Corrigé 10

Exercice 1. (a) (i) Pour $0 \leq y \leq \theta$, la fonction de répartition de $M_n = \max\{Y_1, \dots, Y_n\}$ est

$$F_{M_n}(y) = \Pr(M_n \leq y) = \Pr(Y_1 \leq y, \dots, Y_n \leq y) = \prod_{i=1}^{n} \Pr(Y_i \leq y) = \left(\frac{y}{\theta}\right)^n,$$

par indépendance des $Y_1, \dots, Y_n$. D'autre part, $F_{M_n}(y) = 0$ pour $y \leq 0$ et $F_{M_n}(y) = 1$ pour $y > \theta$. En dérivant $F_{M_n}(y)$ on trouve la fonction de densité

$$f_{M_n}(y) = \frac{n}{\theta}\left(\frac{y}{\theta}\right)^{n-1}, \quad y \in [0, \theta],$$

et $f_{M_n}(y) = 0$ sinon.

(ii) Le biais de l'estimateur M_n est égal à $\mathbb{E}[M_n] - \theta$. Donc pour calculer le biais b_n, on commence par calculer l'espérance de M_n,

$$\mathbb{E}[M_n] = \int_{-\infty}^{\infty} y f_{M_n}(y) dy = \int_0^{\theta} \frac{n}{\theta^n} y^n dy = \frac{n}{n+1}\theta.$$

Le biais est donc $\mathbb{E}[M_n] - \theta = -\theta/(n+1) \neq 0$ et M_n est donc un estimateur biaisé (en particulier M_n sous-estime θ).

(iii) Pour obtenir un estimateur non-biaisé, on doit corriger M_n et poser $\hat{\theta}_1 = \frac{n+1}{n} M_n$. Pour ce nouvel estimateur, on a bien $\mathbb{E}[\hat{\theta}_1] = \theta$ et donc son biais est nul pour tout $\theta > 0$. La variance est $\operatorname{var}(\hat{\theta}_1) = \left(\frac{n+1}{n}\right)^2 \operatorname{var}(M_n)$. On obtient la variance de M_n à l'aide de la fonction de densité trouvée en (i) :

$$\mathbb{E}[M_n^2] = \int_{-\infty}^{\infty} y^2 f_{M_n}(y) dy = \frac{n}{n+2}\theta^2,$$

et donc $\operatorname{var}(M_n) = \mathbb{E}[M_n^2] - \mathbb{E}[M_n]^2 = \frac{n}{(n+1)^2(n+2)}\theta^2$. Finalement, $\operatorname{var}(\hat{\theta}_1) = \frac{1}{n(n+2)}\theta^2$.

(b) (i) Par définition, le premier moment empirique de l'échantillon $Y_1, \dots, Y_n$ est

$$\frac{1}{n}\sum_{i=1}^{n} Y_i$$

et le premier moment théorique est $\mathbb{E}[Y_1] = \theta/2$ (comme tous les Y_i suivent la même distribution uniforme sur $[0, \theta]$). L'estimateur des moments, qu'on dénote $\hat{\theta}_2$, doit résoudre l'équation

$$\frac{\hat{\theta}_2}{2} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \Leftrightarrow \qquad \hat{\theta}_2 = \frac{2}{n}\sum_{i=1}^{n} Y_i.$$

On vérifie que $\mathbb{E}[\hat{\theta}_2] = \theta$ et donc l'estimateur est non-biaisé.

(ii) Sa variance est $\text{var}(\hat{\theta}_2) = \theta^2/(3n)$.

(c) On se retrouve donc avec trois candidats potentiels $(M_n, \hat{\theta}_1, \hat{\theta}_2)$ pour estimer le paramètre inconnu θ. On utilise l'erreur quadratique moyenne $\text{EQM}_\theta(\hat{\theta}) = b(\hat{\theta})^2 + \text{var}(\hat{\theta})$ pour les comparer :

$$\begin{aligned}
\text{EQM}_\theta(M_n) &= [-\theta/(n+1)]^2 + n\theta^2/[(n+1)^2(n+2)] \\
\text{EQM}_\theta(\hat{\theta}_1) &= 0 + \frac{1}{n(n+2)}\theta^2 \\
\text{EQM}_\theta(\hat{\theta}_2) &= 0 + \theta^2/(3n)
\end{aligned}$$

Pour tout $n > 1$, $\text{EQM}_\theta(\hat{\theta}_1)$ est le plus petit des trois et l'estimateur $\hat{\theta}_1$ est donc préférable.

