Monte Carlo Project 2

Samuel Elbaz, Alexandre Zenou,, Sacha Assouly

November 2024

Random Variable simulation with stratification

0.1 Question 9:

In order to apply the accept-reject method algorithm, we first have to find the constant $M > 1$ such that

$$\frac{f(x)}{g(x)} \leq M, \quad \text{(for almost all } x \in \mathbb{R}\text{)}.$$

But here, the integral of $g(x)$ over $\mathbb{R}$ is computed as:

$$\int_{\mathbb{R}} g(x)dx = \int_{D_0} g(x)dx + \sum_{i=1}^{k} \int_{D_i} g(x)dx.$$

On D_0, $g(x) = \frac{1}{1-a} f_1(x)$, so:

$$\int_{D_0} g(x)dx = \frac{1}{1-a} \int_{D_0} f_1(x)dx.$$

On D_i, $g(x)$ is constant, so:

$$\int_{D_i} g(x)dx = \frac{1}{1-a} \left(\sup_{D_i} f_1(x) - a \inf_{D_i} f_2(x) \right) \mathfrak{meas}(D_i),$$

where $\mathfrak{meas}$ represents the Lebesgue measure on $(\mathbb{R}; \mathcal{B}(\mathbb{R}))$. The total integral is:

$$\int_{\mathbb{R}} g(x)dx = \frac{1}{1-a} \int_{D_0} f_1(x)dx + \sum_{i=1}^{k} \frac{1}{1-a} \left(\sup_{D_i} f_1(x) - a \inf_{D_i} f_2(x) \right) \mathfrak{meas}(D_i).$$

Since $\int_{\mathbb{R}} g(x)dx \neq 1$, g is **not** a density.
To make it a density, we normalize it, defining for almost all $x \in \mathbb{R}$:

$$\tilde{g}(x) = \frac{g(x)}{\int_{\mathbb{R}} g(x)dx}.$$

Now we search for $M > 1$, defined previously. For almost all $x \in D_0$:

$$\frac{f(x)}{\tilde{g}(x)} = \int_{\mathbb{R}} g(x)dx \cdot \left(1 - a \frac{f_2(x)}{f_1(x)} \right).$$

For almost all $x \in D_i$:

$$\frac{f(x)}{\tilde{g}(x)} = \int_{\mathbb{R}} g(x)dx \cdot \frac{f_1(x) - a f_2(x)}{\sup_{D_i} f_1(x) - a \inf_{D_i} f_2(x)}.$$

In D_0 :

$$\frac{f_2(x)}{f_1(x)} = \frac{\sqrt{\sigma_1^2}}{\sqrt{\sigma_2^2}} \exp\left(-\frac{x^2}{2\sigma_2^2} + \frac{x^2}{2\sigma_1^2} \right),$$

The dominant term in the exponent is proportional to :

$$\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2},$$

which is negative since $\sigma_1^2 > \sigma_2^2$. Therefore, $\frac{f_2(x)}{f_1(x)} \to 0$ as $x \to \infty$. Then :

$$\frac{f_2(x)}{f_1(x)} \to 0 \quad \text{as } x \to \infty.$$

So :

$$1 - a\frac{f_2(x)}{f_1(x)} \to 1 \quad \text{as } x \to \infty,$$

Finally

$$\sup_{x \in D_0} \frac{f(x)}{\tilde{g}(x)} \leq \int_{\mathbb{R}} g(x)dx.$$

[1] And for almost all $x \in D_i$ with $i \neq 0$:

$$g(x) = \frac{1}{1-a}\left(\sup_{D_i} f_1(x) - a \inf_{D_i} f_2(x)\right).$$

The ratio becomes:

$$\frac{f(x)}{\tilde{g}(x)} = \int_{\mathbb{R}} g(x)dx \cdot \frac{f_1(x) - af_2(x)}{\sup_{D_i} f_1(x) - a \inf_{D_i} f_2(x)}.$$

