LAB_1

Pierre Houdoin, Clément Bonnet

20/10/2020

Lab 1

ML estimation with PDF

Theoretical analysis

Question 1: maximum likelihood estimator?

For n iid observations x_i of the height of the river, the likelihood can be written as following

$$\begin{aligned} L(a; x_1, ..., x_n) &= \prod_{i=1}^{n} f_H(x_i) \\ &= \frac{1}{a^n}\Big(\prod_{i=1}^{n} x_i\Big) e^{-\frac{1}{2a}\sum_{i=1}^{n} x_i^2} \end{aligned}$$

The log-likelihood can be derived from the likelihood as follows

$$\begin{aligned} l(a; x_1, ..., x_n) &= log(L(a; x_1, ..., x_n)) \\ &= \sum_{i=1}^{n} \ln(x_i) - n\ln(a) - \frac{1}{2a}\sum_{i=1}^{n} x_i^2 \\ \frac{\partial l}{\partial a}(a; x_1, ..., x_n) &= -\frac{n}{a} + \frac{1}{2a^2}\sum_{i=1}^{n} x_i^2 \end{aligned}$$

Then one can derive the maximum likelihood estimator by setting the partial derivative to 0.

$$\frac{\partial l}{\partial a}(\hat{a}_n; x_1, ..., x_n) = 0 \iff \boxed{\hat{a}_n = \frac{1}{2n}\sum_{i=1}^{n} x_i^2}$$

Question 2: method of moments estimator?

$$\begin{aligned} E[X] &= \int_0^{+\infty} x f_H(x) dx \\ &= \int_0^{+\infty} \frac{x^2}{a} e^{\frac{-x^2}{2a}} dx \\ &= \int_0^{+\infty} e^{\frac{-x^2}{2a}} dx \\ &= \sqrt{\frac{\pi a}{2}} \end{aligned}$$

One can estimate the expectation using the arithmetic mean. Hence, the method of moments estimator $\bar{a}_n$ is:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \sqrt{\frac{\pi \bar{a}_n}{2}} \iff \boxed{\bar{a}_n = \frac{2}{\pi n^2}\Big(\sum_{i=1}^{n} x_i\Big)^2}$$

Question 3: properties of $\hat{a}_n$?

a) Unbiased?

$$\begin{aligned}
E[\hat{a}_n] &= \frac{1}{2n}\sum_{i=1}^{n} E[X_i^2] \\
&= \frac{1}{2}E[X^2] \\
&= \frac{1}{2}\int_0^{+\infty} \frac{x^3}{a} e^{\frac{-x^2}{2a}}\, dx \\
&= \int_0^{+\infty} x e^{\frac{-x^2}{2a}}\, dx \\
&\boxed{E[\hat{a}_n] = a}
\end{aligned}$$

$\hat{a}_n$ is unbiased.

b) Optimal? Let us derive the variance of the estimator $\hat{a}_n$.

$$\begin{aligned}
Var[\hat{a}_n] &= \frac{1}{4n^2}\sum_{i=1}^{n} Var[X^2] \\
&= \frac{1}{4n}\big(E[X^4] - E[X^2]^2\big)
\end{aligned}$$

$$\begin{aligned}
E[X^4] &= \int_0^{+\infty} \frac{x^5}{a} e^{\frac{-x^2}{2a}}\, dx \\
&= 4\int_0^{+\infty} x^3 e^{\frac{-x^2}{2a}}\, dx \\
&= 4aE[X^2] \\
Thus, Var[\hat{a}_n] &= \frac{1}{4n}\big(4aE[X^2] - E[X^2]^2\big)
\end{aligned}$$

We know from question a) that $E[X^2] = 2E[\hat{a}_n] = 2a$.

$$\boxed{Var[\hat{a}_n] = \frac{a^2}{n}}$$

Let us now compute the Fisher information $I(a)$

$$\frac{\partial^2}{\partial a^2} l(a; x_1, ..., x_n) = \frac{n}{a^2} - \frac{1}{a^3}\sum_{i=1}^{n} x_i^2$$

$$\begin{aligned}
I_n(a) &= E[-\frac{\partial^2}{\partial a^2} l(a; x_1, ..., x_n)] \\
&= \frac{n}{a^2}\big(-1 + \frac{1}{a}E[X^2]\big) \\
I_n(a) &= \frac{n}{a^2}
\end{aligned}$$

$$\boxed{Var[\hat{a}_n] = \frac{1}{I_n(a)}}$$

Its variance equals the Cramer–Rao lower bound and it is unbiased. Hence, $\hat{a}_n$ minimizes the mean squared error. So $\hat{a}_n$ is both optimal.

c) Efficient? Since $\hat{a}_n$ is unbiased and optimal. Therefore, $\hat{a}_n$ is efficient because its variance is equal to the Cramer-Rao lower bound.

d) Asymptotically Gaussian? The maximum likelihood estimator is asymptotically gaussian. Hence, $\hat{a}_n$ is asymptotically gaussian. $I(a) = \frac{I_n(a)}{n} = \frac{1}{a^2}$

$$\boxed{\sqrt{n}(\hat{a}_n - a) \xrightarrow[n\to\infty]{d} N\big(0, a^2\big)}$$

Application on real data

Question 1: p function of a?

