Exercices

à faire chez vous

Exercice 1. Considérons une boîte contenant 6 boules blanches, 3 boules rouges et une boule bleue. Nous tirons de façon aléatoire une boule de la boîte. Soit X une variable aléatoire prenant la valeur 1 si la boule pigée est blanche, 5 si la boule est rouge et 10 si la boule est bleue.

(a) Trouver la fonction de masse de X.

(b) Trouver la fonction de répartition de X.

(c) Représenter graphiquement la fonction trouvée en (b).

Exercice 2. On tire trois boules (sans remise) au hasard d'une boîte contenant $n_1 = 6$ boules rouges et $n_2 = 4$ boules vertes. Soit X la variable aléatoire représentant le nombre de boules rouges parmi les trois boules pigées. Calculer l'espérance et la variance de X.

Exercice 3. Dénotons respectivement par μ et $\sigma^2 > 0$, l'espérance et la variance de la variable aléatoire X. Déterminer $\mathbb{E}\left[\frac{X-\mu}{\sigma}\right]$ et $\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^2\right]$.

Exercice 4. Soient X et Y deux variables aléatoires indépendantes et soient $M_X, M_Y : \mathbb{R} \to \mathbb{R}$ leurs fonctions génératrices de moments respectives. Montrer que la fonction génératrice des moments de la variable aléatoire $Z = X + Y$ est égale à

$$M_Z(t) = M_X(t) \cdot M_Y(t).$$

Exercice 5. Soit Y une variable aléatoire dont la fonction de densité est donnée par

$$g(y) = \begin{cases} cy^2 & \text{si } -1 < y < 1 \\ 0 & \text{sinon.} \end{cases}$$

(a) Déterminer la valeur de la constante c afin que $g(y)$ satisfasse les propriétés d'une fonction de densité.

(b) Trouver la fonction de répartition de Y.

(c) Trouver $\mathbb{P}(0 < Y < 1)$, $\mathbb{P}(0 < Y \leq 3)$ et $\mathbb{P}(Y = 0)$. **Remarque.** On peut répondre à cette question sans calculer aucune intégrale !

(d) Trouver $\mathbb{E}[Y]$ et $\mathrm{Var}[Y]$.

Exercice 6. Soit X une variable aléatoire dont la fonction de densité est donnée par

$$f(x) = \begin{cases} \frac{1}{10} \exp\left(\frac{-x}{10}\right) & \text{si } 0 < x < \infty \\ 0 & \text{sinon.} \end{cases}$$

(a) Trouver la fonction génératrice des moments $M_X(t)$ de X.

(b) En utilisant $M_X(t)$ ou $R_X(t) = \ln(M_X(t))$, déterminer la moyenne et la variance de X.

Exercice 7. Montrer que si $X = \sum_{i=1}^n Y_i$ où $Y_i \overset{iid}{\sim} \mathrm{Bern}(p)$, alors $X \sim \mathrm{Bin}(n, p)$.

Exercice 8. Soit $\{Y_i\}_{i \geq 1}$ une collection infinie de variables aléatoires, où $Y_i \overset{iid}{\sim} \mathrm{Bern}(p)$. Soit $T = \min\{k \in \mathbb{N} : Y_k = 1\} - 1$, montrer que $T \sim \mathrm{Geom}(p)$.

Exercice 9. Montrer que si $X = \sum_{i=1}^r Y_i$ où $Y_i \overset{iid}{\sim} \mathrm{Geom}(p)$, alors $X \sim \mathrm{NegBin}(r, p)$.

Exercice 10. Soient $X_i \overset{iid}{\sim} \text{Poisson}(\lambda)$. Montrer que $Y = \sum_{i=1}^{n} X_i \sim \text{Poisson}(n\lambda)$.

Exercice 11. Soient $X \sim$Poisson(λ) et $Y \sim$Poisson(μ) indépendantes. Montrer que la distribution conditionnelle de X sachant $X + Y = k$ est $\text{Bin}(k, \lambda/(\lambda + \mu))$.

Exercice 12. Soient $X \sim \text{Exp}(\lambda)$ et $t \geq 0$. Montrer que $\mathbb{P}[X \geq x + t | X > t] = \mathbb{P}[X \geq x]$.

Exercice 13. Soient X et Y des variables aléatoires indépendantes qui suivent des distributions exponentielles d'intensité λ_1 et λ_2 respectivement. Montrer que $Z = \min\{X, Y\}$ est une variable aléatoire exponentielle d'intensité $\lambda_1 + \lambda_2$.
Bonus. Montrer que $\mathbb{P}(Z = X) = \lambda_1/(\lambda_1 + \lambda_2)$.

Exercice 14. Montrer que $X \sim \chi^2_2$ si et seulement si $X \sim \text{Exp}(1/2)$.

Exercice 15. Montrer que les distributions suivantes constituent des familles Exponentielles (peut-être lorsqu'un de leurs paramètres est fixé) :

(i) La distribution de Poisson.
(ii) La distribution géométrique.
(iii) La distribution binomiale négative.
(iv) La distribution exponentielle.
(v) La distribution gamma.
(vi) La distribution khi carré.

Exercice 16. Soit $Y \sim Unif(0, 1)$ et soit F une fonction de répartition. Montrer que la fonction de répartition de la variable aléatoire $X = F^{-1}(Y)$ est F, où $F^{-1}(y) = \inf\{t \in \mathbb{R} : F(t) \geq y\}$.

Exercice 17. Soit $X \sim N(\mu, \sigma^2)$, montrer que la fonction de densité de $Y = e^X$ est donnée par

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln y - \mu)^2}{2\sigma^2}\right), \quad 0 < y < \infty.$$

Exercice 18. Prouver le théorème sur les transformations multidimensionnelles (page 45 des diapositives du cours) en utilisant la formule de changement de variables dans une intégrale.

Exercice 19. Soient $X \sim N(\mu_1, \sigma_1^2)$ et $Y \sim N(\mu_2, \sigma_2^2)$ indépendantes. Montrer que $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Exercice 20. Soit Z_1 une variable aléatoire normale standard et Z_2 une variable aléatoire χ^2_n où $n \geq 1$, tels que Z_1 et Z_2 sont indépendantes. À l'aide du théorème 1 du cours (le théorème 1.33 à la page 28 des notes du cours), trouver la densité de la variable aléatoire T, où $T = Z_1/\sqrt{Z_2/n}$. *Indice :* définir $g(Z_1, Z_2) = (T, V) = (T, Z_2)$ pour trouver la densité conjointe de T et V. La densité de T se trouve en intégrant par rapport à V (penser à la distribution *Gamma*).
Remarque. La loi de T s'appelle la loi t de Student avec n degrés de liberté. Elle est très utilisée en statistique et on verra plus tard dans le cours pourquoi. Dans la plupart des cas, n est un nombre entier, mais la distribution est définie pour n'importe quel n réel.

***Exercice 21.** Montrer que la distribution exponentielle est l'unique distribution sans mémoire. Plus précisément, soit X une variable aléatoire telle que $\mathbb{P}(X > 0) > 0$ et

$$\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s), \qquad \forall t, s \geq 0.$$

Montrer qu'il existe un $\lambda > 0$ tel que $X \sim \text{Exp}(\lambda)$.
Indice : Soit $G(t) = \mathbb{P}(X > t)$. L'absence de mémoire implique que $G(t+s) = G(t)G(s)$ pour $t, s \geq 0$ (pourquoi ?). Poser $g(t) = -\ln G(t)$ et $\lambda = g(1)$. Montrer que $g(t) = t\lambda$ pour chaque $t > 0$ rationnel. En déduire (avec justification !) que $g(t) = t\lambda$ pour chaque $t \geq 0$. Quel est le signe de λ ? Enfin, montrer que $\lambda < \infty$ en utilisant le fait que $G(0) > 0$ et la continuité à droite de G.