Clearly :

$$\sup_{D_i} f_1(x) - a \inf_{D_i} f_2(x) \geq f_1(x) - af_2(x),$$

which implies:

$$\frac{f_1(x) - af_2(x)}{\sup_{D_i} f_1(x) - a \inf_{D_i} f_2(x)} \leq 1.$$

Therefore, the ratio:

$$\frac{f(x)}{\tilde{g}(x)} = \int_{\mathbb{R}} g(x)dx \cdot \frac{f_1(x) - af_2(x)}{\sup_{D_i} f_1(x) - a \inf_{D_i} f_2(x)},$$

is again such that :

$$\frac{f(x)}{\tilde{g}(x)} \leq \int_{\mathbb{R}} g(x)dx.$$

Finally, we can claim that

$$\sup_{x \in \mathbb{R}} \frac{f(x)}{\tilde{g}(x)} \leq \int_{\mathbb{R}} g(x)dx.$$

[1] $\mathfrak{meas}(D_i)$ denotes the Lebesgue measure of the set D_i.

From the paper, it is stated that g dominates f almost surely which is a density, then :

$$\int_{\mathbb{R}} f(x)dx = 1 \leq \int_{\mathbb{R}} g(x)dx.$$

So we can define $M := \int_{\mathbb{R}} g(x)dx$.

The following is a description of the "accept-reject" method that one has to use in order to simulate X. :

1. Draw a candidate $Y \sim g$.
2. Draw a uniform random variable $U \sim \mathcal{U}([0,1])$.
3. Accept Y if $U \leq \frac{f(Y)}{M \cdot g(Y)}$, otherwise reject and repeat the process until the condition holds.

The course tells us that the acceptance rate is nothing else that $1/M = \frac{1}{\int_{\mathbb{R}} g(x)dx}$.

0.2 Question 10:

Acceptance Rate

The acceptance rate on a Borel set $A \subseteq \mathbb{R}$ is defined as:

$$\text{Acceptance rate on } A = \frac{\int_A f(x)\,dx}{M \int_A \tilde{g}(x)\,dx},$$

where: $f(x)$ is the target density, $\tilde{g}(x) = \frac{g(x)}{\int_{\mathbb{R}} g(x)\,dx}$ is the normalized version of $g(x)$, and M is the upper bound of $\frac{f}{g}$ as defined previously i.e the integral of g over $\mathbb{R}$.

For the set $\{|x| > N\}$, the acceptance rate becomes:

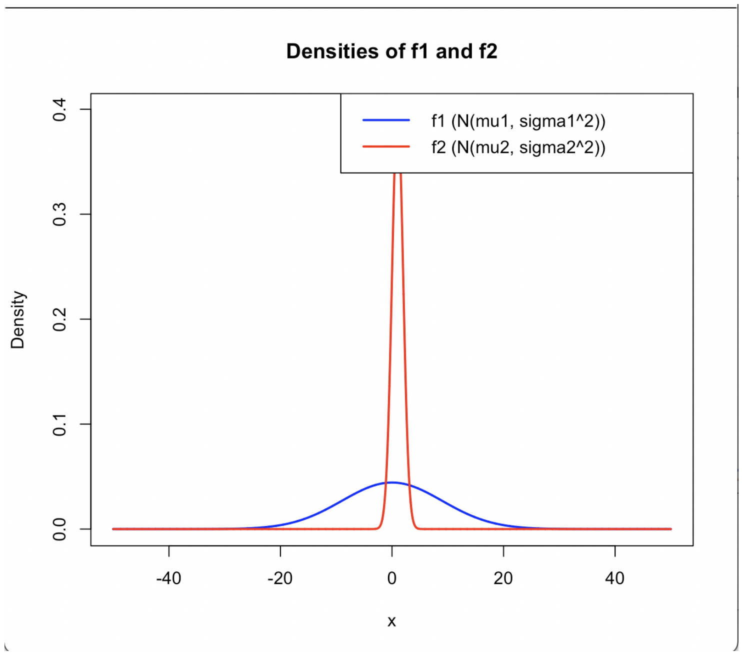
$$\text{Acceptance rate on } \{|x| > N\} = \frac{\int_{|x|>N} f(x)\,dx}{M \int_{|x|>N} \tilde{g}(x)\,dx}.$$

