Method of Maximum Likelihood

Sujit Sandipan Chaugule[1*], **Dr. Amiya Ranjan Bhowmick**[2]

[1*]Department of Pharmaceutical Sciences and Technology, Institute of Chemical Technology, Mumbai

[2]Department of Mathematics, Institute of Chemical Technology, Mumbai

A motivating example

We have a random sample $(X_1, X_2, \ldots, X_n)$ of size n from the Geometric(p) distribution, whose probability mass function is given by

$$P(X = x) = (1-p)^{x-1}p, \quad x \in \{1, 2, \ldots\}.$$

The random variable X represents the number of throws required to obtain the first success if we continue tossing a coin with probability of head p until the first head is observed.
Suppose that a company produces a large number of identical coins. We are interested in estimating the probability of head. The following experimental procedure is planned to be followed to estimate the true probability of head. Out of a large number of coins n coins have been chosen and they are numbered as $\{1, 2, \ldots, n\}$. For each $i \in \{1, 2, \ldots, n\}$, the ith coin has been kept on tossing till the first head appears. X_i denotes the number of throws required to observe the first head for the ith coin.
Suppose that $n = 5$ and the above experiment gave the following observations:

```
[1] 4 6 1 1 2
```

Let us compute the likelihood of observing the above sample as a function of p as follows:

$$P(X_1 = 4, X_2 = 6, X_3 = 1, X_4 = 1, X_5 = 2) = (1-p)^9 p^5.$$

The sample space of $(X_1, \ldots, X_5)$ is given by the following set:

$$\{1, 2, 3, \ldots\}^5 = \{(x_1, x_2, \ldots, x_5) : x_i \in \{1, 2, 3, \ldots\}, 1 \leq i \leq 5\}.$$

Each of the points in this sample space has a positive probability of being included in the random sample of size 5. However, since the given sample is observed, it is not unreasonable to consider that it has significantly larger likelihood. In other words, we ask the question, for what value of $p \in (0, 1)$, the probability of observing the given sample is as large as possible?
To answer this, we can express the likelihood of the given sample as a function of p, which is here given by

$$\mathcal{L}(p) = (1-p)^9 \times p^5, \quad p \in (0, 1).$$

We want to identify that for what value of p, the likelihood of the observed sample is maximum, that is p^*, so that $\mathcal{L}(p^*) \geq \mathcal{L}(p)$ for all $p \in (0, 1)$. It boils down to a maximization problem. Instead of maximizing $\mathcal{L}(\theta)$, we can maximize $\log \mathcal{L}(\theta)$, which are referred to as the likelihood and the log-likelihood function, respectively.

$$l(p) = \log \mathcal{L}(p) = 9 \log(1-p) + 5 \log p.$$

$$\frac{d}{dp}(\log \mathcal{L}(p)) = l'(p) = -\frac{9}{1-p} + \frac{5}{p}.$$

Equating $l'(p) = 0$, we get $p^* = \frac{5}{14}$ and it is easy to verify that $l''(p^*) < 0$.

In [1]:

```
using Plots, Statistics, StatsBase, Random
using Distributions, LaTeXStrings
```

In [2]:

```
p_star = 5/14
p1 = plot(x->(1-x)^9*x^5, 0.01, 0.99, color = "red", lw = 2,
xlabel = "p", ylabel = "l(p)", label = "")
scatter!([p_star],[0], color = "blue", markersize = 6, label = "" )
vline!([p_star], color = "grey", lw = 2, linestyle = :dash, label = "")

p2 = plot(x->9*log(1-x) + 5*log(x), color = "red", lw = 2,
xlabel = "p", ylabel = "l(p)", label = "")
scatter!([p_star],[0], color = "blue", markersize = 6, label = "" )
vline!([p_star], color = "grey", lw = 2, linestyle = :dash, label = "")

plot(p1, p2, layout = (1,2), size = (900, 500))
```

Out[2]:

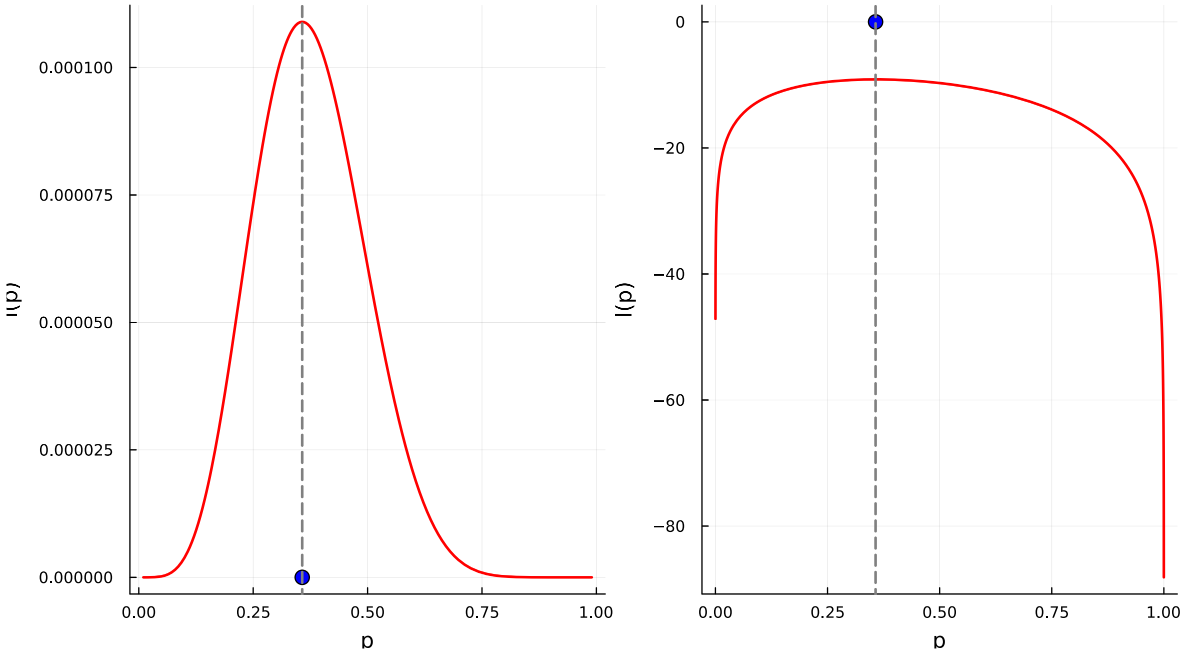


Figure 1: The maximum likelihood estimate of p is $\frac{5}{14}$. The left panel represents the likelihood function $\mathcal{L}(p)$ and the right panel depicts the log-likelihood function $\log \mathcal{L}(p)$. The vertical corresponds to the $p^* = \frac{5}{14}$, at which the function is maximum.

Based on the above data set, we obtained the estimate of the probability as $\frac{5}{14}$. It is intuitively clear that this estimate is subject to uncertainty. By this statement, I essentially mean that if the this experiment would have been carried out by ten different individuals, then we would have obtained ten different samples of size 5. Certainly, the value of p^* is not expected to be the same for all the samples. Therefore, we are interested in computing the risk associated with this estimate. To formalize it further, let us try to obtain some explicit expression for p^*.
Suppose that $x_1, x_2, \ldots, x_n$ are realizations of the random sample $(X_1, X_2, \ldots, X_n)$. Then the likelihood function is given by

$$\mathcal{L}(p) = \prod_{i=1}^{n} P(X_i = x_i) = (1-p)^{\sum x_i - n} p^n, \quad 0 < p < 1,$$

and the log-likelihood function is given by

$$l(p) = \left(\sum x_i - n\right)\log(1-p) + n\log p.$$

$l'(p) = 0$ gives $p^* = \frac{n}{\sum x_i} = (\overline{x_n})^{-1}$. Therefore, the maximum likelihood estimator of p is given by

$$\hat{p_n} = \frac{1}{X_n}.$$

Our goal is to estimate the risk function

$$\mathcal{R}(\hat{p_n}, p) = \mathbb{E}\left(\frac{1}{X_n} - p\right)^2, \quad p \in (0,1).$$

❗ **The likelihood and the log-likelihood**

If $X_1, X_2, \ldots, X_n$ be a random sample of size n from the population with PDF (PMF) $f(x|\theta)$, then the likelihood function $\mathcal{L}_n : \Theta \to [0,\infty)$ is defined as

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i|\theta), \quad \theta \in \Theta.$$

It is important to note that the likelihood function is the joint PDF (PMF), but considered as a function of θ, and certainly the statement $\int_\Theta \mathcal{L}_n(\theta)d\theta = 1$ not necessarily be true. The log-likelihood function is defined by

$$l_n(\theta) = \log \mathcal{L}_n(\theta), \quad \theta \in \Theta.$$

Is the estimator constitent!

A desirable property of an estimator is that as the sample size increases, the sampling distribution of the estimator should be more concentrated about the true parameter value. We should not forget that an estimator is a random quantity which is subject to sampling variation. In this case, before going to some mathematical computation, we first check whether this is indeed happening or not for $\hat{p}_n$, the MLE of p. We employed the following scheme to understand the behavior of $\hat{p}_n$ as $n \to \infty$:

- Fix $p_0 \in (0,1)$
- Fix sample size, $n \in \{1,2,3,\ldots,1000\}$.
- For each $n \in \{1,2,3,\ldots,1000\}$
  - Simulate $X_1,\ldots,X_n \sim \text{Geometric}(p_0)$.
  - Compute X_n.
  - Compute $\hat{p}_n = X_n^{-1}$.
- Plot the pairs $(n, \hat{p}_n)$, $n \in \{1,2,\ldots,1000\}$.
- Do the above experiment for different choices of $p_0 \in (0,1)$.

In [3]:

```julia
p = 0.4
n_vals = 1:1000
pn_hat = zeros(length(n_vals))

for n in n_vals
    x = rand(Geometric(p), n) .+ 1
    pn_hat[n] = 1/mean(x)
end

plot(n_vals, pn_hat, color = "lightgrey", lw = 2, xlabel = "sample size(n)",
ylabel = L"\widehat{p_n}",  label = "")
hline!([p], color = "blue", lw = 3, linestyle = :dash, label = "")
```

Out[3]:

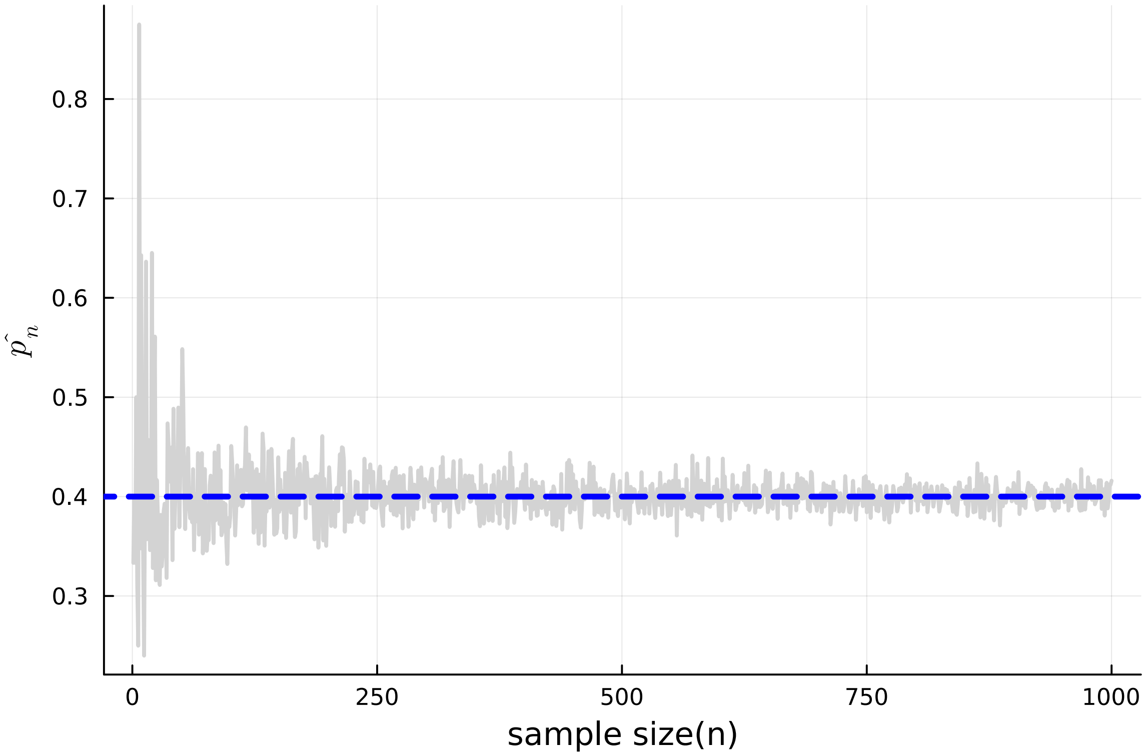


Figure 2: As the sample size increases, the MLE $\hat{p}_n$ converges to the true value of p. The reader is encouraged to perform the simulation experiment for various choices of $p \in (0, 1)$. The horizontal blue dotted line indicates the true probability p.

The statement can also be established theoretically. From the Weak Law of Large Numbers, we know that $\overline{X}_n \rightarrow \mathbb{E}(X) = \frac{1}{p}$ in probability as $n \rightarrow \infty$. If we choose $g(x) = \frac{1}{x}$, which is a continuous function, then

$$g\left(\overline{X}_n\right) = \frac{1}{\overline{X}_n} = \hat{p}_n \overset{P}{\rightarrow} p.$$

! Continuous function and convergence in probability

If $X_n \overset{P}{\rightarrow} X$, then $g(X_n) \overset{P}{\rightarrow} g(X)$, where $g(\cdot)$ is a continuous function.

Simulating the risk function of the MLE of p

The problem is to compute the risk function for different choices of $p \in (0, 1)$,

$$\mathcal{R}\left(X_n^{-1}, p\right), \quad 0 < p < 1.$$

The following steps have been performed using R Programming to approximate the risk function of $\hat{p}_n$ under the squared error loss function, which is also referred to as the Mean Squared Error (MSE) of $\hat{p}_n$.

- Discretize $(0, 1)$ as $(p_1, \ldots, p_L)$.
- Fix sample size n.
- Fix M, the number of replications.
- For each $p \in \{p_1, \ldots, p_L\}$
  - For each $m \in \{1, 2, \ldots, M\}$
    - Simulate $X_1, X_2, \ldots, X_n \sim \text{Geometric}(p)$.
    - Compute loss $l_m = \left(\frac{1}{\bar{X}_n} - p\right)^2$.
  - Compute the risk $\mathcal{R}\left(\frac{1}{\bar{X}_n}, p\right) = \frac{1}{M}\sum_{m=1}^{M} l_m$.
- Plot the pairs of values $\left\{p, \mathcal{R}\left(\frac{1}{\bar{X}_n}, p\right)\right\}, p \in \{p_1, \ldots, p_L\}$.
- Repeat the above exercise for different choices of n.

