TD n°3 – Intervalles et régions de confiance

stoehr@ceremade.dauphine.fr

Exercice 1 ***(À connaître – Un premier exemple simple).*** Pour estimer la précision d'un thermomètre, on réalise $n = 15$ mesures indépendantes de la température d'un liquide maintenu à température constante égale à $\mu = 20$ °C. Compte tenu des erreurs de mesure, on les considère comme des réalisations de variables aléatoires X_i, $i \in [\![1, n]\!]$, de même loi normale $\mathcal{N}(\mu, \sigma^2)$, où $\sigma \in \mathbb{R}_+^*$ est inconnu et caractérise la précision du thermomètre.

1. Donner l'estimateur du maximum de vraisemblance pour σ^2.
2. Préciser la loi de $\sum_{i=1}^n (X_i - \mu)^2 / \sigma^2$ et en déduire un intervalle de confiance bilatère pour σ^2 au niveau de confiance exactement $1 - \alpha$, $\alpha \in]0, 1[$.
3. **Application numérique.** On souhaite déterminer la valeur numérique de cet intervalle au niveau exactement 0.99 sachant que $\sum_{i=1}^n (X_i - \mu)^2 = 0.15$. Donner la commande R permettant de calculer les quantiles souhaités et en déduire la valeur numérique de l'intervalle de confiance.

Exercice 2 ***(À connaître – Un autre exemple d'intervalles de confiance exacts).*** Un coureur automobile doit choisir entre deux types de pneus A et B dont on a étudié l'usure. On note respectivement X et Y le nombre de kilomètres que peut parcourir un pneu de type A et un pneu de type B, avant usure. On suppose la modélisation statistique suivante : $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_p$ sont respectivement des échantillons *i.i.d.* des lois $\mathcal{N}(\mu_A, \sigma_A^2)$ et $\mathcal{N}(\mu_B, \sigma_B^2)$, avec $\mu_A, \mu_B, \sigma_A, \sigma_B \in \mathbb{R}_+^*$.

Partie I – σ_A^2 et σ_B^2 connus.

1. Donner la loi de $\overline{X}_n - \overline{Y}_p$ en fonction de μ_A, μ_B, σ_A et σ_B.
2. En déduire un intervalle de confiance bilatère de niveau exactement $1-\alpha$, $\alpha \in]0, 1[$, pour le paramètre $\mu_A - \mu_B$.
3. **Application numérique :** $\alpha = 0.05$, $n = 9$, $\overline{X}_9 = 20000$ km, $\sigma_A = 2000$ km, $p = 21$, $\overline{Y}_{21} = 25000$ km, $\sigma_B = 5000$ km.

Partie II – $\sigma_A^2 = \sigma_B^2 = \sigma^2$ inconnu.

1. Montrer que l'estimateur du maximum de vraisemblance de σ^2 s'écrit

$$\widehat{\sigma}^2 = \frac{1}{n+p} \left[\sum_{i=1}^n (X_i - \overline{X}_n)^2 + \sum_{i=1}^p (Y_i - \overline{Y}_p)^2 \right].$$

2. Montrer que $(n+p)\widehat{\sigma}^2/\sigma^2$ suit une loi du χ^2 dont on précisera le degré de liberté.

3. En utilisant la définition de la loi de Student, construire un intervalle de confiance bilatère de niveau exactement $1-\alpha$, $\alpha \in]0,1[$, pour le paramètre $\mu_A - \mu_B$.

4. **Application numérique :** $\alpha = 0.05$, $n = 9$, $p = 21$,

$$s_A^2 = \frac{1}{8}\sum_{i=1}^{9}(X_i - \overline{X}_9)^2 = 256 \times 10^4 \quad \text{et} \quad s_B^2 = \frac{1}{20}\sum_{i=1}^{2} 1(Y_i - \overline{Y}_{21})^2 = 25 \times 10^6.$$

Exercice 3. Soit $X = (X_1, \dots, X_n)$ un n-échantillon de loi exponentielle de paramètre $\theta \in \mathbb{R}_+^*$. Les variables X_i, $i \in [\![1,n]\!]$, ne sont pas observées. On observe uniquement les variables aléatoires $Y_1, \dots, Y_n$ définies pour $i \in [\![1,n]\!]$ par $Y_i = \mathbb{1}_{\{X_i > 2\}}$.