Substituting $\tilde{g}(x) = \frac{g(x)}{\int_{\mathbb{R}} g(x)\,dx}$, this simplifies to:

$$\text{Acceptance rate on } \{|x| > N\} = \frac{\int_{|x|>N} f(x)\,dx}{M \cdot \frac{\int_{|x|>N} g(x)\,dx}{\int_{\mathbb{R}} g(x)\,dx}}.$$

$$\text{Acceptance rate on } \{|x| > N\} = \frac{\int_{|x|>N} (f_1(x) - af_2(x))\,dx}{\int_{|x|>N} g(x)\,dx} = 1 - a\frac{\int_{|x|>N} f_2(x)\,dx}{\int_{|x|>N} f_1(x)\,dx}.$$

The integral of f_2 decreases faster than the integral of f_1 as $|x| \to \infty$, due to the smaller variance σ_2^2 of f, as shows the following picture :



Then, we know for sure that when $N > 0$ is large enough, we have that :

$$\frac{\int_{|x|>N} f_2(x)\,dx}{\int_{|x|>N} f_1(x)\,dx} \sim \frac{1-\delta}{a}, \text{(where we set } g = f_1).$$

Indeed, $\frac{1-\delta}{a}$ goes to 0 as δ tends to 1, the ratio $1/a$ is just $1/0.2 = 5$, then it is negligeable when $\delta \to 1$

Substituting this equivalence into the acceptance rate :

$$\text{Acceptance rate on } \{|x| > N\} \sim 1 - a \cdot \frac{1-\delta}{a} = \delta.$$

Thus, for sufficiently large N, the acceptance rate converges to δ. Then, we define $D_0 =]-\infty; -N[\cup]N; +\infty[$ with $N > 0$ being the smallest real such that the previous approximation holds.
$\mathcal{P} = \{D_0, D_1, \ldots, D_k\}$ splits $\mathbb{R}$ such that:

$$\mathbb{R} = D_0 \cup \bigcup_{i=1}^{k} D_i,$$

with D_0 covering the tails of f_1 and $\bigcup_{i=1}^{k} D_i$ being the complement of D_0.

Thus, $\bigcup_{i=1}^{k} D_i$ forms the compact interval:

$$\bigcup_{i=1}^{k} D_i = [-N, N].$$

Note that here, $k \in \mathbb{N}$ is taken arbitrarly, it will be chosen later. Now for each $i > 0$:

$$\text{Acceptance rate on } D_i = \frac{\int_{D_i} f_1(x)\,dx - a\int_{D_i} f_2(x)\,dx}{\mathfrak{meas}(D_i)\cdot\left(\sup_{D_i} f_1(x) - a\inf_{D_i} f_2(x)\right)}.$$

Now,

Let $f \in \mathcal{C}^1(\mathbb{R})$. By the Mean Value Theorem for integrals, for any $y \in \mathbb{R}$ and $h > 0$:

$$\int_y^{y+h} f(x)\,dx = f(\xi)\cdot h, \quad \text{for some } \xi \in]y, y+h[.$$

If h is sufficiently small, we have $\xi \approx y$, so:

$$\int_y^{y+h} f(x)\,dx \approx f(y)\cdot h.$$

Let D_i be a subset of $\mathbb{R}$ containing x. To analyze the acceptance rate, we consider any $x \in D_i$ and argue by synthesis to establish sufficient conditions on D_i such that the acceptance rate exceeds a chosen $\delta > 0$. Shrinking D_i around x, we approximate the integrals as:

$$\int_{D_i} f_1(x)\,dx \approx f_1(x)\cdot\mathfrak{meas}(D_i), \quad \int_{D_i} f_2(x)\,dx \approx f_2(x)\cdot\mathfrak{meas}(D_i).$$