In [4]:

```
using Distributions, Statistics, Random, StatsBase
using Plots, LaTeXStrings
```

In [5]:

```
n_vals = [5, 10, 30, 50, 100, 500]
M = 5000
prob_vals = 0.01:0.01:0.9
plt = plot(layout=(2, 3), size=(800, 500))

for (idx, n) in enumerate(n_vals)
    risk_pn_hat = zeros(length(prob_vals))
    for i in 1:length(prob_vals)
        p = prob_vals[i]
        loss_pn_hat = zeros(M)
        for j in 1:M
            x = rand(Geometric(p), n) .+ 1
            loss_pn_hat[j] = (1/mean(x) .- p).^2
        end
        risk_pn_hat[i] = mean(loss_pn_hat)
    end
    scatter!(prob_vals, risk_pn_hat, color = "red", xlabel = "p",
        ylabel = L"\hat{R}(p_n, p)", label = "", title = "n = $n",
        subplot = idx)
    plot!(prob_vals, prob_vals.^2 .* (1 .- prob_vals) / n,
          color = "blue", lw = 2, label = "", subplot = idx)
end

display(plt)
```

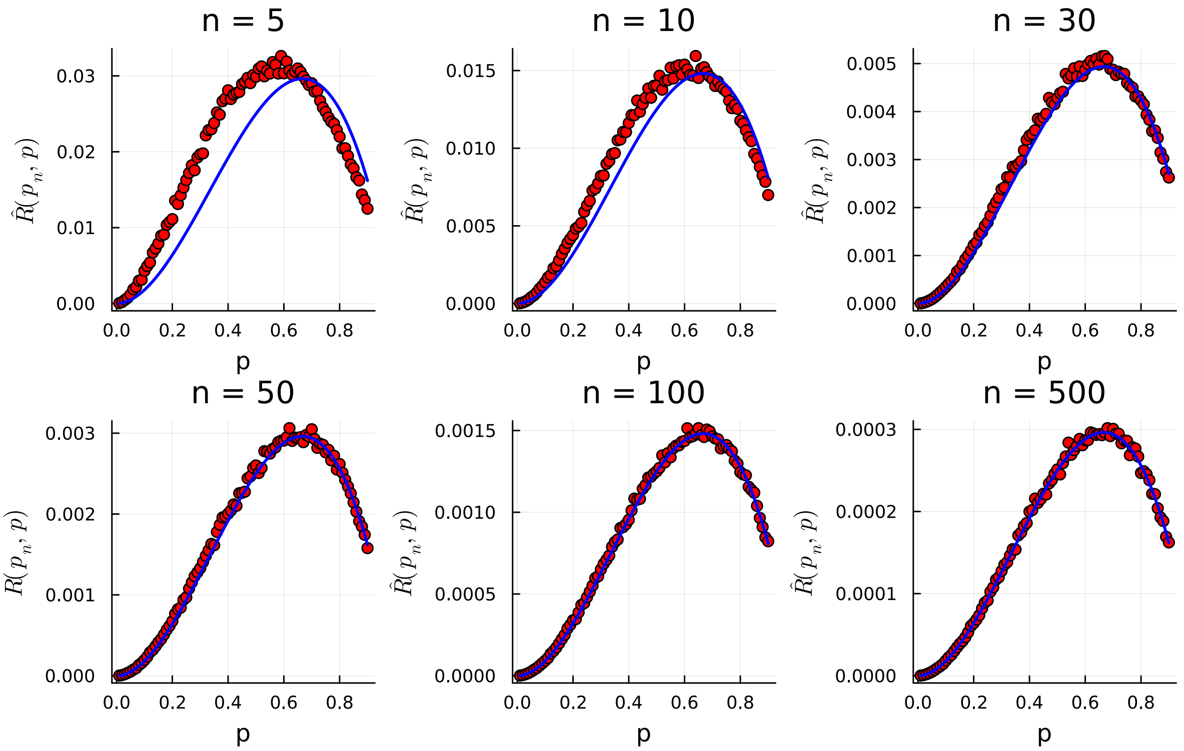


Figure 3: The approximation of the risk function of the MLE of p based on a random sample of size n from the Geometric(p) distribution. The evaluation is carried out based on $M = 1000$ replications under the squared error loss function.

For this problem, it is almost impossible to compute the risk function analytically. Let us try to obtain an approximation of the risk function that performs well for large n. Consider $g(x) = \frac{1}{x}$. Expanding the Taylor's Polynomial of $g(\overline{X}_n)$ about $\frac{1}{p}$ gives

$$\begin{aligned} g(\overline{X}_n) &\approx g\left(\frac{1}{p}\right) + \left(\overline{X}_n - \frac{1}{p}\right) g'\left(\frac{1}{p}\right) \\ &\approx p + \left(\overline{X}_n - \frac{1}{p}\right)(-p^2). \end{aligned}$$

From the Central Limit Theorem, we have

$$\overline{X}_n \approx \mathcal{N}\left(\frac{1}{p}, \frac{1-p}{p^2 n}\right),$$

therefore, for large n,

$$\mathrm{Var}(\overline{X}_n) \approx \frac{1-p}{p^2 n}.$$

Taking expectation on both sides of the Taylor's polynomial, we see that

$$\mathbb{E}\left[g(\overline{X}_n)\right] \approx p.$$

The approximate risk is given by

$$\begin{aligned}\mathcal{R}\left(\frac{1}{\overline{X}_n}, p\right) &= \mathbb{E}\left(\frac{1}{\overline{X}_n} - p\right)^2 \\ &\approx \mathbb{E}\left(\overline{X}_n - \frac{1}{p}\right)^2 (-p^2)^2 \\ &\approx \operatorname{Var}(\overline{X}_n)p^4 \\ &= \frac{(1-p)p^2}{n}.\end{aligned}$$

Therefore

$$\mathcal{R}\left(\frac{1}{\overline{X}_n}, p\right) \approx \frac{(1-p)p^2}{n}, \quad \text{for large } n.$$

In the Fig., the simulated risk function is overlaid with the approximate risk function obtained via Taylor's approximation (first order). It can be noted that for large n, the approximations are remarkably close to the true risk function.
It is interesting to observe that the risk is different at different values of the parameter p. Therefore, if the estimate of the parameter is p^*, then the estimated risk will be

$$\frac{(1-p^*)p^{*2}}{n},$$

which is approximately equal to

$$\frac{0.082}{n} \quad \text{for } p^* = \frac{5}{14} \text{ and } n = 5.$$

However, one may possibly want to get an upper bound of the risk which is independent of the choices of p, which can be obtained by maximizing the risk function with respect to p.

$$\frac{d}{dp}\left[\frac{(1-p)p^2}{n}\right] = \frac{2p-3p^2}{n} = 0$$

gives $p^* = \frac{2}{3}$. Therefore,

$$\max_{p\in(0,1)} \mathcal{R}\left(X_n^{-1}, p\right) = \frac{(1-\frac{2}{3})\left(\frac{2}{3}\right)^2}{n} = \frac{4}{27n}.$$

Therefore, for large n

$$\mathcal{R}\left(X_n^{-1}, p\right) \leq \frac{4}{27n}, \text{ for all } p \in (0,1).$$

One might ask the question that what would be the minimum required sample size so that the maximum risk would be less than some small number $\epsilon = 0.001$ (say). From the inequality, we can obtain $\frac{4}{27n} \leq 0.001$ implies $n \geq \frac{4}{27\times 0.001} = 148.1481$. Therefore, at least a sample size of $n \geq 149$ would be required to ensure the accuracy of the estimate less than 0.001. A general formula can be written as $n \geq \left\lfloor \frac{4}{27\epsilon} \right\rfloor + 1$, for a given accuracy level $\epsilon > 0$, where $[x]$ represents the greatest integer $\leq x$.

A not so common discrete distribution

The discrete uniform distribution on the set $\{1, 2, \ldots, \theta\}$, where $\theta \in \Theta = \{1, 2, 3, \ldots\}$ is given by

$$P(X = x) = \begin{cases} \frac{1}{\theta}, & x \in \{1, 2, \ldots, \theta\} \\ 0, & \text{otherwise.} \end{cases}$$

Suppose, we have a random sample of size n, $(X_1, X_2, \ldots, X_n)$ are available and let $Y_n = \max(X_1, \ldots, X_n)$. We are interested in estimating the parameter θ which can take any value from the countable set $\{1, 2, \ldots\}$. The likelihood function is given by

$$\mathcal{L}(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n, & \theta \in \{y_n, y_n + 1, y_n + 2, \ldots\} \\ 0, & \text{otherwise.} \end{cases}$$

Inaddition, it is easy to understand that the likelihood function attains it's maximum at y_n

In [6]:

```
theta = 6
n = 5
x = sample(1:theta, n, replace=true)
y_n = maximum(x)

function Lik(theta)
    if theta < y_n
        return 0
    end
    if theta >= y_n && theta == floor(theta)  # Ensure theta is an integer
        return float(theta)^(-n)  # Convert to float to allow negative exponent
    end
end

theta_vals = 1:10
Lik_vals = zeros(length(theta_vals))

for i in 1:length(theta_vals)
    Lik_vals[i] = Lik(theta_vals[i])
end

plot(theta_vals, Lik_vals, seriestype=:stem, color=:gray, lw=3,
     xlabel="θ", ylabel="L(θ)", label = "")
scatter!(theta_vals, Lik_vals, color=:blue, markersize=6, label = "")
```

Out[6]:

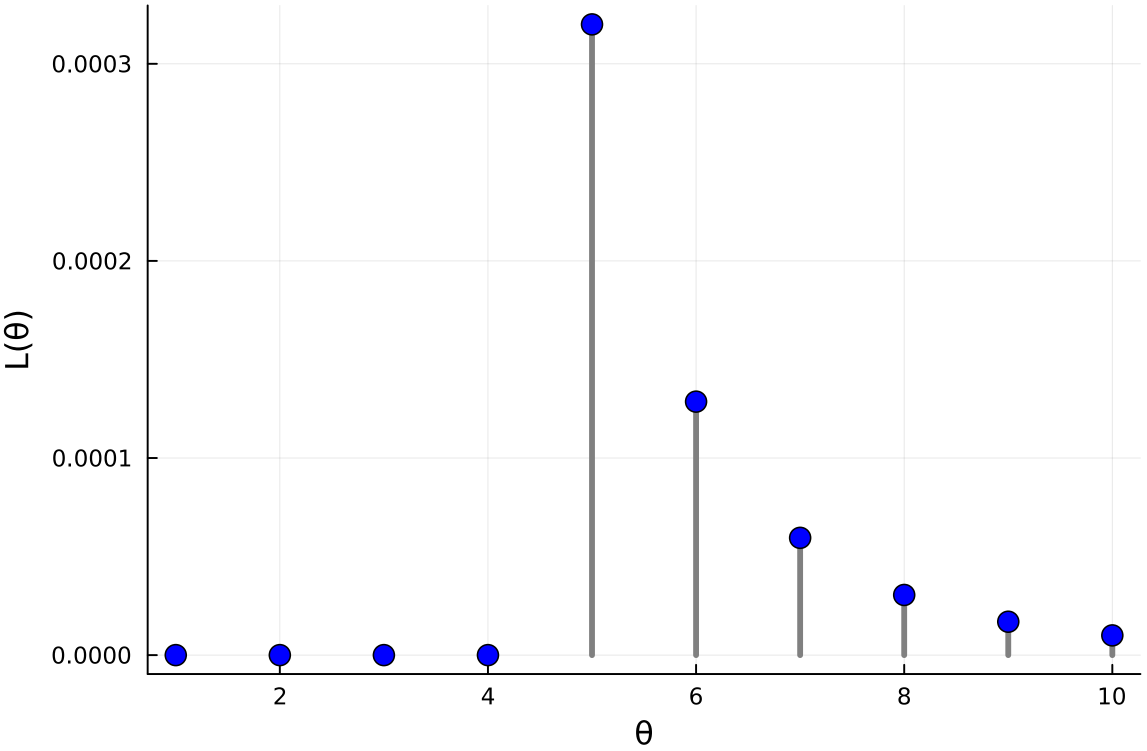


Figure 4: The likelihood function for the parameter θ based on a random sample of size $n = 5$ from the discrete uniform distribution on the set $\{1, 2, 3, \ldots, \theta = 6\}$. The StatBase.sample() function in Julia has been used to simulate the observations from the discrete uniform distribution.

An experiment with the continuous distribution

Suppose that we have a random sample $(X_1, \ldots, X_n)$ of size n from the exponential distribution with rate parameter λ. Let $X \sim \text{Exponential}(\lambda)$ be the population distribution and the parameter space is $\Theta = (0, \infty)$. We are interested in estimating the probability $P(X \leq 1)$.
First, we make the observation that the desired quantity $\psi = P(X \leq 1)$ is a function of λ, which is given by

$$\psi = \int_0^1 \lambda e^{-\lambda x} dx = 1 - e^{-\lambda}.$$

The original parameter $\lambda \in (0, \infty)$, which implies that $\psi \in (0, 1)$. As $\lambda \to \infty$, $\psi \to 1$ and as $\lambda \to 0$, $\psi \to 0$. Our aim is to estimate ψ based on the observations which is a function of λ. Let us understand this problem step by step. First, we need to estimate the parameter λ and then we can use it to approximate ψ.

Computation of MLE of the parameter

Suppose that $(x_1, x_2, \ldots, x_n)$ be the collected sample of size n. The likelihood function is given by

$$\mathcal{L}(\lambda) = f(x_1, x_2, \ldots, x_n | \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}.$$

The function is explicitly written as

$$\mathcal{L}(\lambda) = \begin{cases} \lambda^n e^{-\lambda \sum_{i=1}^n x_i}, & 0 < \lambda < \infty \\ 0, & \text{otherwise.} \end{cases}$$

The log-likelihood function $l(\lambda) = \log \mathcal{L}(\lambda)$ is given by

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

Equating $l'(\lambda) = 0$, we obtain

$$\lambda^* = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}_n},$$

the inverse of the sample mean. It is easy to see that $l''(\lambda) < 0$ for all $\lambda \in (0, \infty)$, therefore, $l''(\lambda^*) < 0$. Thus, the Maximum Likelihood Estimator (MLE) of the parameter λ, based on a sample of size n, is given by

$$\hat{\lambda}_n = (X_n)^{-1}.$$

Here I would like to make an extremely important point which many people miss. It is important to note that in the above discussion, λ^* is a fixed quantity that is computed based on a single realization of the sample $(X_1, \ldots, X_n)$, whereas $\hat{\lambda}_n$ is a random quantity which can be characterized by a probability distribution.

Computation of MLE of the function of parameter

To obtain the estimator of ψ, based on the above sample, a natural choice of the estimator of ψ is given by

$$\hat{\psi}_n = 1 - e^{-\hat{\lambda}_n},$$

where $\hat{\lambda}_n$ is the MLE of λ. The following observations we have:

- The MLE $\hat{\lambda}$ is a function of the sample mean $\bar{X}_n$.
- For the sample mean $\bar{X}_n$, we have large sample normal approximation due to the Central Limit Theorem. The underlying population has finite variance, therefore, CLT holds. It would be interesting to see whether the MLE $\hat{\lambda}_n = \frac{1}{\bar{X}_n}$, a function of $\bar{X}_n$, follows some approximate distribution at least for large n. Also, can this result be extended for $\hat{\psi}_n$, a nonlinear function of the MLE?
- By the WLLN, we know that $\bar{X}_n \overset{P}{\to} \mathbb{E}(X)$. Is it true that $\hat{\lambda}_n \overset{P}{\to} \lambda$ for all $\lambda \in (0, \infty)$? Also, can this result be extended for $\hat{\psi}_n$, a nonlinear function of the MLE?