Exercice 22. Rappelons que pour un échantillon $x_1, \ldots, x_n$ la moyenne échantillonnale est définie par

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

et la médiane échantillonnale par

$$M = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{si } n \text{ est impair,} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{sinon.} \end{cases}$$

Montrer que

(i) la fonction $f(\gamma) = \sum_{i=1}^{n} (x_i - \gamma)^2$ atteint son minimum (uniquement) en $\bar{x}$.

(ii) la fonction $g(\gamma) = \sum_{i=1}^{n} |x_i - \gamma|$ atteint son minimum en M. **Attention :** g n'est pas dérivable au point γ si $\gamma = x_i$ pour un i quelconque.

Exercice 23.

(i) Calculez la moyenne $\bar{x}$ et la médiane M des données suivantes :

9.2	11.5	9.7	11.0	8.5
9.8	10.0	12.1	10.5	10.1

(ii) Refaire votre calcul quand la valeur 12.1 est remplacé par 48.6.

(iii) Comparez les valeurs de $\bar{x}$ et M dans les parties (i) et (ii). Qu'est-ce que vous notez ? Expliquez vos observations.

Exercice 24. Soit $x_1, \ldots, x_n$ un échantillon. Est-ce que c'est possible que la moyenne de cet échantillon est égal la médiane de cet échantillon, mais l'échantillon n'est pas symétrique. Trouvez un exemple.

Exercice 25 (exercice 17)**.** Montrer qu'une formule équivalente pour la variance empirique est $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2$. Expliquer pourquoi cette formule peut être plus utile.

Exercice 26 (exercice 18)**.** Soit un échantillon $x_1, \ldots, x_n$. Quels sont la médiane M et les quartiles Q_1 et Q_3 quand $n = 12$, 13, 14 ou 15 ?
**Bonus (c'est un peu fastidieux)* : trouver des formules générales (pour n quelconque) pour le premier et troisième quartile, Q_1 et Q_3. *Indice* : ces formules seront de la forme

$$\begin{cases} ? & n \equiv 0 \mod 4 \\ ? & n \equiv 1 \mod 4 \\ ? & n \equiv 2 \mod 4 \\ ? & n \equiv 3 \mod 4. \end{cases}$$

Exercice 27 (exercice 19)**.** Les données suivantes représentent les charges maximales (en tonnes) supportées par les câbles fabriqués par une usine :

10.1	12.2	9.3	12.4	13.7	11.1	13.3
10.8	11.6	10.1	11.2	11.4	11.8	7.1
12.2	12.6	9.2	14.2	10.5		

(i) Représenter les données sous la forme d'un histogramme dont la largeur des intervalles est égale à $h = 1$ et l'origine est égale à $\kappa = 10$. Refaire l'histogramme avec $h = 2$ et $\kappa = 11$ et comparer les deux figures.

(ii) Quelle est approximativement la valeur de la charge que les trois quarts des câbles peuvent supporter ?

(iii) Donner le troisième quartile.

(iv) Tracer une boîte à moustaches. Parmi les données, y a-t-il des valeurs aberrantes ? Dans ce diagramme, où visualise-t-on la valeur déterminée au point (ii) ?

Exercice 28 (exercices 70–71)**. (Il serait utile de lire la section 6.5 des notes de cours avant de commencer cet exercice.)**

(i) Soit $X \sim \text{Exp}(\lambda)$ où $\lambda > 0$. Montrer que le $\alpha-$quantile de X est

$$q_\alpha = F_X^-(\alpha) = -\log(1-\alpha)/\lambda,$$

pour $0 < \alpha < 1$.

(ii) Les fonctions quantile déterminent les distributions : soient X et Y des variables aléatoires quelconques avec des fonctions de répartition F_X et F_Y. Supposons que $F_X^-(\alpha) = F_Y^-(\alpha)$ pour tout $\alpha \in]0, 1[$. Montrer que $F_X = F_Y$.

Exercice 29 (exercice 20)**.** Le tableau suivant contient les résultats des matchs de rugby à XV des onzième et douzième journées (novembre 2014) du championnat français de rugby de première ("Top 14") et deuxième ("Pro D2") division. L'équipe jouant à domicile est celle notée à gauche du tiret.

Top 14		D2	
Montpellier – Brive	10–25	Albi – Agen	22–9
Castres – Toulon	22–14	Béziers – Aurillac	14–19
Clermont – Stade Français	51–9	Colomiers – Pau	50–10
Grenoble – Lyon	34–30	Montauban – Tarbes	31–13
Oyonnax – La Rochelle	37–9	Biarritz – Massy	21–3
Racing Métro – Bayonne	27–10	Dax – Narbonne	12–3
Bordeaux Bègles – Toulouse	20–21	Perpignan – Bourgoin	42–0
		Carcassonne – Mont-de-Marsan	17–28
Toulon – Clermont	27–19	Biarritz – Agen	42–18
Castres – Racing Métro	9–14	Albi – Carcassonne	34–22
La Rochelle – Bayonne	19–19	Aurillac – Colomiers	20–13
Lyon – Montpellier	23–20	Bourgoin – Montauban	14–20
Oyonnax – Bordeaux Bègles	28–23	Massy – Dax	50–13
Toulouse – Grenoble	22–25	Mont-de-Marsan – Béziers	32–18
Stade Français – Brive	20–17	Narbonne – Tarbes	36–23
		Pau – Perpignan	22–19

(i) Nous voulons comparer le comportement des équipes en première et en deuxième division. Pour ce faire, calculer pour chacune des divisions quelques statistiques pertinentes (la moyenne, la médiane, les quartiles et l'écart interquartile) pour la différence de points entre le club jouant à domicile et le club visiteur et pour le somme des points par match.

(ii) Représenter côte à côte, sous forme de deux boîtes à moustaches, la somme de points par match en première et en deuxième divison. Faire de même pour la différence de points. Quelles conclusions peut-on en tirer ?

Exercice 30 (exercice 21)**.** Soient $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. Montrer que $T(X_1, \ldots, X_n) = X_{(n)}$ est une statistique exhaustive pour θ, et trouver sa distribution d'échantillonage.

Exercice 31 (exercice 22)**.** Soient $X_1, \ldots, X_n \overset{iid}{\sim} Pois(\lambda)$. Montrer que $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ est une statistique exhaustive pour λ, et trouver sa distribution d'échantillonage.

Exercice 32 (le théorème 2.9 (p. 54) du livre)**.** Prouver que si $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, alors

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

où t_{n-1} représente la distribution de Student avec $n - 1$ degrés de libérte.

Exercice 33 (Une preuve alternative de la proposition 2.7, p. 51)**.** Soient $X_1, \ldots, X_n \overset{iid}{\sim}$

$N(\mu, \sigma^2)$. Définir

$$\begin{aligned}
\mathbf{a}_1 &= \frac{1}{\sqrt{n}}(1, 1, \ldots, 1)', \\
\mathbf{a}_2 &= \frac{1}{\sqrt{2}}(1, -1, 0, \ldots, 0)', \\
\mathbf{a}_3 &= \frac{1}{\sqrt{6}}(1, 1, -2, 0, \ldots, 0)', \\
&\vdots \\
\mathbf{a}_n &= \frac{1}{\sqrt{n(n-1)}}(1, 1, \ldots, 1, -(n-1))'.
\end{aligned}$$

(i) Définir la $n \times n$ matrice $\mathbf{A} = [\mathbf{a}_1 : \mathbf{a}_2 : \ldots : \mathbf{a}_n]$. Montrer que $\mathbf{A}$ est un matrice orthogonale, c'est-à-dire $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}_n$, où $\mathbf{I}_n$ est la $n \times n$ matrice d'identité.

