CPES 3 : Examen Statistique 2020-2021

Documents et calculatrice interdits.

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 $k \geq 0$, $\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$. . Écrire la fonction de vraisemblance associée à ces observations.

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

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

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

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.

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. Écrire $\widehat{\mu}_1$ et $\widehat{\mu}_2$ les estimateurs usuels de μ_1 et μ_2.

2. 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.

3. En déduire que, sous $\mathcal{H}_0$,

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

   Quelle est la loi de T_1 sous $\mathcal{H}_1$. Peut-on utiliser T_1 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) μ_1

   (b) $\mu_1 - \mu_2$.

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

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

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

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

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) En déduire la loi de T_2.

   (b) Pour quelle(s) valeur(s) de u_0 a-t'on

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

   (c) En déduire un test au niveau α de

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

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

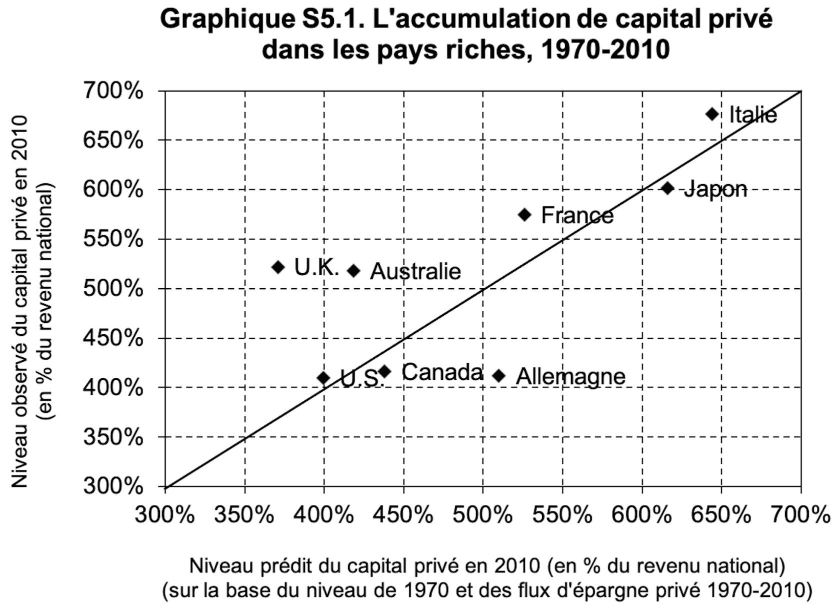

Figure 1: Représentation des données du Tableau 1.

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.

	$p = 95\%$	$p = 97.5\%$	$p = 99\%$	$p = 99.5\%$
$n = 6$	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: Tableau des quantiles $q^t_{n,p}$ d'ordre p de la loi de Student à n degrés de liberté.

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.

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

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

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

   (b) 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.

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

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

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

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

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

   (a) À partir des résultats présentés ci-dessus et des valeurs des quantiles du 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.

   (b) 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 ?