MIDO - L3 Math. Appliquées 2022–2023

Statistical modelling

Examen partiel du 2 Novembre 2022

DURÉE 2H00 – DOCUMENTS ET CALCULATRICE NON AUTORISÉS

French – English Lexicon

- *i.i.d.* : *independent and identically distributed*
- échantillon : *sample*
- fonction de répartition : *cumulative distribution function*
- fonction de densité : *probability distribution function*
- fonction génératrice des moments : *moment-generating function*
- famille exponentielle : *exponential family*
- espace naturel des paramètres : *natural parameter space*

Exercise 1

...... /7

For the following statements, give the correct answer(s). There is no negative point but any incorrect/missing answer or incorrect/missing justification give no point.

1. Consider the function f defined, for $x \in \mathbb{R}$, by $f(x) = (x^2 - x - 2)\mathbb{1}_{\{x \in]-1,2[\}}$. For the Lebesgue measure

(a) $x \mapsto 9f(x)/2$ is a density,

(b) $x \mapsto 2f(x)/9$ is a density,

(c) $x \mapsto -2f(x)/9$ is a density,

(d) None of the other answers.

(c) The function f is a convex polynomial function such that $f(-1) = f(2) = 0$. Thus f is negative on $]-1,2[$, which excludes (a) and (b). Moreover

$$\int_{\mathbb{R}} f(x)\mathrm{d}x = \left[\frac{x^3}{3} - \frac{x^2}{2} - 2x\right]_{-1}^{2} = -\frac{9}{2}.$$

2. Given X a random variable with density on $\mathbb{R}$ (with respect to the Lebesgue measure) proportional to

$$g(x) = \exp(-b|x|), \quad b \in \mathbb{R}_+^*.$$

The characteristic function of X at point $t \in \mathbb{R}$ is given by

(a) $2b/(b^2 + t^2)$,

(b) $4/(b^2 + t^2)$,

(c) $b^2/(b^2 + t^2)$,

(d) None of the other answers.

Hint. for any $t \in \mathbb{R}$, $\exp(-bx \pm itx) \underset{x \to +\infty}{\longrightarrow} 0$.

(c) g is a positive function such that

$$\int_{\mathbb{R}} g(x)\mathrm{d}x = 2\int_0^{+\infty} \exp(-bx)\mathrm{d}x = \frac{2}{b}.$$

Then the density of X is $f(x) = bg(x)/2$, for all $x \in \mathbb{R}$. The characteristic function of X at point $t \in \mathbb{R}$ is

$$\phi_X(t) = \int_{-\infty}^{0} \frac{b}{2}\exp(itx + bx)\mathrm{d}x + \int_0^{+\infty} \frac{b}{2}\exp(itx - bx)\mathrm{d}x = \frac{b}{2}\left[\frac{\exp(itx + bx)}{it + b}\right]_{-\infty}^{0} + \frac{b}{2}\left[\frac{\exp(itx - bx)}{it - b}\right]_{0}^{+\infty}.$$

Using the **Hint**, we get

$$\phi_X(t) = \frac{b}{2}\frac{1}{it + b} - \frac{b}{2}\frac{1}{it - b} = \frac{b^2}{b^2 + t^2}.$$

3. A sequence $(Y_n)_{n\in\mathbb{N}^*}$ of random variables is said to c-converge to a random variable Y if, for any $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}\{|Y_n - Y| > \varepsilon\} < \infty.$$

(a) c-convergence implies convergence in distribution.

(b) almost sure convergence implies c-convergence.

(c) convergence in probability implies c-convergence

(d) c-convergence implies convergence in probability.

(e) c-convergence implies almost sure convergence.

(f) none of the other answers holds in full generality

(a, d, e) If X_n c-converges, the Borel Cantelli lemma shows that X_n converges almost surely. Then we use the result that almost sure convergence implies convergence in probability and in distribution.

4. Let $(X_n)_{n\in\mathbb{N}^*}$ be a sequence of *i.i.d.* discrete random variables such that

$$\mathbb{P}[X_n = 0] = \frac{n-1}{n} \quad \text{and} \quad \mathbb{P}[X_n = \sqrt{n}] = \frac{1}{n}.$$

Then, when n goes to $+\infty$,

(a) the sequence converges in L^1 (convergence in mean),

(b) the sequence converges in L^2 (convergence in quadratic mean),

(c) for any continuous function g, we have $\mathbb{E}[g(X_n)]$ converges to 0,

(d) converges in distribution,

(e) does not converge at all.