(ii) Définir la transformation $Y_i = \mathbf{a}_i'(\mathbf{X} - \mathbf{m}), i = 1, 2, \ldots, n$, où $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ et $\mathbf{m} = (\mu, \mu, \ldots, \mu)'$. Trouvez la densité conjointe de $Y_1, Y_2, \ldots, Y_n$. Sont-ils indépendants ? Quelle est la distribution de Y_i pour chaque i ?

(iii) Montrer que

$$Y_1 = \sqrt{n}(\bar{X} - \mu) \quad \& \quad \sum_{i=2}^{n} Y_i^2 = (n-1)S^2.$$

Indice : Puisque $\mathbf{A}$ est une matrice orthogonale, $\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n (X_i - \mu)^2$.

(iv) Utilisez la partie (iii) pour montrer que $\bar{X}$ et S^2 sont indépendants. Montrer aussi que $\bar{X} \sim N(\mu, \sigma^2/n)$ et $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Exercice 34. Soient $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(0, 1)$. Montrer que $X_i^2 / \sum_{j=1}^n X_j^2$ et $\sum_{j=1}^n X_j^2$ sont indépendants pour chaque $i = 1, 2, \ldots, n$.

Exercice 35 (exercice 23)**.** Soient $X_1, \ldots, X_n \overset{iid}{\sim} f$, où f est de la forme d'une famille exponentielle, exprimée dans la paramétrisation usuelle comme $f(x) = \exp[\eta(\theta)T(x) - d(\theta) + S(x)]$, $\theta \in \Theta \subseteq \mathbb{R}$ ouvert. Montrer que :

(i) Si η est k-fois continûment dérivable ($k \geq 1$) et inversible avec la dérivée jamais nulle, alors d est aussi k-fois continûment dérivable.

(ii) Si η est deux fois continûment dérivable et inversible avec la dérivée jamais nulle, alors

$$\mathbb{E}[\tau(X_1, \ldots, X_n)] = n\frac{d'(\theta)}{\eta'(\theta)} \quad \& \quad \mathrm{Var}[\tau(X_1, \ldots, X_n)] = n\frac{d''(\theta)\eta'(\theta) - d'(\theta)\eta''(\theta)}{[\eta'(\theta)]^3},$$

où $\tau(X_1, \ldots, X_n) = \sum_{i=1}^n T(X_i)$.

Indice : utiliser le théorème de la fonction inverse (théorème 6.2, p. 162).

Exercice 36 (loi des événements rares, exercice 24)**.** Soit $\{X_n\}_{n \geq 1}$ une séquence de variables aléatoires $\mathrm{Bin}(n, p_n)$, telle que $p_n = \lambda/n$, pour une certaine constante $\lambda > 0$. Montrer que $X_n \xrightarrow{d} Y$, où $Y \sim \mathrm{Poisson}(\lambda)$.

Indice : (1) montrer que pour $k \in \mathbb{N} \cup \{0\}$, $\mathbb{P}(X_n = k) \to \mathbb{P}(Y = k)$. (2) Déduire que $\mathbb{P}(X_n \leq k) \to \mathbb{P}(Y \leq k)$. (3) Conclure.

Exercice 37 (de la distribution exponentielle à la géométrique et inversement)**.**

(i). Soit $X \sim \text{Exp}(\lambda)$ pour $\lambda > 0$. Montrer que $\lfloor X \rfloor \sim \text{Geom}(p)$ pour un p approprié à trouver. (On définit $\lfloor t \rfloor = \max\{n \in \mathbb{Z} : \ n \leq t\}$, pour $t \in \mathbb{R}$.)

(ii). Soit $\{X_n\}_{n=1}^{\infty}$ une suite de variables aléatoires avec $X_n \sim \text{Geom}\left(\frac{\lambda}{n}\right)$ et soit $Z \sim \text{Exp}(\lambda)$, pour un certain $\lambda > 0$. Montrer que $\frac{X_n}{n} \xrightarrow{d} Z$, lorsque $n \to \infty$.

Exercice 38 (exercice 25)**.** On dit qu'une suite de variables aléatoires X_n converge vers une variable aléatoire Y *en probabilité* (p. 60) si

$$\forall \epsilon > 0 \qquad \lim_{n\to\infty} \mathbb{P}[|X_n - Y| > \epsilon] = 0.$$

Dans ce cas on écrit $X_n \xrightarrow{p} Y$.

Soit $\{X_n\}_{n=1}^{\infty}$ une suite de variables aléatoires avec

$$X_n = (-1)^n X, \quad \mathbb{P}(X = -1) = \mathbb{P}(X = 1) = \frac{1}{2}.$$

Montrer que $X_n \xrightarrow{d} X$, mais que $X_n \overset{p}{\nrightarrow} X$.

Exercice 39 (exercise 27)**.** Soient $X_1, \ldots, X_n \overset{iid}{\sim} Pois(\lambda)$, où $\lambda \in (0, \infty)\backslash\{1\}$ et considérons la probabilité $\pi = \mathbb{P}(X_i = 1) = \lambda e^{-\lambda}$. Nous voulons estimer π par $\hat{\pi}_n = \hat{\lambda}_n e^{-\hat{\lambda}_n}$ où $\hat{\lambda}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Montrer que

$$\frac{\sqrt{n}(\hat{\pi}_n - \pi)}{\sqrt{\hat{\lambda}_n} e^{-\hat{\lambda}_n}(1 - \hat{\lambda}_n)} \xrightarrow{d} Y,$$

où $Y \sim N(0, 1)$. Indication : vous aurez besoin du théorème limite central, de la méthode delta, de la loi faible des grands nombres ainsi que du théorème de Slutsky.

Exercice 40 (exercice 28)**.** Soient $x_1, \ldots, x_n$ des réalisations indépendantes d'une variable aléatoire X ayant une fonction de densité f continue. Soit $y \in \mathbb{R}$, montrer que la fonction $hist_{x_1,\ldots,x_n}(y)$ converge en probabilité vers $f(y)$, lorsque $n \to \infty$, $h_n \to 0$ et $nh_n \to \infty$. Indication : le nombre d'observation tombant dans l'intervalle I_{j_n}, donné par $N_n = \sum_{i=1}^{n} 1_{\{x_i \in I_{j_n}\}}$, suit une loi $\text{Bin}(n, p_n)$ où $p_n = \int_{I_{j_n}} f(x)\text{d}x$. Vous aurez besoin d'utiliser le fait que

$$\left|\frac{N_n}{nh_n} - f(y)\right| \leq \left|\frac{N_n}{nh_n} - \frac{p_n}{h_n}\right| + \left|\frac{p_n}{h_n} - f(y)\right|,$$

ainsi que l'inégalité de Chebyshev (lemme 6.4, p. 163).

***Exercice 41** (exercice 26)**.** Prouver le lemme 2.20 (p. 60) du livre.
(*L'étoile est la notation standard dans les livres de mathématiques pour des exercices plus difficiles.*)

Exercice 42 (exercise 29)**.** Nous allons traiter la question de l'existence d'estimateurs non biaisés.
Soit $Y \sim \text{Bin}(n, p)$, où $p \in]0, 1[$.

(i) Montrer que Y/n est un estimateur non biaisé pour p.

(ii) Montrer qu'il n'existe pas d'estimateur non biaisé pour $1/p$.