Convergence of Estimators $\widehat{\lambda}_n$ and $\widehat{\psi}_n$ for Large n

Using the Weak Law of Large Numbers, we know that

$$\bar{X}_n \overset{P}{\to} \mathbb{E}(X) = \frac{1}{\lambda}.$$

Our claim is that

$$\hat{\lambda}_n \overset{P}{\to} \lambda, \quad \hat{\psi}_n \overset{P}{\to} \psi.$$

Before going into the theoretical justifications, let us see by computer simulations how these estimators behave as the sample size n increases. We implement the following algorithm:

- Fix $\lambda \in (0, \infty)$
- Fix $n \in \{1, 2, \dots, 1000\}$
- For each n
    - Simulate $X_1, X_2, \dots, X_n \sim \text{Exponential}(\lambda)$
    - Compute $\bar{X}_n$
    - Compute $\hat{\lambda}_n = \bar{X}_n^{-1}$
    - Compute $\hat{\psi}_n = 1 - e^{-\hat{\lambda}_n}$
- Plot the pairs $(n, \bar{X}_n)$, $n \in \{1, 2, \dots, 1000\}$
- Plot the pairs $\left(n, \hat{\lambda}_n\right)$, $n \in \{1, 2, \dots, 1000\}$
- Plot the pairs $\left(n, \hat{\psi}_n\right)$, $n \in \{1, 2, \dots, 1000\}$

In [7]:

```julia
using Distributions, Statistics, StatsBase, Random
using Plots, LaTeXStrings
```

In [8]:

```julia
lambda = 2
psi = 1-exp(-lambda)

n_vals = 1:1000
sample_means = zeros(length(n_vals))
lambda_hat = zeros(length(n_vals))
psi_hat = zeros(length(n_vals))

for n in n_vals
    x = rand(Exponential(1/lambda), n)
    sample_means[n] = mean(x)
    lambda_hat[n] = 1/sample_means[n]
    psi_hat[n]= 1 .- exp.(-lambda_hat[n])
end

p1 = plot(n_vals, sample_means, color = "red", lw = 2 ,
    xlabel = "sample size (n)", title = L"\bar{X}_n",label = "")
hline!([1/lambda], color = "blue", lw = 2, linestyle = :dash, label = "" )

p2 = plot(n_vals, lambda_hat, color = "red", lw = 2,
    xlabel = "sample size (n)", title = L"\widehat{\lambda}", label = "" )
hline!([lambda], color = "blue", lw = 2, linestyle = :dash, label = "" )

p3 = plot(n_vals, psi_hat, color="red", xlabel="Sample size (n)",
    title=L"\widehat{\psi}_n = 1 - e^{-\lambda_n}", label="")
hline!([psi], color = "blue", lw = 2, linestyle = :dash, label = "" )

plot(p1, p2, p3 , layout = (2,2))
```

Out[8]:

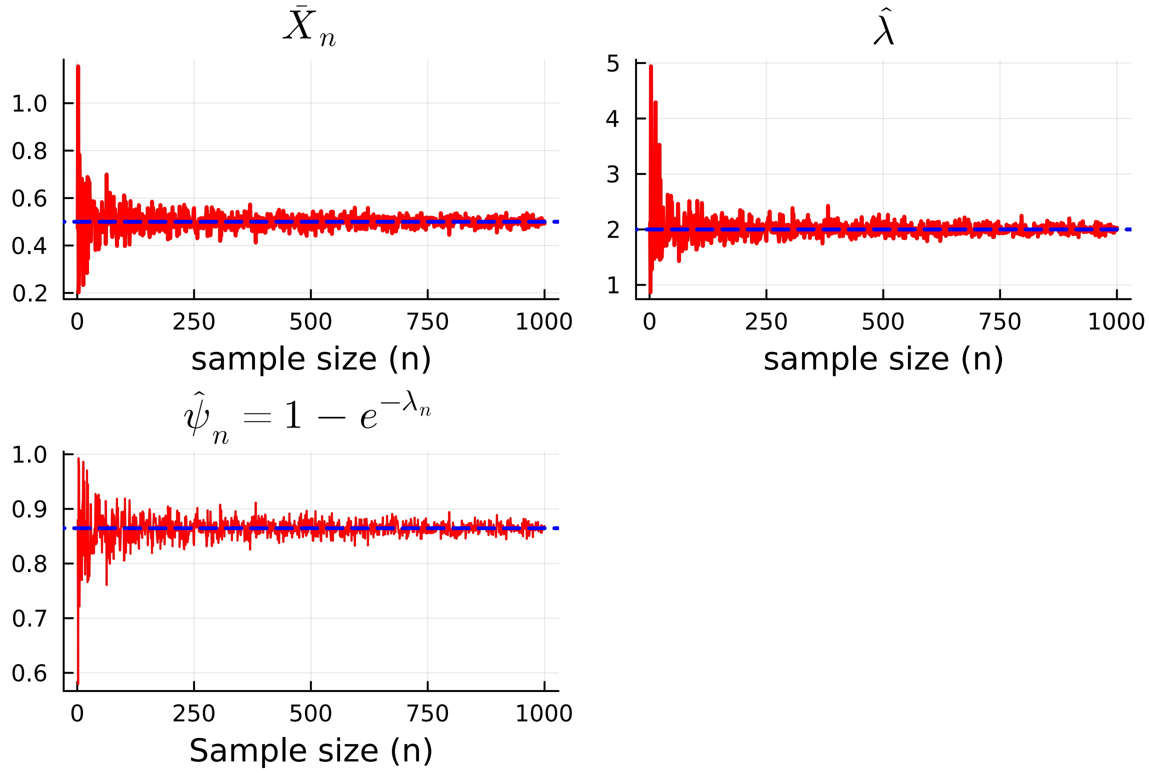


Figure 5: As the sample size n increases, the sample mean converges to the population mean (WLLN) (left panel), the MLE converges to the true value of the parameter (middle panel), and the function of the MLE converges to the function of the parameter (rightmost panel).

The figures clearly suggest that all the estimators converge to the true value as $n \to \infty$. This is due to the fact that both $\widehat{\lambda}_n$ and $\widehat{\psi}_n$ are continuous transformations of $\bar{X}_n$. Let us now look into the large sample approximation of the sampling distributions of $\widehat{\lambda}_n$ and $\widehat{\psi}_n$.

Large Sample Approximations

We recall that as the sample size $n \to \infty$, $\bar{X}_n$ can be well approximated by the normal distribution. More specifically, in our context, for large n,

$$\bar{X}_n \overset{a}{\sim} \mathcal{N}\left(\frac{1}{\lambda}, \frac{1}{\lambda^2 n}\right)$$

as for the given population $\mathbb{E}(X) = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$. Before going into some theoretical justifications, let us visualize the sampling distribution of $\widehat{\lambda}_n$ and $\widehat{\psi}_n$ for different choices (ascending order) of n. The simulation scheme is already demonstrated for the Geometric(p) distribution in the previous section.

In [9]:

```
using Distributions, Statistics, StatsBase, Random
using Plots, LaTeXStrings
```

In [10]:

```
M = 1000  # no. of replications
n = 500   # sample size
lambda = 2 # true rate parameter

sample_means = zeros(M)       # store sample means
lambda_hat = zeros(M)         # store MLEs for lambda
psi_hat = zeros(M)            # store MLEs for psi

for i in 1:M
```

```
    x = rand(Exponential(1/lambda), n)
    sample_means[i] = mean(x)               # compute sample means
    lambda_hat[i] = 1/mean(x)               # compute MLE of lambda
    psi_hat[i] = 1 .- exp.(-lambda_hat[i])  # compute MLE of psi
end

p1 = histogram(sample_means, normalize = true, xlabel = L"\bar{X}_n",
ylabel = "density", title = "n = $n", label = "")
plot!(x->pdf.(Normal(1/lambda, 1/sqrt(lambda^2*n)),x), color = "red",
    lw = 2, label = "")

p2 = histogram(lambda_hat, normalize = true, xlabel = L"\widehat{\lambda}_n",
ylabel = "density", title = "n = $n", label = "")
plot!(x->pdf.(Normal(lambda, sqrt(lambda^2/n)),x), color = "red",
    lw = 2, label = "")

p3 = histogram(psi_hat, normalize = true, xlabel = L"g(\widehat{\lambda}_n)",
ylabel = "density", title = "n = $n", label = "")
scatter!([psi],[0], color = "red", markersize = 8, label = "")
plot!(x->pdf.(Normal(psi, sqrt(lambda^2*exp(-2*lambda)/n)),x), color = "red",
    lw = 2, label = "")

plot(p1, p2, p3, layout = (2,2))
```

Out[10]:

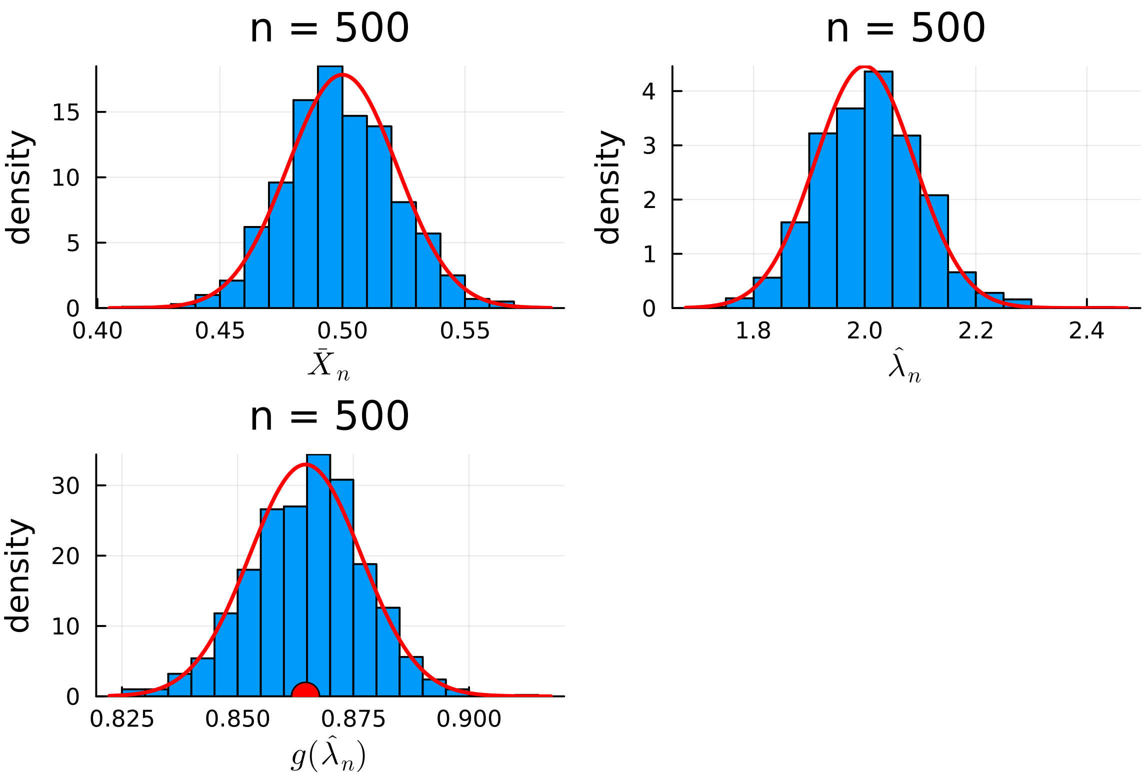


Figure 6: The sampling distribution for X_n, λ_n, and ψ_n are obtained based on $M = 1000$ replications. The value of n is kept at $n = 500$, and it can be observed that the sampling distributions take on a bell-shaped nature. The parameter $\lambda = 2$ is considered for the simulation purpose. Therefore, the true value of $\psi = 1 - e^{-2} = 0.8646647$, which is marked as a red dot in the rightmost panel.

The figure clearly suggests that as the sample size n increases, the sampling distributions of the MLE and the function of the MLE are well approximated by the normal distribution. Therefore, if we can compute the mean and variance of the estimators, at least for large n, we can get the normal approximation of these histograms.

Normal approximation of $\widehat{\lambda}_n$

Consider $\widehat{\lambda}_n = g(\bar{X}_n) = \frac{1}{\bar{X}_n}$. Taking the first-order Taylor expansion about the expected value of $\bar{X}_n = \frac{1}{\lambda}$, we get

$$g(\bar{X}_n) \approx g\left(\frac{1}{\lambda}\right) + \left(\bar{X}_n - \frac{1}{\lambda}\right) g'\left(\frac{1}{\lambda}\right).$$

Since $\mathbb{E}(\bar{X}_n) = \frac{1}{\lambda}$, we get

$$\mathbb{E}\left(\widehat{\lambda}_n\right) = \mathbb{E}\left(\frac{1}{\bar{X}_n}\right) \approx g\left(\frac{1}{\lambda}\right) = \lambda.$$

Therefore, the MLE $\widehat{\lambda}_n$ is an asymptotically unbiased estimator of λ. To compute the variance, we observe that

$$\mathrm{Var}\left(\widehat{\lambda}_n\right) \approx \mathbb{E}\left(\widehat{\lambda}_n - \lambda\right)^2 \approx \mathbb{E}\left(\bar{X}_n - \frac{1}{\lambda}\right)^2 \left(g'\left(\frac{1}{\lambda}\right)\right)^2 = \frac{\lambda^4}{\lambda^2 n} = \frac{\lambda^2}{n}.$$

Therefore, we can overlay a normal distribution on the simulated histograms approximating the sampling distribution of $\widehat{\lambda}_n$ and check whether

$$\widehat{\lambda}_n \overset{a}{\sim} \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right).$$

In the next phase, we can approximate the mean and variance of $\widehat{\psi}_n = g\left(\widehat{\lambda}_n\right)$, (say), where $g(x) = 1 - e^{-x}$. Considering the first-order Taylor approximation of $g\left(\widehat{\lambda}_n\right)$ about the expected value of $\widehat{\lambda}_n$, which is approximately equal to λ for large n, we obtain:

$$g\left(\widehat{\lambda}_n\right) \approx g(\lambda) + \left(\widehat{\lambda}_n - \lambda\right) g'(\lambda).$$

As $n \to \infty$, $\mathbb{E}\left(\widehat{\psi}_n\right) = \mathbb{E}\left[g\left(\widehat{\lambda}_n\right)\right] \approx g(\lambda) = 1 - e^{-\lambda} = \psi$, therefore, $\widehat{\psi}_n$ is an asymptotically unbiased estimator of ψ. The approximate variance of $\widehat{\psi}_n$, for large n, is obtained as

$$\mathrm{Var}\left(\widehat{\psi}_n\right) \approx \mathbb{E}\left(\widehat{\psi}_n - \psi\right)^2 = \mathbb{E}\left[g\left(\widehat{\lambda}_n\right) - g(\lambda)\right]^2 \approx \mathbb{E}\left(\widehat{\lambda}_n - \lambda\right)^2 (g'(\lambda))^2.$$

This gives $\mathrm{Var}\left(\widehat{\psi}_n\right) \approx \frac{\lambda^2 e^{-2\lambda}}{n}$. Therefore, we can claim that

$$\widehat{\psi}_n = 1 - e^{-\widehat{\lambda}_n} \overset{a}{\sim} \mathcal{N}\left(1 - e^{-\lambda}, \frac{\lambda^2 e^{-2\lambda}}{n}\right).$$

In [11]:

```
n_vals = [5, 10, 50, 100, 500, 1000]  # Sample sizes
M = 1000  # Number of replications
lambda = 2  # True lambda value
psi = 1 - exp(-lambda)  # True psi value


plot_layout = (6, 3)
p_list = []

for n in n_vals
```

```
    sample_means = zeros(M)
    lambda_hat = zeros(M)
    psi_hat = zeros(M)

    for i in 1:M
        x = rand(Exponential(1/lambda), n)
        sample_means[i] = mean(x)
        lambda_hat[i] = 1 / mean(x)
        psi_hat[i] = 1 - exp(-lambda_hat[i])
    end

    # Histogram for sample means
    p1 = histogram(sample_means, normalize = true, bins=30, title="n= $n",
                   xlabel=L"\bar{X}_n", legend=false)
    plot!(x -> pdf.(Normal(1/lambda, 1/sqrt(lambda^2 * n)), x),
          color= "red", lw = 2)

    # Histogram for lambda_hat
    p2 = histogram(lambda_hat, normalize= true, bins=30, title="n= $n",
                   xlabel=L"\hat{\lambda}_n", legend = false)
    plot!(x -> pdf.(Normal(lambda, sqrt(lambda^2 / n)), x),
          color= "red", lw = 2)

    # Histogram for psi_hat
    p3 = histogram(psi_hat, normalize = true , bins=30,
        title="n= $n", xlabel=L"g(\hat{\lambda}_n)", legend=false)
    scatter!([psi], [0], color = "red", markersize = 4)
    plot!(x -> pdf.(Normal(psi, sqrt(lambda^2 * exp(-2*lambda) / n)), x),
          color= "red", lw = 2)

    append!(p_list, [p1, p2, p3])
end

plot(p_list..., layout=plot_layout, size=(1200, 1800))
```