Let p the probability that a disaster happens during one year.

$$\begin{aligned} p &= 1 - F_H(6) \\ &= \int_6^{\infty} \frac{x}{a} e^{-\frac{x^2}{2a}} dx \\ &= \Big[- e^{-\frac{x^2}{2a}} \Big]_6^{\infty} \end{aligned}$$

$$\boxed{p = e^{-\frac{18}{a}}}$$

Question 2: probability of at most one disaster?

During one thousand years, if at most one disaster happened, it means either there was no disasters, or there was only one. Let us derive p_1, the probability that at most one disaster happens during one thousand years.

$$\boxed{p_1 = (1-p)^{999} = (1 - e^{-\frac{18}{a}})^{999}}$$

Question 3: estimation of the probability?

```
X = c(2.5, 1.8, 2.9, 0.9, 2.1, 1.7, 2.2, 2.8)
n = length(X)
a = sum(X^2)/(2*n)
p = (1-exp(-18/a))^999
```

Regarding the set of 8 observations, one can estimate $\hat{a} = 2.42$. The probability p_1 can be estimated: $\boxed{p_1 = 0.557}$.

Exercise 1: Rayleigh distribution

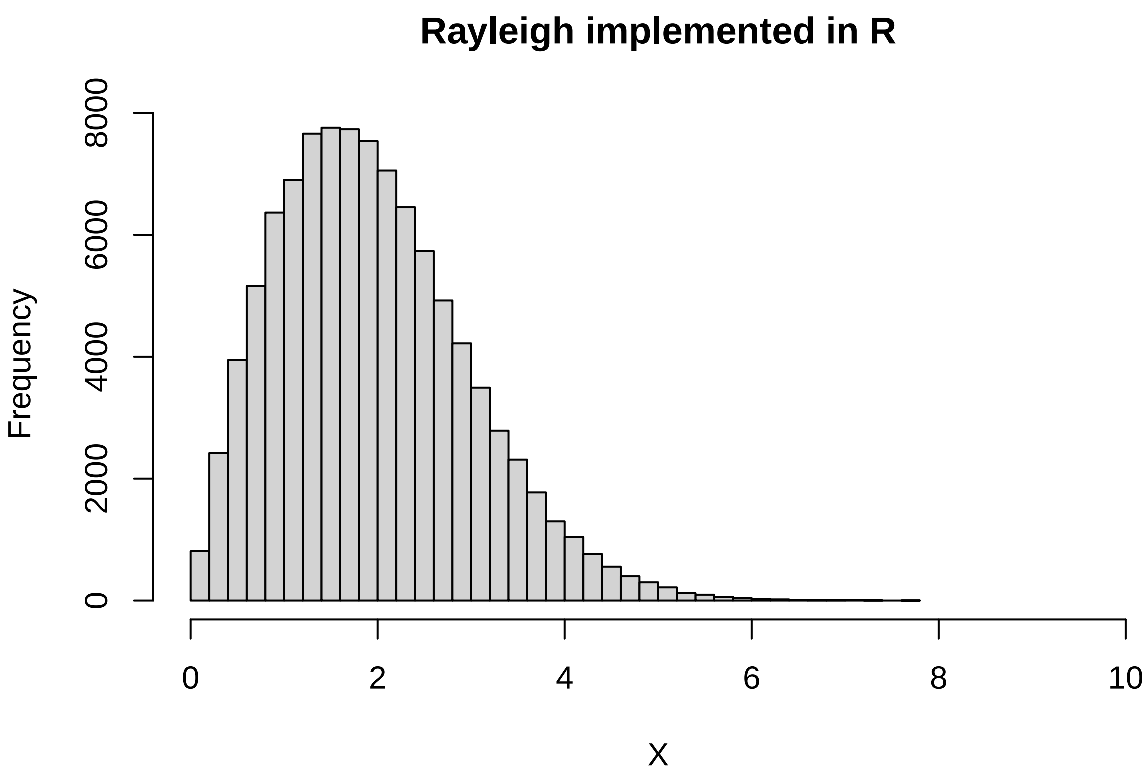
(a)

The parameter a of the Rayleigh distribution was estimated with the maximum likelihood estimator $\hat{a}_n$. It was found that $a \approx 2.42$.

(b)

One can generate more samples following a Rayleigh distribution by using the Rayleigh distribution function implemented in R. One has to be careful that the scale σ used in R corresponds to $\sqrt{a}$.

```
n = 100000
X = rrayleigh(n, scale=sqrt(a))
hist(X, nclass=50, xlim = c(0,10), main = "Rayleigh implemented in R")
```



If one has only access to uniform distribution and would like to output a Rayleigh distribution, one can use the inverse distribution function.