(iii) Montrer qu'il n'existe pas d'estimateur non biaisé pour le paramètre naturel $\phi = \log\left(\frac{p}{1-p}\right)$.
Remarque : ϕ s'appelle le *log odds ratio* ou de manière moins anglophone le *log du rapport des chances*.

Exercice 43. Soient $X_1, \dots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$. Définissons les estimateurs $\widehat{\lambda}_n = \overline{X}_n = \sum_{i=1}^n X_i/n$ et $S_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.
Montrer que $\operatorname{Var} S_n^2 \geq \operatorname{Var} \widehat{\lambda}_n$.
Indice : la borne de Cramér–Rao peut s'averer utile.

Exercice 44. Soient $X_1, \dots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$, où $n > 2$.

(i) Montrer que l'estimateur $\hat{\lambda}_n = (\overline{X})^{-1}$ est consistent pour λ.

(ii) Montrer que $\mathbb{E}_\lambda(\hat{\lambda}_n) = \lambda n/(n-1)$, et trouver un estimateur $\hat{\lambda}_n^{\text{NB}}$ non biaisé de λ.
Indice : utiliser le fait que $Z = \sum_{i=1}^n X_i \sim Gamma(n, \lambda)$.

(iii) Montrer que $\operatorname{Var}_\lambda(\hat{\lambda}_n) = n^2\,\lambda^2 / \big((n-1)^2\,(n-2)\big)$.

(iv) L'estimateur $\hat{\lambda}_n^{\text{NB}}$ atteint-il la borne inférieure de Cramér–Rao ?

Exercice 45. Soient $X_1, \dots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$.

(i) Montrer que l'estimateur du maximum de vraisemblance $\hat{\lambda}_n$ de λ est consistant et non-biaisé.

(ii) Donner un estimateur (par exemple une simple modification de $\hat{\lambda}_n$) qui est consistant, mais néanmoins biaisé.

Exercice 46 (exercice 31)**.** Soient $X_1, \dots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$, où $n > 2$.

(i) Trouver l'estimateur du maximum de vraisemblance $\hat{\lambda}_n$.

(ii) Déterminer l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{\text{MV}}$ et la borne de Cramér–Rao associés au paramètre $\theta = 1/\lambda$. Peut-on utiliser la proposition 3.17 ?

(iii) Comparer $\widehat{\lambda}_n$ et $\widehat{\theta}_n^{\text{MV}}$ avec les bornes de Cramér–Rao correspondantes. *Attention* : quand l'estimateur est biaisé, le nominateur de la borne de Cramér–Rao n'est pas 1.

Exercice 47 (exercice 33)**.** Un malheureux époux bavarde souvent à son téléphone portable afin d'oublier ses misères. On sait que la longueur de ses jasettes téléphoniques suit une loi exponentielle de paramètre $\lambda > 0$. Longtemps gênée par les conversation de son époux, la femme de ce monsieur malchanceux se mit à mesurer la longueur de celles-ci ; ayant un nombre infini d'observations, elle connaît la valeur précise du paramètre λ.
Lors d'une dispute avec son mari et afin d'avoir un argument plus concret, la femme montra à son époux un échantillon $t_1, \dots, t_n$ des longueurs de n de ses conversations téléphoniques, et ce, afin de lui prouver qu'il placote au téléphone de manière excessive.
L'homme, tout méfiant, ne croit guère sa femme ; connaissant celle avec laquelle il vit déjà depuis quelques décennies, il la soupçonne d'avoir choisi l'échantillon de manière aléatoire, mais uniquement à partir des conversations qui duraient plus longtemps que la moyenne (théorique) de la longueur des conversations. En supposant ceci, le bavard s'attaque au problème d'estimer le paramètre λ, dont seule son épouse connaît la valeur véritable.

Trouver l'estimateur de maximum de vraisemblance de λ à partir de l'échantillon $t_1, \ldots, t_n$, mais sous l'hypothèse que le monsieur a raison. *Attention* : comme à l'exemple 3.20 (du livre), le support de la distribution dépend de l'état de la nature, c'est-à-dire de la vraie valeur de λ.

Exercice 48 (exercice 35)**.** Soient $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ où les deux paramètres sont inconnus ($n > 1$). On peut estimer σ^2 par

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

ou bien par l'estimateur de maximum de vraisemblance $\widehat{\sigma}_n^2 = (n-1)S_n^2/n$ (cf. l'exemple 3.16, p. 75).

(i) Lequel de ces estimateurs est meilleur au sens de l'erreur quadratique moyenne ?
Indication : on a $(n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$ (cf. proposition 2.7, p. 51).

(ii) Considérons les estimateurs de la forme aS_n^2 où $a \in \mathbb{R}$. Quelle est la meilleure valeur de a au sens de l'erreur quadratique moyenne ?

Exercice 49. Soient $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$, où $\theta > 0$. Soit $\widehat{\theta}_n$ l'estimateur de maximum de vraisemblance. Trouver $\widehat{\theta}_n$ et montrer que $n(\theta - \widehat{\theta}_n)$ converge en distribution vers une distribution à trouver.

Exercice 50 (exercices 36 et 37)**.**

(i) Considérons la représentation usuelle d'une famille exponentielle

$$f(x;\theta) = \exp(\eta(\theta)T(x) - d(\theta) + S(x)), \qquad x \in \mathcal{X}, \quad \theta \in \Theta,$$

où $\Theta \subseteq \mathbb{R}$ est un ouvert et η est deux fois continûment dérivable et inversible avec la dérivée jamais nulle. Soient $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. Montrer que

$$\mathbb{E}\left[\frac{\partial}{\partial\theta} \log f(X_1, \ldots, X_n; \theta)\right] = 0, \quad \text{et}$$

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f(X_1, \ldots, X_n; \theta)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(X_1, \ldots, X_n; \theta)\right]. \tag{1}$$

Indication : ce n'est pas pour rien qu'on a fait l'exercice 35.

(ii) * Soit $f(x;\theta)$ un modèle paramétrique régulier (pas forcément une famille exponentielle !) tel que

$$\mathcal{X} = \{x \in \mathbb{R} : f(x;\theta) > 0\}$$

ne dépend pas de θ, et que f est doublement dérivable par rapport à θ. Soient en plus $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. Montrer que l'égalité (1) est équivalente à une condition de régularité qui dit que l'on peut interchanger la dérivée et l'intégrale.
Indication : il faut absolument se rendre compte que pour chaque fonction $g : \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{E}[g(X)] = \int_{\mathcal{X}^n} g(\vec{x}) f(\vec{x};\theta)\, \mathrm{d}\vec{x} \qquad \text{quand cette intégrale existe} \quad (\vec{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n).$$

Exercice 51. Soit la variable aléatoire X, dont la densité est donnée par

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

où $\theta > 0$ est un paramètre inconnu. Trouver, sans calculer aucune intégrale, $\mathbb{E}[\log X]$ et $\mathbb{E}[(\log X)^2]$.
Remarque. Cette méthode est beaucoup moins laborieuse que de calculer explicitement

$$\int_0^1 \theta x^{\theta-1} \log x \,\mathrm{d}x \qquad \text{et} \qquad \int_0^1 \theta x^{\theta-1} (\log x)^2 \,\mathrm{d}x.$$

Exercice 52. Soit $X \sim \mathrm{Exp}(\lambda)$, où $\lambda > 0$. Montrer que $Y = aX \sim \mathrm{Exp}(\lambda/a)$ pour $a > 0$.