Out[11]:

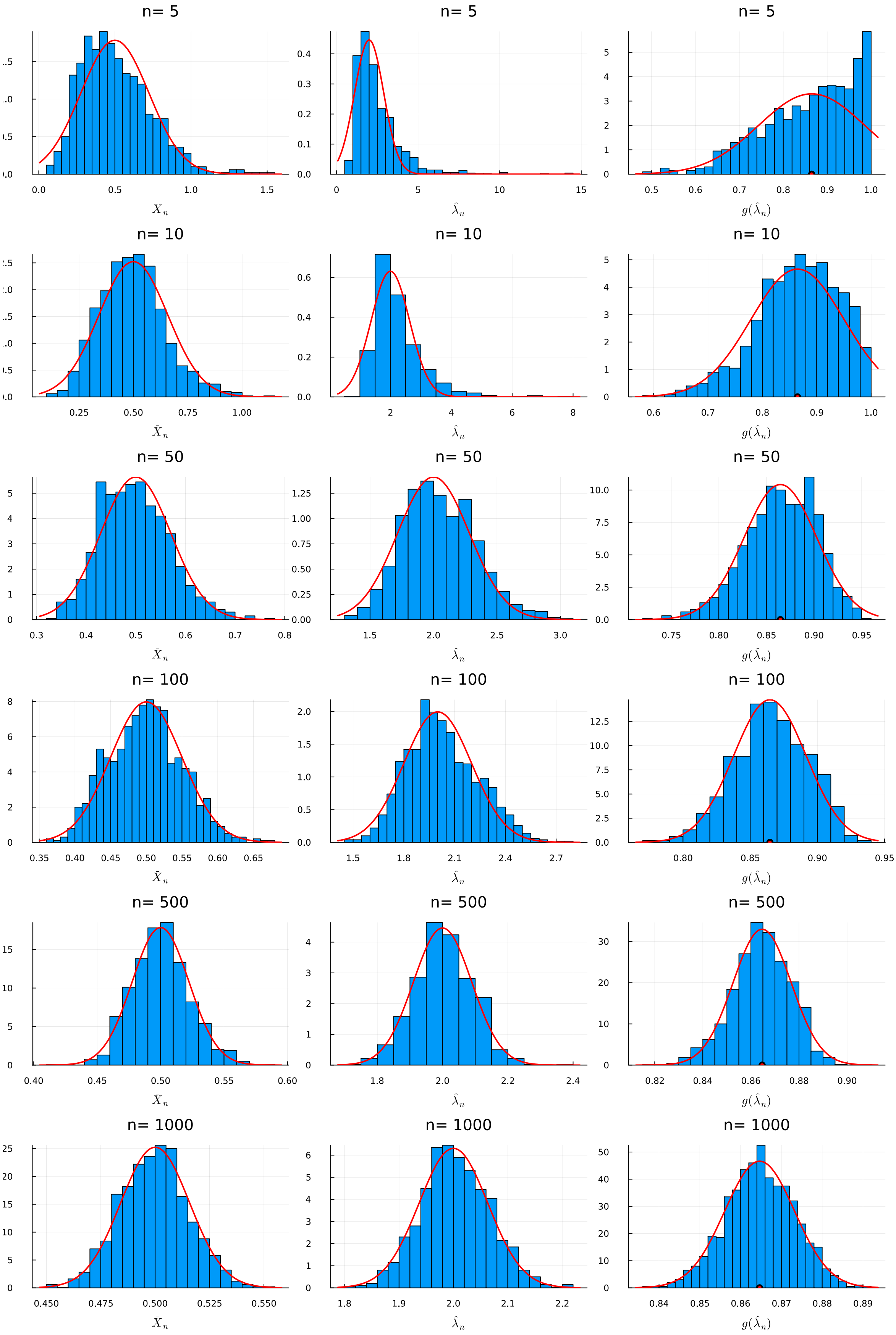


Figure 7: As the sample size n increases, the sampling distributions of the functions of the data points are well approximated by the normal distributions.

Invariance property of MLE

Suppose $X_1, \ldots, X_n$ be a random sample from the population PDF (PMF) $f(x|\theta)$. If $\hat{\theta}_n$ be the MLE of θ, then for any function $g(\theta)$, the MLE of $g(\theta)$ is $g(\hat{\theta}_n)$.

Approximation of Risk Function

In the language of risk function, we can write that for large n,

$$\mathcal{R}\left(\widehat{\lambda_n}, \lambda\right) = \mathbb{E}\left(\widehat{\lambda_n} - \lambda\right)^2 \approx \frac{\lambda^2}{n}.$$

However, for the estimator $\widehat{\psi_n}$, we have computed the risk as a function of λ only. By expressing λ in terms of ψ (it is a monotone transformation from $\lambda \rightarrow \psi$) as $\lambda = -\log(1-\psi)$, we obtain the approximate risk function as

$$\mathcal{R}\left(\widehat{\psi_n}, \psi\right) \approx \frac{\left(-(1-\psi)\log(1-\psi)\right)^2}{n}, \quad \text{for large } n, \quad \psi \in (0,1).$$

The verification of the above expression can be performed by computer simulation. Instead of discretizing the range of λ, discretize the possible parameter space corresponding to $\psi \in (0,1)$.

In [12]:

```julia
using Statistics, StatsBase, Random, Distributions
using Plots, LaTeXStrings
```

In [13]:

```julia
n_vals = [5, 10, 30, 50, 100, 500]
M = 5000
psi_vals = 0.01:0.05:0.9
plt = plot(layout=(2, 3), size=(800, 500))

for (idx, n) in enumerate(n_vals)
    risk_psi_hat = zeros(length(psi_vals))
    for i in 1:length(psi_vals)
        psi = psi_vals[i]                # true probability
        lambda = - log.(1 .- psi)        # rate for exponential
        loss_psi_hat = zeros(M)
        for j in 1:M
            x = rand(Exponential(1/lambda), n)
            psi_hat = 1-exp(-1/mean(x))
            loss_psi_hat[j] =( psi_hat .- psi).^2
        end
        risk_psi_hat[i] = mean(loss_psi_hat)
    end
    scatter!(psi_vals, risk_psi_hat, color = "red", xlabel = L"\psi",
    ylabel = L"R(\widehat{\psi}_n, \psi)", label = "", markersize = 6,
        title = "n = $n", subplot = idx)
    plot!(psi_vals, (-(1 .- psi_vals).*log.(1 .- psi_vals)).^2/n, color = "blue",
    lw = 3, linestyle = :dash,label = "", subplot = idx)
end

display(plt)
```

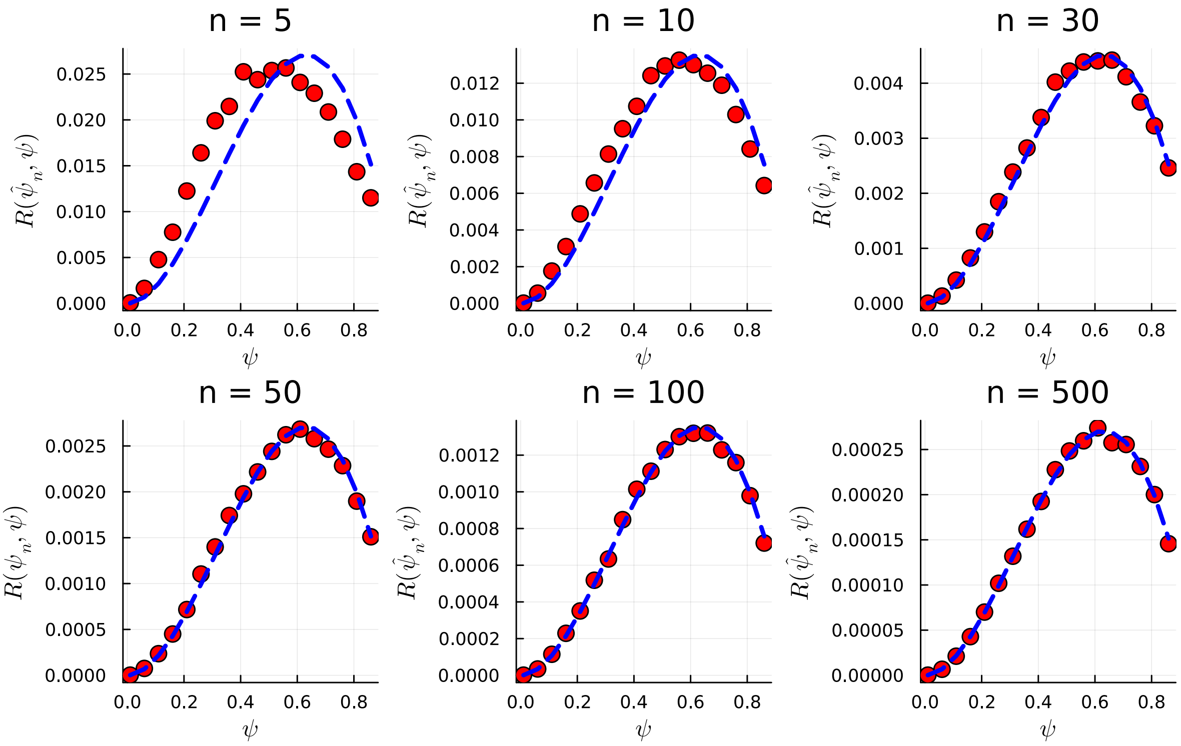


Figure 8: As the sample size n increases, the risk function $\mathcal{R}\left(\widehat{\psi_n}, \psi\right)$ can be well approximated by the approximate risk obtained by the first-order Taylor's polynomial approximation. The approximated risk function based on $M = 1000$ replications is shown in red color, which is the approximation of the true risk function based on simulation. The approximation of the true risk function by the first-order Taylor's polynomial is shown in a blue dotted line. It is evident that as the sample size increases, the approximation is very close.

! Delta Method

Let W_n be a sequence of random variables that satisfies

$$\sqrt{n}(W_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

in distribution. For a given function $g(\cdot)$ and a specific value of $\theta \in \Theta$, suppose that $g'(\theta)$ exists and $g'(\theta) \neq 0$. Then

$$\sqrt{n}\left[g(W_n) - g(\theta)\right] \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\left[g'(\theta)\right]^2\right),$$

in distribution.

The basic idea is that as the sample size increases, the function of asymptotically normally distributed random variables can also be approximated by the normal distribution. The mean and variance can be easily computed using the first-order Taylor's polynomial approximation.

! Comparing Risk Functions

Suppose in the same problem of estimating $\psi = P(X \geq 1)$, another alternative estimator

$$\hat{\xi}_n = \frac{1}{n}\sum_{i=1}^{n} I(X_i \geq 1)$$

is suggested, where the quantity $I(X \geq 1)$ is defined as

$$I(X \geq 1) = \begin{cases} 1, & X \geq 1 \\ 0, & X < 1 \end{cases}$$

Therefore, $\hat{\xi}_n$ basically represents the proportion of observations that are greater than or equal to 1. It is easier to check that $\mathbb{E}(\hat{\xi}_n) = \psi$ and $\mathrm{Var}(\hat{\xi}_n) = \frac{\psi(1-\psi)}{n}$. Using computer simulation, simulate the risk function $\mathcal{R}(\hat{\xi}_n, \psi)$ and compare it with $\mathcal{R}(\widehat{\psi}_n, \psi)$ at different values of $\psi \in (0,1)$ for different sample sizes n. Write your conclusion in simple English.

Multi-parameter optimization

In the previous sections, we have learnt that how one can obtain the sampling distribution of the MLE and also tested whether the MLE converges in probability to the true parameter value by using computer simulations. Many real-life problems are modeled by probability distributions which are parameterized by more than one parameter, for example, $\mathcal{N}(\mu, \sigma^2)$, $\mathcal{B}(a,b)$, $\mathcal{G}(a,b)$, etc. In this section, we consider the computation of the MLE for multiparameter probability distributions.

Suppose that we have a random sample $(X_1, X_2, \ldots, X_n)$ of size n from the normal distribution with parameters μ and σ^2. Our goal is to obtain the Maximum Likelihood Estimators of μ and σ^2 based on the sample observations and study properties of these estimators. The parameter space is

$$\Theta = \left\{(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma < \infty\right\}$$

In [14]:

```
n = 30
mu = 3
sigma2 = 1.5
x = rand(Normal(mu,sqrt(sigma2)),n)
histogram(x, normalize = true, xlabel = "x", ylabel = "density",
    title = "histogram of x", label = "")
```

Out[14]:

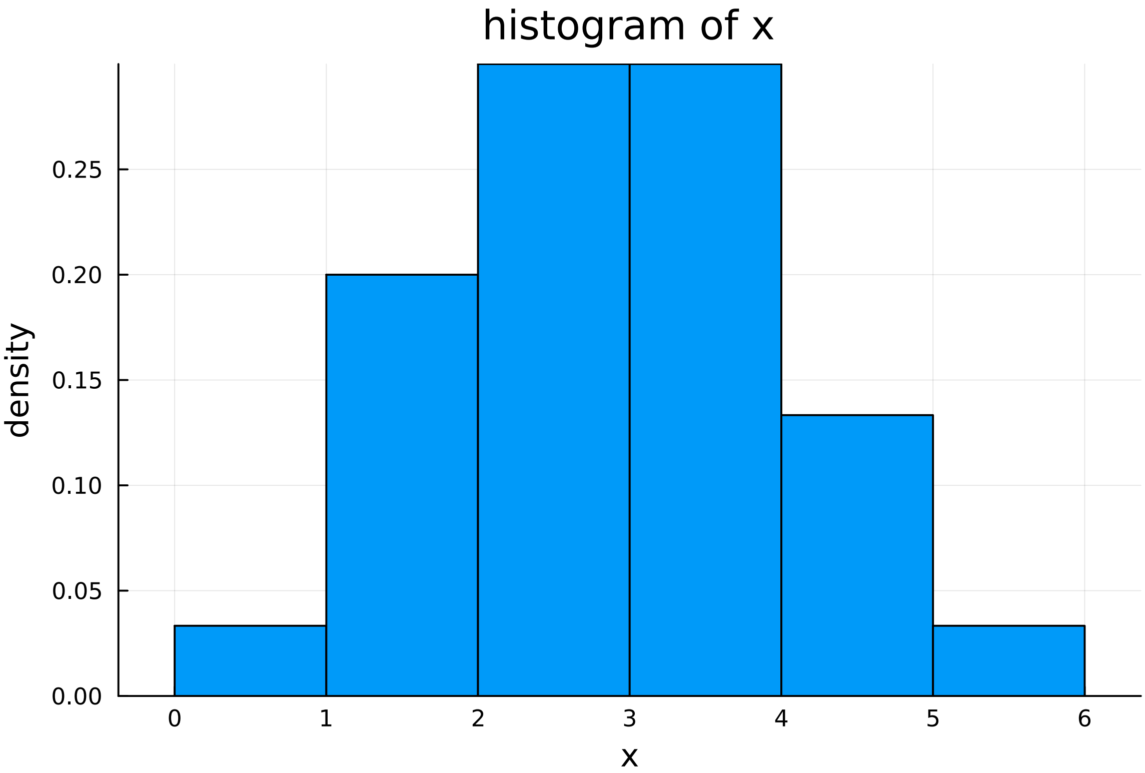


Figure 9: A random sample of size $n = 30$ has been simulated from the normal distribution with mean $\mu = 3$ and variance $\sigma^2 = 1.5$ for demonstration. The corresponding histogram is used for graphical display of the data.

