T.D. 4 Statistique

Théorème de Student

Exercice 1. Dans l'atmosphère, le taux d'un gaz nocif, pour un volume donné, suit une loi normale d'espérance μ et de variance σ^2 inconnues. On effectue n prélèvements conduisant aux valeurs $X_1, X_2, \ldots, X_n$.

(a) Sur un échantillon de taille $n = 10$, on observe que $\overline{X} = 50$ et $S^2 = 100$ où $S^2 = \frac{1}{9}\sum_{i=1}^{10}(X_i - \overline{X})^2$.

(b) Quel est l'intervalle de confiance de niveau de confiance 95 % du taux moyen μ de gaz dans l'atmosphère ?

(c) Construire un intervalle de confiance pour σ^2 de niveau de confiance 95%.

Exercice 2. On s'interroge sur la comparaison des tailles moyennes des garçons et des filles de 6 ans dans une population; pour cela on a pris comme échantillon, jugé représentatif de cette tranche d'âge, une classe d'école primaire (niveau CP en France), et on a observé :
– $n_1 = 16$ garçons : moyenne $126,5$ cm, écart-type 12,9 cm
– $n_2 = 15$ filles : moyenne 136,9 cm, écart-type 11,9 cm.
On admet que la distribution des tailles dans chacune des sous-populations (garçons, filles) suit une loi gaussienne.

(a) Donner des intervalles de confiance pour les tailles moyennes des garçons et des filles.

(b) Donner un intervalle de confiance pour l'écart type de la taille des garçons. Même question pour les filles.

(c) Sur la base de la réponse à la question précédente, on suppose que la variance est la même dans les deux populations et vaut σ^2. Montrer que

$$\hat{\sigma}^2 = \frac{1}{n_1 + n_2 - 2}\Big(\sum_{i=1}^{n_1}(Y_i - \bar{Y})^2 + \sum_{i=1}^{n_2}(X_i - \bar{X})^2\Big)$$

est un estimateur sans biais de σ^2.

(d) Montrer que $\hat{\sigma}^2$ est indépendant de $\bar{X} - \bar{Y}$.

(e) En déduire que la variable aléatoire

$$T = \big(\frac{1}{n_1} + \frac{1}{n_2}\big)^{-1/2}\frac{\bar{X} - \bar{Y} - \Delta}{\hat{\sigma}}$$

suit une loi de Student dont on précisera le nombre de degrés de liberté.

(f) Construire un intervalle de confiance de niveau de confiance 95% pour la différence entre la taille moyenne des filles et celle des garçons.

Test statistique

Exercice 3 (Modèle binaire)**.** Soit $(\Omega, \mathcal{F})$ un espace mesurable et $X : \Omega \to \mathbb{R}^n$ un vecteur aléatoire à valeurs dans $\mathbb{R}^n$. On suppose que P_{θ_0} et P_{θ_1} sont deux probabilités sur $(\Omega, \mathcal{F})$. On observe une réalisation de X sous P_θ où $\theta \in \{\theta_0, \theta_1\}$ est inconnu. On s'intéresse au problème de test suivant: au vu de l'observation X on veut décider si $\theta = \theta_0$ ou si $\theta = \theta_1$.
Une fonction de test est une fonction $\varphi : \mathbb{R}^n \to \{0, 1\}$ à laquelle on associe la procédure de décision suivante : si on observe $\varphi(X) = 1$ on décide que $\theta = \theta_0$, si on observe $\varphi(X) = 0$ on décide que $\theta = \theta_0$.
La question est: comment choisir φ? A une fonction de test φ on associe deux probabilités d'erreur:

$$\alpha(\varphi) = P_{\theta_0}\big(\varphi(X) = 1\big) \quad \text{et} \quad \beta(\varphi) = P_{\theta_1}\big(\varphi(X) = 0\big).$$

On va construire une fonction de test qui minimise $\alpha(\varphi)+\beta(\varphi)$ ou plus généralement $\gamma\alpha(\varphi)+\beta(\varphi)$ où $\gamma > 0$ est un paramètre donné, en faisant l'hypothèse suivante: sous P_{θ_0} le vecteur aléatoire X admet comme densité $f_0 : \mathbb{R}^n \to \mathbb{R}_+$, sous P_{θ_1} le vecteur aléatoire X admet comme densité $f_1 : \mathbb{R}^n \to \mathbb{R}_+$.