Exercice 53. Nous avons montré une sorte de théorème centrale limite pour les familles exponentielles (théorème 3.23, p. 81 ; corollaire 3.27, p. 84). Nous verrons dans cet exercice deux exemples de ce qui se passe en dehors du cadre des familles exponentielles.
Considérons $\widehat{\lambda}_n$, l'estimateur de l'exercice 47. Trouver une suite de nombres réels a_n telle que $a_n(\lambda - \widehat{\lambda}_n)$ converge en distribution vers une distribution non dégénérée.
Indication : utiliser l'exercice 13 et l'exercice 52.

Exercice 54. (i). Soit $X = (x_1, \ldots, x_n)^T$ une image de dimension 1. Supposons que l'on puisse uniquement observer une version de cette image sur laquelle il y a du bruit numérique, i.e, que l'on observe $Y = (y_1, \ldots, y_n)^T$, où chaque pixel s'écrit comme

$$y_i = x_i + \varepsilon_i,$$

où $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Trouver une estimation de l'image originale X par la méthode du maximum de vraisemblance.

(ii). Supposons maintenant que l'on vous donne une information supplémentaire sur l'allure de l'image : l'image est en fait une ligne, où chaque pixel satisfait la relation

$$y_i = a + bx_i + \varepsilon_i.$$

Calculer l'estimateur du maximum de vraisemblance des paramètres a et b.

Exercice 55. Soient $x_1, \ldots, x_n \overset{iid}{\sim} Gamma(r, 1)$. Trouver l'estimateur des moments $\widehat{r}^{\mathrm{mom}}$ de r, et la loi limite de

$$\sqrt{n}\left(\widehat{r}^{\mathrm{mom}} - \frac{(\log\Gamma(\cdot))'(\widehat{r}^{\mathrm{mom}}) - \overline{\log X}}{(\log\Gamma(\cdot))''(\widehat{r}^{\mathrm{mom}})} - r\right)$$

où $\overline{\log X} = n^{-1}\sum \log x_i$.

Exercice 56. Soit $X_1, \ldots, X_n$ un échantillon i.i.d. tiré d'une distribution de densité

$$f(x;\theta) = \begin{cases} 3\theta^3 x^{-4}, & \text{si } x \geq \theta, \\ 0, & \text{sinon,} \end{cases}$$

où $\theta > 0$.

(i) Trouver l'estimateur $\hat{\theta}_n^{\mathrm{MoM}}$ de θ par la méthode des moments.

(ii) Trouver l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{\mathrm{MV}}$ de θ.

(iii) Montrer que $\hat{\theta}_n^{\mathrm{MoM}}$ est non-biaisé, tandis que $\hat{\theta}_n^{\mathrm{MV}}$ est un estimateur biaisé.

(iv) Calculer l'erreur quadratique moyenne de $\hat{\theta}_n^{\mathrm{MoM}}$ et de $\hat{\theta}_n^{\mathrm{MV}}$. Quel estimateur est le meilleur au sens de l'erreur quadratique moyenne ?

Exercice 57. Soit $X_1, \ldots, X_n$ un échantillon i.i.d. tiré de la distribution uniforme sur $[0, \theta]$ où le paramètre $\theta > 0$ est inconnue. Dans les exercices précédentes on a trouvé l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{\mathrm{MV}} = X_{(n)}$.

(i) Trouver l'estimateur $\hat{\theta}_n^{\mathrm{MoM}}$ de θ par la méthode des moments. Montrer qu'il est non-biaisé.

(ii) Modifier l'estimateur $\hat{\theta}_n^{\mathrm{MV}}$, par example en multipliant par un constant, pour le rendre non-biaisé. Dénoter cet estimateur $\hat{\theta}_n^{\mathrm{MV,modif}}$.

(iii) Calculer l'erreur quadratique moyenne de $\hat{\theta}_n^{\mathrm{MoM}}$ et de $\hat{\theta}_n^{\mathrm{MV,modif}}$. Quel estimateur est le meilleur au sens de l'erreur quadratique moyenne ?

(iv) Commenter la vitesse de convergence de l'erreur quadratique moyenne de ces deux estimateur.

Exercice 58. Soit $X_1, \ldots, X_n$un échantillon i.i.d. tiré de la distribution binomial avec les deux paramètres m et p inconnues. Trouver $\hat{m}$, $\hat{p}$ les estimateurs des m et p par la méthode des moments. Montrer que cela peut arriver que $\hat{m} \notin \{0, 1, \ldots\}$ ou $\hat{p} \notin (0, 1)$.

***Exercice 59.** (un exercice théorétique)
Soit $f(x; \theta) = \exp(T(x)\eta(\theta) - d(\theta) + S(x))$ une famille exponentielle non dégénérée, où l'espace des paramètres Θ est ouvert, et soit $x_1, x_2, \ldots, x_n$ un échantillon iid tiré de $f(x; \theta_0)$ pour un certain θ_0. Soit α_n n'importe quel estimateur tel que $\sqrt{n}(\alpha_n - \theta_0) \to V$ pour une variable aléatoire V. Imaginons qu'on cherche à approximer l'estimateur de maximum de vraisemblance $\widehat{\theta}_n$ avec une seule itération de Newton–Raphson,

$$\beta_n = \alpha_n - \frac{\ell_n'(\alpha_n)}{\ell_n''(\alpha_n)}.$$

En supposant que $\eta \in C^3(\Theta)$, montrer que

$$\sqrt{n}(\beta_n - \theta_0) \to N\left(0, \frac{1}{I(\theta_0)}\right),$$

où $I(\theta_0)$ est l'information de Fisher, et commenter ce résultat.
Indice : faire une développement de Taylor d'ordre 2 de ℓ_n' autour de θ_0, et remarque que cette fonction (aléatoire !) est une somme de variables aléatoires iid.

Exercice 60 (exercice 40)**.** Pour chacun des scénarios suivants, trouver les hypothèses à tester ainsi que les deux types d'erreurs qu'on peut commettre. Sur la base de ces informations, décider quelle hypothèse devrait être l'hypothèse nulle H_0 et laquelle devrait être l'alternative H_1.

(i) Une physicienne travaille sur une expérience dont le but est de détecter des particules de matières noires. Elle aimerait tester si ses données indiquent la présence de matière noire.

(ii) Un fêtard voudrait savoir s'il est en mesure de conduire après un apéro. Il aimerait donc tester si le taux d'alcool dans son sang est supérieur à celui autorisé par la loi.

(iii) Barack Obama et Mitt Romney étaient les deux candidats principaux à l'élection présidentielle de 2012 aux États-Unis. Le directeur de campagne de M. Obama aimerait savoir si M. Obama est en tête dans l'état d'Iowa afin de décider s'il doit allouer ou non plus de ressources financières pour la campagne dans cet état. Il faut donc tester si M. Obama est en tête dans l'état d'Iowa. De quelle façon le test changerait-il si on était à la place du directeur de campagne de M. Romney ?

(iv) Un scientifique travaillant pour une compagnie pharmaceutique a pu développer un nouveau médicament afin de réduire la pression artérielle trop élevée. Il voudrait tester si le médicament produit l'effet attendu.

Exercice 61 (**tests d'hypothèses intuitifs**, exercice 48)**.** Le but de cet exercice est de donner une motivation intuitive aux tests d'hypothèses. Soient $X_1, \dots, X_n$ iid avec la fonction de densité

$$f_X(x) = \frac{1}{48}\lambda^5 x^{3/2} e^{-\lambda\sqrt{x}}, \qquad x > 0,$$

où $\lambda > 0$ est un paramètre. On aimerait tester l'hypothèse $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda = \lambda_1$, où $\lambda_0 > \lambda_1$.

(i) Trouver l'estimateur du maximum de vraisemblance $\widehat{\lambda}_n$.