Likelihood Function

If $(x_1, x_2, \ldots, x_n)$ be a fixed observed sample of size n, then the likelihood function is given by

$$\mathcal{L}(\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}, \quad -\infty < \mu < \infty, \quad 0 < \sigma < \infty.$$

The log-likelihood function is given by

$$l(\mu, \sigma^2) = \log \mathcal{L}(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Compute $\frac{\partial l}{\partial \mu}$ and $\frac{\partial l}{\partial \sigma^2}$ and set them equal to zero.Solve the above system simultaneously and obtain the critical points μ^* and σ^{2*}. Show that the following matrix

$$H = \begin{bmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l}{\partial \mu \partial \sigma^2} & \frac{\partial^2 l}{\partial (\sigma^2)^2} \end{bmatrix}$$

is negative definite when the partial derivatives are evaluated at (μ^*, σ^{2*}). This can be checked by ensuring the following two conditions:

$$H_{11} < 0, \quad \text{and} \quad \det(H) = H_{11}H_{22} - (H_{12})^2 > 0.$$

In the following, we plot the likelihood function and the log-likelihood function as a function of the parameters μ and σ^2.

In [15]:

```julia
using Statistics, StatsBase, Random, Distributions
using Plots, LaTeXStrings
```

In [16]:

```julia
using Distributions, Plots

n = 30
mu = 3
sigma2 = 1.5

x = rand(Normal(mu, sqrt(sigma2)), n)

histogram(x, normalize=true, xlabel="x", ylabel="Density", label="Histogram")

function Likelihood(mu, sigma2)
    return (1 / (sigma2 * 2 * pi)^(n / 2)) * exp(-sum((x .- mu).^2) / (2 * sigma2))
end

function LogLikelihood(mu, sigma2)
    return log(Likelihood(mu, sigma2) + 0.5)
end

mu_vals = collect(2:0.1:4)
sigma2_vals = collect(1:0.1:1.5)

Lik_vals =  Matrix{Float64}(undef, length(mu_vals), length(sigma2_vals))
LogLik_vals =  Matrix{Float64}(undef, length(mu_vals), length(sigma2_vals))

for i in 1:length(mu_vals)
    for j in 1:length(sigma2_vals)
        Lik_vals[i, j] = Likelihood(mu_vals[i], sigma2_vals[j])
        LogLik_vals[i, j] = LogLikelihood(mu_vals[i], sigma2_vals[j])
    end
end

zticks = collect(extrema(Lik_vals))

plot1 = surface(mu_vals, sigma2_vals, Lik_vals, xlabel="μ",
    ylabel="σ²", zlabel="L(μ,σ²)",
    title="Likelihood Surface", c=:viridis,
    xticks=2:0.5:4, yticks=1:0.1:1.5, zticks=zticks)

plot2 = surface(mu_vals, sigma2_vals, LogLik_vals,
    xlabel="μ", ylabel="σ²", zlabel="log L(μ,σ²)",
    title="Log-Likelihood Surface", c=:viridis,
    xticks=2:0.5:4, yticks=1:0.1:1.5, zticks=zticks)

plot(plot1, plot2, layout = (1,2))
```

Out[16]:

From the computation, we observed that the MLE of the μ and σ^2 are given by

$$\hat{\mu}_n = \bar{X}_n, \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Given an estimators, the following questions immediately appear, such as (a) are these estimators unbiased? (b) are these estimators consistent, (c) are these estimators are the best choices under some given loss function? We shall not address all these questions here only,

but certainly show how we basic properties of these estimators can be checked and when feasible, we shall also compute the exact sampling distribution of these estimators.
At the first step, let us check whether as the sample size increases, the MLEs converge to the true parameter values. Mathematically, whether the following statements true:

$$\bar{X}_n \xrightarrow{P} \mu, \quad \text{and} \quad \hat{\sigma}^2_n \xrightarrow{P} \sigma^2.$$

The following simulation will help us to understand whether the above statements are true.

- Fix sample size $n \in \{2, 3, \ldots, 1000\}$.
- Fix $\mu = \mu_0$ and $\sigma^2 = \sigma^2_0$.
- For each n
  - Simulate $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu_0, \sigma^2_0)$.
  - Compute $\bar{X}_n$ and $\hat{\sigma}^2_n$.
- Plot the pairs $\left(n, \bar{X}_n\right)$ and $\left(n, \hat{\sigma}^2_n\right)$.
- Add a horizontal straight line to the plots at μ_0 and σ^2_0, respectively.

In [17]:

```
n_vals = 2:1000
mu_hat = zeros(length(n_vals))
sigma2_hat = zeros(length(n_vals))

for i in 1:length(n_vals)
    n = n_vals[i]
    x = rand(Normal(mu, sqrt(sigma2)), n)
    mu_hat[i] = mean(x)
    sigma2_hat[i] = (1/n)*sum((x .- mean(x)).^2)
end

p1 = plot(n_vals, mu_hat, color = "grey", lw = 2, xlabel = "sample size (n)",
ylabel = L"\widehat{\mu}_n", label = "")
hline!([mu], color = "blue", lw = 2, linestyle = :dash, label = "")

p2 = plot(n_vals, sigma2_hat, color = "grey", lw = 2, xlabel = "sample size (n)",
ylabel = L"\widehat{{\sigma}^2}_n", label = "")
hline!([sigma2], color = "blue", lw = 2, linestyle = :dash, label = "")

plot(p1, p2, layout = (1,2), size = (900, 500))
```

Out[17]:

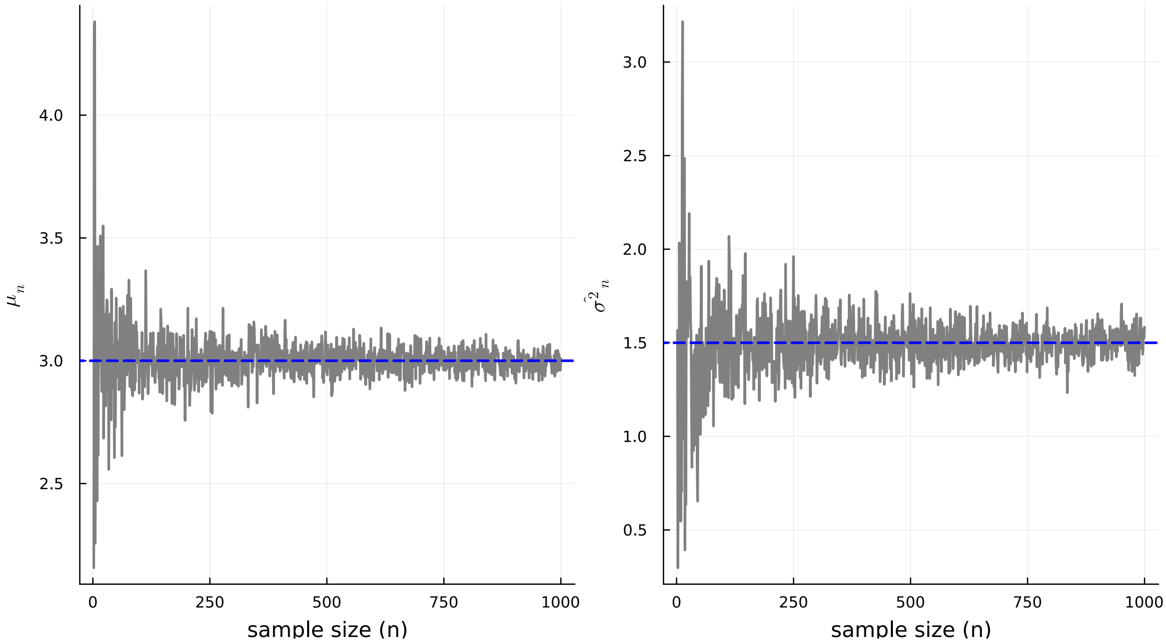


Figure 10: The simulations clearly indicate that as the sample size increases, MLEs are getting highly concentrated around the true values, which is a signature of the convergence in probability. The simulation has been carried out by fixing the population distribution as $\mathcal{N}(\mu = 3, \sigma^2 = 1.5)$.

To quantify the uncertainty associated with the MLE, we can either obtain the exact sampling distributions of $\overline{X}_n$ and $\widehat{\sigma}^2_n$, or we can approximate the standard error of these estimators at different true values by computer simulations by fixing n. Let us fix the value of n and simulate $M = 1000$ realizations from the sampling distributions of the MLEs and visualize their distribution through histograms.

In [18]:

```
using SpecialFunctions # for gamma function
```

In [19]:

```
n = 10
M = 1000
mu_hat = zeros(M)
sigma2_hat = zeros(M)

for i in 1:M
    x = rand(Normal(mu, sqrt(sigma2)), n)
    mu_hat[i] = mean(x)
    sigma2_hat[i] = (1/n)*sum((x .- mean(x)).^2)
end

p1 = histogram(mu_hat, normalize = true, xlabel = L"\widehat{\mu}_n",
ylabel = "density", title = "n = $n" ,label = "")
plot!(x -> pdf.(Normal(mu, sqrt(sigma2/n)), x), color = "red", lw = 2,
label = "")

using SpecialFunctions  # For gamma function

function dist_sigma2_hat(x, n, sigma2)
    num = exp.(-n .* x / (2 * sigma2)) .*
              (n .* x / sigma2).^((n - 1) / 2 - 1) .* n .* (x .> 0)
    den = gamma((n - 1) / 2) * 2^((n - 1) / 2) * sigma2
    return num ./ den
end

p2 = histogram(sigma2_hat, normalize = true, xlabel = L"\widehat{{\sigma}^2}_n",
ylabel = "density", title = "n = $n",label = "")
plot!(x->dist_sigma2_hat(x, n, sigma2), color = "red", lw = 2, label = "")

plot(p1, p2, layout = (1,2), size = (900, 500))
```

Out[19]:

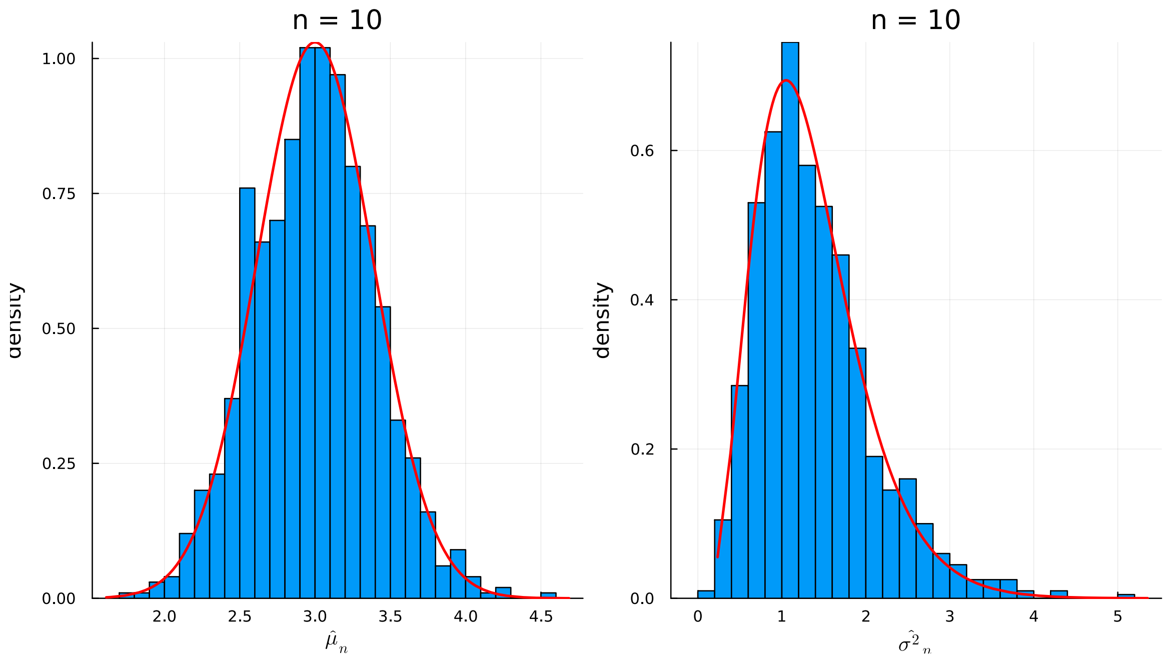


Figure 11: The sampling distribution of the MLEs of μ and σ^2 have been visualized through histograms based on $M = 1000$ replications. The reader is encouraged to update the code and visualize the histograms for different choices of n. The exact sampling distributions are overlaid on the histograms.

The sample mean follows a normal distribution due to the Central Limit Theorem when the sample size (n) is large. However, if the data is drawn from a normal distribution, the sample mean has an exact normal distribution with mean μ and variance $\frac{\sigma^2}{n}$ for any n, not just for large n. This can be proven using the moment-generating function (MGF) technique, as shown below:

$$M_{\bar{X}_n}(t) = e^{\mu t + \frac{1}{2}\left(\frac{\sigma^2}{n}\right)t^2}, \quad -\infty < t < \infty.$$

For the sample mean, applying the MGF technique is straightforward. However, deriving the maximum likelihood estimator (MLE) for σ^2 requires a more detailed calculation. The following subsection presents key results related to sampling from a normally distributed population.

! Sampling from the normal distribution

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the $\mathcal{N}(\mu, \sigma^2)$ population distribution. Then:

- $\bar{X}_n$ and S_n^2 are independent random variables.
- $\bar{X}_n$ follows $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
- $\frac{(n-1)S_n^2}{\sigma^2}$ follows a χ^2_{n-1} distribution with $n-1$ degrees of freedom.

For detailed proofs, different approaches have been explored. One approach uses a transformation formula to derive the joint distribution of $\bar{X}_n$ and S_n^2, while another elegantly applies the moment-generating function technique.