(a) Soit $\gamma > 0$. Montrer que pour toute fonction de test $\varphi : \mathbb{R}^n \to \{0, 1\}$ on a

$$\gamma\alpha(\varphi) + \beta(\varphi) = 1 + \int_{\mathbb{R}^n} \varphi(x)\big(\gamma f_0(x) - f_1(x)\big)dx.$$

En déduire une fonction de test φ_γ telle que $\gamma\alpha(\varphi_\gamma) + \beta(\varphi_\gamma) \leq \gamma\alpha(\varphi) + \beta(\varphi)$ pour tout φ.

(b) On suppose que $n = 1$ et $X \sim \mathcal{N}(\theta, 1)$ sous P_θ. Calculer φ_1.

(c) Soit $\alpha \in [0, 1]$. On suppose qu'il existe $\gamma_\alpha > 0$ tel que $\alpha(\varphi_{\gamma_\alpha}) = \alpha$. Montrer que pour tout φ telle que $\alpha(\varphi) \leq \alpha$ on a $\beta(\varphi_{\gamma_\alpha}) \leq \beta(\varphi)$.

(d) On observe $X_1, \ldots, X_n$ i.i.d., tous de loi $\mathcal{N}(\theta, 1)$. Pour $\alpha \in]0, 1[$, déterminer φ_{γ_α}.

Exercice 4. On observe $X_1, \ldots, X_n$ i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$ où $\mu \in \mathbb{R}$ est inconnue et $\sigma^2 > 0$ est connu. On s'intéresse au problème de test pour les hypothèses suivantes:
$H_0 : \mu \leq \mu_0$ contre $H_1 : \mu > \mu_0$ où $\mu_0 \in \mathbb{R}$ est donnée. H_0 est donc l'hypothèse nulle et par définition un test de niveau $\alpha \in [0, 1]$ est une fonction $\varphi : \mathbb{R}^n \to \{0, 1\}$ telle que $\sup_{\mu \leq \mu_0} P_\mu(\varphi(X_1, \ldots, X_n) = 1) \leq \alpha$.

(a) Soit $\alpha \in]0, 1[$. Déterminer un test φ^* de niveau α tel que pour tout autre test φ de niveau α on a pour tout $\mu > \mu_0$: $P_\mu(\varphi^* = 0) \leq P_\mu(\varphi = 0)$.

(b) On suppose que σ^2 est aussi inconnu. On considère la statistique

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\hat{\sigma}_n}.$$

Soit $\alpha \in]0, 1[$. Déterminer la constante C_α tel que $P_{\mu_0}(T > C_\alpha) = \alpha$. Montrer que le test $\varphi(X_1, \ldots, X_n) = 1_{\{T > C_\alpha\}}$ est de niveau α.

Intervalle de confiance

Exercice 5. Un gramme de carbone extrait d'un organisme mort contient un nombre θ d'atomes de carbone C14. Ces atomes de C14 ne sont pas stables et on modélise leur durée de vie par des variables aléatoires $X_1, X_2, \ldots, X_\theta$ i.i.d. de loi commune la loi exponentielle de paramètre 1. Dans la situation où θ est inconnu et que l'on veut l'estimer, on compte le nombre de désintégrations sur l'intervalle de temps $[0, t_0]$ où $t_0 > 0$ est une constante choisie, autrement dit on observe la variable aléatoire

$$S = \sum_{i=1}^{\theta} 1_{\{X_i \leq t_0\}}.$$

(a) Déterminer la loi de S c'est-à-dire calculer $P(S = k)$ pour tout entier k.

