Corrigé 9

Exercice 1. (i).

$$\begin{aligned}
\mathbb{E}[S_n] &= \mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{E}[X_i] = n\,\mu, \\
\mathrm{Var}(S_n) &= \mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i) = n\,\sigma^2.
\end{aligned}$$

(ii). Avec $b_n = n\mu$ et $a_n = 1/(\sqrt{n}\sigma)$ on a

$$\begin{aligned}
\mathbb{E}[Z_n] = \mathbb{E}\left[\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}\right] = \frac{1}{\sqrt{n}\sigma}\left(\sum_{i=1}^n \mathbb{E}[X_i] - n\mu\right) = 0, \\
\mathrm{Var}(Z_n) = \mathrm{Var}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}\right) = \frac{1}{n\sigma^2}\sum_{i=1}^n \mathrm{Var}(X_i) = 1.
\end{aligned}$$

(iii).

$$\begin{aligned}
\mathbb{E}[\overline{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \mu, \\
\mathrm{Var}\left(\overline{X}_n\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\sigma^2}{n}.
\end{aligned}$$

(iv). Avec $d_n = \mu$ et $c_n = \sqrt{n}/\sigma$ on a

$$\begin{aligned}
\mathbb{E}[Z_n] &= \mathbb{E}\left[\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma}\right] = \frac{\sqrt{n}}{\sigma}\left(\mathbb{E}[\overline{X}_n] - \mu\right) = 0, \\
\mathrm{Var}(Z_n) &= \mathrm{Var}\left(\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma}\right) = \frac{n}{\sigma^2}\,\mathrm{Var}(\overline{X}_n) = 1.
\end{aligned}$$

(v). On ne connaît pas la loi des X_i. Bien que l'on connaisse $\mathbb{E}[X_i]$ et $\mathrm{Var}(X_i)$, ce n'est pas assez pour déterminer la loi de X_i.

(vi). Puisqu'on ne connaît pas la loi de X_i, on ne sait pas quelle est la probabilité $\Pr(X_i \le x)$, pour $x \in \mathbb{R}$.

(vii). Par le théorème central limite, la limite de la distribution de Z_n si $n \to \infty$ est $\mathcal{N}(0,1)$. Donc pour n grand, la distribution de Z_n est approximativement $\mathcal{N}(0,1)$.

(viii). D'après la partie ((vii)), on a $\Pr(Z_n \le x) \approx \Phi(x)$, où $\Phi(x)$ est la fonction de répartition de la loi $\mathcal{N}(0,1)$. Cela montre la grande puissance du théorème central limite et le rôle central de la loi $\mathcal{N}(0,1)$.

Exercice 2. La variable aléatoire X représente le résultat du jet d'un dé équilibré. Elle peut donc prendre comme valeur $\{1, 2, 3, 4, 5, 6\}$. La fonction de masse de X est

$$f_X(x) = 1/6,$$

pour tout x dans $\{1,2,3,4,5,6\}$ et 0 sinon. Par définition,

$$\mathbb{E}(X) = \sum_{x=1}^{6} x f_X(x) = (1+2\cdots+6)/6 = 7/2.$$

Par ailleurs, on trouve $E(X^2) = \sum_{x=1}^{6} x^2 f_X(x) = 91/6$ et donc $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 91/6 - 49/4 = 35/12$. Soit X_i le résultat du i-ème dé. On cherche

$$\Pr\left(30 \leq \sum_{i=1}^{10} X_i \leq 40\right) = \Pr\left(30/10 \leq \sum_{i=1}^{10} X_i/10 \leq 40/10\right) = \Pr(3 \leq \bar{X}_n \leq 4).$$

Par le théoréme centrale limite, on sait que $\bar{X}_n (= \sum_{i=1}^{10} X_i/10)$ suit approximativement une loi normale $\mathcal{N}(\mu, \sigma^2/10)$, avec $\mu = \mathbb{E}(X_i) = 7/2$ et $\sigma^2 = \mathrm{Var}(X_i) = 35/12$. En centrant et en réduisant :

$$\begin{aligned}
\Pr\left(30 \leq \sum_{i=1}^{10} X_i \leq 40\right) &= \Pr(3 \leq \bar{X}_n \leq 4) \\
&= \Pr\left(\frac{3-\mu}{\sqrt{\sigma^2/10}} \leq \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/10}} \leq \frac{4-\mu}{\sqrt{\sigma^2/10}}\right) \\
&\approx \Phi(\sqrt{6/7}) - \Phi(-\sqrt{6/7}) \\
&\approx 0.65.
\end{aligned}$$