Considering

$$U_n = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1},$$

the sampling distribution of the MLE $V_n = \hat{\sigma}_n^2 = \frac{\sigma^2}{n} \times U_n$ can be easily obtained. The CDF of V_n is obtained as

$$F_{V_n}(v) = P(V_n \leq v) = P\left(U_n \leq \frac{nv}{\sigma^2}\right).$$

Therefore,

$$f_{V_n}(v) = \frac{d}{dv}F_{V_n}(v) = f_{U_n}\left(\frac{nv}{\sigma^2}\right) \cdot \frac{n}{\sigma^2}, \quad 0 < v < \infty.$$

For every n, the sampling distribution of the MLE of σ^2, $\hat{\sigma}_n^2$ is given by:

$$f_{\hat{\sigma}_n^2}(v) = \frac{e^{-\frac{nv}{2\sigma^2}}\left(\frac{nv}{\sigma^2}\right)^{\frac{n-1}{2}-1}}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}} \cdot \frac{n}{\sigma^2}, \quad 0 < v < \infty.$$

In [20]:

```
mu = 3
sigma2 = 1.5
n_vals = [3,5,10,50,100, 200]
plt = plot(layout=(2, 3), size=(800, 500))
M = 1000

for (idx, n) in enumerate(n_vals)
    mu_hat = zeros(M)
    sigma2_hat = zeros(M)
    for i in 1:M
        x = rand(Normal(mu, sqrt(sigma2)), n)
        mu_hat[i] = mean(x)
        sigma2_hat[i] = (1/n)*sum((x .- mean(x)).^2)
    end
    histogram!(sigma2_hat, normalize = true, xlabel = L"\widehat{{\sigma}^2}_n",
    ylabel = "density", title = "n = $n", label = "", subplot = idx)
    plot!(x->dist_sigma2_hat(x, n, sigma2), color = "red", lw = 2, label = "",
    subplot = idx)
    scatter!([sigma2],[0], color = "red", markersize = 14, label = "",
    subplot = idx)
    scatter!([mean(sigma2_hat)],[0], color = "blue", markersize = 9,
        label = "", subplot = idx )
end

display(plt)
```

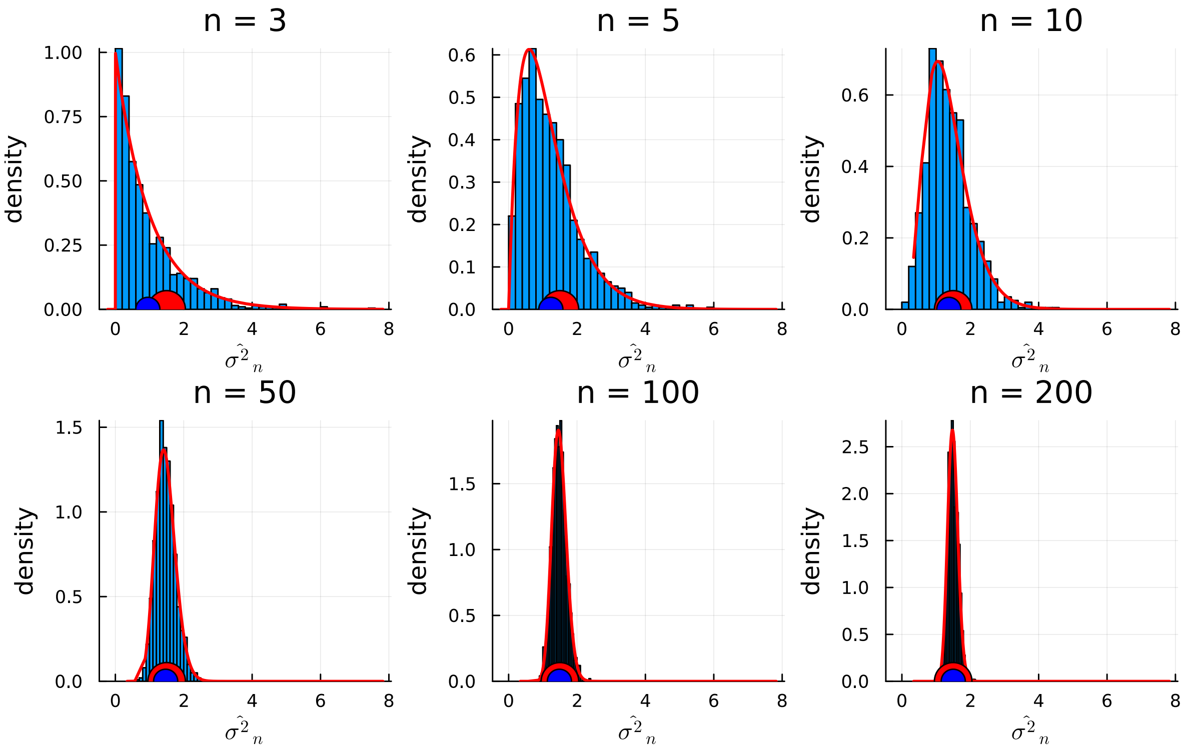


Figure 12: The sampling distribution of the MLE of σ^2 is obtained through simulation for different sample sizes. The exact sampling distribution is shown using the red curve. The red dot indicates the true value of σ^2, whereas the blue circle indicates $\frac{1}{M}\sum_{m=1}^{M}\widehat{\sigma^2_n}^{(m)}$, which is the approximate expectation of $\widehat{\sigma^2_n}$ based on M replications. As the sample size increases, the circle engulfs the red dot, which is a signature of asymptotic unbiasedness. In other words, it ensures that $\mathrm{Bias}(\widehat{\sigma^2_n}, \sigma^2) \to 0$ as $n \to \infty$.

By performing the above exercises for different choices of n, we observed that this sampling distribution is valid for every n. However, as n increases, the shapes of the sampling distributions eventually tend to the normal distribution. We need to give some reasoning for this which is intuitively appealing.

Let us explore the connections between $U_n = \frac{n-1}{n}S^2_n$ and $V_n = \widehat{\sigma^2_n} = \frac{\sigma^2}{n} \times U_n$ and their approximation by the normal distribution for large n. Suppose that we are given the fact that for every n, $U_n \sim \chi^2_{n-1}$ distribution. The PDF of V_n has been obtained by using the transformation formula. The shapes of the PDF of V_n in fact show a normally distributed pattern.

behavior for large n values. Since the chi-squared distribution belongs to the $\mathcal{G}(\cdot,\cdot)$ family, V_n is a constant multiple of a $\mathcal{G}\left(\alpha = \frac{n-1}{2}, \beta = 2\right) \equiv \chi^2_{n-1}$.

Therefore, if we can show that for large n, χ^2_n distribution can be well approximated by the normal distribution, we are done. In the following, we show that for large n, χ^2_n PDF is well approximated by the normal distribution with mean n and variance $2n$. We investigate the MGF of $\frac{X-n}{\sqrt{2n}}$ and see that it is approximately $e^{\frac{t^2}{3}}$ for large n.

$$\log\left[M_{\frac{X-n}{\sqrt{2n}}}(t)\right] = -\sqrt{\frac{n}{2}}t + \frac{n}{2}\log\left(1 - \frac{t}{\sqrt{\frac{n}{2}}}\right) = \frac{t^2}{2} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right).$$

Therefore,

$$\frac{X-n}{\sqrt{2n}} \overset{a}{\sim} \mathcal{N}(0,1) \implies X \overset{a}{\sim} \mathcal{N}(n,2n).$$

Therefore,
$U_n \sim \mathcal{N}(n-1, 2(n-1))$ for large n. Hence,

$$V_n = \widehat{\sigma_n^2} = \frac{\sigma^2}{n} \times U_n \overset{n\to\infty}{\sim} \mathcal{N}\left(\frac{(n-1)\sigma^2}{n}, \frac{2(n-1)\sigma^4}{n^2}\right).$$

In the following, Julia Codes, the exact sampling distribution of $\widehat{\sigma_n^2}$ and the normal approximation is shown graphically, and the approximations are remarkably accurate for large n values.

In [21]:

```
n_vals = [3, 5, 10, 30, 50, 100]
plt = plot(layout=(2, 3), size=(800, 500))

for (idx, n) in enumerate(n_vals)
    lower_lim = sigma2*(n-1)/n - 5*sqrt(sigma2^2 * 2 * (n-1) / n^2)
    upper_lim = sigma2*(n-1)/n + 5*sqrt(sigma2^2 * 2 * (n-1) / n^2)

    plot!(x -> dist_sigma2_hat(x,n, sigma2), color="red", lw=2, label="",
        xlims=(lower_lim, upper_lim), ylab="f(x)", title = "n = $n",
        subplot=idx)
    plot!(x -> pdf(Normal(sigma2*(n-1)/n, sqrt(sigma2^2 * 2 * (n-1) / n^2)), x),
        color="blue", lw=2, linestyle=:dash, label="", subplot=idx)
end

display(plt)
```

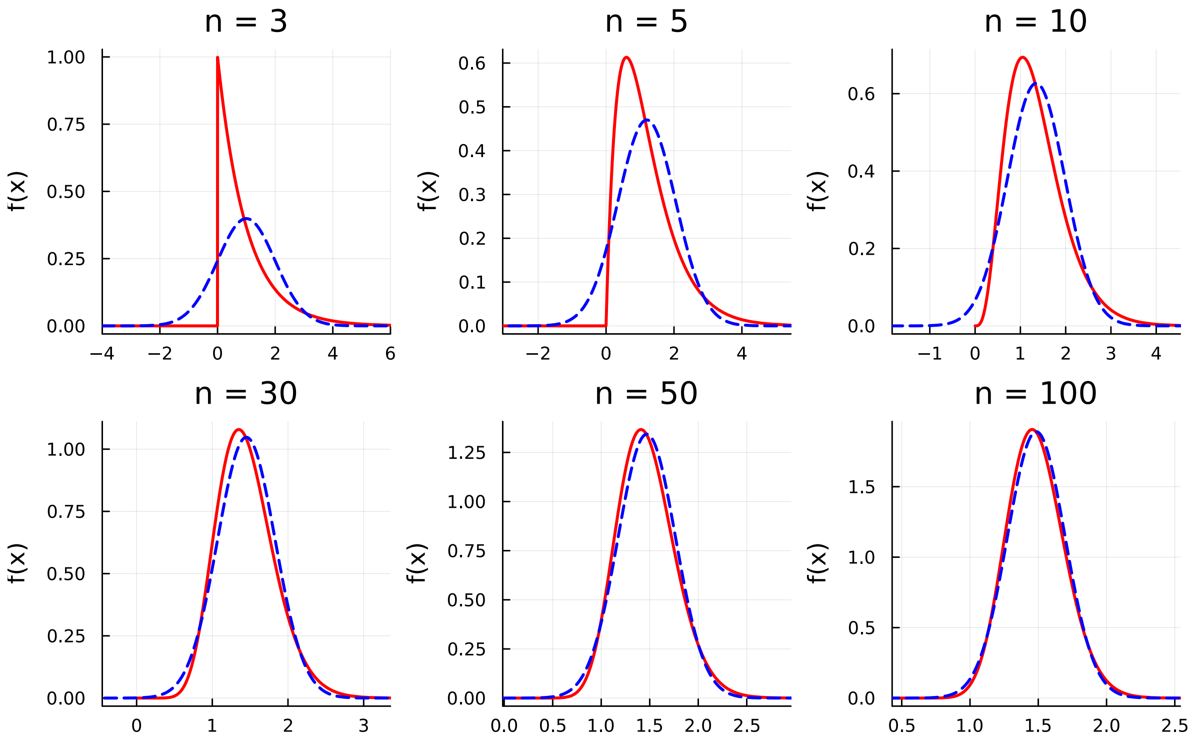


Figure 13: The exact sampling distribution of the MLE of σ_n^2 is shown for different choices of n (red curve). The normal approximation with mean $\frac{(n-1)\sigma^2}{n}$ variance $\frac{2(n-1)\sigma^4}{n^2}$ is shown using the blue dotted line. It is evident that as the sample size increases, the exact sampling distribution can be well approximated by the normal distribution.

Connection between the Hessian and Fisher Information

In the examples discussed above, the MLEs have always appeared to be approximately normal at least for large n and centered about the true value of the parameter. However, we did not explore yet the computation of the variance of the MLEs. In fact, some analytical computation for the variance can be carried out and established as well by computer simulation. We consider some definition first

❗ **Score Function and Fisher Information**

If $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the PDF (PMF) $f(x|\theta)$, where $\theta \in \Theta$,
the **Score function** is defined as:

$$S(\mathbf{X}; \theta) = \frac{\partial \log f(\mathbf{X}|\theta)}{\partial \theta}.$$

It can be easily shown that:

$$\mathbb{E}_\theta[S(\mathbf{X}; \theta)] = 0.$$

📌 **Fisher Information**

The **Fisher Information** is defined as:

$$I_n(\theta) = \mathrm{Var}_\theta\left(\sum_{i=1}^{n} S(X_i; \theta)\right) = \sum_{i=1}^{n} \mathrm{Var}_\theta(S(X_i; \theta)) = \sum_{i=1}^{n} \mathbb{E}_\theta(S(X_i; \theta))^2.$$

It can also be expressed as:

$$I_n(\theta) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log f(\mathbf{X}|\theta)}{\partial \theta^2}\right) = -\int_{\mathbb{R}^n}\left(\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}\right) f(x)\, dx.$$

A natural question arises: how to interpret the expectation of the score function? Here is an interpretation in terms of simulation:

- Fix parameter $\theta = \theta_0$
- Fix the sample size n
- Fix the number of replications M
- For each $m \in \{1, 2, \ldots, M\}$:
  - Simulate $\mathbf{X} = (X_1, \ldots, X_n) \sim f(x|\theta_0)$.
  - Compute $S(\mathbf{X}; \theta_0) = S_m$ (say).
- As $M \to \infty$,

$$\frac{1}{M}\sum_{m=1}^{M} S_m \approx \mathbb{E}\left(S(\mathbf{X}; \theta_0)\right) = 0.$$

Visualization of the Score function

In the following, we visualize the score function for the Poisson distribution. We simulate multiple sets of random samples of size $n = 10$ from the $\mathrm{Poisson}(\lambda_0)$ distribution and plot the score function as a function of λ.

In [22]:

```julia
using Statistics, StatsBase, Distributions, Random
using Plots, LaTeXStrings
```

In [23]:

```julia
using Random, Distributions, Plots

lambda_0 = 3   # true value
n = 10         # sample size
x = rand(Poisson(lambda_0), n)

function score_poisson(lambda)
     -n + sum(x) / lambda
end

lambda_vals = collect(1:0.01:5)
score_vals = zeros(length(lambda_vals))

for i in eachindex(lambda_vals)
    score_vals[i] = score_poisson(lambda_vals[i])
end

plot(lambda_vals, score_vals, color = "grey", linewidth=2,
     xlabel=L"\lambda", ylabel=L"S(\mathbf{X}, \lambda)", label = "")
scatter!([lambda_0],[0], color = "red", markersize = 6, label = "")
hline!([0], color = "blue", lw = 2, linestyle = :dash, label = "")
scatter!([mean(x)],[0], color = "blue", markersize = 10, label = "")
```

Out[23]:

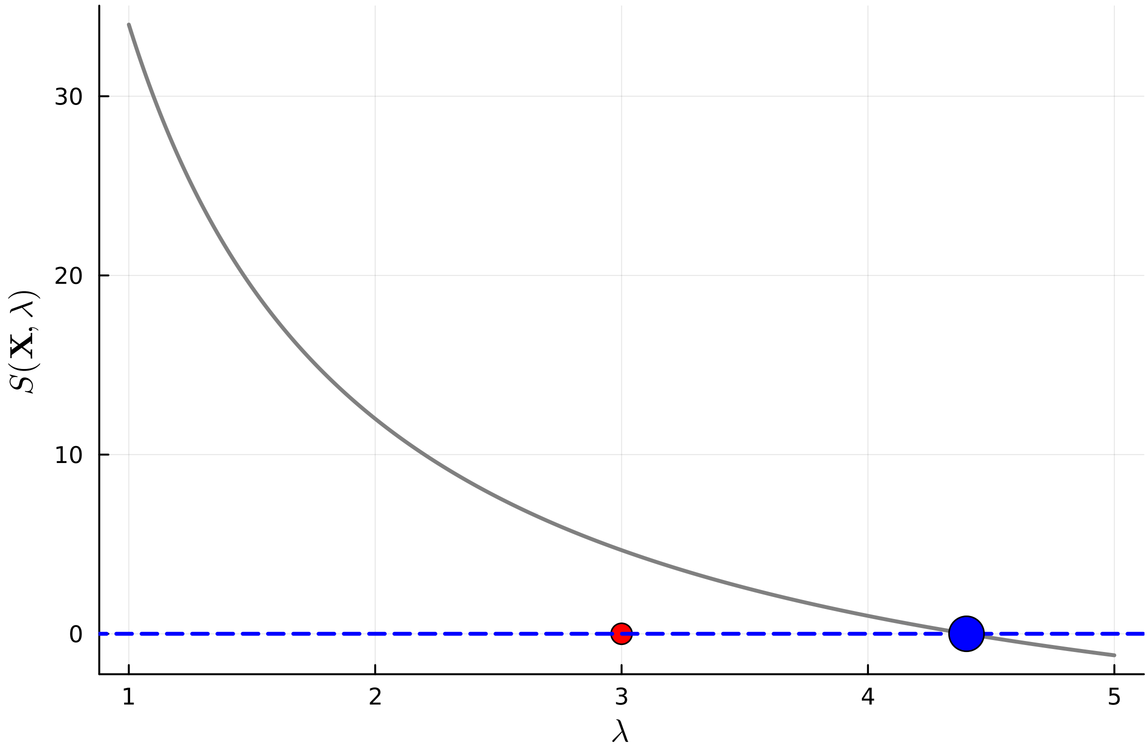


Figure 14: The shape of the score function is shown for a simulated sample of size $n = 10$ from the Poisson($\lambda_0 = 3$) distribution. The blue dot represents the MLE of λ and the red dot is the true value.