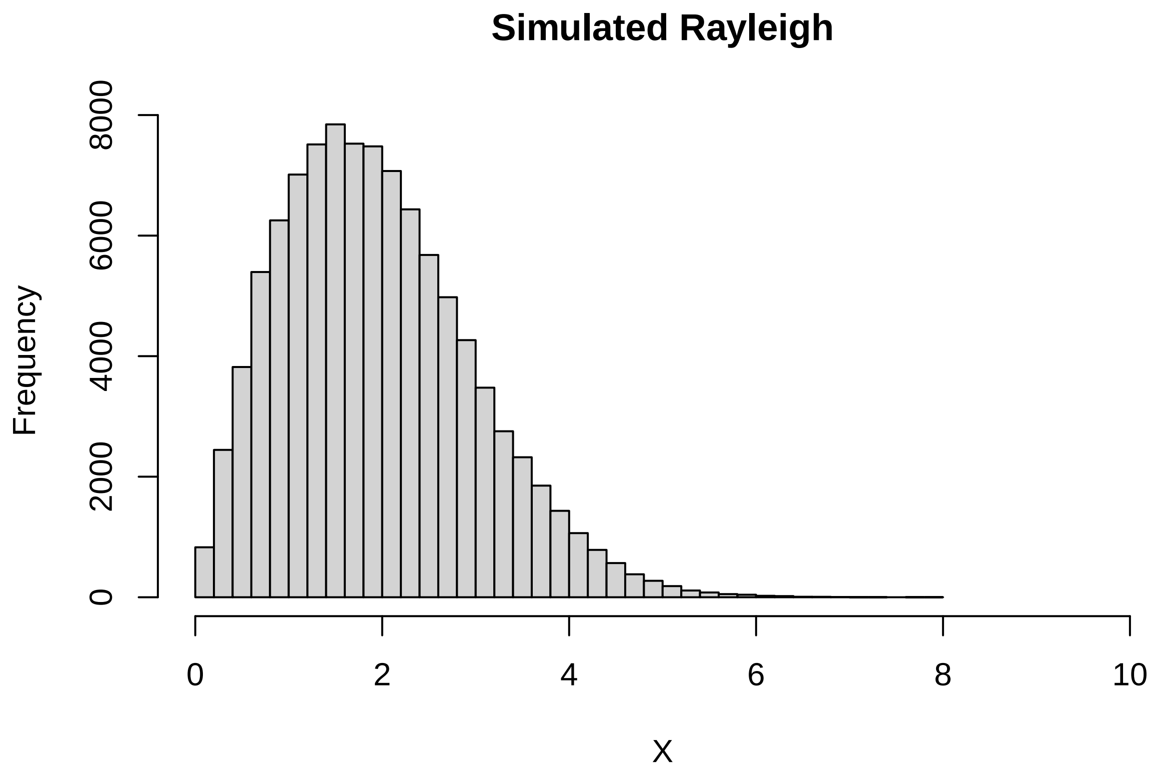
$$\begin{aligned} F(x) &= \int_0^x \frac{t}{a} e^{-\frac{t^2}{2a}} dt \\ &= 1 - e^{-\frac{x^2}{2a}} \end{aligned}$$

$$F(x) = u \iff x = \sqrt{-2a\ln(1-u)}$$

If U follows a uniform distribution on $[0, 1]$, one can generate samples following the Rayleigh distribution using the uniform distribution.

$$U \sim U[0,1] \implies \sqrt{-2a\ln(U)} \sim Rayleigh(a)$$

```
n = 100000
U = runif(n)
X = sqrt(-2*a*log(U))
hist(X, nclass=50, xlim = c(0,10), main = "Simulated Rayleigh")
```



(c)

Empirically, one can verify that the MLE is unbiased by averaging N samples of the MLE $\hat{a}_{n,1}, ..., \hat{a}_{n,N}$ with whatever value for n. For computing resources reasons, let's take $n = 10$ and average over $N = 100000$ samples of n observations.

$$\begin{aligned} E[\hat{a}_n - a] &\approx \frac{1}{N}\sum_{k=1}^{N}(\hat{a}_{n,k} - a) \\ &= \frac{1}{N}\sum_{k=1}^{N}(\frac{1}{2n}\sum_{i=1}^{n}x_{i,k}^2 - a) \end{aligned}$$

```r
N = 100000
n = 10
E = 0
for (k in 1:N){
  X = rrayleigh(n, scale=sqrt(a))
  E = E + 1/(2*n)*sum(X^2) - a
}
E = E/N
E
```

```
## [1] -0.002336315
```

$\frac{E[\hat{a}_n - a]}{a} \approx -9.7 \times 10^{-4} \ll 1$. Hence, empirically, the estimator is unbiased.

(d)

Empirically, one can verify the efficiency of the MLE estimator by computing its variance and compare it to the inverse of the Fisher information. One needs an unbiased estimator of the variance, knowing that the mean is a.

$$\begin{aligned} Var[\hat{a}_n] &\approx \frac{1}{N}\sum_{k=1}^{N}(\hat{a}_{n,k}-a)^2 \\ &= \frac{1}{N}\sum_{k=1}^{N}(\frac{1}{2n}\sum_{i=1}^{n}x_{i,k}^2-a)^2 \end{aligned}$$

```
N = 100000
n = 10
I = n/a^2  # Fisher information
V = 0
for (k in 1:N){
  X = rrayleigh(n, scale=sqrt(a))
  dV = 1/(2*n)*sum(X^2) - a
  V = V + dV^2
}
V = V/N
V
```

```
## [1] 0.5887242
```

$$\frac{1}{I_n(a)} = \frac{a^2}{n} \approx 0.5847$$

$$\frac{Var[\hat{a}_n] - \frac{1}{I_n(a)}}{\frac{1}{I_n(a)}} \approx 0.00683 \ll 1$$