(a, d) For any continuous and bounded function g, we have

$$\mathbb{E}\left[g(X_n)\right] = \frac{n-1}{n} g(0) + \frac{1}{n} g(\sqrt{n}).$$

Since g is bounded, the second term in the above sum converges to 0 when n goes to $+\infty$ and thus

$$\mathbb{E}\left[g(X_n)\right] \xrightarrow[n\to+\infty]{\text{a.s.}} g(0).$$

Then, (X_n) converges in distribution to 0. Moreover, if it converges in L^p, $p \in \mathbb{N}^*$, it is necessarily to 0. We have $\mathbb{E}[X_n] = 1/\sqrt{n}$ and $\mathbb{E}\left[X_n^2\right] = 1$. Thus, X_n converges in L^1 to 0, but not in L^2.

5. Given a vector $\mathtt{x} = (x_1, \ldots, x_n)$, which of the following codes compute a vector $\mathtt{u} = (u_1, \ldots, u_n)$ such that for $i \in \{1, \ldots, n\}$,

$$u_i = \sum_{k=1}^{i} \frac{x_k}{i} \quad ?$$

(a) `mean(x)`

(b) `cumsum(x)/cumsum(rep(1, length(x)))`

(c) `cumsum(x)/seq_len(length(x))`

(d) `cumsum(x)/(1:length(x))`

(e) `sum(x)/length(x)`

(f) `sum(x)/sum(length(x))`

(b, c, d) We have

$$(u_1, \ldots, u_n) := \frac{(x_1, x_1 + x_2, \ldots, x_1 + \ldots + x_n)}{(1, 2, \ldots, n)}.$$

The function `cumsum(x)` returns $(x_1, x_1 + x_2, \ldots, x_1 + \ldots + x_n)$. Functions `1:length(x)`, `seq_len(length(x))` and `cumsum(rep(1, length(x)))` return $(1, 2, \ldots, n)$.

6. Consider an exponential family defined by $f(x \mid \theta) = c(\theta) h(x) \exp[\eta(\theta) T(x)]$, for $x \in \mathbb{R}$ and $\theta \in \mathbb{R}$. Assume that η is a bijective function from $\mathbb{R}$ to $\mathbb{R}$, and denote $\tau \mapsto \eta^{-1}(\tau)$ its inverse. If X has density $f(\cdot \mid \theta)$, the moment generating function of $T(X)$ is defined for $t \in \mathbb{R}$ by

(a) $c(\theta)/c(\theta + t)$

(b) $c\left(\eta^{-1}(\tau)\right)/c\left(\eta^{-1}(\tau + t)\right)$

(c) $c\left(\eta^{-1}(\tau)\right)/c\left(\eta^{-1}(\tau) + t\right)$

(d) $c(\theta)/c\left(\eta^{-1}(\eta(\theta) + t)\right)$

(b, d) The canonical form of the family corresponds to the parametrisation $\tau = \eta(\theta)$, that is

$$f(x \mid \tau) = c\left(\eta^{-1}(\tau)\right) h(x) \exp[\tau T(x)].$$

The moment generating function of $T(X)$ is then defined for $t \in \mathbb{R}$ by

$$\frac{c\left(\eta^{-1}(\tau)\right)}{c\left(\eta^{-1}(\tau + t)\right)} = \frac{c(\theta)}{c\left(\eta^{-1}(\eta(\theta) + t)\right)}.$$

7. Consider the density (with respect to the Lebesgue measure on $\mathbb{R}$) parametrised by an **unknown** $(k, \lambda) \in \mathbb{N}^* \times \mathbb{R}_+^*$ and defined by

$$f(x \mid k, \lambda) = \frac{\lambda^k x^{k-1} \exp(-\lambda x)}{(k-1)!} \mathbb{1}_{x \geq 0}.$$

(a) It constitutes a minimal and canonical exponential family.

(b) It constitutes a minimal exponential family but is not in a canonical form.

(c) It constitutes an exponential family that is weither minimal nor canonical.

(d) None of the other answers.