Substitute these approximations into the expression for the acceptance rate:

$$\text{Acceptance rate on } D_i = \frac{\mathfrak{meas}(D_i)\cdot(f_1(x) - af_2(x))}{\mathfrak{meas}(D_i)\cdot\left(\sup_{D_i} f_1(x) - a\inf_{D_i} f_2(x)\right)} = \frac{f_1(x) - af_2(x)}{\sup_{D_i} f_1(x) - a\inf_{D_i} f_2(x)}.$$

For any sufficiently small D_i, there exists $\varepsilon > 0$ such that:

$$|f_1(x) - \sup_{D_i} f_1(x)| < \varepsilon \quad \text{and} \quad |f_2(x) - \inf_{D_i} f_2(x)| < \varepsilon.$$

The denominator of the acceptance rate $\sup_{D_i} f_1(x) - a\inf_{D_i} f_2(x)$ satisfies:

$$\sup_{D_i} f_1(x) - a\inf_{D_i} f_2(x) \leq \left|\sup_{D_i} f_1(x) - f_1(x)\right| + (f_1(x) - af_2(x)) + a\left|f_2(x) - \inf_{D_i} f_2(x)\right|.$$

Substituting the bounds for the differences:

$$\sup_{D_i} f_1(x) - a\inf_{D_i} f_2(x) \leq \varepsilon + (f_1(x) - af_2(x)) + a\varepsilon.$$

Then, the acceptance rate on D_i is given by:

$$\text{Acceptance rate on } D_i = \frac{f_1(x) - af_2(x)}{\sup_{D_i} f_1(x) - a\inf_{D_i} f_2(x)}.$$

Using the upper bound found before, we get:

$$\text{Acceptance rate on } D_i \geq \frac{f_1(x) - af_2(x)}{\varepsilon + (f_1(x) - af_2(x)) + a\varepsilon}.$$

We must have :

$$\frac{f_1(x) - af_2(x)}{\varepsilon + (f_1(x) - af_2(x)) + a\varepsilon} \geq \delta.$$

Finally, solve for ε:

$$\varepsilon \leq \frac{(1-\delta)\cdot(f_1(x) - af_2(x))}{\delta\cdot(1+a)}.$$

To ensure this inequality holds for all D_i, we let:

$$\varepsilon \leq \frac{(1-\delta)\cdot\left(\sup_{D_i} f_1 - a\inf_{D_i} f_2\right)}{\delta\cdot(1+a)}.$$

Then, one has to take x in $[-N; N]$ randomly, as close as possible from 0, as there is the largest values of our functions f_1 and f_2, and such that the D_i must be as smallest as possible. For example, take $x = 0$ and let $D_1 = [-N; N]$. Taking a random ε verifying the inequality before, if this holds :

$$|f_1(x) - \sup_{D_i} f_1(x)| < \varepsilon \quad \text{and} \quad |f_2(x) - \inf_{D_i} f_2(x)| < \varepsilon,$$

then it implies that we do not need anymore to reduce D_1 nor find another set from the partition as both of them will give an acceptance rate larger than the chosen δ. Otherwise, one has to find the very first real d_1 such that $D_1 = [-d_1; d_1]$ and such that all of the previous inequalities holds. Then, we repeat again, treating each case by setting $D_2 = [-N; -d_1[$ and $D_3 =]-d_1; N]$ and applying the same reasoning for the new generated D_i's.

