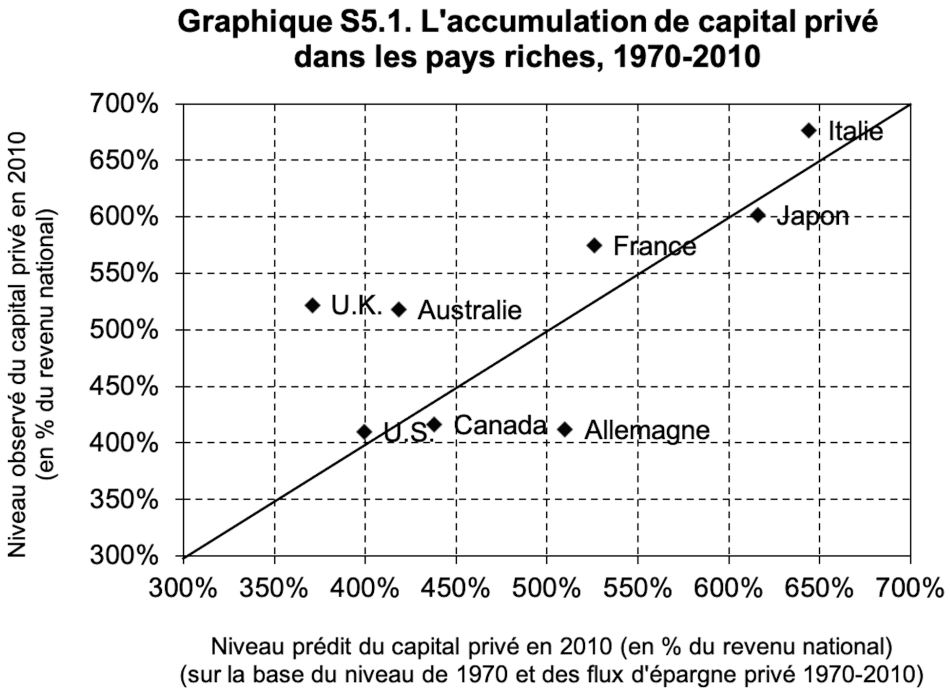

Figure 1: Représentation des données du Tableau 1.

	$p = 95\%$	$p = 97.5\%$	$p = 99\%$	$p = 99.5\%$
$n = 6$	1.94	2.45	3.14	3.71
$n = 7$	1.89	2.36	3.00	3.50
$n = 8$	1.86	2.31	2.90	3.36

Table 2: Tableaux des quantiles $q^t_{n,p}$ d'ordre p de la loi de Student à n degrés de liberté.

1. (1 point) Nous supposons que, pour tout $i = 1, \dots, n$, $\mathbb{E}[P_i] = O_i$. Que vaut β ?

$$\mathbb{E}[P_i] = \mathbb{E}[\beta O_i + \varepsilon_i] = \beta O_i + \mathbb{E}[\varepsilon_i] = \beta O_i = O_i.$$

Prenons par exemple $i = 1$, d'après le tableau 1, $O_1 \neq 0$ donc $\beta = 1$.

2. (a) (1 point) Quelles conclusions sur les données peut-on déduire d'un test ayant pour hypothèses nulle et alternative

$$\mathcal{H}_0 : \beta = 1 \text{ contre } \mathcal{H}_1 : \beta \neq 1?$$

Si nous rejetons $\mathcal{H}_0$ alors nous pouvons conclure que les prédictions ne sont pas en adéquation avec les observations.

(b) (1 point) Écrire, en fonction de β les hypothèses nulles et alternatives d'un test permettant de vérifier que les prédictions ont tendance à surestimer la valeur réelle des observations.

$$\mathcal{H}_0 : \beta = 1 \text{ contre } \mathcal{H}_1 : \beta > 1.$$

3. Les calculs faits à partir des observations nous donnent les valeurs suivantes :

$$\bar{P} = \frac{1}{8}\sum_{i=1}^{8} P_i = 490.50, \quad \bar{O} = \frac{1}{8}\sum_{i=1}^{8} O_i = 516.25$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{8}(O_i - \bar{O})(P_i - \bar{P})}{\sum_{i=1}^{8}(O_i - \bar{O})^2} = 1.04 \quad (\pm 0.01).$$

$$\widehat{\sigma}^2 = \frac{1}{6}\sum_{i=1}^{8}(P_i - \widehat{\beta} O_i)^2 = 7219.509.$$

$$T = \frac{\widehat{\beta} - 1}{\sqrt{\widehat{\sigma}^2 / \sum_{i=1}^{8}(O_i - \bar{O})^2}} = 18.67 \quad (\pm 0.01).$$

(a) (1 point + 0.5 bonus si le degré de liberté de la loi est correct) À partir des résultats présentés ci-dessus et des valeurs des quantiles présentés dans le tableau 2, donner la conclusion des deux tests de la question précédente. Aucune justification mathématique n'est demandée dans cette question, vous expliciterez simplement les valeurs numériques sur lesquelles vous vous appuyez pour donner votre conclusion.

Sous $\mathcal{H}_0$, T suit une loi de Student à $n-1 = 7$ degrés de liberté, nous comparons $|T|$ au quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n-1$ degrés de liberté pour le test de la question 2.(a) et T au quantile d'ordre $1-\alpha$ de la loi de Student à $n-1$ degrés de liberté pour le test de la question 2.(b). La valeur de $|T|$ étant supérieure aux quantiles d'ordre 99.5% de la loi, nous rejetons $\mathcal{H}_0$ au niveau $\alpha = 1\%$ pour le test de la question 2.(a). De même, nous rejetons $\mathcal{H}_0$ au niveau $\alpha = 0.5\%$ pour le test de la question 2.(b).

(b) (1 point) Comment changent les valeurs des quantités $\bar{P}$, $\bar{O}$, $\widehat{\beta}$, $\widehat{\sigma}^2$ et T lorsque les données ne sont pas exprimées en pourcentages (c'est-à-dire 3 au lieu de 300%) ? Les conclusions des tests changent-elles ?

Si nous n'exprimons pas les données en pourcentage, toutes les valeurs numériques du tableau sont divisées par 100, cela ne change pas la valeur de $\widehat{\beta}$, par contre $\widehat{\sigma}^2$ est divisé par 100^2, donc $\widehat{\sigma}$ est divisé par 100. La valeur de T ne change pas et sa loi est la même donc les conclusions du test sont inchangées.