(a) The density writes as

$$f(x \mid k, \lambda) = \frac{\lambda^k}{(k-1)!} \frac{1}{x} \mathbb{1}_{x \geq 0} \exp\left[k \log(x) - \lambda x\right].$$

Then it constitutes a canonical exponential family with natural parameter (k, λ) and natural statistic $T(x) = (\log(x), -x)$. Moreover for $(\alpha_1, \alpha_2) \in \mathbb{R}^* \times \mathbb{R}^*$ and $c \in \mathbb{R}$, the set $\{x \in \mathbb{R}_+ ; \alpha_1 \log(x) - \alpha_2 x - c = 0\}$ contains at most 2 elements (maximal number of intersections between an affine function and $x \mapsto \log(x)$) and hence has measure zero (null set) for the Lebesgue measure. The family is minimal.

8. We run an experiment where we measure how much time n different customers spend on a specific page of a website. Our observations $x_1, \ldots, x_n$ are stored in a vector `x`. We assume that the underlying statistical model is a Gamma distribution with parameter (α, β). Which one among the following command lines does return the first quartile of the Gamma model with parameter $(1, 2)$?

(a) `rgamma(0.25, 1, 2)`

(b) `pgamma(0.25, 1, 2)`

(c) `dgamma(0.25, 1, 2)`

(d) `qgamma(0.25, 1, 2)`

(e) `quantile(x, 0.25, 1, 2)`

(f) `quantile(0.25, 1, 2)`

(d) In order to get the theoretical quantiles of a distribution we use the prefix `q` and the name of the distribution. The function `quantile` is only for computing empirical quantiles of a sample.

9. When given a sample `x` of size n from F and considering the median $\mathrm{med}(X)$ as the quantity of interest, a bootstrap approximation of a 95% interval of variability of the empirical median is given by

(a) `quantile(median(matrix(sample(x,n*m,rep=TRUE),m)),c(.025,.975))`

(b) `quantile(apply(matrix(sample(x,n*m,rep=TRUE),m),1,median),c(.025,.975))`

(c) `quantile(matrix(sample(median(x),n*m,rep=TRUE),m),c(.035,.985))`

(d) `median(apply(matrix(sample(x,n*m,rep=TRUE),m),1,sum),prob=.95)`

(e) `quantile(0.25, 1, 2)`

(b) In order to that interval we need to

1. generate m bootstrap samples of length n by sampling uniformly with replacement in `x` (done with function `sample` and stored in a matrix),
2. compute the median of each bootstrap sample using the function `median` (`apply` is used to apply the median to each row of the matrix),
3. taking the quantiles of order 2.5% and 97.5% of that sample of medians.

Exercise 2

...... /7.5

Suppose that Z_1 and Z_2 are *i.i.d.* random variables distributed according to the standard normal distribution $\mathcal{N}(0,1)$. We define the general Rayleigh distribution with **unknown** scale parameter $b \in \mathbb{R}_+^*$ as the density of the random variable $X = b\sqrt{Z_1^2 + Z_2^2}$.

Reminder. The probability density function of the normal $\mathcal{N}(0,1)$ distribution is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

1. Show that the cumulative distribution function of X is defined for all $t \in \mathbb{R}_+^*$ by

$$F(t) = 1 - \exp\left(-\frac{t^2}{2b^2}\right).$$

Since Z_1 and Z_2 are independent the joint distribution of (Z_1, Z_2) is the product of the marginal distributions of Z_1 and Z_2 and we have

$$F(t) = \mathbb{P}[X \le t] = \mathbb{P}\left[Z_1^2 + Z_2^2 \le \frac{t^2}{b^2}\right] = \int_{\mathbb{R}^2} \frac{1}{2\pi} \exp\left(-\frac{z_1^2 + z_2^2}{2}\right) \mathbb{1}_{\{z_1^2 + z_2^2 \le t^2/b^2\}} \mathrm{d}z_1 \mathrm{d}z_2.$$

Using the change of variables in polar coordinates, which is a $\mathscr{C}^1$-diffeormorphism between $\mathbb{R}_+ * \times]-\pi,\pi[$ and $\mathbb{R}^2 \backslash]-\infty, 0]$, we get

$$F(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \int_{\mathbb{R}_+*} \exp\left(-\frac{r^2}{2}\right) \mathbb{1}_{\{r^2 \le t^2/b^2\}} r \mathrm{d}r \mathrm{d}\theta = \int_0^{\frac{t}{b}} \exp\left(-\frac{r^2}{2}\right) \mathbb{1}_{\{r^2 \le t^2/b^2\}} r \mathrm{d}r = 1 - \exp\left(-\frac{t^2}{2b^2}\right).$$

2. Compute the density $f(\cdot \mid b)$ of X and show that it can define an exponential family with a natural statistic $T(\cdot)$ that is positive. Precise its canonical form and its natural parameter space. Is the family regular and minimal?

