Probabilités et statistique

Série 10

Estimateurs et intervalles de confiance (solutionnaire)

Exercice 1. Quelle est la différence entre un paramètre et un estimateur ?

Solution. Un paramètre est une caractéristique réelle de la population, en particulier c'est un nombre. Un estimateur est une variable aléatoire qui tente d'estimer la vraie valeur du paramètre à partir d'un échantillon.

Exercice 2. Pour chacun des paramètres des lois suivantes, proposer un estimateur sans biais.

a) $N(\mu, \sigma^2)$
b) exponentielle(λ)
c) Poisson(θ)
d) binomiale(n,p) (n connu)

Exercice 3. Soit $X_1, \dots, X_n$ iid suivant une loi $N(\mu_1,\sigma^2)$ et $Y_1, \dots, Y_n$ iid suivant un loi $N(\mu_2,\sigma^2)$. On souhaite estimer $\mu_1 + \mu_2$.

a) L'estimateur $\hat{\theta} = \frac{1}{n+m}(X_1 + \cdots + X_n + Y_1 + \cdots + Y_m)$ est-il sans biais ?

b) On pose $\hat{\mu}_1 = \overline{X}$ et $\hat{\mu}_2 = \overline{Y}$. Vérifier que $\hat{\mu}_1 + \hat{\mu}_2$ est un estimateur sans biais et calculer son erreur quadratrique moyenne, où $\hat{\mu}$ désigne l'estimateur de la moyenne habituelle.

Solution. a) Non. On trouve $\mathbb{E}(\hat{\theta}) = \frac{1}{n+m}(n\mu_1 + m\mu_2) \neq \mu_1 + \mu_2$. On peut dire un peu plus : on voit que

$$\frac{n}{n+m}\mu_1 + \frac{m}{n+m}\mu_2 \leq \frac{n}{n+0}\mu_1 + \frac{m}{0+m}\mu_2 = \mu_1 + \mu_2,$$

donc l'estimateur à tendance à sous-estimer le paramètre, en moyenne.

b) On a $\mathbb{E}(\hat{\mu}_1 + \hat{\mu}_2) = \mu_1 + \mu_2$, donc il est sans biais.

Pour l'erreur quadratique moyenne, on a $\mathbb{V}\mathrm{ar}(\hat{\mu}_1) = \frac{\sigma^2}{n}$ et $\mathbb{V}\mathrm{ar}(\hat{\mu}_2) = \frac{\sigma^2}{n}$, donc

$$\mathrm{EQM}(\hat{\mu}_1 + \hat{\mu}_2) = \mathbb{V}\mathrm{ar}(\hat{\mu}_1 + \hat{\mu}_2) = \mathbb{V}\mathrm{ar}(\mu_1) + \mathbb{V}\mathrm{ar}(\hat{\mu}_2) = \frac{2\sigma^2}{n}.$$

Remarque : On a bien que $\mathrm{EQM}(\hat{\theta}) = \mathbb{V}\mathrm{ar}(\hat{\theta})$ lorsqu'un estimateur est sans biais, ce qui est le cas ici.

Exercice 4. On reprend l'exemple précédent. On cherche un estimateur de σ^2, la variance commune aux deux populations. Soit $\hat{\sigma}_1 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \hat{\mu}_1)^2$ et $\hat{\sigma}_2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \hat{\mu}_2)^2$ les estimateurs sans biais de σ^2.

a) Est-ce que $\hat{\theta} = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{2}$ est un estimateur sans biais de σ^2 ?

b) Vérifier que $\tilde{\sigma}^2 = \frac{(n-1)\hat{\sigma}_1^2 + (m-1)\hat{\sigma}_2^2}{n+m-2}$ est un estimateur sans biais de σ^2.

c) Calculer leur erreur quadratique moyenne. *Remarque*. Pour calculer $\mathbb{V}\mathrm{ar}(\hat{\sigma}^2)$, remarquez que $(k-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{k-1}$ et on sait que la variance la loi du χ^2_{k-1} est $2(k-1)$. Notez que cela fonctionne seulement dans le cas où les v.a. sont normalement distribuées et iid.

Solution. c) D'abord, on sait que $\mathbb{V}\mathrm{ar}\big[(k-1)\frac{\hat{\sigma}^2}{\sigma^2}\big] = 2(k-1)$, donc $\mathbb{V}\mathrm{ar}(\hat{\sigma}^2) = \frac{2\sigma^4}{k-1}$.