Question 11:

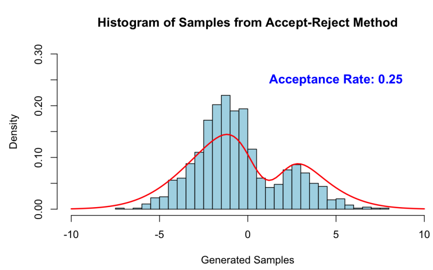


Figure 1: Histogram of the generated samples from the accept-reject method

Cumulative Density Function

Question 13

Let's recall that f is of the form:

$$f(x) = \frac{1}{1-a}\left(f_1(x) - a \cdot f_2(x)\right), \quad \forall x \in \mathbb{R}$$

Thus, its cumulative distribution function F is:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \frac{1}{1-a}\left(\int_{-\infty}^{x} f_1(t)\,dt - a\int_{-\infty}^{x} f_2(t)\,dt\right) \quad \forall x \in \mathbb{R}$$

Since $f_1 \sim \mathcal{N}(0,9)$ and $f_2 \sim \mathcal{N}(1,1)$, we have that:

$$F(x) = \frac{1}{1-a}\left(\psi\left(\frac{x}{3}\right) - a \cdot \psi(x-1)\right), \quad \forall x \in \mathbb{R}$$

where ψ is the CDF of a $\mathcal{N}(0,1)$.

Finally, substituting the values $a = 0.2$, we get the following closed-form expression for the F:

$$F(x) = 1.25 \cdot \psi\left(\frac{x}{3}\right) - 0.25 \cdot \psi(x-1), \quad \forall x \in \mathbb{R}$$

Moreover, for all $x \in \mathbb{R}$:

$$F(x) = P(X \leq x) = \mathbb{E}[\mathbb{1}_{X \leq x}].$$

Thus, a Monte Carlo estimator of $F(x)$, for $X_1, \dots, X_n$ following the distribution of X, is given by:

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}.$$

Question 14

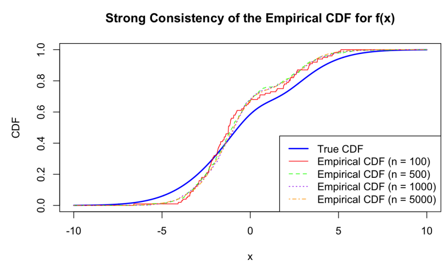
The X_i are i.i.d., and the random variables $\mathbb{1}_{X_i \leq x}$ belong to L^2 since $\mathbb{E}[(\mathbb{1}_{X_i \leq x})^2] = \mathbb{E}[\mathbb{1}_{X_i \leq x}] = F(x) < \infty$. And one has $\mathbb{E}[\mathbb{1}_{X_i \leq x}] = F(x)$, so by the **strong** law of large numbers:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq x} \xrightarrow{\text{a.s.}} \mathbb{E}[\mathbb{1}_{X \leq x}] = F(x) \quad \text{as } n \to \infty.$$

Meaning that the convergence is almost sure, due to the fact that the X_i are iid and squared-integrable.

Question 15

We generate n i.i.d random variables using the accept-reject algorithm and we get strong consistency of the empirical cdf :



Question 16

Let $X_1, \ldots, X_n$ be i.i.d. random variables such that $\mathbb{E}[X_1^2] < \infty$. Then:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_1]\right) \xrightarrow{\mathcal{L}} N(0, \sigma^2),$$

where $\sigma^2 = \mathrm{Var}(X_1)$.

Now, we apply this result to the sequence $\mathbb{1}_{X_i \leq x}$, $i = 1, \ldots, n$. We then have:

$$\sqrt{n}\,(F_n(x) - F(x)) \xrightarrow{\mathcal{L}} N(0, F(x)(1 - F(x))),$$

where $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}$, and $F(x)$ is the cumulative distribution function.

Since $F_n(x) \xrightarrow{\mathcal{L}} F(x)$ as $n \to \infty$ $x \mapsto \frac{1}{\sqrt{x(1-x)}}$, we have:

$$\sqrt{n}\frac{F_n(x) - F(x)}{\sqrt{F_n(x)(1 - F_n(x))}} \xrightarrow{\mathcal{L}} N(0, 1),$$

by Slutsky's theorem.