(b) Calculer l'espérance et la variance de S en fonction de θ et de t_0.

(c) Démontrer que pour tout $\varepsilon > 0$, pour tout $\theta \in \mathbb{N}$ on a

$$P_\theta\Big(|\frac{S}{1-e^{-t_0}} - \theta| > \varepsilon\sqrt{\theta}\Big) \leq \frac{e^{-t_0}}{1-e^{-t_0}}\frac{1}{\varepsilon^2}.$$

(d) En déduire un intervalle de confiance pour θ de niveau d'erreur $\alpha \in]0, 1[$.

Exercice 6. On dispose de l'observation de n_1 variables aléatoires $X_1, \ldots, X_{n_1}$ indépendantes, de loi de Bernoulli de paramètre $p_1 \in [0, 1]$. On observe aussi $Y_1, \ldots, Y_{n_2}$ indépendantes, de loi de Bernoulli de paramètre $p_2 \in [0, 1]$, et indépendantes des X_i.

On souhaite construire un intervalle de confiance pour $p_1 - p_2$, les paramètres p_1 et p_2 étant inconnus.

(a) Construire un estimateur T satisfaisant pour tout $\varepsilon > 0$:

$$P\Big(\big|T - p_1 + p_2\big| \geq \varepsilon\Big) \leq \frac{1}{4\varepsilon^2}(\frac{1}{n_1} + \frac{1}{n_2}).$$

En déduire la construction d'un intervalle de confiance, noté $I_1(\alpha)$, de niveau d'erreur $\alpha \in]0, 1[$ pour $p_1 - p_2$.

(b) A l'aide de l'inégalité de Hoeffding (*qui est rappelée plus bas*), construire un estimateur S tel que pour tout $\varepsilon > 0$:

$$P\Big(\big|S - p_1 + p_2\big| \geq \varepsilon\Big) \leq 2\exp\Big(-2\varepsilon^2 n_1 n_2/(n_1 + n_2)\Big).$$

En déduire la construction d'un intervalle de confiance, noté $I_2(\alpha)$, de niveau d'erreur $\alpha \in]0, 1[$ pour $p_1 - p_2$.

(c) On a obtenu (avec $n_1 = 30$, $n_2 = 25$) les statistiques

$$\frac{1}{n_1}\sum_{i=1}^{n_1} X_i = 0,6 \qquad \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i = 0,75$$

Comparer les longueurs de $I_1(\alpha)$ et $I_2(\alpha)$ pour $\alpha = 30\%$.

Inégalité de Hoeffding :
Soit $\xi_1, \ldots, \xi_n$ des variables aléatoires réelles indépendantes telles que pour tout $i \in \{1, \ldots, n\}$ on a $E(\xi_i) = 0$ et il existe des constantes $a_i, b_i \in \mathbb{R}$ satisfaisant $a_i \leq \xi_i \leq b_i$. Alors on a

$$\forall t \geq 0 \quad P(|\xi_1 + \cdots + \xi_n| \geq t) \leq 2\exp\big(-2t^2 \,/\, \sum_{i=1}^{n}(b_i - a_i)^2\big).$$

Exercice 7. Soient $X_1, \ldots, X_n$, i.i.d. de loi

$$\nu_\theta(dx) = \exp(\theta - x)1_{[\theta,+\infty[}(x)dx, \;\; \theta > 0.$$

(a) Ecrire le modèle statistique associé. Est-il identifiable ?
(b) Calculer $\mathrm{E}_\theta^n\{X_1\}$ et en déduire un estimateur convergent de θ que l'on notera $\hat{\theta}_n$.
(c) Etudier le risque quadratique $\mathrm{E}_\theta^n\{(\hat{\theta}_n - \theta)^2\}$ de l'estimateur $\hat{\theta}_n$. En déduire un intervalle de confiance I_n pour θ au niveau de risque $0 < \alpha < 1$.
(d) Montrer que l'estimateur $\theta_n^\star := \min_{1\leq i\leq n} X_i$ est meilleur que $\hat{\theta}_n$ au sens du risque quadratique.
(e) Construire à l'aide de $\theta_n^\star$ un intervalle de confiance pour θ au niveau de risque α, dont la longueur est plus petite que celle de I_n.