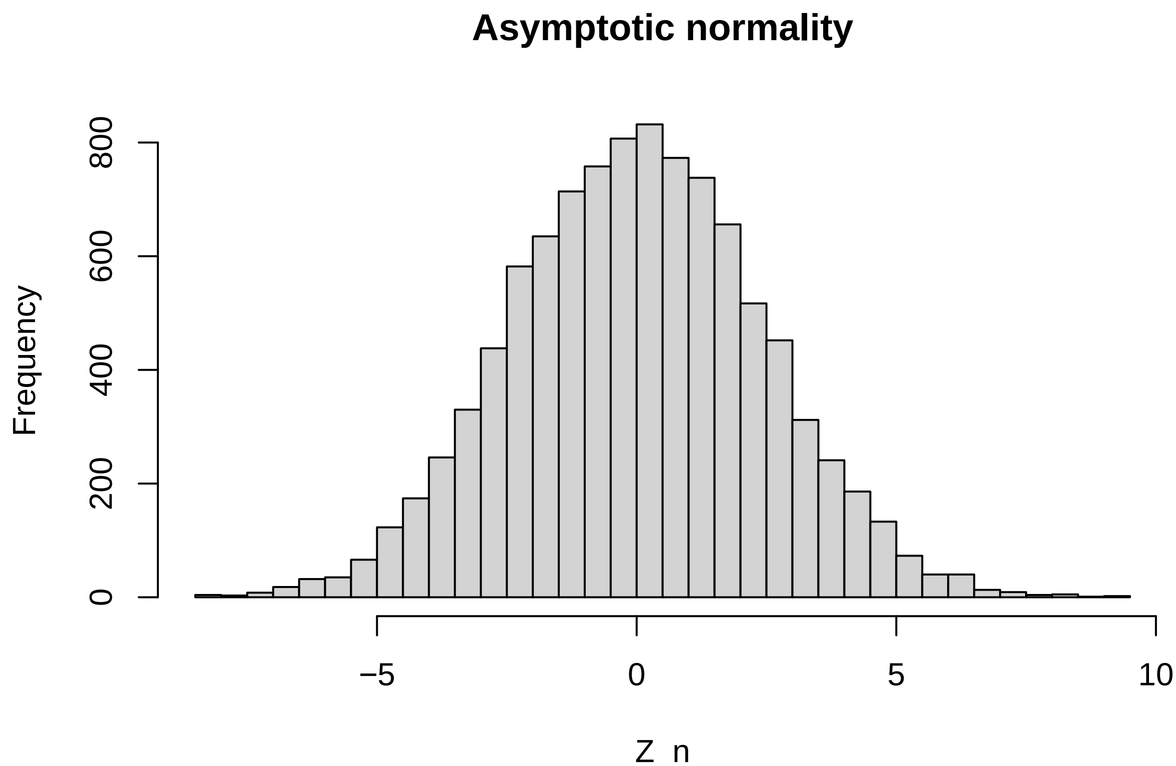
Hence, one can say that $Var[\hat{a}_n] = \frac{1}{I_n(a)}$ and the estimator is efficient empirically.

(e)

$$\boxed{\sqrt{n}(\hat{a}_n - a) \xrightarrow[n\to\infty]{d} N\big(0, a^2\big)}$$

The asymptotic normality means that for n large, $\sqrt{n}(\hat{a}_n - a) \sim N\big(0, a^2\big)$. Thus one can plot several samples of the random variable $Z_n = \sqrt{n}(\hat{a}_n - a)$ and check whether the distribution looks Gaussian.

```
n = 10000 # Size of the observations for each a_n
N = 10000   # Number of samples of Z_n
Z_n = rep(0,N)
for (k in 1:N){
  X = rrayleigh(n, scale=sqrt(a))
  a_n = 1/(2*n)*sum(X^2)
  Z_n[k] = sqrt(n)*(a_n - a)
}
hist(Z_n, breaks=40, main = "Asymptotic normality")
```



The MLE estimator is asymptotically normal. One can verify the standard deviation $a = 2.418$.

ML estimation with PMF

Statistical modelling and theoretical analysis

Question 1: Belonging to exponential family

We study the random variable X that follows a geometric distribution with parameter $q \in]0;1[$. We have :

$$\forall k \in \mathbb{N}^\star, P(X=k) = q(1-q)^{k-1}$$

The likelihood function can be written as :

$$\begin{aligned} L(x,q) &= \mathbf{1}_{\mathbb{N}^\star}(x) \times q(1-q)^{x-1} \\ &= \mathbf{1}_{\mathbb{N}^\star}(x) \times q \times e^{(x-1)\ln(1-q)} \\ &= \mathbf{1}_{\mathbb{N}^\star}(x) \times \frac{q}{1-q} \times e^{xln(1-q)} \end{aligned}$$

We can notice that the model is dominated and the distribution domain where $L(x,q) > 0$ is $\mathbb{N}^\star$ which does not depend on q. Thus the distribution domain is homogeneous. We then define :

$$\begin{aligned} h &: x \mapsto \mathbf{1}_{\mathbb{N}^\star}(x) \\ \phi &: q \mapsto \frac{q}{1-q} \\ Q &: q \mapsto \ln(1-q) \\ S &: x \mapsto x \end{aligned}$$

We can then write the likelihood like :

$$\boxed{L(x,q) = h(x)\phi(q)\exp\left(Q(q)S(x)\right)}$$

We can conclude that a geometric distribution belongs to the exponential family and S is a sufficient statistic.Since S is linearly independent with itself, we also deduce that the model is identifiable.