Thus, we can deduce a confidence interval:

$$I_{1-\alpha}(x) = \left[F_n(x) - q_{\frac{\alpha}{2}}\sqrt{\frac{F_n(x)(1 - F_n(x))}{n}},\ F_n(x) + q_{\frac{\alpha}{2}}\sqrt{\frac{F_n(x)(1 - F_n(x))}{n}}\right],$$

where $q_{\frac{\alpha}{2}}$ is the quantile of order $\frac{\alpha}{2}$ of $N(0,1)$. Here, we take $\alpha = 0.05$ to obtain a 95% confidence interval.

Question 17

We are looking for the smallest n such that:

$$2q_{\frac{\alpha}{2}}\sqrt{\frac{F_n(x)(1-F_n(x))}{n}} \leq \varepsilon,$$

where ε is a small precision parameter.

Rearranging the inequality, we obtain:

$$n > \frac{4q_{\frac{\alpha}{2}}^2}{\varepsilon^2} \cdot F_n(x)(1-F_n(x)).$$

To ensure n is an integer, we take the ceiling of the result:

$$n_2 = \left\lceil \frac{4q_{\frac{\alpha}{2}}^2}{\varepsilon^2} \cdot F_n(x)(1-F_n(x)) \right\rceil.$$

Another way to do it is by using the well-known "*Bienaymé-Tshebyshev*" inequality.

For any $\varepsilon > 0$, we have

$$\mathbb{P}\left(|F_n(x) - F(x)| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(F_n(x))}{\varepsilon^2},$$

where $\mathrm{Var}(F_n(x)) = \frac{F(x)(1-F(x))}{n}$.

Thus:

$$\mathbb{P}\left(|F_n(x) - F(x)| \geq \varepsilon\right) \leq \frac{F(x)(1-F(x))}{n\varepsilon^2}.$$

As we are supposed to estimate F, we cannot use it but we have that for all $p \in [0;1]$:

$$p(1-p) \leq \frac{1}{4}$$

Then :

$$\mathbb{P}\left(|F_n(x) - F(x)| \geq \varepsilon\right) \leq \frac{\frac{1}{4}}{n\varepsilon^2}.$$

And we require:

$$\frac{1}{4n\varepsilon^2} \leq \delta,$$

where δ is the desired probability bound.

To ensure the probability is less than 0.05, we find:

$$n = \left\lceil \frac{1}{0.2\varepsilon^2} \right\rceil = \left\lceil \frac{5}{\varepsilon^2} \right\rceil.$$

While this formula gives the minimum n, it may lack precision due to the loose bound $F(x)(1-F(x)) \leq \frac{1}{4}$ which hides the values of F that vary depending what is x.

The following algorithm is used to determine the value of n:
1. Set a maximum number of iterations for n. 2. Compute n_2 using the formula above. 3. If $n = n_2$, stop the iteration. 4. Otherwise, continue updating n until convergence.

Here is an example of values for x and n (for $\varepsilon = 0.02$):

x	n
1	7309
-15	100

We notice that n strongly decreases as x is far from 0.

Empirical Quantile Function Q_n

Question 18

The order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(N)}$ are the sorted values of the sample $X_1, \ldots, X_N$ in increasing order. To ensure their existence, we prove that all X_i are distinct almost surely (P-a.s.).

The X_i are i.i.d., so for each $i \neq j$ in $1, ..., n$, X_i and $-X_j$ are independent. The difference $X_i - X_j$ has a continuous density because it is the convolution of the densities of X_i and $-X_j$, which are both f. A continuous density for Lebesgue measure does not charge singletons, so $\mathbb{P}(X_i - X_j = 0) = 0$, i.e $\mathbb{P}(X_i = X_j) = 0$.