Exercice 8. Pour $a > 0$ et $b > 0$, la densité $\gamma_{(a,b)}$ de la loi Gamma(a, b) est définie par

$$\gamma_{(a,b)}(x) = \frac{1}{\Gamma(a)} b^a \exp(-bx)x^{a-1}1_{\mathbb{R}_+^*}(x) \quad \text{où} \quad \Gamma(a) = \int_0^\infty e^{-x}x^{a-1}dx.$$

(a) Soit X de loi Gamma(a, b), calculer $\mathrm{E}(X)$ et $\mathrm{Var}(X)$.
(b) Calculer la transformée de Laplace $t \rightsquigarrow \mathrm{E}(e^{-tX})$ de la loi Gamma.
(c) On considère deux variables aléatoires indépendantes X_1 et X_2, de loi Gamma(a_1, b) et Gamma(a_2, b). Déduire du calcul de la transformée de Laplace de $X_1 + X_2$ la loi de $X_1 + X_2$.
(d) Soit $(X_1, \ldots, X_n)$ un n-échantillon de la loi $\mathcal{N}(0, 1)$. Quelle est la densité de $X_1^2 + \ldots + X_n^2$? (On pourra d'abord reconnaître la loi de X_1^2 via le calcul de sa transformée de Laplace.)

Exercice 9. La durée de vie d'une ampoule électrique est une variable aléatoire de loi $\mathcal{E}(\lambda) = \text{Gamma}(1, \lambda)$, la loi exponentielle de paramètre $\lambda \in \mathbb{R}_+^*$. On mesure la durée de vie de n ampoules, on obtient donc une réalisation des variables aléatoires $X_1, \ldots, X_n$, que l'on supposera indépendantes.

(a) Que vaut l'espérance de la durée de vie d'une ampoule ? Calculer le risque quadratique de la statistique $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ en tant qu'estimateur de $1/\lambda$.
(b) On veut estimer λ. Calculer le risque quadratique de l'estimateur $1/\bar{X}$.
(c) Soit $\alpha \in]0, 1[$. Construire un intervalle de confiance asymptotique pour λ de niveau d'erreur α.

Exercice 10. Soient X une variable aléatoire gaussienne centrée réduite, indépendante de la suite de variables aléatoires (Y_n) où Y_n suit la loi $\chi^2(n)$. On pose $T_n = \frac{X}{\sqrt{\frac{Y_n}{n}}}$.

(a) Montrer que la densité t_n de T_n est donnée par

$$t_n(x) = \frac{1}{\sqrt{\pi}} \frac{\Gamma((n+1)/2)}{\sqrt{n}\Gamma(n/2)} \frac{1}{(1+t^2/n)^{(n+1)/2}}.$$

(b) Montrer que pour tout $x \in \mathbb{R}$ $\lim_{n\to\infty} t_n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. On pourra tout d'abord montrer que $\Gamma(a+1/2)/\Gamma(a) \sim \sqrt{a}$ quand $a \to \infty$ en remarquant que $\Gamma(a+1/2)/\Gamma(a) = E[\sqrt{U}]$ où U suit une loi $Gamma(a,1)$.

(c) Montrer que (T_n) converge en loi et identifier sa limite.

Exercice 11. Dans l'atmosphère, le taux d'un gaz nocif, pour un volume donné, suit une loi normale d'espérance μ et de variance σ^2 inconnues. On effectue n prélèvements conduisant aux valeurs $X_1, X_2, \ldots, X_n$.

(a) Sur un échantillon de taille $n = 10$, on observe que $\overline{X} = 50$ et $S^2 = 100$ où $S^2 = \frac{1}{9}\sum_{i=1}^{10}(X_i - \overline{X})^2$.

(b) Quel est l'intervalle de confiance de niveau de confiance 95 % du taux moyen μ de gaz dans l'atmosphère ?