Question 2: Computation of the Fisher Information Matrix

We saw in question 1 that the model was dominated and the distribution domain was homogeneous. We can also easily show that $L(x,q)$ is twice differentiable for variable q and integrable, since it is a polynomial function. Thus the model is regular. We note $l(x,q)$ the log-likelihood of the model :

$$\begin{aligned}\forall x \in \mathbb{N}^\star, l(x,q) &= \ln(q(1-q)^{x-1}) \\ &= \ln(q) + (x-1)\ln(1-q)\end{aligned}$$

We can now deduce the score function:

$$\begin{aligned}\forall x \in \mathbb{N}^\star, s_q(x) &= \frac{\partial}{\partial q} l(x,q) \\ &= \frac{1}{q} - \frac{x-1}{1-q}\end{aligned}$$

We can note that the score function is an affine transform of X, thus it is square-integrable because X is, so the Fisher Information Matrix is well-defined. We showed previously that the model was regular, thus we have :

$$\begin{aligned}I(q) &= E_q\left(s_q(X)^2\right) \\ &= -E_q\left(\frac{\partial^2}{\partial q^2} l(X,q)\right) \\ &= -E_q\left(\frac{-1}{q^2} - \frac{X-1}{(1-q)^2}\right) \\ &= \frac{1}{q^2} + \frac{E_q(X)-1}{(1-q)^2} \qquad and \qquad E(X) = \frac{1}{q} \\ &= \frac{1}{q^2} + \frac{\frac{1}{q}-1}{(1-q)^2} \\ &= \frac{(1-q)^2}{q^2(1-q)^2} + \frac{q-q^2}{q^2(1-q)^2} \\ &= \boxed{\frac{1}{q^2(1-q)}}\end{aligned}$$

Question 3: Maximum likelihood estimator

Let $X_1, ..., X_n$ a n-sample following the same distribution as X. The likelihood of the model is :

$$\begin{aligned}L(x_1,...,x_n,q) &= \prod_{i=1}^{n} q(1-q)^{x_i-1} \\ &= \left(\frac{q}{1-q}\right)^n (1-q)^{\sum_{i=1}^n x_i} \\ l(x_1,...,x_n,q) &= nln\left(\frac{q}{1-q}\right) + \ln(1-q)\sum_{i=1}^{n} x_i\end{aligned}$$

We look for $\widehat{q_n}$ that maximizes the likelihood of the n-sample. Thus it satisfies two equations :

$$\frac{\partial}{\partial q} l(x_1, ..., x_n, \widehat{q_n}) = 0 \qquad (1)$$
$$\frac{\partial^2}{\partial q^2} l(x_1, ..., x_n, \widehat{q_n}) < 0 \qquad (2)$$

From equation (1) we deduce :

$$\begin{aligned}
& n\left(\frac{1}{\widehat{q_n}} + \frac{1}{1-\widehat{q_n}}\right) - \frac{\sum_{i=1}^n x_i}{\widehat{q_n}} = 0 \\
\iff & \frac{n}{\widehat{q_n}(1-\widehat{q_n})} - \frac{\sum_{i=1}^n x_i}{1-\widehat{q_n}} = 0 \\
\iff & \frac{n}{\widehat{q_n}} = \sum_{i=1}^n x_i \\
\iff & \frac{1}{\widehat{q_n}} = \frac{1}{n}\sum_{i=1}^n x_i
\end{aligned}$$

The maximum likelihood estimator $\widehat{q_n}$ is $\boxed{\frac{1}{\overline{X_n}} \quad where \quad \overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i}$

Question 4: Asymptotic behavior of the estimator

Now we are going to study the asymptotic behavior of the estimator. According to the Central limit theorem, we have :

$$\begin{aligned}
& \sqrt{n}\left(\overline{X_n} - \frac{1}{q}\right) \xrightarrow{L} \mathcal{N}\left(0, V(X)\right) \qquad where \qquad V(X) = \frac{1-q}{q^2} \\
\iff & \sqrt{n}\left(\overline{X_n} - \frac{1}{q}\right) \xrightarrow{L} \mathcal{N}\left(0, \frac{1-q}{q^2}\right)
\end{aligned}$$

Then, we use the delta method. We define : $g : x \mapsto \frac{1}{x}$ which is differentiable in $\frac{1}{q}$. We have:

$$\begin{aligned}
& \sqrt{n}\left(g\left(\overline{X_n}\right) - g\left(\frac{1}{q}\right)\right) \xrightarrow{L} \mathcal{N}\left(0, g'\left(\frac{1}{q}\right)^2 \frac{1-q}{q^2}\right) \\
\iff & \sqrt{n}\left(\widehat{q_n} - q\right) \xrightarrow{L} \mathcal{N}\left(0, (1-q)q^2\right) \\
\iff & \boxed{\sqrt{n}\left(\widehat{q_n} - q\right) \xrightarrow{L} \mathcal{N}\left(0, I^{-1}(q)\right)}
\end{aligned}$$

This estimator is asymptotically normal and its asymptotic variance is the Cramer Rao bound, thus the estimator is asymptotically efficient. This was expected because this is a maximum likelihood estimator.