To verify that any X_i cannot be equal, consider:

$$\mathbb{P}\left(\bigcup_{i \neq j} \{X_i = X_j\}\right) \leq \sum_{i \neq j} \mathbb{P}(X_i = X_j).$$

Since $\mathbb{P}(X_i = X_j) = 0$ for all $i \neq j$, this sum equals 0:

$$\mathbb{P}\left(\bigcup_{i \neq j} \{X_i = X_j\}\right) = 0.$$

Thus, almost surely, all X_i are distinct, ensuring the order statistics $X_{(1)}, \ldots, X_{(N)}$ exist.

The empirical quantile function $Q_n(u)$ is defined as:

$$Q_n(u) = \inf\{x \in \mathbb{R} : u \leq F_n(x)\},$$

where $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i \le x}$ is the empirical cumulative distribution function. To determine when $F_n(x)$ exceeds u, we solve:

$$F_n(x) = \frac{k}{n} \ge u,$$

where $k = |\{(X_i \le x), i \in \{1, ..., n\}\}|$ is the number of observations less than or equal to x. The smallest k satisfying $\frac{k}{n} \ge u$ is $k = \lceil nu \rceil$.

Thus, $Q_n(u)$ corresponds to the $\lceil nu \rceil$-th smallest value in the sample, i.e.,

$$Q_n(u) = X_{\lceil nu \rceil}.$$

This concludes the derivation of $Q_n(u)$ using order statistics and the empirical CDF.

Question 19

One has to derive a CLT for Q_n.

Let $X_1, \ldots, X_n$ be an independent and identically distributed (i.i.d.) sample drawn from X. We recall that F is continuous and strictly increasing, such that it's derivative : f, never vanishes and remains strictly positive. By the global inversion theorem, F is invertible on $\mathbb{R}$ and we identify it's generalized inverse by Q, where for all $u \in (0, 1)$:

$$Q(u) = \inf\{x \in \mathbb{R} : u \le F(x)\}.$$

Define, for any $t \in \mathbb{R}$, for any $j \in \{1, ..., n\}$:

$$Y_{j,n} := \mathbb{1}_{X_j < Q(u) + \frac{t}{\sqrt{n}} \frac{\sqrt{u(1-u)}}{f(Q(u))}}$$

The statistic S_n is defined as the sum: (it follows a $\mathrm{Bin}(n; \mathbb{P}\left(X_j < Q(u) + \frac{t}{\sqrt{n}} \frac{\sqrt{u(1-u)}}{f(Q(u))}\right))$, as it is the sum of iid Bernoulli's..)

$$S_n = \sum_{j=1}^n Y_{j,n}.$$

The expectation of $Y_{j,n}$ is:

$$\mathbb{E}[Y_{j,n}] = \mathbb{P}\left(X_j < Q(u) + \frac{t}{\sqrt{n}} \frac{\sqrt{u(1-u)}}{f(Q(u))}\right).$$

Using the first-order Taylor expansion of $F(x)$ around $Q(u)$, we have:

$$F\left(Q(u) + \frac{t}{\sqrt{n}} \frac{\sqrt{u(1-u)}}{f(Q(u))}\right) = F(Q(u)) + F'(Q(u)) \cdot \frac{t}{\sqrt{n}} \frac{\sqrt{u(1-u)}}{f(Q(u))} + o\left(\frac{1}{\sqrt{n}}\right).$$

Since $F(Q(u)) = u$ and $f(Q(u)) = F'(Q(u))$ by the Fundamental analysis theorem. Thus :

$$\mathbb{E}[Y_{j,n}] = u + \frac{t}{\sqrt{n}}\sqrt{u(1-u)} + o\left(\frac{1}{\sqrt{n}}\right).$$