(ii) Comme expliqué au chapitre 3 du livre, $\widehat{\lambda}_n$ est un bon estimateur. Ainsi, il est en un certain sens naturel de rejeter H_0 si λ_0 n'est pas « compatible »avec $\widehat{\lambda}_n$. Dans notre cas, cela voudrait dire : rejeter H_0 lorsque $\widehat{\lambda}_n$ est petit. (Si $\widehat{\lambda}_n > \lambda_0$, on préferera certainement H_0 et non H_1.) Quelle forme prendra donc la fonction de test ? Donner-la à une constante D près.

(iii) Maintenant, il faut trouver la fonction de test précise. Pour cela, il faudrait choisir une borne en dessous de laquelle on juge $\widehat{\lambda}_n$ suffisamment petit pour rejeter H_0. Pour un seuil $\alpha \in]0, 1[$ donné, on voudrait que la probabilité de commettre une erreur de type I soit α. À partir de là, décrire la relation entre α et D.

(iv) Nous voilà un test au niveau α. On peut ensuite se demander s'il est le meilleur test. Aurons-nous pu faire mieux, c'est-à-dire trouver un test au niveau α mais plus puissant ? Montrer que la réponse est négative, en montrant que notre fonction de test est exactement la même que celle décrite par le lemme de Neyman–Pearson. (On peut supposer que la valeur Q du lemme existe ; ce résultat sera démontrée ultérieurement.)

(v) Trouver une formule, la plus simple possible, pour la fonction de test $\delta(X_1, \dots, X_n)$. *Indice :* $\widehat{\lambda}_n$ contient une somme dont chaque élément suit une distribution qu'on a déjà vu.

Exercice 62 (exercice 41)**.** Soit $X_1, \dots, X_n$ un échantillon iid provenant d'une distribution $N(\mu, 1)$. On va tester l'hypothèse nulle $H_0 : \mu = 0$ vs. l'hypothèse alternative $H_1 : \mu \neq 0$ en utilisant la statistique de test

$$T_n(X_1, \dots, X_n) = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

et la fonction de test

$$\delta(X_1, \ldots, X_n) = \begin{cases} 1, & \text{si } |T_n(X_1, \ldots, X_n)| \geq Q, \\ 0, & \text{sinon,} \end{cases}$$

où $Q > 0$.

(i) Trouver la probabilité de commettre une erreur de type I.

(ii) Trouver la probabilité de commettre une erreur de type II.

(iii) Comment se comportent ces deux probabilités lorsqu'on augmente la valeur de Q ?

(iv) Trouver la plus petite valeur de Q pour laquelle le seuil de signification du test est $\alpha \in]0, 1[$. Quelle est cette valeur lorsque $\alpha = 0.05$ et $n = 10$? Trouver le supremum de la probabilité de commettre une erreur de type II pour cette valeur de Q.

Exercice 63 (exercice 45)**.** Soient $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ avec $\sigma^2 > 0$ connue. Trouver le test le plus puissant pour tester $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$ avec $\mu_0 < \mu_1$ à un seuil de signification $\alpha \in (0, 1)$.

Exercice 64 (exercice 46)**.** Pour un échantillon $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim}$ Bernoulli(p), on veut tester

$$H_0 : \ p = 0.49 \ \text{ vs } \ H_1 : \ p = 0.51.$$

Déterminez approximativement la taille de l'échantillon pour laquelle la probabilité de commettre une erreur de type I et la probabilité de commettre une erreur de type II sont approximativement égales à 0.01. Utilisez une fonction de test qui rejette H_0 si $\sum_i X_i$ est grande.
Indice : Utilisez le théorème centrale limite pour approximer la distribution de $n^{-1} \sum_{i=1}^n X_i$ *par une loi normale. Vous avez aussi besoin du fait que* $z_{0.99} \approx 2.33$*, où* $z_{0.99}$ *est le 0.99-quantile de la loi* $N(0, 1)$*.*

Exercice 65 (exercice 47)**.** Soient $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Unif(0, \theta)$ et considérez $H_0 : \theta = \theta_0$ et $H_1 : \theta = \theta_1$ avec $\theta_1 < \theta_0$.

(i) Trouvez le test le plus puissant de H_0 vs. H_1 à un seuil de signification $\alpha = (\theta_1/\theta_0)^n$. Considérez le comportement de ce seuil, comme fonction de θ_0, θ_1 et n. Quelle est la puissance de ce test ? Est-ce qu'on peut définir un test optimal de type Neyman–Pearson pour d'autres valeurs de α ?

(ii) Considérez un test (pas nécessairement optimal) de seuil de signification $\alpha < (\theta_1/\theta_0)^n$ qui rejette H_0 quand $X_{(n)} < k$. Trouvez la valeur appropriée de k. Quelle est la puissance de ce test ?

Exercice 66 (exercice 49)**.** Un laboratoire de traitement d'images a développé une nouvelle méthode pour scanner le cerveau. Le laboratoire prétend qu'ils sont capables de scanner le cerveau en moins de 20 minutes. Voici un échantillon de temps de 12 scans de cerveau :

$$\mathbb{X} = \{21, 18, 19, 16, 18, 24, 22, 19, 24, 26, 18, 21\}.$$

(i) Supposons que la durée de scan suit $\mathcal{N}(\mu, 3^2)$. Testez si la durée moyenne de scan est moins de 20 minutes, i.e., testez $H_0 : \ \mu \leq \mu_0$ vs $H_1 : \ \mu > \mu_0$ avec $\mu_0 = 20$ à un seuil de signification $\alpha = 0.05$.

(ii) Pourriez-vous faire la même analyse sachant que la variance de la loi normale est inconnue ? *Indice : Utilisez* $\delta = \mathbf{1}\left(\frac{\sqrt{n}(\bar{X}-\mu_0)}{S} \geq t_{n-1,1-\alpha}\right)$ *comme fonction de test. Ici* $t_{n-1,1-\alpha}$ *est le* $1-\alpha$ *quantile de la loi Student avec* $n-1$ *degrés de liberté.*

Exercice 67 (exercice 50)**.** Soient $Y_1, \ldots, Y_4$ des variables aleatoires independantes et identiquement distribues selon une loi normale $\mathcal{N}(\mu, 4^2)$. On veut montrer que μ est plus grand que $\mu_0 = 10$. Par conséquent, on effectue un test au niveau $\alpha = 5\%$ de l'hypothèse nulle $H_0 : \mu \leq 10$.

(i) Calculez la puissance du test pour des vraies valeurs de μ égales à 13 et 11.

(ii) Pour augmenter la chance de détection, déterminez le nombre d'observations nécessaires pour obtenir une puissance de 90% dans le cas $\mu = 13$.

Exercice 68 (exercice 51, **test apparié**)**.** Une compagnie pharmaceutique veut vérifier si son nouveau produit amaigrissant ABALGRA est efficace. Pour ce faire, le poids (en kilo) de 10 hommes choisis de façon aléatoire a été recueilli juste avant la première prise du médicament ainsi qu'à la fin du traitement, 7 semaines plus tard. Soit X_i le poids du i^{e} homme avant le traitement et soit Y_i son poids à la fin du traitement. On peut donc supposer que X_i sont iid, puisque les différentes personnes ont été choisies au hasard. De même pour Y_i, car chaque personne a reçu le même traitement. Soient $\mu_1 = \mathbb{E}X_i$ et $\mu_2 = \mathbb{E}Y_i$.
On s'intéresse donc aux différences $d_i = Y_i - X_i$. Celles-ci sont indépendantes et on suppose qu'elles suivent une loi normale $\mathcal{N}(\mu_2 - \mu_1, 5)$. Tester à l'aide des données du tableau ci-dessous si le médicament semble entraîner une perte de poids au seuil $\alpha = 0.05$.

i	1	2	3	4	5	6	7	8	9	10
X_i	55.5	75	63.8	54.7	62.7	71	68.3	56	74.4	65
Y_i	52.8	73.7	62.7	55	59.3	70.2	67.1	55.4	71.9	65.2

Remarque. Puisque X_1 et Y_1 proviennent de la même personne, il est irréaliste de les supposer indépendantes. Dans ce contexte, on parle d'un *test apparié* (angl. « paired test »).
Bonus. Expliquer le nom ABALGRA.