(c) Construire un intervalle de confiance pour σ^2 de niveau de confiance 95%.

Exercice 12. On cherche à évaluer la proportion $p \in [0,1]$ de personnes téléchargeant illégalement sur internet dans une population donnée. On choisit au hasard n individus dans cette population. La procédure de sondage est la suivante. On demande à chaque individu interrogé de tirer au préalable (sans être vu) une boule dans une urne, et de répondre ensuite par "OUI" ou "NON" à l'une des deux questions : Question 1 : "Est-ce que vous avez déjà téléchargé illégalement ?" si la boule tirée est blanche, Question 2 : "Est-ce que vous n'avez jamais téléchargé illégalement?" si la boule tirée n'est pas blanche. (bien sûr, seul l'individu interrogé a connaissance de la couleur de la boule, afin que personne d'autre ne sache à quelle question il a répondu). La proportion φ de boules blanches dans l'urne est connue et différente de 1/2. On note $X_1, \ldots, X_n$ les variables aléatoires qui correspondent aux réponses des n personnes intérrogées (ce sont donc des variables aléatoires à valeurs dans $\{Oui, Non\}$). Soit R le nombre de réponses "OUI" obtenues.

(a) Quelle est la loi de X_1 ?

(b) Quelle la loi de R ? Calculer $E(R)$.

(c) En déduire un estimateur sans biais $\hat{p}$ de p.

(d) Calculer le risque quadratique de l'estimateur $\hat{p}$. Commenter les cas $\varphi = 0$, $\varphi = 1$, $\varphi \to 1/2$.

(e) Construire à partir de $\hat{p}$ un intervalle de confiance pour p de niveau α.

Histogramme

Exercice 13. On observe $X_1, \ldots, X_n$ à valeurs dans $[0,1]$, tous de densité $f : [0,1] \to \mathbb{R}_+$. Pour $p \in \mathbb{N}^*$, on introduit les classes

$$A(p,k) = [(k-1)/p, k/p[, \quad 1 \le k \le p-1, \quad A(p,p) = [(p-1)/p, 1]$$

et l'histogramme associé

$$\hat{f}_{n,p}(x) = \frac{p}{n} \sum_{k=1}^{p} N(n,p)_k \, 1_{A(p,k)}(x), \quad x \in [0,1],$$

$$\text{où} \quad N(n,p)_k = \sum_{i=1}^{n} 1_{\{X_i \in A(p,k)\}}.$$

On suppose dans la suite que f est continûment dérivable.

(a) Montrer que $E_f^n[\hat{f}_{n,p}(x)]$ ne dépend de p. On le notera $\bar{f}_p(x)$.

(b) En utilisant que

$$\lim_{p\to\infty} p \max_{1\le k\le p} \sup_{x\in A(k,p)} \left| f(x) - f((k-1)/p) - f'((k-1)/p)(x-(k-1)/p) \right| = 0$$

montrer qu'il existe une constante $C_{f'}$ qui ne dépend que de f' telle que

$$p^2 \int_0^1 \left(f(x) - \bar{f}_p(x) \right)^2 dx \to C_{f'} \quad \text{quand } p \to \infty.$$

(c) Montrer que

$$\frac{n}{p} \int_0^1 \mathrm{Var}_{P_f^n}(\hat{f}_{n,p}(x)) dx \to 1 \quad \text{quand } (n,p) \to (\infty,\infty).$$

(d) On note $\mathcal{R}_{\hat{f}_{n,p}}(f)$ le risque quadratique de $\hat{f}_{n,p}$. Soit a et b strictement positifs. On pose $p_n = an^b$. Montrer que si f n'est pas constante, il existe $\alpha > 0$ tel que $n^\alpha \mathcal{R}_{\hat{f}_{n,p_n}}(f)$ converge vers une limite strictement positive.
Pour quelle valeur de b l'exposant α est-il le plus grand ? Pour ce choix de b, quelle est la meilleure valeur de a ?