Since $Y_{j,n}$ is a Bernoulli random variable:

$$\mathrm{Var}(Y_{j,n}) = \mathbb{E}[Y_{j,n}] \cdot (1 - \mathbb{E}[Y_{j,n}]).$$

Using the approximation for $\mathbb{E}[Y_{j,n}]$, we get:

$$\mathrm{Var}(Y_{j,n}) \approx u(1-u).$$

Summing over n, the total variance is:

$$\sum_{j=1}^{n} \mathrm{Var}(Y_{j,n}) \approx n \cdot u(1-u).$$

Therefore, $s_n := \sqrt{\sum_{j=1}^{n} \mathrm{Var}(Y_{j,n})} \approx \sqrt{n} \to +\infty$, so that all of the Lindeberg-Levy-Theorem assumptions hold. Finally :

$$Z_n = \frac{S_n - n \cdot \mathbb{E}[Y_{j,n}]}{\sqrt{\sum_{j=1}^{n} \mathrm{Var}(Y_{j,n})}}.$$

Substituting the expressions for $\mathbb{E}[Y_{j,n}]$ and $\mathrm{Var}(Y_{j,n})$, we get:

$$Z_n = \frac{S_n - n\left(u + \frac{t}{\sqrt{n}}\sqrt{u(1-u)}\right)}{\sqrt{n \cdot u(1-u)}}.$$

Reorganizing the numerator:

$$S_n - n\left(u + \frac{t}{\sqrt{n}}\sqrt{u(1-u)}\right) = S_n - n \cdot u - t\sqrt{n \cdot u(1-u)}.$$

Thus:

$$Z_n = \frac{S_n - n \cdot u - t\sqrt{n \cdot u(1-u)}}{\sqrt{n \cdot u(1-u)}} = \frac{S_n - nu}{\sqrt{n \cdot u(1-u)}} - t.$$

Using the Lindeberg-Levy Central Limit Theorem, we conclude that if $t_n \to t$, the, :

$$\mathbb{P}(Z_n < t_n) \xrightarrow[n\to\infty]{} F_X(t),$$

where $F_X(t)$ is the cumulative distribution function of the standard normal distribution. This implies:

$$Z_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

We replace t with $Q(u)$ and $t_n := Q_n(u)$ therefore, we end up with :

$$\sqrt{n}\,(Q_n(u) - Q(u)) \xrightarrow{\mathcal{L}} N(0, \frac{u(1-u)}{f(Q(u))^2})$$

Question 20

Let $n \in \mathbb{N}$,
$$Q_n(u) = \inf\{x \in \mathbb{R} : F_n(x) \geq u\}.$$

So, when $u \to 0$:

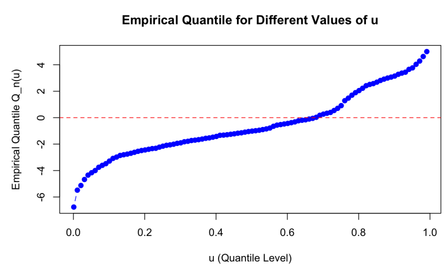
$$Q_n(u) \underset{u \to 0}{\longrightarrow} -\infty,$$

and

$$Q_n(u) \underset{u \to 1}{\longrightarrow} +\infty.$$

Question 21

Graphically, we get the following result for the value of $Q_n(u)$ for $u \longrightarrow 0$ or $u \longrightarrow 1$



We see that it agrees with the intuition of the previous question.

Question 22

To have a 95% confidence interval we use the Central Limit Theorem from Question 19 so we then have :

$$I_{1-\alpha}(u) = \left[Q_n(u) - \frac{q_{\frac{\alpha}{2}}}{f(Q(u))}\sqrt{\frac{u(1-u)}{n}},\, F_n(x) + \frac{q_{\frac{\alpha}{2}}}{f(Q(u))}\sqrt{\frac{u(1-u)}{n}} \right],$$

Thus, we use the same method as in Question 17 to compute the number of simulations n.

u	n
0.5000	2692
0.9000	1609
0.9900	511
0.9990	23
0.9999	2

Table 1: Number of simulations n for different values of u