Exercice 3. (i). Notons p le pourcentage recherché, et considérons $p \in (0,1)$. Si on choisit par hasard une personne parmi les étudiant-e-s de l'EPFL/UNIL, celle-ci sera une femme avec la probabilité p et un homme avec la probabilité $1-p$. On peut définir une variable aléatoire

$$X = \begin{cases} 1 & \text{si la personne choisie est une femme,} \\ 0 & \text{si la personne choisie est un homme.} \end{cases}$$

La loi de cette variable est $\mathcal{B}(p)$.

(ii). Le paramètre d'intérêt est p.

(iii). Puisqu'il serait difficile d'observer toutes les personnes qui étudient à l'EPFL/UNIL, on va observer un sous-ensemble. Ce sous-ensemble doit être représentatif, par exemple on peut observer un certain nombre d'étudiant-e-s qui mangent dans une grande cafétéria pendant la pause de midi.

(iv). Un choix intuitif est le pourcentage de femmes dans le sous-ensemble observé.

(v). Même si on connaissait la valeur de p, on ne connaîtrait pas en avance la valeur de l'estimateur. Si l'on va dans la même cafétéria deux jours différents et l'on observe le même nombre d'étudiant-e-s, ce ne seront pas exactement les mêmes étudiant-e-s, donc on n'obtiendra pas le même résultat.

(vi). On suppose que $p = 0.4$ et $n = 100$. D'après la partie ((i)), on peut supposer que les observations $x_1, \ldots, x_{100}$ constituent une réalisation de $X_1, \ldots, X_{100} \overset{iid}{\sim} \mathcal{B}(p)$. L'estimateur proposé dans la partie ((iv)) s'écrit $\hat{p}_{100} = \bar{X}_{100} = (\sum_{i=1}^{100} X_i)/100$.

$$\begin{aligned}
\mathbb{E}[\hat{p}_{100}] &= \mathbb{E}\left(\frac{1}{100}\sum_{i=1}^{100} X_i\right) = \mathbb{E}[X_1] = p, \\
\mathrm{Var}[\hat{p}_{100}] &= \mathrm{Var}\left(\frac{1}{100}\sum_{i=1}^{100} X_i\right) = \frac{1}{100}\mathrm{Var}[X_1] = \frac{p(1-p)}{100}, \\
\mathrm{b}(\hat{p}_{100}) &= \mathbb{E}[\hat{p}_{100}] - p = 0.
\end{aligned}$$

L'estimateur $\hat{p}_n$ est non-biaisé. Si la taille de l'échantillon augmente, la variance diminue alors que l'espérance ne change pas. Donc, avec un plus grand échantillon, on estime le pourcentage avec une plus grande précision (on s'attend à être plus proche de la vraie valeur).

(vii). Remarquons tout d'abord que nous sommes ici dans la même situation que dans l'xercice 1. Les variables $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées, d'espérance $\mu = p$ et de variance $\sigma^2 = p(1-p)$. Nous pouvons donc utiliser le théorème central limite pour approximer la loi de

$$Z_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} = \sqrt{n}\,\frac{\hat{p}_n - p}{\sqrt{p(1-p)}}.$$

Pour trouver n tel que $\Pr(\hat{p}_n < 0.5) \geq 0.95$ on calcule (avec $p = 0.4$)

$$\begin{aligned}
& \Pr(\hat{p}_n < 0.5) \geq 0.95 \\
\Rightarrow\ & \Pr\left(\sqrt{n}\,\frac{\hat{p}_n - 0.4}{\sqrt{0.4 \times 0.6}} < \sqrt{n}\,\frac{0.5 - 0.4}{\sqrt{0.4 \times 0.6}}\right) \geq 0.95 \\
\Rightarrow\ & \Phi\left(0.204\sqrt{n}\right) \geq 0.95 \\
\Rightarrow\ & \sqrt{n} \geq \frac{\Phi^{-1}(0.95)}{0.204} \\
\Rightarrow\ & n \geq 65.42.
\end{aligned}$$

Donc on a besoin d'observer au moins 66 personnes.