(d) Notons tout d'abord que les bornes de l'intervalle ne dépendent pas de θ mais uniquement des données $Y_1, \ldots, Y_n$.
Comme $M_n \leq \theta$ forcément, on a pour tout $\theta > 0$, en utilisant également (a)(i) :

$$\Pr(M_n \leq \theta \leq M_n\alpha^{-1/n}) = \Pr(M_n \geq \theta\alpha^{1/n}) = 1 - \Pr(M_n \leq \theta\alpha^{1/n}) = 1 - \frac{(\theta\alpha^{1/n})^n}{\theta^n} = 1 - \alpha.$$

On a donc bien un intervalle de confiance de niveau $1 - \alpha$.

Exercice 2. (a) Si $X_1, \ldots, X_n$ sont iid $\mathcal{N}(\mu, \sigma^2)$, alors

$$Z_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

On sait que

$$\Pr(-z < Z_n < z) = 2\,\Phi(z) - 1,$$

pour toute constante $z > 0$. Donc si on choisit $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ (le $(1-\alpha/2)-$quantile de la loi $\mathcal{N}(0,1)$), on obtient que

$$\begin{aligned}
1 - \alpha &= \Pr(-z_{1-\alpha/2} < Z_n < z_{1-\alpha/2}) \\
&= \Pr\left(-z_{1-\alpha/2} < \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} < z_{1-\alpha/2}\right) \\
&= \Pr\left(\bar{X}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}}\,\sigma < \mu < \bar{X}_n + \frac{z_{1-\alpha/2}}{\sqrt{n}}\,\sigma\right).
\end{aligned}$$

L'intervalle

$$\left(\bar{X}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}}\,\sigma,\ \bar{X}_n + \frac{z_{1-\alpha/2}}{\sqrt{n}}\,\sigma\right)$$

couvre donc la vraie valeur de μ avec la probabilité $1 - \alpha$.
Dans notre cas, $\sigma^2 = 3.5$, $n = 12$, $\bar{x}_{12} = 13.31$, et $\alpha = 0.05$. On peut trouver dans le tableau de la fonction de répartition de la loi normale que $\Phi^{-1}(0.975) = 1.96$. L'intervalle cherché est donc $(12.25,\ 14.37)$.

(b) Si $X_1, \ldots, X_n$ sont iid $\mathcal{N}(\mu, \sigma^2)$, alors

$$T_{n-1} = \sqrt{n}\,\frac{\bar{X}_n - \mu}{S_n} \sim \mathrm{t}_{n-1},$$

où $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$, et t_ν est la loi de Student avec ν degrés de liberté. La loi de Student est symétrique autour de zéro de la même manière que la loi $\mathcal{N}(0,1)$. On a donc

$$\mathrm{Pr}_{\mu,\sigma^2}(|T| > t_{n-1,1-\alpha/2}) = \alpha,$$

où $t_{n-1,1-\alpha/2}$ est le $(1-\alpha/2)$-quantile de la loi t_{n-1}.
De la même manière que dans la partie (a) on obtient l'intervalle de confiance sous la forme

$$\left(\bar{X}_n - \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}}\,S_n,\ \bar{X}_n + \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}}\,S_n\right).$$

Dans notre situation on a $n = 12$, $\alpha = 0.05$, $s^2 = 3.69$, et on peut trouver dans le tableau que $t_{11,0.975} = 2.201$. L'intervalle cherché est $(12.09,\ 14.53)$. On note que cet intervalle est plus large que celui de la partie (a). Cela vient du fait que nous avons maintenant deux paramètres à estimer et donc que l'incertitude est plus grande que dans le cas où un seul paramètre est à estimer.

(c) On trouve que $t_{11}(0.95) = 1.796$, et que l'intervalle cherché est $(12.31, 14.31)$. Cet intervalle est plus étroit que celui calculé dans la partie (b), car son seuil de confiance est plus petit. Plus on veut être confiant qu'un intervalle couvre la vraie valeur de μ, plus cet intervalle doit être large (et vice-versa).

(d) Récolter plus de données. Plus on a de données, plus l'incertitude est faible.