Ensuite, on a

$$\mathbb{V}\mathrm{ar}(\hat{\theta}) = \frac{1}{4}\left(\frac{2\sigma^4}{n-1} + \frac{2\sigma^4}{m-1}\right) = \frac{\sigma^4}{2}\left(\frac{m+n-2}{(n-1)(m-1)}\right).$$

Comme $\hat{\theta}$ est sans biais, on sait que $EQM(\hat{\theta}) = \mathbb{V}\mathrm{ar}(\hat{\theta})$.

Pour $\tilde{\sigma}^2$, on a

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\tilde{\sigma}^2) &= \frac{(n-1)^2\frac{2\sigma^4}{(n-1)} + (m-1)^2\frac{2\sigma^4}{(m-1)}}{(n+m-2)^2} \\ &= 2\sigma^4\frac{(n-1)+(m-1)}{(n+m-2)^2} \\ &= \frac{2\sigma^4}{n+m-2}.\end{aligned}$$

Remarque : On peut montrer que $EQM(\tilde{\sigma}^2) \leq EQM(\hat{\theta})$, avec égalité ssi $m = n$.

Exercice 5. La fonction de densité de la loi $\Gamma(k,\theta)$ (se lit « gamma ») est donnée par

$$f_{k,\theta}(x) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\Gamma(k)\theta^k},$$

où $\Gamma(k)$ est une constante.

a) Lorsque $k = 2$, on a $\Gamma(2) = 1$ et la densité devient $f_{2,\theta}(x) = \frac{xe^{-\frac{x}{\theta}}}{\theta^2}$. Trouver un estimateur de θ par la méthode du maximum de vraisemblance. (Si vous êtes à l'aise, vous pouvez passer directement au b.)

b) En supposons que k est connu, trouver un estimateur de θ par la méthode du maximum de vraisemblance.

Solution. b) Par définition, on a

$$L(\theta) = \prod_{k=1}^{n} f_{k,\theta}(X_k) = \frac{(X_1 \cdots X_n)^{k-1}e^{-\frac{1}{\theta}\sum_{k=1}^n X_k}}{\Gamma(k)^n\theta^{nk}}.$$

On peut remplacer $\sum X_k$ par $n\overline{X}$. Ensuite, on prend le log et on obtient

$$\mathcal{L}(\theta) = (k-1)\sum_{k=1}^{n}\log(X_k) - \frac{n}{\theta}\overline{X} - n\log\big(\Gamma(k)\big) - nk\log\theta.$$

Ensuite, on dérive par rapport à θ :

$$\mathcal{L}'(\theta) = \frac{n}{\theta^2}\overline{X} - \frac{nk}{\theta}.$$

On veut donc résoudre $\frac{n}{\theta^2}\overline{X} - \frac{nk}{\theta} = 0$, c'est-à-dire $\overline{X} - k\theta = 0$. Ainsi $\theta = \frac{\overline{X}}{k}$ est un point critique. On a même

$$\mathcal{L}'(\theta) \begin{cases} > 0, & \text{si } \theta < \frac{\overline{X}}{k}; \\ = 0, & \text{si } \theta = \frac{\overline{X}}{k}; \\ < 0, & \text{si } \theta > \frac{\overline{X}}{k}; \end{cases}$$

il s'ensuit que $\theta = \frac{\overline{X}}{k}$ est un maximum.

Exercice 6. La fonction de densité de la loi de Weibull(k,λ) est

$$f_{k,\lambda}(x) = \frac{k}{\lambda^k} x^{k-1} e^{-(x/\lambda)^k}.$$

(Si $k = 1$, cela donne une loi exponentielle de paramètre $\frac{1}{\lambda}$.) Trouver un estimateur de λ par la méthode de vraisemblance, en supposant que k est connu.

Exercice 7. Lorsqu'on calcule l'intervalle de confiance d'une moyenne, quelle est la différence si on connaît la variance ou non ?

Solution. Si on connaît le vrai écart-type, on utilise un intervalle de confiance de Wald, puisque la variable $Z = \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}}$ suit une loi normale centrée réduite. Si on remplace l'écart-type par son estimateur non biaisé, on utilise un intervalle de confiance de Student, puisque $T = \frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}$ suit une loi de Student avec $n-1$ degrés de liberté.