The density is given by

$$f(x \mid b) = F'(x) = \frac{x}{b^2} \exp\left(-\frac{x^2}{2b^2}\right) \mathbb{1}_{\{\mathbb{R}_+\}}(x) = c(b) h(x) \exp(\eta(b) T(x)), \quad \text{where} \quad \begin{cases} c(b) = \frac{1}{b^2}, & h(x) = x\mathbb{1}_{\{\mathbb{R}_+\}}(x), \\ \eta(b) = -\frac{1}{2b^2}, & T(x) = x^2. \end{cases}$$

The associated canonical form is obtained with parameter $\theta = \eta(b)$ and partition function $a(\theta) = -2\theta$, that is

$$f(x \mid \theta) = -2\theta x \mathbb{1}_{\{\mathbb{R}_+\}}(x) \exp\left(\theta x^2\right).$$

The natural parameter space is $\Theta = \{\theta \in \mathbb{R};\, a(\theta) > 0\} = \mathbb{R}_-^*$. Alternatively, we can say that

$$\Theta = \left\{\theta \in \mathbb{R};\, \int_0^{+\infty} x \exp(\theta x^2) \mathrm{d}x < \infty\right\} = \mathbb{R}_-^*,$$

since the integrand is integrable in $+\infty$ solely for $\theta < 0$. The family is regular since Θ is an open set, and minimal, since it is a one dimensional family.

3. Show that the moment generating function of X^2 is defined by

$$M_{X^2}(t) = \frac{1}{1 - 2b^2 t}, \quad \text{for} \quad t \in \left(-\infty, \frac{1}{2b^2}\right).$$

The moment generating function is defined for $t \in \mathbb{R}$ such that $\theta + t \in \Theta$, that is for $t \in]-\infty, -\theta[$ or equivalently for $t \in]-\infty, 1/(2b^2)[$. Using the canonical form, it is then given by

$$M_{X^2}(t) = \frac{a(\theta)}{a(\theta + t)} = \frac{\theta}{\theta + t} = \frac{1}{1 - 2b^2 t}, \quad \text{using that} \quad \theta = \frac{-1}{2b^2}.$$

4. Given a sequence $(X_n)_{n \in \mathbb{N}^*}$ of *i.i.d.* random variables distributed according to $f(\cdot \mid b)$, show that

$$\widehat{b}_n^2 = \frac{1}{2n} \sum_{k=1}^n X_k^2 \xrightarrow[n \to +\infty]{\mathbb{P}} b^2.$$

The first moment of the natural statistic $T(X) = X^2$ is given by

$$\mathbb{E}\left[X_1^2\right] = -\frac{\mathrm{d}}{\mathrm{d}\theta} \log a(\theta) = -\frac{1}{\theta} = 2b^2.$$

$(X_n^2/2)_{n \in \mathbb{N}^*}$ is a sequence of *i.i.d.* random variables (as a measurable transform of *i.i.d.* random variables) that are integrable. The Law of Large Numbers gives

$$\frac{1}{2n} \sum_{k=1}^n X_k^2 \xrightarrow[n \to +\infty]{\mathbb{P}} \mathbb{E}\left[\frac{1}{2} X_1^2\right] = b^2.$$

5. Show that

$$\frac{\sqrt{n}}{\widehat{b}_n^2}\left(\widehat{b}_n^2 - b^2\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,1).$$

The second moment of the natural statistic $T(X) = X^2$ is given by

$$\mathbb{V}\mathrm{ar}\left[X_1^2\right] = -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log a(\theta) = \frac{1}{\theta^2} = 4b^4.$$