Question 5: Asymoptotic confidence interval

Finally, we build an asymptotic confidence interval for q. On pose :

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

Then, we know that : $\frac{\sqrt{n}\left(\overline{X_n}-\frac{1}{q}\right)}{S_n} \sim t(n-1)$, where t is the student law. Since the law is symmetric, we can write : $-t^{n-1}_{\alpha/2} \leq \frac{\sqrt{n}\left(\overline{X_n}-\frac{1}{q}\right)}{S_n} \leq t^{n-1}_{\alpha/2}$ where t^k_α is the unique real number that verifies $P(t(k) < t^k_\alpha) = 1-\alpha$. Finally, we have:

$$
\begin{aligned}
& -t^{n-1}_{\alpha/2} \leq \frac{\sqrt{n}\left(\overline{X_n}-\frac{1}{q}\right)}{S_n} \leq t^{n-1}_{\alpha/2} \\
\iff & -\frac{S_n}{\sqrt{n}} t^{n-1}_{\alpha/2} \leq \overline{X_n} - \frac{1}{q} \leq \frac{S_n}{\sqrt{n}} t^{n-1}_{\alpha/2} \\
\iff & \overline{X_n} - \frac{S_n}{\sqrt{n}} t^{n-1}_{\alpha/2} \leq \frac{1}{q} \leq \overline{X_n} + \frac{S_n}{\sqrt{n}} t^{n-1}_{\alpha/2} \\
\iff & \boxed{\frac{1}{\overline{X_n} + \frac{S_n}{\sqrt{n}} t^{n-1}_{\alpha/2}} \leq q \leq \frac{1}{\overline{X_n} - \frac{S_n}{\sqrt{n}} t^{n-1}_{\alpha/2}}}
\end{aligned}
$$

Application on real data

Question 1: Estimation of the fraud probability

```
X = c(44, 9, 11, 59, 81, 19, 89, 10, 24, 07, 21, 90, 38, 01, 15, 22, 29, 19, 37, 26, 219, 2, 57, 11, 34
n = length(X)
p = (n)/sum(X)
```

The probability of fraud p_{fraud} can be estimated with the estimator studied above: $\boxed{p_{fraud} = 0.026}$. We take $1-\alpha = 0.95$, we can deduce a confidence interval:

```
t_alpha = 2.021 #quantile for student law for n = 40, closest value available.
Xn=mean(X)
Sn=sqrt(mean((X-Xn)**2)*n/(n-1))
a = 1/(Xn+Sn*t_alpha/sqrt(n))
b = 1/(Xn-Sn*t_alpha/sqrt(n))
```

With a 95% confidence, we know that p_{fraud} is between $\boxed{0.019 \; and \; 0.043}$. If we have $n_0 = 20000$ validated tickets, we can estimate the number n_{fraud} of fraudsters. For 1000 users, there are 26 fraudsters and 974 honest users. Thus we have $\boxed{\frac{n_0}{1-p_{fraud}} = 534}$ fraudsters.

Exercise 2: Geometric distribution

(a)

The parameter p_{fraud} of the geometric distribution was estimated with the maximum likelihood estimator $\hat{p}_n$. It was found that $\boxed{p_{fraud} \approx 0.026}$.

(b)

One has only access to uniform distribution and would like to output a geometric distribution. We start by randomly drawing a number q between 0 and 1, according to the uniform distribution. Then, we realize the following segmentation of the interval $[0,1]$:

$$
\begin{aligned}
[0,1] &= \bigcup_{k=1}^{+\infty} \left[\sum_{i=1}^{k-1} p_{fraud}(1-p_{fraud})^{i-1}, \sum_{i=1}^{k} p_{fraud}(1-p_{fraud})^{i-1} \right] \\
&= \bigcup_{k=1}^{+\infty} \left[1-(1-p_{fraud})^{k-1}, 1-(1-p_{fraud})^{k} \right]
\end{aligned}
$$

Thus, if we draw q, we look for k that satisfies:

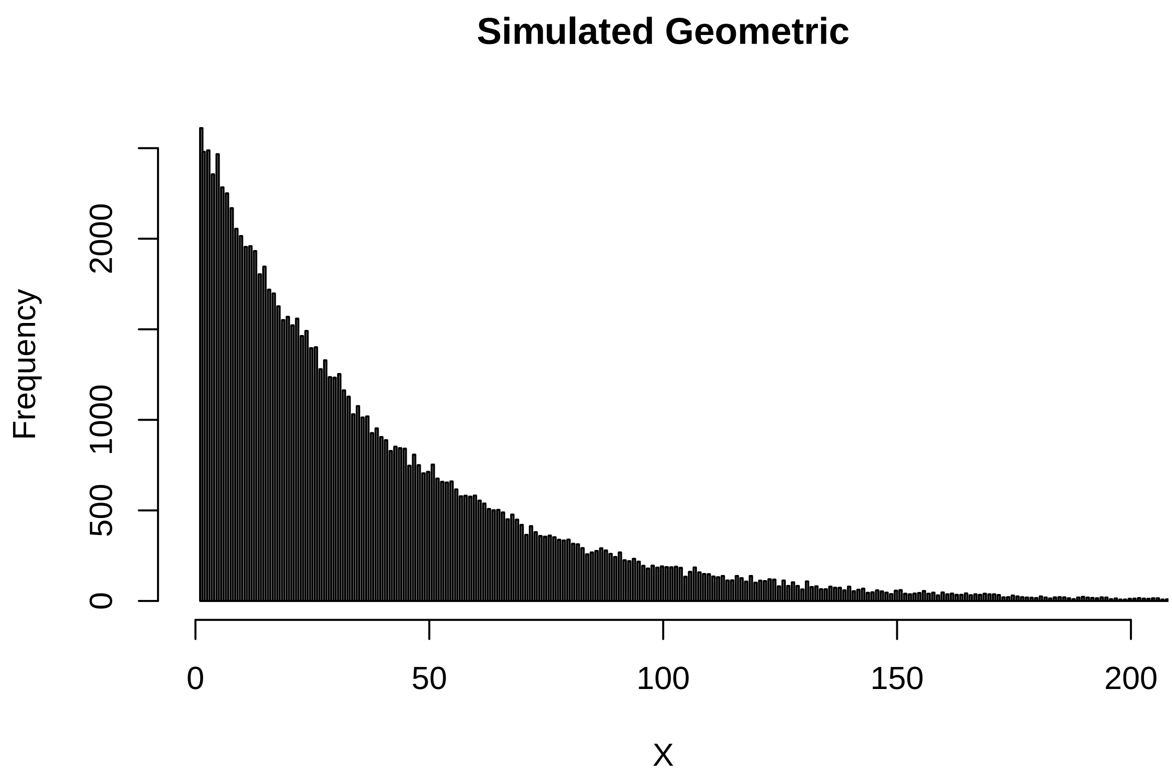
$$
\begin{aligned}
&1-(1-p_{fraud})^{k-1} \leq q \leq 1-(1-p_{fraud})^{k} \\
&\iff -(1-p_{fraud})^{k-1} \leq q-1 \leq -(1-p_{fraud})^{k} \\
&\iff (1-p_{fraud})^{k} \leq 1-q \leq (1-p_{fraud})^{k-1} \\
&\iff kln(1-p_{fraud}) \leq \ln(1-q) \leq (k-1)\ln(1-p_{fraud}) \\
&\iff k-1 \leq \frac{\ln(1-q)}{\ln(1-p_{fraud})} \leq k
\end{aligned}
$$

Since the probability to draw an integer is 0, we can choose $k = \lceil \frac{\ln(1-q)}{\ln(1-p_{fraud})} \rceil$

If U follows a uniform distribution on $[0,1]$, one can generate samples following the geometric distribution using the uniform distribution.

$$U \sim U[0,1] \implies \lceil \frac{\ln(1-U)}{\ln(1-p_{fraud})} \rceil \sim G(p_{fraud})$$

```
n = 100000
U = runif(n)
X = ceiling(log(1-U)/log(1-p))
hist(X, nclass=800, xlim = c(0,200), main = "Simulated Geometric")
```



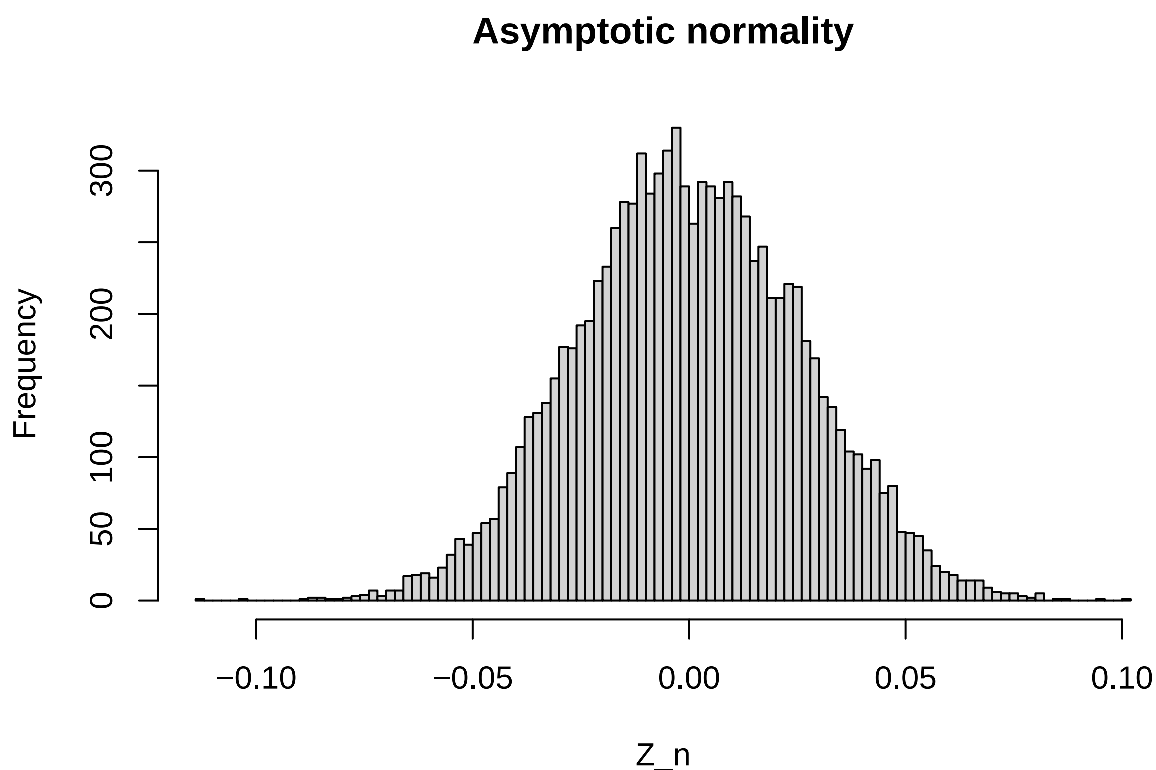
(c)