Exercice 8. On a mesuré la hauteur (en cm) d'une espèce de plante dans une culture de six mois :

$$24 \qquad 18 \qquad 25 \qquad 19 \qquad 20 \qquad 19.$$

Calculer un intervalle de confiance pour la moyenne avec niveau de confiance 95%. On suppose que la hauteur de cette plante suit une loi normale.

Solution. On calcule

$$\overline{x} = \frac{24+18+25+19+20+19}{6} = \frac{125}{6} = 20.8\overline{3},$$

$$s^2 = \frac{1}{5}\Big((24-\overline{x})^2 + (18-\overline{x})^2 + (25-\overline{x})^2 + (19-\overline{x})^2 + (20-\overline{x})^2 + (19-\overline{x})^2\Big)$$
$$= \frac{42.8\overline{3}}{5} = 8.5\overline{6}$$

$$s = \sqrt{s^2} = \sqrt{8.5\overline{6}} = 2.9269$$

On pose $T = \frac{\overline{X}-\mu}{\hat{\sigma}/\sqrt{n}}$. Comme on utilise l'estimateur de la variance, T suit une loi de Student avec 5 degrés de liberté. On cherche $t_{0,975}$ dans la table de la loi de Student et on trouve $t_{0,975} = 2{,}57$.

La borne inférieur est $\overline{x} - t_{0,975}\frac{s}{\sqrt{n}} = 20{,}8\overline{3} - 2{,}57 \times \frac{8{,}5\overline{6}}{\sqrt{5}} = 10{,}987$.

Pour la borne inférieur, on trouve 30,679. Ainsi, l'intervalle est $[10{,}987; 30{,}679]$.

Exercice 9. Calculer un intervalle de confiance pour la moyenne avec un niveau de confiance de 99% pour les données suivantes

$$12 \qquad 20 \qquad 17 \qquad 19.$$

On suppose que les données sont normalement distribuées.

Solution. Moyenne empirique : $\overline{x} = \frac{12+20+17+19}{4} = \frac{68}{4} = 17$

Variance empirique : $s^2 = \frac{(12-17)^2+(20-17)^2+(17-17)^2+(19-17)^2}{3} = \frac{38}{3} = 12{,}\overline{6}$

Écart-type empirique : $s = \sqrt{12{,}\overline{6}} = 3{,}559$

Quantile : $100\% - 99\% = 1\%$ et $\frac{1\%}{2} = 0{,}5\%$, ensuite $99\% + 0{,}5\% = 0{,}995\%$, donc avec $\alpha = 1 - 0{,}995 = 0{,}005$, on trouve $t_{0,995} = 5{,}8409$ dans la table de Student (colonne 0,005 et ligne 3).

Borne inf $= \overline{x} - t_{0,995}\frac{s}{\sqrt{n}} 17 - 5{,}8409 \times \frac{3{,}559}{\sqrt{4}} = 6{,}606$

Borne sup $= \overline{x} - t_{0,995}\frac{s}{\sqrt{n}} 17 + 5{,}8409 \times \frac{3{,}559}{\sqrt{4}} = 27{,}304$

Conclusion : l'intervalle de confiance de Student pour la moyenne avec un niveau de confiance 99% est $[6{,}606; 27{,}304]$.

Exercice 10. On veut estimer la longueur moyenne d'une population d'éperlans. On suppose que la longueur d'un éperlan moyenne est normalement distribuée. On obtient les données suivantes (en cm)

$$\begin{array}{ccccc} 15 & 5 & 16 & 8 & 12 \\ 11 & 30 & 15 & 18 & 6 \end{array}.$$

Calculer un intervalle de confiance pour la moyenne μ avec un niveau de confiance de 97,5%.