Let us repeat the above process and plot the score functions in a single plot

```
In [24]: using Plots, LaTeXStrings, Random, Distributions

         n = 10
         M = 50
         lambda_0 = 3

         for i in 1:M
             data = rand(Poisson(lambda_0), n)
             if i == 1
                 plot(x -> -n + sum(data) / x, 1, 5, color = "grey", lw = 2,
                     xlabel=L"\lambda", ylabel=L"S(\mathbf{X}, \lambda)", label = "")
             else
                 plot!(x -> -n + sum(data) / x, 1, 5, color= "grey",label = "")
             end
             scatter!([mean(data)], [0], color= "blue", markersize = 5, label = "")
         end

         scatter!([lambda_0], [0], color= "red", markersize=7, label= "")
         hline!([0], color = "blue", linestyle=:dash, lw =2, label = "")
         vline!([lambda_0], color = "magenta", linestyle=:dash, lw = 3,
             label = "")

         display(plot!())
```

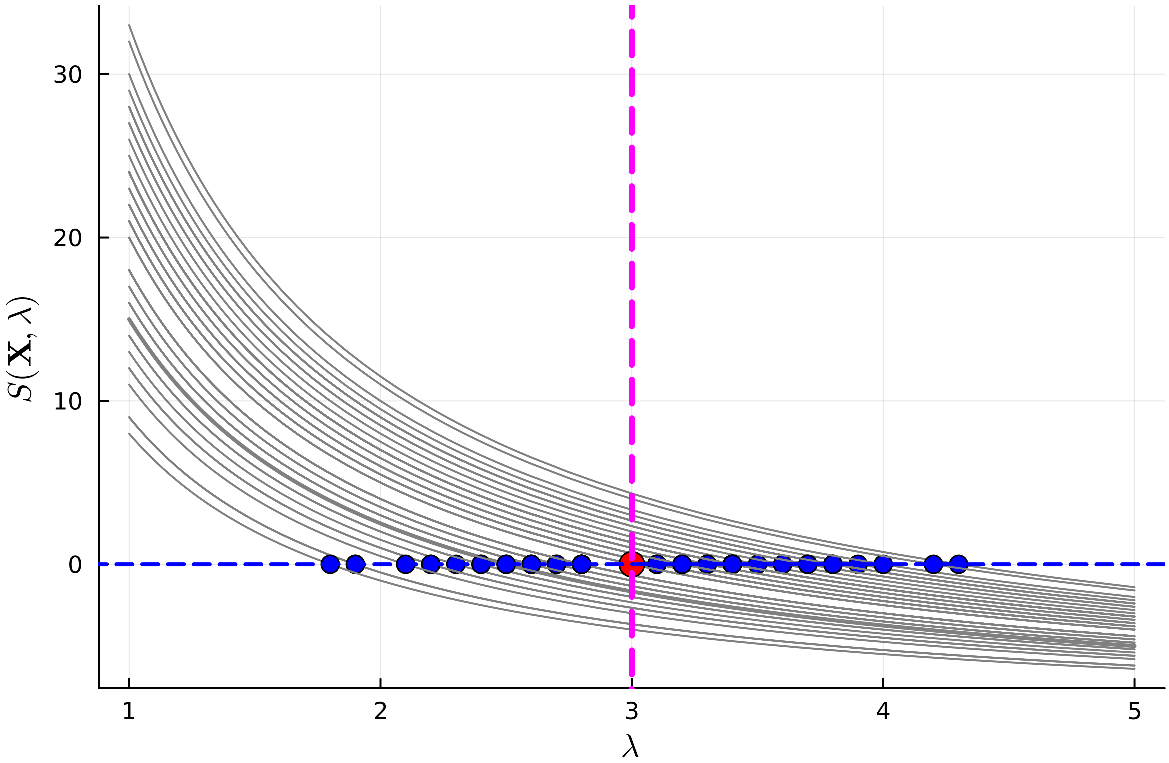


Figure 15: Shapes of the score function for different sets of samples each of size $n = 10$. The blue dot represents the MLE of λ which is $\overline{X}_n$. The red dot is the true value. The vertical magenta line at λ_0 intersects the score function at $S(\mathbf{X}, \lambda_0)$ and the average of these values would be close to zero. $\frac{1}{M}\sum_{m=1}^{M} S(\mathbf{X}^{(m)}, \lambda_0) \approx \mathbb{E}S(\mathbf{X}, \lambda_0) = 0.$

Let us do the same experiment for the Cauchy distribution whose score function is given by

$$S(\mathbf{X}; \mu) = 2\sum_{i=1}^{n} \frac{(x_i - \mu)}{[1 + (x_i - \mu)^2]}$$

In [25]:

```julia
using Random, Distributions

mu_0 = 3   # True value
n = 10
x = rand(Cauchy(mu_0), n)   # Simulate from Cauchy distribution

score_cauchy(mu) =  2 * sum((x .- mu) ./ (1 .+ (x .- mu).^2))

mu_vals = 1:0.01:5
score_vals = zeros(length(mu_vals))

for i in 1:length(mu_vals)
    score_vals[i] = score_cauchy(mu_vals[i])
end

plot(mu_vals, score_vals, color = "grey", lw = 2, xlabel = L"\mu",
ylabel = L"S(X, \mu)", label = "" )
scatter!([mu_0],[0], color = "red", markersize = 6, label = "")
hline!([0], color = "blue", lw = 2, linestyle = :dash, label = "")
```

Out[25]:

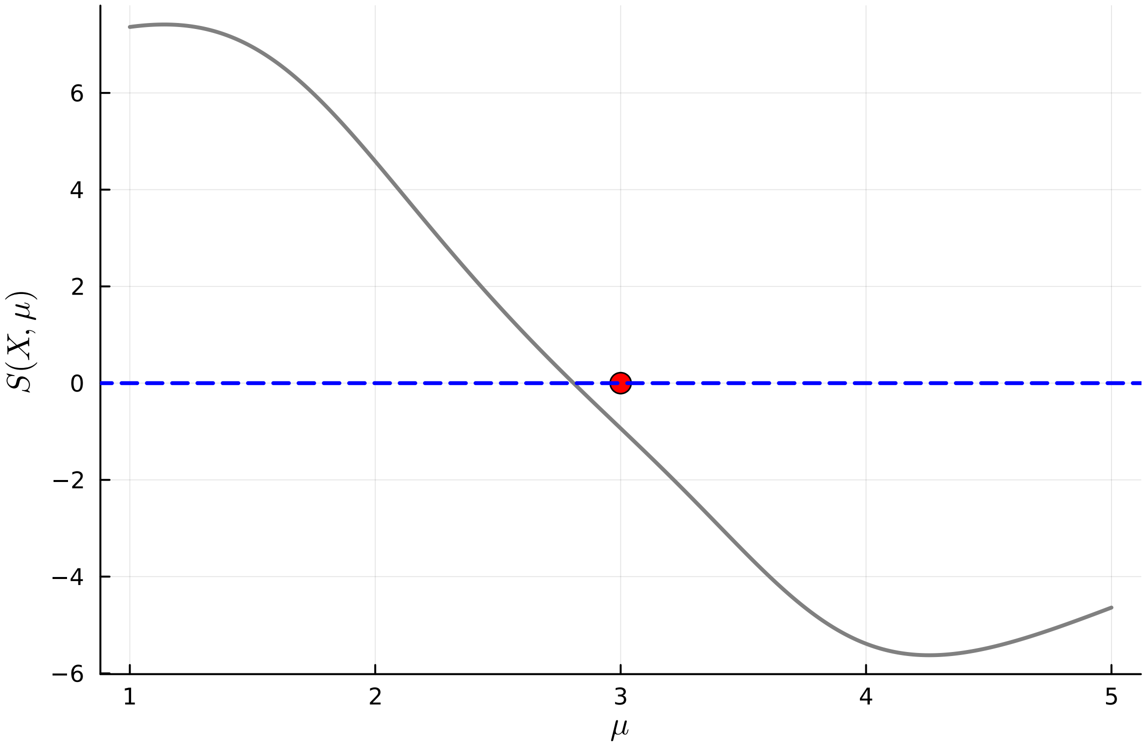


Figure 16: The shapes of the score function for different samples of size $n = 10$. The true value of $\mu = \mu_0 = 3$ has been used for simulation purposes (marked as red dot). The reader is encouraged to execute the following code multiple times and check the shapes for different choices of μ_0 and sample size n.

Asymptotic distribution of the MLE

Under certain regularity conditions for the studied population distribution $f(x|\theta)$, if $\hat{\theta}_n$ be the MLE of θ and $I_n(\theta)$ be the Fisher Information. Then

$$\mathrm{SE}\left(\hat{\theta}_n\right) = \sqrt{\mathrm{Var}\left(\hat{\theta}_n\right)} \approx \sqrt{\frac{1}{I_n(\theta)}}.$$

In addition,

$$\frac{\hat{\theta}_n - \theta}{\text{SE}\left(\hat{\theta}_n\right)} \to \mathcal{N}(0,1), \quad \text{in distribution.}$$

In addition,

$$\widehat{\text{SE}}\left(\hat{\theta}_n\right) = \sqrt{\widehat{\text{Var}}\left(\hat{\theta}_n\right)} \approx \sqrt{I_n(\theta_n)^{-1}}$$

and

$$\frac{\hat{\theta}_n - \theta}{\widehat{\text{SE}}\left(\hat{\theta}_n\right)} \to \mathcal{N}(0,1), \quad \text{in distribution.}$$

In the following simulation experiment, we verify the above results for the Poisson distribution. The MLE for the parameter λ is $\hat{\lambda}_n = \overline{X}_n$.

$$\text{SE}\left(\hat{\lambda}_n\right) = \sqrt{\frac{\lambda}{n}}, \quad \text{and} \quad \widehat{\text{SE}}\left(\hat{\lambda}_n\right) = \sqrt{\frac{\hat{\lambda}_n}{n}} = \sqrt{\frac{\overline{X}}{n}}.$$

In [26]:

```
lambda_0 = 3
M = 1000
n_vals = [3, 5, 10, 35, 50, 100]
plot_layout = (6, 2)
p_list = []

for n in n_vals
    U = zeros(M)
    V = zeros(M)

    for i in 1:M
        x = rand(Poisson(lambda_0), n)
        U[i] = (mean(x) .- lambda_0) ./ (sqrt(lambda_0/n))
        V[i] = (mean(x) .- lambda_0) ./ (sqrt(mean(x)/n))
    end

    # histogram for U
    p1 = histogram(U, normalize = true, title = "n = $n", bins = 30,
    xlabel = L"U_n", ylabel = "density", label = "")
    plot!(x -> pdf.(Normal(0,1), x), lw = 2, color = "red", label = "")

    # histogram for V
    p2 = histogram(V, normalize = true, title = "n = $n", bins = 30,
    xlabel = L"V_n", ylabel = "density", label = "")
    plot!(x -> pdf.(Normal(0,1), x), lw = 2, color = "red", label = "")

    append!(p_list, [p1, p2])
end

plot(p_list..., layout=plot_layout, size=(1200, 1800))
```

Out[26]:

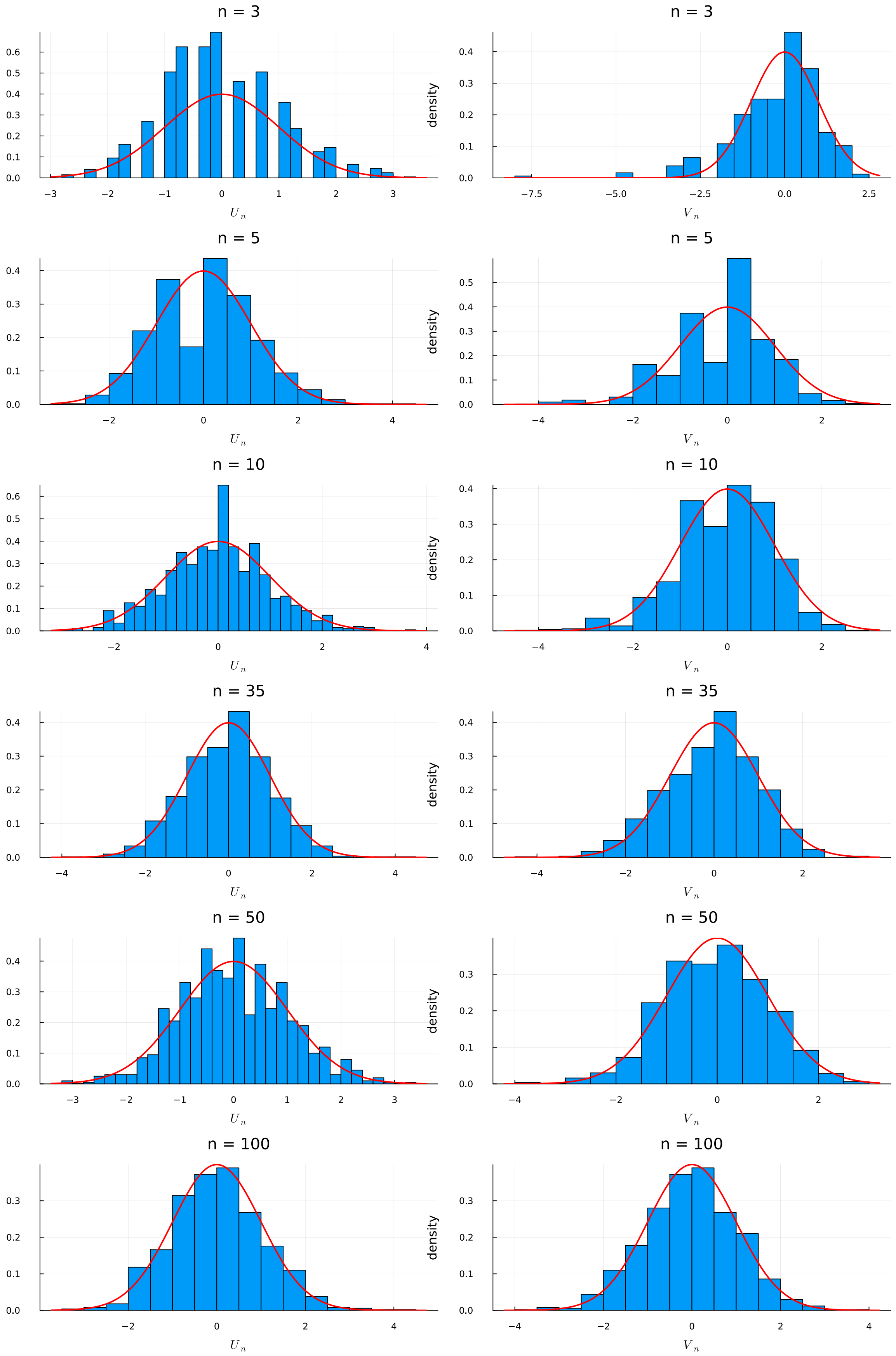


Figure 17: As the sample size increases, the standardized estimators of the MLE of λ is approximately $\mathcal{N}(0,1)$ distributed.