Exercice 69 (exercice 52, **test de variance pour la loi gaussienne**)**.**

(i) Soit $X_1, \ldots, X_n$ un échantillon iid tiré d'une distribution normale $\mathcal{N}(\mu, \sigma^2)$, où les paramètres μ et σ^2 sont inconnus. Montrer que la fonction de test du test du rapport de vraisemblance pour les hypothèses $H_0 : \sigma^2 = \sigma_0^2$ et $H_1 : \sigma^2 \neq \sigma_0^2$ à un seuil α est de la forme $\mathbf{1}\{W > c_1\} + \mathbf{1}\{W < c_2\}$, où $W = (1/\sigma_0^2)\sum_{i=1}^n (X_i - \bar{X})^2$ et où c_1 et c_2 sont tels que $c_1^{-n} e^{c_1} = c_2^{-n} e^{c_2}$.
Indice : écrire le rapport de vraisemblance comme une fonction de W et étudier la forme de cette fonction.

(ii) En pratique, on choisit c_1 et c_2 tel que $\mathbb{P}_{H_0}(W > c_1) = \mathbb{P}_{H_0}(W < c_2) = \alpha/2$. (Le test obtenu n'est donc pas un test du rapport de vraisemblance.) Trouver les valeurs de c_1 et c_2 lorsque $\alpha = 0.05$, et effectuer ce test pour $\sigma_0^2 = 4$ sur les données suivantes :

$$0.449,\ -3.421,\ -2.841,\ 0.829,\ -0.941,\ 1.789,\ 0.889,\ 1.109,\ 0.969,\ 1.169$$

(Noter que $\bar{X} = 0$.)

Exercice 70. La brasserie québécoise Unibroue produit des bières mondialement reconnues [1]. Elle souhaite vérifier si les bouteilles de bière qu'elle produit contiennent bien 341 ml, comme indiqué à l'étiquette. En effet, si la quantité était inférieure à 341 ml, la brasserie risquerait un mécontentement de la part de sa fidèle clientèle, ainsi que des problèmes juridiques. En revanche, une quantité supérieure à 341 entrainerait des pertes financières. Afin d'effectuer cette vérification, la quantité de bière dans $n = 100$ bouteilles a été mesurée, et les valeurs $x_1, \dots, x_n$ ont été observées. On suppose que les observations x_i sont indépendantes et tirées d'une loi normale $\mathcal{N}(\mu, \sigma^2)$ dont les deux paramètres sont inconnus. Les observations obtenues sont de moyenne $\bar{x} = 337$ et de variance échantillonnale $S^2 = 40$. Tester à un niveau $\alpha = 0.05$ si les bouteilles produites contiennent en moyenne 341 ml.
Indice : consulter l'exemple 4.22 (p. 119).
Est-ce que la conclusion changerait si n était égal à 10 ?

Exercice 71 (exercice 54)**.**

(i) Soit $X_1, \dots, X_n$ un échantillon tiré d'une distribution de Poisson de paramètre θ. Nous voulons tester $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. Trouver un test du rapport de vraisemblance approximatif permettant de tester ces deux hypothèses.
Indice : utiliser le théorème 4.23.

(ii) Supposons que nous ayons observé $n = 100$ observations de moyenne $\bar{x} = 2.1$. Tester à un seuil de signification $\alpha = 0.05$ les hypothèses H_0 et H_1 définies ci-dessus pour $\theta_0 = 2$.

Exercice 72 (exercice 55)**.** Soit un échantillon iid $X_1, \dots, X_n$ issu d'une loi $N(0, \sigma^2)$ où la variance σ^2 est inconnue. Construire un test de Wald approximatif (de niveau α) afin de tester l'hypothèse $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$ pour $\sigma_0^2 > 0$ fixé. Comparer avec le test du rapport de vraisemblance.

Exercice 73 (exercice 56)**.** Soit un échantillon iid $X_1, \dots, X_n$ issu d'une loi Bernoulli de paramètre p inconnu. Construire un test de Wald approximatif (de niveau α) afin de tester l'hypothèse $H_0 : p = p_0$ versus $H_1 : p \neq p_0$ pour $p_0 \in]0, 1[$ fixé. Comparer avec le test de rapport du vraisemblance.

Exercice 74 (exercice 53, **test non apparié**)**.** Soit un échantillon $X_1, \dots, X_n$, $Y_1, \dots, Y_m$ de $n + m$ variables aléatoires indépendantes, où $X_i \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma^2)$ et $Y_i \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma^2)$, où σ^2 est inconnue (mais la même pour les X et les Y). Le but de cet exercice est de trouver le test du rapport de vraisemblance permettant de tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$.

(i) Définir la fonction de vraisemblance du paramètre $\theta = (\mu_1, \mu_2, \sigma^2)$.

(ii) En remarquant que $\Theta_0 = \{(\mu, \mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$ et que $\Theta_1 = \{(\mu_1, \mu_2, \sigma^2) : -\infty < \mu_1 \neq \mu_2 < \infty, 0 < \sigma^2 < \infty\}$, montrer que

$$\sup_{\theta \in \Theta_0} L(\theta) = \left(\frac{e^{-1}}{2\pi \hat{\sigma}^2_{\Theta_0}} \right)^{(m+n)/2},$$

où $\hat{\sigma}^2_{\Theta_0} = \frac{1}{n+m} \left(\sum_{i=1}^n (X_i - \hat{\mu})^2 + \sum_{j=1}^m (Y_j - \hat{\mu})^2 \right)$, avec $\hat{\mu} = \frac{1}{n+m} \left(\sum_{i=1}^n X_i + \sum_{j=1}^m Y_j \right)$.
Montrer aussi que

$$\sup_{\theta \in \Theta_1} L(\theta) = \left(\frac{e^{-1}}{2\pi \hat{\sigma}^2_{\Theta_1}} \right)^{(m+n)/2},$$

1. http ://www.unibroue.com/fr/unibroue/medals

où $\hat{\sigma}^2_{\Theta_1} = \frac{1}{n+m}\left(\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2\right)$.

(iii) En utilisant le fait que $\sum_{i=1}^n (X_i - \hat{\mu})^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + \frac{nm^2(\bar{X}-\bar{Y})^2}{(n+m)^2}$ et que $\sum_{j=1}^m (Y_j - \hat{\mu})^2 = \sum_{j=1}^m (Y_j - \bar{Y})^2 + \frac{mn^2(\bar{X}-\bar{Y})^2}{(n+m)^2}$, montrer que

$$\Lambda(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = \left(1 + \frac{t^2}{m+n-2}\right)^{(n+m)/2},$$

où

$$t = \frac{\sqrt{\frac{nm}{n+m}}(\bar{X} - \bar{Y})}{\sqrt{\frac{1}{n+m-2}[(n-1)S_X^2 + (m-1)S_Y^2]}},$$

avec $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ et $S_Y^2 = \frac{1}{m-1}\sum_{j=1}^m (Y_j - \bar{Y})^2$.