$X_1^2/2$ has then a finite variance and the Central Limit theorem applies to $(X_n^2/2)_{n\in\mathbb{N}^*}$:

$$\sqrt{n}\left(\widehat{b}_n^2 - b^2\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0, \mathbb{V}\mathrm{ar}\left[0.5X_1^2\right]) \equiv \mathcal{N}(0, b^4). \tag{1}$$

Using the previous question and the continuity of $x \mapsto x^{-1}$ at $b^2 > 0$, we get

$$\frac{1}{\widehat{b}_n^2} \xrightarrow[n\to+\infty]{\mathbb{P}} \frac{1}{b^2}. \tag{2}$$

It follows from Slutsky's theorem that

$$\frac{\sqrt{n}}{\widehat{b}_n^2}\left(\widehat{b}_n^2 - b^2\right) \xrightarrow[n\to+\infty]{d} \frac{1}{b^2}\mathcal{N}(0, b^4) \equiv \mathcal{N}(0,1).$$

6. Show that there exists a sequence c_n, whose expression depends on known constants and $\widehat{b}_n^2$, such that

$$c_n\left(\sqrt{\widehat{b}_n^2} - b\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,1).$$

Starting from (1) and using the delta method with the function $g : x \mapsto \sqrt{x}$ that is differentiable in $b^2 > 0 : g'(b^2) = 1/(2b)$, we have

$$\sqrt{n}\left(\sqrt{\widehat{b}_n^2} - b\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}\left(0, \frac{b^2}{4}\right). \tag{3}$$

(2), (3) and Slutsky's theorem give

$$2\sqrt{\frac{n}{\widehat{b}_n^2}}\left(\sqrt{\widehat{b}_n^2} - b\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,1).$$

7. Find a function g such that

$$\sqrt{n}\left[g\left(\sqrt{\frac{2}{\pi}}\overline{X}_n\right) - g(b)\right] \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,1), \quad \text{where} \quad \overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Since X_1^2 is integrable, we have that X_1 is integrable and

$$\mathbb{E}[X_1] = \int_0^{+\infty} \frac{x^2}{b^2}\exp\left(-\frac{x^2}{2b^2}\right)\mathrm{d}x = \frac{\sqrt{2\pi}}{2b}\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}b}x^2\exp\left(-\frac{x^2}{2b^2}\right)\mathrm{d}x = b\sqrt{\frac{\pi}{2}}.$$

Applying the Central Limit theorem to $(X_n)_{n\in\mathbb{N}^*}$ yields that

$$\sqrt{n}\left(\sqrt{\frac{2}{\pi}}\overline{X}_n - b\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}\left(0, \frac{2}{\pi}\mathbb{V}\mathrm{ar}[X_1]\right), \quad \text{where} \quad \mathbb{V}\mathrm{ar}[X_1] = \mathbb{E}\left[X_1^2\right] - \mathbb{E}[X_1]^2 = b^2\left(2 - \frac{\pi}{2}\right). \tag{4}$$

Then using the delta method for a function g, differentiable at b, leads to

$$\sqrt{n}\left[g\left(\sqrt{\frac{2}{\pi}}\overline{X}_n\right)-g(b)\right] \xrightarrow[n\to+\infty]{d} \mathcal{N}\left(0, b^2\left(\frac{4}{\pi}-1\right)[g'(b)]^2\right).$$

The function g is then solution of

$$g'(x)=\frac{\sqrt{\pi}}{x\sqrt{4-\pi}}.$$

We can take for instance

$$g : x \mapsto \frac{\sqrt{\pi}}{\sqrt{4-\pi}}\log(x).$$

8. In order to approximate b, is it better to use an approximation based on $\overline{X}_n$ or an approximation based on $\widehat{b}_n^2$?

Hint. $4/\pi \approx 1.27$.

The asymptotic variance from (4) is larger than the asymptotic variance from (3). It is better to use $\widehat{b}_n^2$.

Exercise 3

...... /7.5

Let f be a probability density function on $\mathbb{R}$ such that

$$f(x)=c\tilde{f}(x), \quad \forall x\in\mathbb{R},$$

where the positive function $\tilde{f}$ is known and computable (for instance, by an R function `df(x)`), and the constant c is unknown. This setting is called a *missing normalising constant* problem.