Multiparameter setting and Score function

Suppose that $X_1, X_2, \ldots, X_n$ be a random sample of size n from the population PDF (PMF) $f(x|\theta)$ and $\theta \in \Theta \subseteq \mathbb{R}^p$ and $p > 1$. Let $\theta = (\theta_1, \ldots, \theta_p)$ and the corresponding MLE is expressed as

$$\hat{\theta}_n = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p).$$

MLE is computed by solving the system of p equations given by

$$\frac{\partial l_n(\theta)}{\partial \theta_j} = 0, \quad j \in \{1, 2, \ldots, p\}$$

We define

$$H_{jj} = \frac{\partial^2 l_n(\theta)}{\partial \theta_j^2}, \quad \text{and} \quad H_{jk} = \frac{\partial^2 l_n(\theta)}{\partial \theta_j \partial \theta_k},$$

for $i, j \in \{1, 2, \ldots, p\}$. Similar to the one variance case, here we will have the Fisher Information Matrix defined as

$$I_n(\theta) = \begin{bmatrix} \mathbb{E}_\theta(H_{11}) & \mathbb{E}_\theta(H_{12}) & \ldots & \mathbb{E}_\theta(H_{1p}) \\ \mathbb{E}_\theta(H_{21}) & \mathbb{E}_\theta(H_{22}) & \ldots & \mathbb{E}_\theta(H_{2p}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}_\theta(H_{p1}) & \mathbb{E}_\theta(H_{p2}) & \ldots & \mathbb{E}_\theta(H_{pp}) \end{bmatrix}$$

When the partial derivatives are evaluated at the MLE $\hat{\theta}_n$, then we obtain the Observed Fisher Information Matrix.

❗ **Fisher Information Matrix and Multivariate Normality of MLE**

Let $J_n = I_n^{-1}(\theta)$, the inverse of the expected Fisher Information Matrix. Under appropriate regularity conditions

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{n\to\infty} \text{MVN}_p\left(0, J_n(\theta)\right).$$

In particular

$$\frac{\sqrt{n}\left(\hat{\theta}_j - \theta_j\right)}{\widehat{\text{SE}}(\hat{\theta}_j)} \xrightarrow{n\to\infty} \mathcal{N}(0, 1)$$

where $\widehat{\text{SE}}(\hat{\theta}_j) = J_n(j, j)$, the jth diagonal entry of J_n, evaluated at $\hat{\theta}$. Also

$$\text{Cov}\left(\hat{\theta}_j, \hat{\theta}_k\right) \approx J_n(j, k).$$

The famous regularity conditions

Throughout the discussions in the statistical computing lectures, we have observed only nice things about the MLE of model parameter(s) θ, two prominent properties are:
(a) MLEs are consistent estimators, that means as the sample size goes to infinity, the sampling distribution of the MLEs will be highly concentrated at the true parameter value θ, or in other words, the limiting distribution of the MLE is degenerate at the true value.

In fact, this property is true for any probability density (or mass) functions $f(x|\theta)$, $\theta \in \Theta$ which belongs to the family satisfying the following four conditions:

- The random sample $X_1, X_2, \ldots, X_n$ are independent and identically distributed (IID) following $f(x|\theta)$.
- The parameter is identifiable, that is, if $\theta_1 \neq \theta_2$, then $f(x|\theta_1) \neq f(x|\theta_2)$.
- For every θ, the density functions $f(x|\theta)$ have common support, and $f(x|\theta)$ is differentiable in θ.
- The parameter space Ω contains an open set ω of which the true parameter value θ_0 is an interior point.

In many examples in this document, we also observed that MLEs appeared to be asymptotically normal and asymptotically efficient as well. In addition to the above four conditions, the following two conditions are needed to ensure the above two properties.

- For every $x \in \chi$, support of the PDF (PMF) $f(x|\theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ, and $\int f(x|\theta)dx$ can be differentiated three times under the integral sign.

- For any $\theta_0 \in \Omega$, there exists a positive number c and a function $M(x)$ (both of which depend on θ_0) such that

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x|\theta)\right| \leq M(x) \quad \text{for all } x \in \chi, \quad \theta_0 - c < \theta < \theta_0 + c,$$

  with $\mathbb{E}_\theta[M(X)] < \infty$.

These conditions are typically known as the regularity conditions.

Some more examples

In [27]:

```
using CSV, DataFrames, Plots, QuadGK

JJ = CSV.read("JohnsonJohnson.csv", DataFrame)
JJ_data = JJ[:, 1]
plot(JJ_data, label = "")
```

Out[27]:

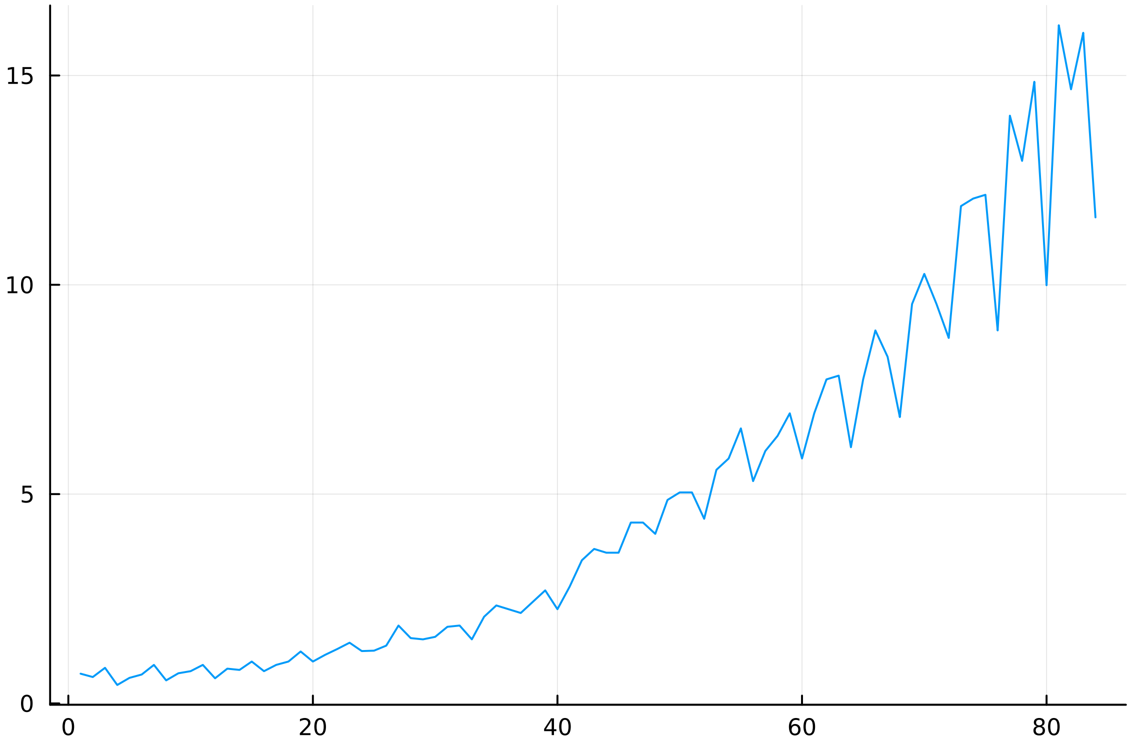


Figure 18: The plot displays the time series of Johnson & Johnson's quarterly earnings data, showing fluctuations over time.

In [28]:

```
histogram(JJ_data, normalize = true, xlabel="Quarterly earnings (dollars)",
ylabel = "density",label = "")
```

Out[28]:

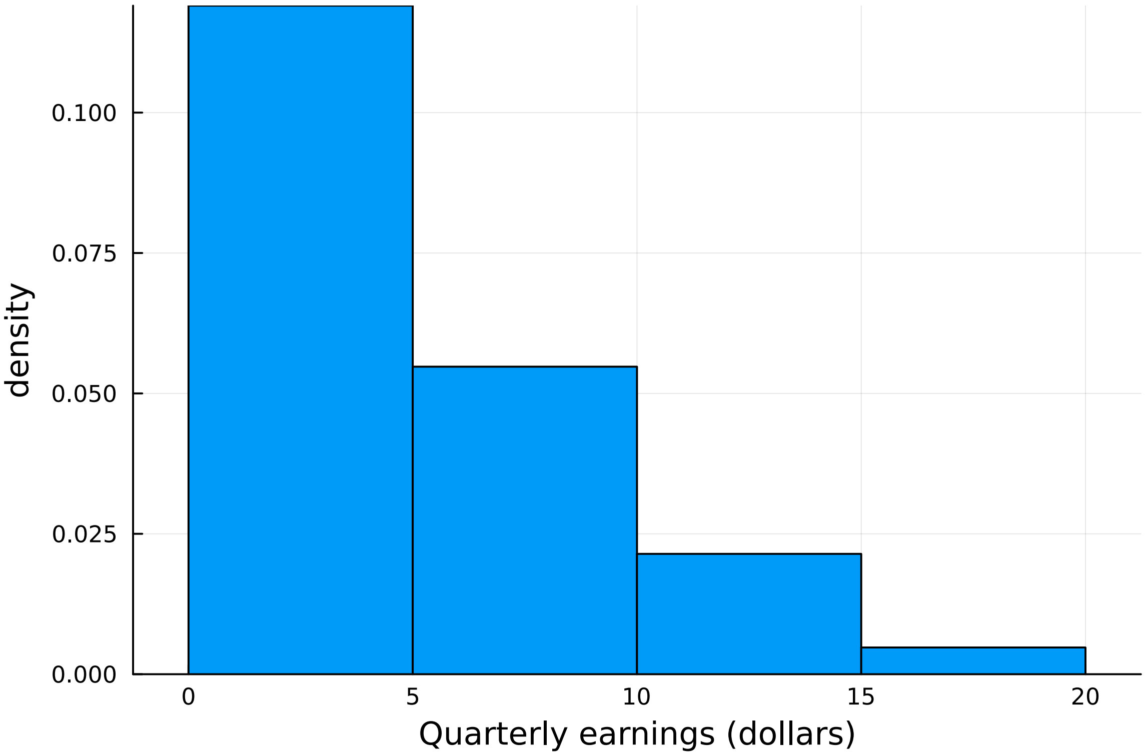


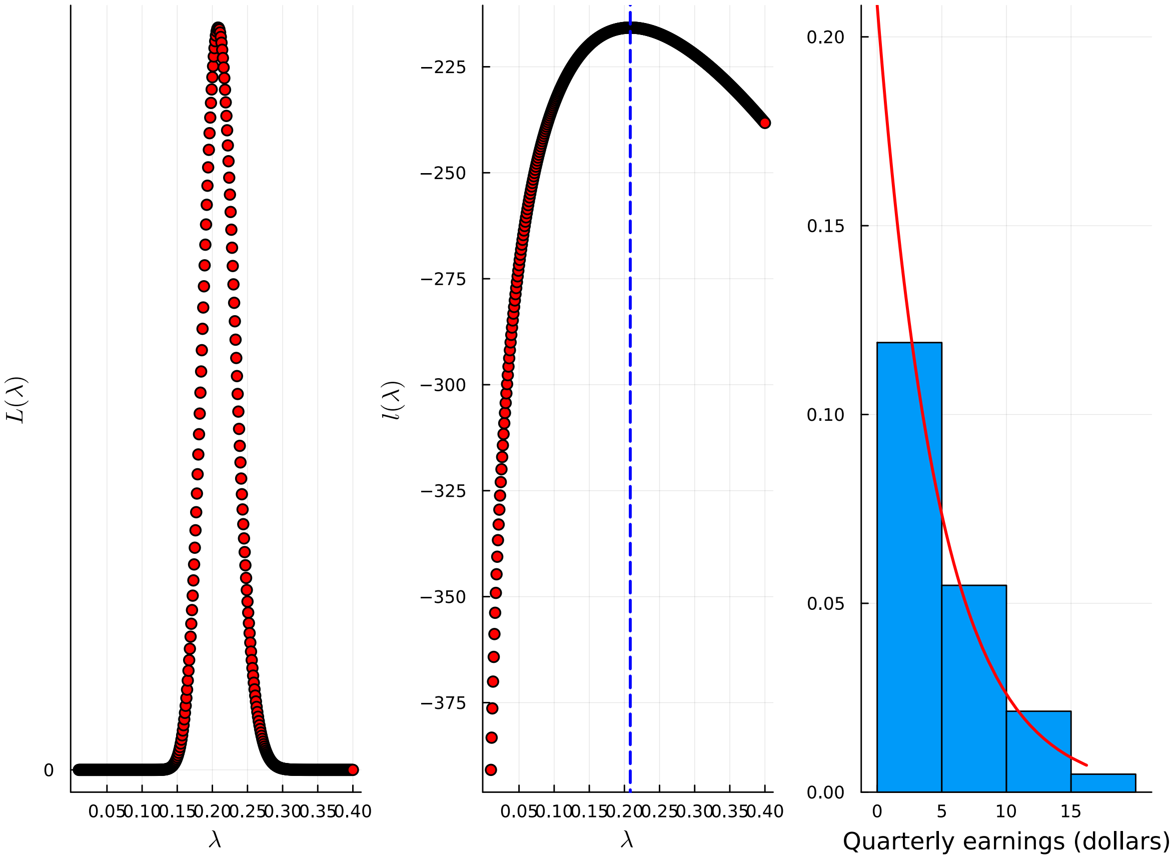
Figure 19: A histogram of the quarterly earnings data is presented, where the x-axis represents earnings in dollars. The distribution provides insight into the spread and frequency of the earnings.

In [32]:

```
JJ = CSV.read("JohnsonJohnson.csv", DataFrame)
JJ_data = JJ[:, 1]
```

```
n = length(JJ_data)
lambda_hat = n / sum(JJ_data)

lambda_vals = collect(0.01:0.001:0.4)
Lik_vals = [ (lambda^n) * exp(-lambda * sum(JJ_data)) for lambda in lambda_vals ]

p1 = scatter(lambda_vals, Lik_vals, markershape=:circle, color="red",
    xlabel=L"\lambda", ylabel=L"L(\lambda)", label="",
    xticks=0.05:0.05:0.4, yticks=0:0.5:maximum(Lik_vals))

p2 = scatter(lambda_vals, log.(Lik_vals), markershape=:circle, color="red",
    xlabel=L"\lambda", ylabel=L"l(\lambda)", label="",
    xticks=0.05:0.05:0.4, yticks=:auto)
vline!([lambda_hat], color="blue", linestyle=:dash, lw=2, label="")

p3 = histogram(JJ_data, normalize=true, xlabel="Quarterly earnings (dollars)",
    xticks=0:5:maximum(JJ_data), yticks=0:0.05:0.5, label="")
plot!(x -> lambda_hat * exp(-lambda_hat * x), 0, maximum(JJ_data), color="red",
    lw=2, label="")

plot(p1, p2, p3, layout=(1,3), size=(800, 600))
```

Out[32]:



(a) The likelihood function $L(\lambda)$ is plotted for different values of λ. The red points indicate computed likelihood values, while the blue dot represents the maximum likelihood estimate (MLE), $\hat{\lambda}$. (b)The logarithm of the likelihood function, $l(\lambda)$, is displayed. The vertical dashed blue line marks the MLE $\hat{\lambda}$, where the log-likelihood reaches its maximum. (c)A histogram of the quarterly earnings data is overlaid with an exponential probability density function (red curve) fitted using the MLE $\hat{\lambda}$.

In [30]:

```
function f(x)
    return lambda_hat * exp(-lambda_hat * x) * (x > 0)
end

P_X_greater_18 = exp(-18 * lambda_hat)
```

Out[30]:

```
0.023513371465180056
```

In [31]:

```julia
integral_value, integral_error = quadgk(f, 18, Inf)
println("Integral: ", integral_value)
```

```
Integral: 0.023513371465180066
```