(iv) En utilisant le fait que le test de niveau α dont la fonction de test est donnée par $\mathbf{1}\{\Lambda(X_1, \ldots, X_n, Y_1, \ldots, Y_m) > Q\}$ est le même que celui dont la fonction de test est $\mathbf{1}\{|t| > Q'\}$ où Q' est tel que $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(|t| > Q') = \alpha$, énoncer le test du rapport de vraisemblance, i.e. trouver la loi de t sous H_0 et par le fait même la valeur de Q'.
Indice : si $A \sim \chi^2_a$ et $B \sim \chi^2_b$ sont indépendantes, alors $A + B \sim \chi^2_{a+b}$.

Exercice 75 (*exercice 57)**.** Soient $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. Supposons que l'on veut tester $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ en utilisant la fonction de test δ_α de la forme

$$\delta_\alpha(T(X_1, \ldots, X_n)) = \mathbf{1}\{T(X_1, \ldots, X_n) > q_{1-\alpha}\} \text{ ou } \delta_\alpha(T(X_1, \ldots, X_n)) = \mathbf{1}\{T(X_1, \ldots, X_n) < q_\alpha\},$$

où q_α est le α-quantile de G_0, la fonction de distribution de $T(X_1, \ldots, X_n)$ quand $\theta = \theta_0$. Supposons que G_0 est une fonction continue. Montrer que sous H_0, la valeur-p suit la distribution uniforme sur $[0, 1]$.
Indice : utiliser le lemme 4.30.

Exercice 76 (exercice 60, **intervalle bilatéral optimal**)**.** Afin de construire un intervalle de confiance bilatéral pour la moyenne d'une distribution normale (dont la variance est connue), nous avons choisi $z_{\alpha/2}$ et $z_{1-\alpha/2}$ comme bornes de l'intervalle (cf. exemple 5.3). L'on peut se demander pourquoi ne pas choisir par exemple $z_{\alpha/3}$ et $z_{1-2\alpha/3}$.
Il est vrai qu'on aime les intervalles plus « naturels » ou symétriques, mais la raison de ce choix est la suivante :

(i) Soient $Z \sim N(0,1)$ et $\alpha \in]0, 1[$. Montrer que l'intervalle $I = [L, U]$ ayant la plus petite longueur et tel que $\mathbb{P}(I \ni Z) \geq 1 - \alpha$ est donné par $L = z_{\alpha/2}$ et $U = z_{1-\alpha/2}$.

(ii) Soient $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ où la variance σ^2 est connue. Trouver l'intervalle $I_n = [A_n, B_n]$ ayant la plus petite longueur et tel que $\mathbb{P}(I_n \ni \mu) \geq 1 - \alpha$.

(iii) *Peut-on généraliser ce résultat ?

Exercice 77 (exercice 61, **différence de moyennes**)**.**

(i) Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu_X, \sigma^2)$ et $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu_Y, \sigma^2)$ deux échantillons indépendants, où μ_X, μ_Y et σ^2 sont inconnus. Trouver un intervalle de confiance bilatéral pour le paramètre $\theta = \mu_X - \mu_Y$ avec un seuil de confiance $1 - \alpha$.

(ii) On veut comparer la durée d'efficacité de deux nouveaux médicaments, M_1 et M_2, contre la lombalgie [2]. On a donc administré chaque médicament à un groupe de 15 patients, et ensuite mesuré (en heures) la période sans douleur après la prise du médicament. On obtient la moyenne du temps d'efficacité $\overline{X_1} = 7.5$ pour M_1 et $\overline{X_2} = 6.3$ pour M_2. On a aussi les écart-types estimés $S_1 = 1.1$ et $S_2 = 1.3$ pour M_1 et M_2 respectivement. En supposant que les observations des groupes 1 et 2 sont indépendantes et suivent des lois $N(\mu_1, \sigma^2)$ et $N(\mu_2, \sigma^2)$ respectivement, donner l'intervalle de confiance à 95% pour la différence $\mu_1 - \mu_2$. Que peut-on constater sur l'efficacité relative de M_1 et M_2 ?

Exercice 78 (*exercice 62)**.** Soient $T_k \sim \mathbf{t}_k$ et soit $Z \sim N(0,1)$. Montrer que $T_k \xrightarrow{d} Z$ lorsque $k \to \infty$.
Indice : s'inspirer des exemples 5.3 et 5.7.

Exercice 79 (exercice 63)**.** En utilisant la même notation que celle de la proposition 5.8 du cours, prouver que le tableau suivant contient les intervalles de confiance approximatifs avec seuil $(1-\alpha)$ pour θ :

Confiance approximative $1-\alpha$	$L(X_1, \dots, X_n)$	$U(X_1, \dots, X_n)$
Bilatéral	$\hat{\theta}_n - z_{1-\alpha/2} \hat{J}_n^{-1/2}$	$\hat{\theta}_n + z_{1-\alpha/2} \hat{J}_n^{-1/2}$
Unilatéral à gauche	$\hat{\theta}_n - z_{1-\alpha} \hat{J}_n^{-1/2}$	$+\infty$
Unilatéral à droite	$-\infty$	$\hat{\theta}_n + z_{1-\alpha} \hat{J}_n^{-1/2}$

Indice : si $Z_n \xrightarrow{d} Z$, où Z est une variable aléatoire continue, alors $\mathbb{P}[Z_n = a] \to 0$ pour chaque $a \in \mathbb{R}$.

Exercice 80 (*exercice 64, **pivot général**)**.**

(i) Soient $X_1, \dots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$ et $T_n(X_1, \dots, X_n)$ une statistique continue. Soit $Y_n = F_{T_n}(T_n;\theta)$, où $F_{T_n}(t;\theta) = \mathbb{P}_\theta[T_n \leq t]$ est la fonction de répartition de T_n. Supposons que $F_{T_n}(t;\theta)$ est pour chaque t une fonction continue de θ. Montrer que $Y_n \sim U(0,1)$ et donc que Y_n est un pivot. Comment peut on utiliser ce résultat pour trouver un intervalle de confiance pour θ ?

(ii) Soit $f(x;\theta) = e^{-(x-\theta)} \mathbf{1}_{[\theta,\infty)}(x)$. Utiliser la partie a) et la statistique $T_n = \min\{X_1, \dots, X_n\}$ pour trouver un intervalle de confiance pour θ avec un seuil $1-\alpha$.

Exercice 81 (exercice 65)**.** Soient $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, où σ^2 est connu. Trouver une expression pour l'intervalle de confiance unilatéral à gauche avec seuil $1-\alpha$ pour μ.

Exercice 82 (exercice 66)**.** Soient $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$. Avec l'aide d'une statistique exhaustive $\tau_n(X_1, \dots, X_n)$ pour p, trouver une expression pour l'intervalle de confiance unilatéral à gauche pour p avec seuil approximatif $1-\alpha$, en inversant le test

$$H_0 : p \leq p_0 \qquad \text{vs} \qquad H_1 : p > p_0.$$

Utiliser une fonction de test qui rejette H_0 lorsque τ_n prend une valeur (strictement) plus grande qu'une certaine valeur critique. Les bornes de cet intervalle ne seront malheureusement pas si explicitent qu'à l'exercice précédente.

2. C'est ce qu'a eu Pierre Brochant dans le film *le dîner des cons*. Il n'est pas le seul : on estime qu'entre 40 et 70% de la population en sera touché au cours de la vie.

Indice : suivre la proposition 5.14. Hélas, une des conditions de cette proposition n'est pas satisfaite (laquelle ?). Ainsi, pour la plupart des valeurs de p, la probabilité de couverture de l'intervalle sera seulement approximativement $1 - \alpha$.