Exercice 4. Soit X_i le nombre de clients qui entrent le $i$ème jour. Comme X_i est une variable aléatoire suivant la loi de Poisson avec paramètre $\lambda = 12$, on a $\mathbb{E}[X_i] = 12$ et $\mathrm{Var}(X_i) = 12$. Le nombre total de clients entrants n jours est $S_n = \sum_{i=1}^{n} X_i$.

Par le théorème central limite, on sait que la variable standardisée

$$Z_n = \frac{S_n - n \times 12}{\sqrt{n \times 12}}$$

suit approximativement la loi $\mathcal{N}(0,1)$ pour n grand.

(i). Avec $n = 22$ nous pouvons calculer

$$\begin{aligned}
\Pr(S_{22} \geq 250) &= \Pr\left(\frac{S_{22} - 22 \times 12}{\sqrt{22 \times 12}} \geq \frac{250 - 22 \times 12}{\sqrt{22 \times 12}}\right) \\
&\approx 1 - \Phi\left(\frac{250 - 22 \times 12}{\sqrt{22 \times 12}}\right) = 1 - \Phi\left(-\frac{14}{\sqrt{264}}\right) = \Phi(0.862) = 0.805.
\end{aligned}$$

(ii). Maintenant on cherche n tel que $\Pr(S_n \geq 250) = 0.975$. On veut donc que

$$\begin{aligned}
&\Pr(S_n \geq 250) = 0.975 \\
&\Leftrightarrow \Pr\left(\frac{S_n - n \times 12}{\sqrt{n \times 12}} \geq \frac{250 - n \times 12}{\sqrt{n \times 12}}\right) = 0.975 \\
&\Leftrightarrow 1 - \Phi\left(\frac{250 - n \times 12}{\sqrt{n \times 12}}\right) = 0.975 \\
&\Leftrightarrow \Phi\left(\frac{250 - n \times 12}{\sqrt{n \times 12}}\right) = 0.025 \\
&\Leftrightarrow \frac{250 - n \times 12}{\sqrt{n \times 12}} = \Phi^{-1}(0.025) \\
&\Leftrightarrow \frac{250 - n \times 12}{\sqrt{n \times 12}} = -\Phi^{-1}(0.975) \\
&\Leftrightarrow \frac{250 - n \times 12}{\sqrt{n \times 12}} = -1.96 \\
&\Leftrightarrow -n \times 12 + 1.96\sqrt{12}\sqrt{n} + 250 = 0.
\end{aligned}$$

Cela donne lieu à une équation du second degré en termes de $\sqrt{n}$, dont les solutions sont

$$\frac{-1.96\sqrt{12} \pm \sqrt{1.96^2 \times 12 + 4 \times 12 \times 250}}{-2 \times 12} = \begin{cases} 4.86 \\ -4.29 \end{cases}.$$

Puisque $\sqrt{n} \geq 0$, on obtient que $\sqrt{n} = 4.86$ et $n = 23.6$. Si le magasin ouvre ses portes pendant au moins 24 jours, il reçoit au moins 250 clients avec une probabilité plus grande que 0.975.

Exercice 5. (i). On sait que $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$. Donc

$$1 = \int_0^1 c\,x^{\theta-1}\,\mathrm{d}x = c\left[\frac{x^\theta}{\theta}\right]_0^1 = \frac{c}{\theta},$$

et on voit bien que $c = \theta$. On a donc la densité

$$f(x) = \begin{cases} \theta\,x^{\theta-1} & \text{si } x \in (0,1) \\ 0 & \text{sinon.} \end{cases}$$

(ii). On a

$$\mathbb{E}[X_1] = \int_{-\infty}^{\infty} x\,f(x)\,\mathrm{d}x = \int_0^1 x\,\theta\,x^{\theta-1}\,\mathrm{d}x = \theta\int_0^1 x^\theta\,\mathrm{d}x = \theta\left[\frac{x^{\theta+1}}{\theta+1}\right]_0^1 = \frac{\theta}{\theta+1}.$$