1. Écrire la vraisemblance des observations $(Y_1, \dots, Y_n)$ et calculer l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ de θ.

2. Montrer que $\widehat{\theta}_n$ est un estimateur consistant de θ.

3. Donner la loi limite de $\sqrt{n}(\widehat{\theta}_n - \theta)$ lorsque n tend vers $+\infty$.

4. En déduire un intervalle de confiance asymptotique bilatère pour θ de niveau exactement 95%.

Exercice 4. Soient $X \sim \mathscr{B}(n, p_1)$ et $Y \sim \mathscr{B}(n, p_2)$ des variables aléatoires indépendantes. On veut construire un intervalle de confiance pour $\Delta = p_1 - p_2$.

1. Calculer l'estimateur du maximum de vraisemblance pour p_1 et p_2, notés respectivement $\widehat{p}_1$ et $\widehat{p}_2$.

2. Déterminer la loi limite de $\left(\sqrt{n}(\widehat{p}_1 - p_1), \sqrt{n}(\widehat{p}_2 - p_2)\right)$ lorsque n tend vers $+\infty$.

3. En déduire la loi limite de $\sqrt{n}(\widehat{p}_1 - \widehat{p}_2 - (p_1 - p_2))$ lorsque n tend vers $+\infty$.

4. En déduire un intervalle de confiance asymptotique bilatère pour Δ au niveau exactement $1-\alpha$, $\alpha \in]0,1[$.

Exercice 5 *(**À connaître – Modèle linéaire gaussien**)*. La concentration d'ozone dans l'atmosphère dépend de la température à midi et de la nébulosité. On suppose ici que le lien entre la concentration d'ozone et ces deux quantités est affine. Si l'on note Y, X_2 et X_3 respectivement : la concentration d'ozone, la température et la nébulosité, on a donc la relation approchée suivante :

$$Y \approx \beta_1 + \beta_2 X_2 + \beta_3 X_3$$

pour certains paramètres inconnus $\beta_1, \beta_2, \beta_3 \in \mathbb{R}$. Cette relation est approximative parce qu'on imagine assez facilement qu'il peut y avoir des erreurs (de mesure par exemple). On modélise alors le problème en introduisant un terme d'erreur aléatoire que l'on suppose suivre une loi gaussienne centrée :

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Pour estimer les paramètres $(\beta_i)_{1\leq i\leq 3}$ (et σ^2), on dispose d'un échantillon $(x_i, y_i)_{1\leq i\leq n}$ avec $y_i \in \mathbb{R}$ et $x_i = (x_{i2}, x_{i3}) \in \mathbb{R}^2$ tels que, pour $i = 1,\dots,n$

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2). \tag{1}$$

Pour simplifier on suppose que les x_i sont déterministes (*i.e.*, non aléatoires). On suppose également que le niveau de bruit σ^2 est inconnu. Le vecteur de paramètre est donc $\theta = (\beta_1, \beta_2, \beta_3, \sigma^2)$. On va utiliser les notations suivantes :

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & x_{12} & x_{13} \\ \vdots & \vdots & \vdots \\ 1 & x_{n2} & x_{n3} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \quad \text{et} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

On suppose que les colonnes de la matrice $\mathbb{X}$ sont linéairement indépendantes.

1. Écrire le système (1) matriciellement et la vraisemblance du modèle.

2. Calculer l'estimateur du maximum de vraisemblance $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\sigma}^2_n)$ de θ.

3. Déterminer la loi de $\widehat{\beta}_n$.

4. Donner la loi de $\|Y - \mathbb{X}\widehat{\beta}_n\|^2 / \sigma^2$.

5. Donner la loi de $\|\mathbb{X}(\widehat{\beta}_n - \beta)\|^2 / \sigma^2$. En déduire une région de confiance pour β au niveau exactement $1-\alpha$, $\alpha \in]0,1[$, quand σ est connu.

6. Donner une région de confiance pour β au niveau exactement $1-\alpha$, $\alpha \in]0,1[$ dans le cas (plus réaliste) où σ est inconnu.