1. Show that that the constant c is uniquely defined by $\tilde{f}$.

Since f is a density, its integral on $\mathbb{R}$ is equal to 1 and thus

$$c=\left(\int_{\mathbb{R}}\tilde{f}(x)\mathrm{d}x\right)^{-1}.$$

Until further notice, consider the special case of an interval $]a,b[$ such that

$$\forall x\notin]a,b[, \quad \tilde{f}(x)=0 \quad \text{and} \quad \sup_{x\in\mathbb{R}}\tilde{f}(x)=M<\infty.$$

2.a. Defining the rectangle $\mathscr{R}=]a,b[\times]0,M[$ and the subset of $\mathscr{R}$

$$\mathscr{S}=\{x=(x_1,x_2)\in\mathscr{R}\,;\, x_2\le\tilde{f}(x_1)\},$$

give the ratio of the surfaces of $\mathscr{S}$ and of $\mathscr{R}$.

The surface of $\mathscr{S}$ is

$$\int_{\mathbb{R}^2}\mathbb{1}_{x_2\le\tilde{f}(x_1)}\mathrm{d}x_2\mathrm{d}x_1=\int_a^b\int_0^{\tilde{f}(x_1)}\mathrm{d}x_2\mathrm{d}x_1=\int_a^b\tilde{f}(x_1)\mathrm{d}x_1=\frac{1}{c}.$$

The ratio of the surfaces of $\mathscr{S}$ and of $\mathscr{R}$ is then $[cM(b-a)]^{-1}$.

2.b. Considering $U = (U_1, U_2)$ a Uniform random point on $\mathscr{R}$, compute the probability $\mathbb{P}[U_2 \leq \tilde{f}(U_1)]$.

We have

$$\mathbb{P}[U_2 \leq \tilde{f}(U_1)] = \frac{1}{M(b-a)} \int_a^b \int_0^{\tilde{f}(u_1)} \mathrm{d}u_2 \mathrm{d}u_1 = \frac{1}{M(b-a)} \int_a^b \tilde{f}(u_1)\mathrm{d}u_1 = \frac{1}{cM(b-a)}.$$

2.c. Given an *i.i.d.* sequence $U^1, \ldots, U^n$ of Uniform random points on $\mathscr{R}$, and exploiting the Law of Large Numbers, construct a converging (with n) approximation of c^{-1}.

For $k \in \mathbb{N}^*$, let set

$$X_k = M(b-a)\mathbb{1}_{U_2^k \leq \tilde{f}(U_1^k)}.$$

$X_1, \ldots, X_n$ is a sequence of *i.i.d.* random variables (as a measurable transform of *i.i.d.* random variables) and integrable (they are bounded). The Law of Large Numbers then gives that

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^n X_k = \frac{M(b-a)}{n}\sum_{k=1}^n \mathbb{1}_{U_2^k \leq \tilde{f}(U_1^k)} \xrightarrow[n\to+\infty]{\mathbb{P}} \mathbb{E}[X_1] = M(b-a)P[U_2^1 \leq \tilde{f}(U_1^1)] = \frac{1}{c}.$$

2.d. Derive a converging estimator of c.

The function $x \mapsto x^{-1}$ is continuous on $\mathbb{R}_+^*$. Then

$$\frac{1}{\overline{X}_n} = \frac{n}{M(b-a)\sum_{k=1}^n \mathbb{1}_{U_2^k \leq \tilde{f}(U_1^k)}} \xrightarrow[n\to+\infty]{\mathbb{P}} c.$$

2.e. Write an R code calling df() that produces this estimator of c.

```r
1/(M * (b - a) * mean(runif(n, 0, M) <= df(runif(n, a, b))))
```

From now and till the end of the Exercise, consider two simultaneous missing normalising constant densities

$$f_1(x) = c_1\tilde{f}_1(x) \quad \text{and} \quad f_2(x) = c_2\tilde{f}_2(x),$$

where c_1, c_2 are unknown and $\tilde{f}_1$, $\tilde{f}_2$ are known (with associated R functions `df1` and `df2`).