(iii). Les variables X_i sont continues, donc la fonction de vraisemblance est

$$L(\theta) = f_1(x_1;\theta) \times f_2(x_2;\theta) \times \ldots \times f_n(x_n;\theta),$$

où $f_i(x_i;\theta) = f(x_i;\theta)$ est la densité pour chaque X_i. On trouve

$$L(\theta) = \theta\,x_1^{\theta-1}\theta\,x_2^{\theta-1}\ldots\theta\,x_n^{\theta-1} = \theta^n\left(\prod_{i=1}^n x_i\right)^{\theta-1}.$$

Donc

$$\ell(\theta) = \log(L(\theta)) = n\log(\theta) + (\theta - 1)\sum_{i=1}^{n} \log(x_i).$$

Pour trouver la valeur de θ qui maximise $\ell(\theta)$ on résout

$$\begin{aligned} & \ell'(\theta) = 0 \\ \Leftrightarrow\ & \frac{n}{\theta} + \sum_{i=1}^{n} \log(x_i) = 0 \\ \Leftrightarrow\ & \frac{1}{\theta} = -\frac{1}{n}\sum_{i=1}^{n} \log(x_i) \\ \Leftrightarrow\ & \theta = -\frac{n}{\sum_{i=1}^{n} \log(x_i)}. \end{aligned}$$

Il s'agit bien d'un maximum puisque

$$\ell''(\theta) = -\frac{n}{\theta^2} < 0,$$

pour tout $\theta > 0$. Donc la valeur $\theta = -\frac{n}{\sum_{i=1}^{n}\log(x_i)}$ maximise la fonction $L(\theta)$ et $-\frac{n}{\sum_{i=1}^{n}\log(X_i)}$ est l'estimateur du maximum de vraisemblance, $\hat{\theta}_{ML} = -\frac{n}{\sum_{i=1}^{n}\log(X_i)}$. Remarquons que puisque $x_i \in (0,1)$, on a $\log(x_i) < 0$ et par conséquent $-\frac{n}{\sum_{i=1}^{n}\log(x_i)} > 0$.

(iv). Pour la méthode des moments on doit résoudre l'équation

$$\overline{X}_n = \frac{\widehat{\theta}}{\widehat{\theta} + 1}.$$

Ceci donne $\widehat{\theta}_{MOM} = \overline{X}_n/(1 - \overline{X}_n)$.

Exercice 6. (i). Les variables X_i sont discrètes, donc la fonction de vraisemblance est

$$L(p) = f_1(x_1; p) \times f_2(x_2; p) \times \ldots \times f_n(x_n; p),$$

où $f_i(x_i; p) = P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$ est la fonction de fréquences pour chaque X_i. On trouve

$$L(p) = p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2} \ldots p^{x_n}(1-p)^{1-x_n} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}.$$

(ii). Comme $\mathbb{E}[X_i] = p$, l'estimateur des moments se trouve en résolvant l'équation $\widehat{p}_{MOM} = \overline{X}_n$. Cette équation est déjà résolue, donc $\widehat{p}_{MOM} = \overline{X}_n$.

(iii). L'estimateur du maximum de vraisemblance est la valeur de p qui maximise $L(p)$, ou, de manière équivalente, la valeur qui maximise la fonction $\ell(p) = \log(L(p))$.

On a

$$\ell(p) = \sum_{i=1}^{n} x_i \log(p) + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p).$$

Pour trouver le maximum on résout

$$
\begin{aligned}
& \ell'(p) = 0 \\
\Leftrightarrow\quad & \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} = 0 \\
\Leftrightarrow\quad & (1-p)\sum_{i=1}^n x_i - p\left(n - \sum_{i=1}^n x_i\right) = 0 \\
\Leftrightarrow\quad & \sum_{i=1}^n x_i = pn \\
\Leftrightarrow\quad & p = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}_n.
\end{aligned}
$$

Il s'agit bien d'un maximum, étant donné que

$$\ell''(p) = -\frac{\sum_{i=1}^n x_i}{p^2} - \frac{n - \sum_{i=1}^n x_i}{(1-p)^2} < 0,$$

quel que soit $p \in (0,1)$. Donc la valeur $p = \bar{x}_n$ maximise la fonction $L(p)$ et $\bar{X}_n$ est l'estimateur du maximum de vraisemblance, $\hat{p}_{ML} = \bar{X}_n$.