Exercice 3. (a) On compte le nombre de succès ("succès" = pièce défectueuse) parmi 100 essais. Donc $Y \sim \mathcal{B}(100, p)$, où p est le vrai pourcentage de pièces défectueuses dans un paquet. On observe une réalisation de l'échantillon $Y_1, \ldots, Y_{18} \sim \mathcal{B}(100, p)$.

(b) La fonction de vraisemblance est

$$L(p) = \prod_{i=1}^{18} P(Y_i = y_i) = \prod_{i=1}^{18} \binom{100}{y_i} p^{y_i}(1-p)^{100-y_i} = \left(\prod_{i=1}^{18}\binom{100}{y_i}\right) p^{\sum_{i=1}^{18} y_i}(1-p)^{1\,800-\sum_{i=1}^{18} y_i}.$$

Pour maximiser cette fonction, on peut maximiser son logarithme,

$$\ell(p) = \log(L(p)) = \log\left(\prod_{i=1}^{18}\binom{100}{y_i}\right) + \sum_{i=1}^{18} y_i \log(p) + \left(1\,800 - \sum_{i=1}^{n} y_i\right)\log(1-p).$$

Ainsi, on obtient que

$$\frac{\partial}{\partial p}\ell(p) = \sum_{i=1}^{18} y_i/p - \left(1\,800 - \sum_{i=1}^{18} y_i\right)/(1-p),$$

et donc que

$$\frac{\partial}{\partial p}\ell(p) = 0 \Leftrightarrow p = \frac{1}{1\,800}\sum_{i=1}^{18} y_i = \bar{y}/100.$$

On peut vérifier que $\frac{\partial^2}{\partial p^2}\ell(p) < 0$ pour tout $p \in (0,1)$ et donc il s'agit d'un maximum. Ainsi, l'estimateur du maximum de vraisemblance est

$$\hat{p} = \frac{1}{1\,800}\sum_{i=1}^{18} y_i = \bar{y}/100.$$

(c) On a

$$\hat{p} = \bar{y}/100 = 0.00833.$$

Ici on ne peut pas donner d'intervalle de confiance exact, vu que les observations ne sont pas issues d'une loi normale. Cependant, on a vu en cours que l'intervalle avec limites $\hat{p} \pm z_{1-\alpha/2} J(\hat{p})^{-1/2}$ est un intervalle de confiance approximatif de niveau α pour p. On trouve que pour $\alpha = 0.05$, $z_{1-\alpha/2} = 1.96$. De plus

$$J(\hat{p}) = -\frac{\partial^2}{\partial p^2}\ell(\hat{p}) = \frac{18\bar{y}}{\hat{p}^2} + \frac{18\,(100-\bar{y})}{(1-\hat{p})^2},$$

et en insérant les valeurs numériques, $J(0.00833)^{-1/2} = 0.00214$. L'intervalle approximatif à 95 % est donc $(0.00413, 0.01253)$.

Exercice 4. On a pour $t > 0$

$$\Pr(\lambda n X_{(1)} > t) = \Pr(X_{(1)} > t/(n\lambda)) = \prod_{i=1}^{n}\Pr(X_i > t/(n\lambda)) = [\exp(-\lambda t/(n\lambda))]^n = \exp(-t).$$

Donc pour $t > 0$, $\Pr(\lambda n X_{(1)} \le t) = 1 - \exp(-t)$ qui est la fonction de répartition d'une variable aléatoire Exp(1). Comme la distribution Exp(1) ne dépend pas du paramètre λ, $\lambda n X_{(1)}$ est bien un pivot. Soit q_β le β-quantile de la loi Exp(1). Alors

$$1-\alpha = \Pr_\lambda\left(q_{\alpha/2} \le n\lambda X_{(1)} \le q_{1-\alpha/2}\right) = \Pr_\lambda\left(\frac{q_{\alpha/2}}{nX_{(1)}} \le \lambda \le \frac{q_{1-\alpha/2}}{nX_{(1)}}\right)$$

pour tout $\lambda > 0$. On a donc l'intervalle de confiance

$$\left[\frac{q_{\alpha/2}}{nX_{(1)}}, \frac{q_{1-\alpha/2}}{nX_{(1)}}\right].$$

Remarque. Dans ce cas les quantiles sont explicites $q_\beta = -\log(1-\beta)$.