3.a. Let X_1 be a random variable with density $f_1(x)$. Show that

$$\mathbb{E}\left[\frac{\tilde{f}_2(X_1)}{\tilde{f}_1(X_1)}\right] = \frac{c_1}{c_2}.$$

We have

$$\mathbb{E}\left[\frac{\tilde{f}_2(X_1)}{\tilde{f}_1(X_1)}\right] = \int \frac{\tilde{f}_2(x)}{\tilde{f}_1(x)} f_1(x)\mathrm{d}x = \int \frac{\tilde{f}_2(x)}{\cancel{\tilde{f}_1(x)}} c_1 \cancel{\tilde{f}_1(x)}\mathrm{d}x = c_1 \int \tilde{f}_2(x)\mathrm{d}x = \frac{c_1}{c_2} \quad \text{(using question 2.a.).} \tag{5}$$

3.b. Given an *i.i.d.* sequence $X_{11},\dots,X_{1n}$ of random variable with density f_1, deduce from question 3.a. an approximation of c_1/c_2 that is converging with n.

The Law of Large Numbers applied to the sequence of *i.i.d.* random variables $(\tilde{f}_2(X_{1k})/\tilde{f}_1(X_{1k}))$ shows that

$$\frac{1}{n}\sum_{k=1}^{n}\frac{\tilde{f}_2(X_{1k})}{\tilde{f}_1(X_{1k})}\xrightarrow[n\to+\infty]{\mathbb{P}}\mathbb{E}\left[\frac{\tilde{f}_2(X_1)}{\tilde{f}_1(X_1)}\right]=\frac{c_1}{c_2}.$$

3.c. Let $\alpha(\cdot)$ be a positive function such that

$$\int \alpha(x)\tilde{f}_1(x)\tilde{f}_2(x)\mathrm{d}x<+\infty.$$

Show that, if X_1 is a random variable with density f_1 and X_2 is a random variable with density f_2,

$$\frac{\mathbb{E}\left[\alpha(X_1)\tilde{f}_2(X_1)\right]}{\mathbb{E}\left[\alpha(X_2)\tilde{f}_1(X_2)\right]}=\frac{c_1}{c_2}.$$

We have

$$\mathbb{E}\left[\alpha(X_1)\tilde{f}_2(X_1)\right]=\int \alpha(x)\tilde{f}_2(x)f_1(x)\mathrm{d}x=\frac{1}{c_2}\int \alpha(x)f_2(x)f_1(x)\mathrm{d}x.$$

In the same way,

$$\mathbb{E}\left[\alpha(X_2)\tilde{f}_1(X_2)\right]=\frac{1}{c_1}\int \alpha(x)f_1(x)f_2(x)\mathrm{d}x.$$

The result directly follows.

3.d. Deduce from question 3.c. a converging approximation of c_1/c_2 based on two sequences $X_{11},\dots,X_{1n}$ and $X_{21},\dots,X_{2n}$ of *i.i.d.* random variables with density f_1 and f_2 respectively.

The Law of Large Numbers applied to $(\alpha(X_{1k})\tilde{f}_2(X_{1k}))$ and $(\alpha(X_{2k})\tilde{f}_1(X_{2k}))$ gives that

$$\frac{1}{n}\sum_{k=1}^{n}\alpha(X_{1k})\tilde{f}_2(X_{1k})\xrightarrow[n\to+\infty]{\mathbb{P}}\mathbb{E}\left[\alpha(X_1)\tilde{f}_2(X_1)\right]\quad\text{and}\quad\frac{1}{n}\sum_{k=1}^{n}\alpha(X_{2k})\tilde{f}_1(X_{2k})\xrightarrow[n\to+\infty]{\mathbb{P}}\mathbb{E}\left[\alpha(X_2)\tilde{f}_1(X_2)\right].$$

Thus,

$$\frac{\sum_{k=1}^{n}\alpha(X_{1k})\tilde{f}_2(X_{1k})}{\sum_{k=1}^{n}\alpha(X_{2k})\tilde{f}_1(X_{2k})}\xrightarrow[n\to+\infty]{\mathbb{P}}\frac{c_1}{c_2}.$$

3.e. Assuming there exist simulation functions `rf1(n)` and `rf2(n)` to simulate from f_1 and f_2, write an R code that produces this estimator of c_1/c_2.

```
x_1 <- rf1(n)
x_2 <- rf2(n)
sum(alpha(x_1) * df2(x_1)) / sum(alpha(x_2) * df1(x_2))
```