(iv). On a $\hat{p}_{MOM} = \hat{p}_{ML} = \bar{X}_n$. Donc

$$\mathbb{E}[\hat{p}_{MOM}] = \mathbb{E}[\hat{p}_{ML}] = \mathbb{E}[\bar{X}_n] = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = p,$$

parce que les variables X_i sont toutes $\mathcal{B}(p)$. Donc les estimateurs sont non-biaisés. Pour la variance on a

$$
\begin{aligned}
&\mathrm{Var}[\hat{p}_{MOM}] = \mathrm{Var}[\hat{p}_{ML}] = \mathrm{Var}[\bar{X}_n] = \\
&\qquad = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[X_i] = \frac{p(1-p)}{n},
\end{aligned}
$$

parce que les variables X_i sont indépendantes et toutes $\mathcal{B}(p)$.

Exercice 7. (i). $\mathbb{E}(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy = 1/\lambda$, et $\mathrm{Var}(Y) = \int_{-\infty}^{\infty} y^2 f_Y(y)dy - (\mathbb{E}(Y))^2 = 1/\lambda^2$.

(ii). On sait que $\mathbb{E}(\bar{Y}_n) = \mathbb{E}(Y) = 1/\lambda$ et $\mathrm{Var}(\bar{Y}_n) =\mathrm{Var}(Y)/n$. Par le théorème central limite,

$$\Pr\left(\frac{\bar{Y}_n - \mathbb{E}(\bar{Y}_n)}{\sqrt{\mathrm{Var}(\bar{Y}_n)}} \le z\right) = \Pr\left(\sqrt{n}\frac{\bar{Y}_n - \mathbb{E}(Y)}{\sqrt{\mathrm{Var}(Y)}} \le z\right) = \Pr\left(\sqrt{n}\frac{\bar{Y}_n - 1/\lambda}{1/\lambda} \le z\right) \to \Pr(Z \le z),$$

où $Z \sim \mathcal{N}(0,1)$.

(iii). La vraisemblance

$$L(\lambda) = \prod_{i=1}^n f_Y(y_i;\lambda) = \prod_{i=1}^n \lambda\exp(-\lambda y_i) = \lambda^n \exp(-\sum_{i=1}^n \lambda y_i).$$

La log-vraisemblance est

$$l(\lambda) = n \log(\lambda) - \sum_{i=1}^{n} y_i \lambda$$

On a

$$\frac{dl}{d\lambda}(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i$$

et

$$\frac{d^2 l}{d\lambda^2}(\lambda) = -\frac{n}{\lambda^2} < 0$$

pour tout $\lambda > 0$. L'estimateur du maximum de vraisemblance satisfait donc

$$\frac{dl}{d\lambda}(\hat{\lambda}) = 0 \Leftrightarrow \hat{\lambda} = 1/\bar{Y}_n.$$

C'est bien un maximum car la deuxième dérivée est négative.

(iv). On applique la méthode delta avec $g(x) = 1/x$. On a

$$\overline{Y}_n \overset{\text{approx.}}{\sim} \mathcal{N}(\frac{1}{\lambda}, \frac{1}{\lambda^2}\frac{1}{n}) = \mathcal{N}(\mu, \sigma^2/n),$$

avec $\mu = 1/\lambda$ et $\sigma^2 = 1/\lambda^2$. Ainsi $g(\mu) = 1/\mu = \lambda$ et $g'(\mu) = -1/\mu^2 = -\lambda^2$ de sorte que

$$\hat{\lambda}_{ML} = g(\overline{Y}_n) \overset{\text{approx.}}{\sim} \mathcal{N}(g(\mu), g'(\mu)^2\sigma^2/n) = \mathcal{N}(\lambda, \lambda^2/n).$$