We have shown that:

$$\boxed{\sqrt{n}(\hat{q}_n - q) \xrightarrow{d} N\left(0, q^2(1-q)\right)}$$

The asymptotic normality means that for n large, $\sqrt{n}(\hat{a}_n - a) \sim N\big(0, a^2\big)$. Thus one can plot several samples of the random variable $Z_n = \sqrt{n}(\hat{q}_n - q)$ and check whether the distribution looks Gaussian.

```
n = 10000 # Size of the observations for each q_n
N = 10000   # Number of samples of Z_n
Z_n = rep(0,N)
for (k in 1:N)
  {
  U = runif(n)
  X = ceiling(log(1-U)/log(1-p))
  p_n = 1/mean(X)
  Z_n[k] = sqrt(n)*(p_n - p)
  }
hist(Z_n, breaks=100, main = "Asymptotic normality")
```




```
var_emp = mean(Z_n**2)
var_theo = p^2*(1-p)
```

The MLE estimator is asymptotically normal.An estimation of the variance gives $\sigma = 6.72 \times 10^{-4}$, which is close to the value : $p_{fraude}^2 * (1 - p_{fraude}) = 6.67 \times 10^{-4}$

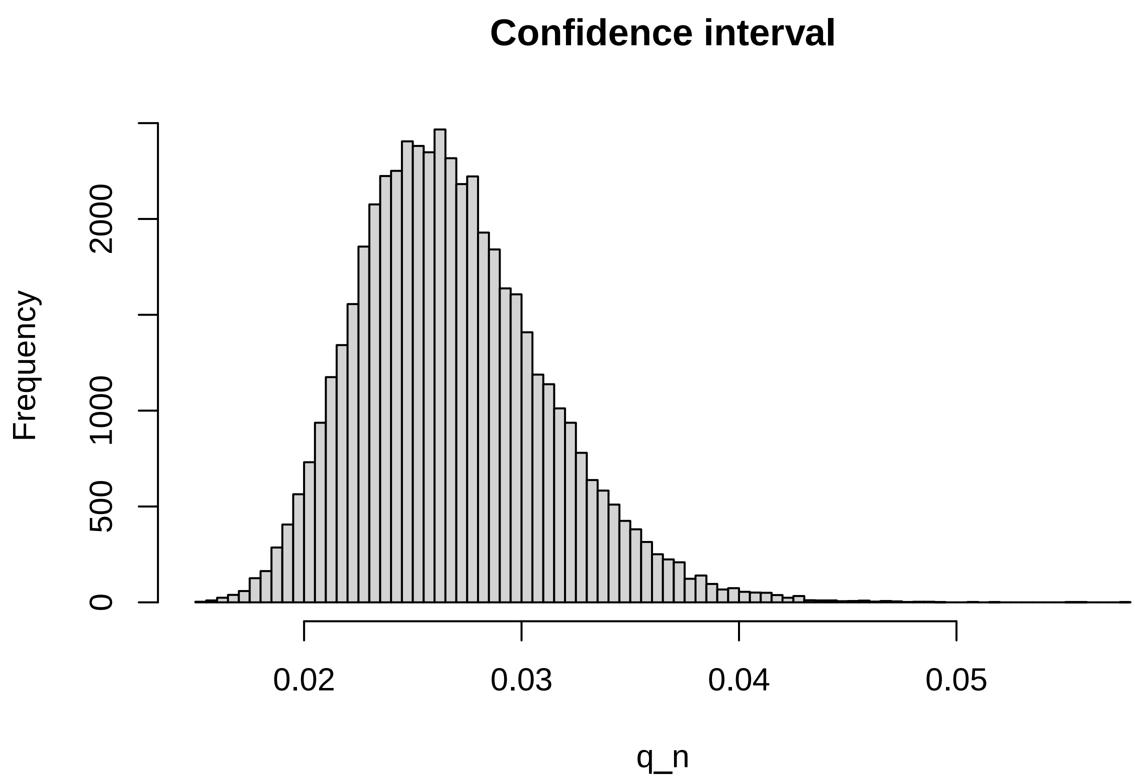
(d)

We have shown that a 95% confidence interval is :

$$\frac{1}{\overline{X_n} + \frac{S_n}{\sqrt{n}} t_{\alpha/2}^{n-1}} \leq q \leq \frac{1}{\overline{X_n} - \frac{S_n}{\sqrt{n}} t_{\alpha/2}^{n-1}}$$

For a 95% confidence interval. On obtain: $\frac{1}{\overline{X_n}+2.021\frac{S_n}{\sqrt{n}}} \leq q \leq \frac{1}{\overline{X_n}-2.021\frac{S_n}{\sqrt{n}}}$. We use the 39 values given and we simulated 5000 times a 39 dataset, we estimate q and compute the % of q in the confidence interval.

```
n = 39 # Size of the observations for each q_n
N = 50000    # Number of samples of Z_n
count = 0
q_n = rep(0,N)
for (k in 1:N)
  {
  U = runif(n)
  X = ceiling(log(1-U)/log(1-p))
  p_n = 1/mean(X)
  q_n[k] = p_n
  if (a<p_n & p_n<b)
  {
    count = count + 1
  }
  }
hist(q_n, breaks=100, main = "Confidence interval")
```



We obtain that 0.986 of the simulations give an estimation of q in the confidence interval.