Solution. Pour la moyenne empirique $\overline{x}$, on trouve

$$\overline{x} = \frac{1}{10}\Big[15 + 5 + 16 + 8 + 12 + 11 + 30 + 15 + 18 + 6\Big] = \frac{136}{10} = 13{,}6.$$

Ensuite,pour la variance empirique s^2, on trouve

$$\begin{aligned} s^2 = \frac{1}{9}\Big[&(15 - 13,6)^2 + (5 - 13,6)^2 + (16 - 13,6)^2 \\ &+ (8 - 13,6)^2 + (12 - 13,6)^2 + (11 - 13,6)^2 \\ &+ (30 - 13,6)^2 + (15 - 13,6)^2 + (18 - 13,6)^2 + (6 - 13,6)^2\Big] \end{aligned}$$

$$= \frac{470,4}{9}$$
$$= 52,2\overline{6}.$$

On pose $X = \frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{10}}$, qui suit une loi de Student avec 9 degrés de liberté. (En effet, comme on a remplacé σ par $\hat{\sigma}$, la variable X suit une loi de Student au lieu d'une loi normale.) On cherche le bon quantile. On a $100\% - 97{,}5\% = 2{,}5\%$ et $\frac{2{,}5\%}{2} = 1{,}25\%$. Enfin, on a $97{,}5\% + 1{,}25\% = 98{,}75\%$. (Je vous encourage à faire les dessins comme on a fait en classe. Je ne peux malheureusement pas les faire dans le solutionnaire.) Ainsi, on cherche le quantile 0,9875-quantile, c'est-à-dire on cherche $t_{0,9875}$ tel que $\mathbb{P}(T < t_{0,9875}) = 0{,}9875$.

[Remarque : on peut également trouver le bon quantile avec le calcul suivant

$$\begin{aligned}\mathbb{P}(-t_{0,9875} \leq X \leq t_{0,9875}) = 0{,}975 \quad &\Leftrightarrow \quad 2\mathbb{P}(0 \leq X \leq t_{0,9875}) = 0{,}975 \\ &\Leftrightarrow \quad 2\left[\mathbb{P}(X \leq t_{0,9875}) - \tfrac{1}{2}\right] = 0{,}975 \\ &\Leftrightarrow \quad \mathbb{P}(X \leq t_{0,9875}) = 0{,}9875.\end{aligned}$$

Utilisez la méthode que vous préférez.]

En utilisant une table de valeur, on trouve $t_{0,9875} = 2{,}6850$. On calcule ensuite les bornes :

$$\text{borne inf} = \overline{x} - t_{0,9875}\frac{s}{\sqrt{n}} = 13{,}6 - 2{,}6850 \times \frac{\sqrt{52{,}2\overline{6}}}{\sqrt{10}} = 7{,}4616$$

$$\text{borne sup} = \overline{x} + t_{0,9875}\frac{s}{\sqrt{n}} = 13{,}6 + 2{,}6850 \times \frac{\sqrt{52{,}2\overline{6}}}{\sqrt{10}} = 24{,}73.$$

On trouve donc l'intervalle de confiance $[7{,}4616; 19{,}7384]$ pour μ avec un niveau de confiance 97,5%.

Exercice 11. Soit μ la moyenne de la population pour lequel on cherche un intervalle de confiance avec niveau de confiance α. Soit $\hat{\mu}$ l'estimateur habituel. Soit $I = \left(\overline{x} - \frac{\sigma}{\sqrt{n}}a, \overline{x} + \frac{\sigma}{\sqrt{n}}b\right)$ l'intervalle de confiance.

Il y a plusieurs a et b possibles, car il faut simplement que $\mathbb{P}(a \leq \hat{\mu} \leq b) = \alpha$. On pose $L(a,b) = 2\frac{\sigma}{\sqrt{n}}(b-a)$, la longueur de l'intervalle de confiance. Le but est de montrer que si la longueur est minimale, alors $f(a) = f(b)$.

a) On suppose que b dépend de a. En dérivant $\mathbb{P}(a \leq \hat{\mu} \leq b) = \alpha$, montrer que $\frac{\mathrm{d}b}{\mathrm{d}a} = \frac{f(b)}{f(a)}$.

b) En supposant que a et $b(a)$ est un minimum de L, montrer que $\frac{\mathrm{d}b}{\mathrm{d}a} = 1$.

c) Déduire que $f(b) = f(a)$.

d) Montrer que le minimum est atteint en $a = -x_{\frac{1+\alpha}{2}}$, $b = x_{\frac{1+\alpha}{2}}$, où $x_{\frac{1+\alpha}{2}}$ est le $\left(\frac{1+\alpha}{2}\right)$-quantile de la distribution, c'est-à-dire $F\left(X \leq x_{\frac{1+\alpha}{2}}\right) = \frac{1+\alpha}{2}$.