Theoretical Foundations of Large Data Sets

List 1

Paulina Podgórska

Let $X_1, ..., X_n$ be the simple random sample from the beta distribution $\beta(\alpha+1, 1)$ with the density $f(x, \alpha) = (\alpha+1)x^\alpha$, for $x \in (0, 1)$, $\alpha > -1$.

Maximum likelihood estimator

Let us find the maximum likelihood estimator $\hat{\alpha}_{MLE}$ of the parameter α. The likelihood function is given by

$$L(\alpha, x) = \prod_{i=1}^{n}(1+\alpha)x_i^\alpha.$$

From this, we obtain the log-likelihood function as

$$logL(\alpha, x) = l(\alpha, x) = nlog(1+\alpha) + \alpha\sum_{i=1}^{n} logx_i.$$

The maximum likelihood estimator can be found using the following relation:

$$\hat{\alpha} = argmax_\alpha L(\alpha, x) \Leftrightarrow max_\alpha L(\alpha, x) = L(\hat{\alpha}, x).$$

Taking the derivative with respect to α and setting it to zero, we have

$$\frac{\partial l}{\partial \alpha} = \frac{n}{1+\alpha} + \sum_{i=1}^{n} logx_i = 0,$$

from which the MLE for α is

$$\hat{\alpha}_{MLE} = -\frac{n}{\sum\limits_{i=1}^{n} log(x_i)} - 1.$$

In the last step we need to check if the estimator truly corresponds to a maximum in the log-likelihood function. To do that we need to inspect the second derivative of $logL(\alpha, x)$:

$$\frac{\partial l}{\partial \alpha^2} = -\frac{n}{(1+\alpha)^2} \overset{\alpha=\hat{\alpha}}{=} -\frac{n}{(1+\hat{\alpha})^2}.$$

The above value is smaller than zero because $\alpha > -1$.

Fisher Information and the asymptotic distribution

Let's calculate the Fisher information for the parameter α.

$$logf(x,\alpha) = log(\alpha+1) + \alpha log(x)$$
$$\frac{\partial logf(x,\alpha)}{\partial \alpha} = \frac{1}{\alpha+1} + log(x)$$
$$\frac{\partial^2 logf(x,\alpha)}{\partial \alpha^2} = -\frac{1}{(\alpha+1)^2}$$
$$I(\alpha) = -E\frac{\partial^2 logf(x,\alpha)}{\partial \alpha^2} = \frac{1}{(\alpha+1)^2}$$

Now let us find the asymptotic distribution of the MLE estimator. We know that

$$\sqrt{n}(\hat{\alpha}_n - \alpha) \xrightarrow{D} N(0, \frac{1}{I(\alpha)}),$$

from which we can obtain

$$(\hat{\alpha}_n - \alpha) \xrightarrow{D} N(0, \frac{1}{nI(\alpha)})$$

Ultimately we get

$$\hat{\alpha}_n \xrightarrow{D} N(\alpha, \frac{(\alpha+1)^2}{n})$$

Knowing the distribution of $\hat{\alpha}_{MLE}$ we can determine the MSE of this estimator:

$$MSE(\hat{\alpha}) = Var(\hat{\alpha}) + (bias(\hat{\alpha}))^2 = \frac{(\alpha+1)^2}{n} + \alpha - \alpha = \frac{(\alpha+1)^2}{n}$$

The moment estimator $\hat{\alpha}_{mom}$

Given that our distribution follows $\beta(\alpha+1, 1)$ the expected value for this distribution is given by:

$$E(X) = \frac{\alpha}{\alpha+\beta}.$$

So, for our specific distribution:

$$\mu_1 = EX = \frac{\alpha+1}{\alpha+2}.$$

The estimate of μ_1 is:

$$\mu_1 \to \hat{\mu_1} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Using the method of moments, we have:

$$\hat{\mu}_1 = \frac{\hat{\alpha}_{mom}+1}{\hat{\alpha}_{mom}+2}$$

Upon solving for $\hat{\alpha}_{mom}$ we obtain:

$$\hat{\alpha}_{mom} = \frac{1-2\hat{\mu_1}}{\hat{\mu}_1 - 1}.$$

Calculating estimators, bias and MSE

Next, we will calculate estimators $\hat{\alpha}_{MLE}$ and $\hat{\alpha}_{mom}$ as well as bias $(\alpha - \hat{\alpha})$ and MSE $(\alpha - \alpha)^2$ with fixed $\alpha = 5$ and $n = 20$.

Table 1: Values of estimators, bias and MSE

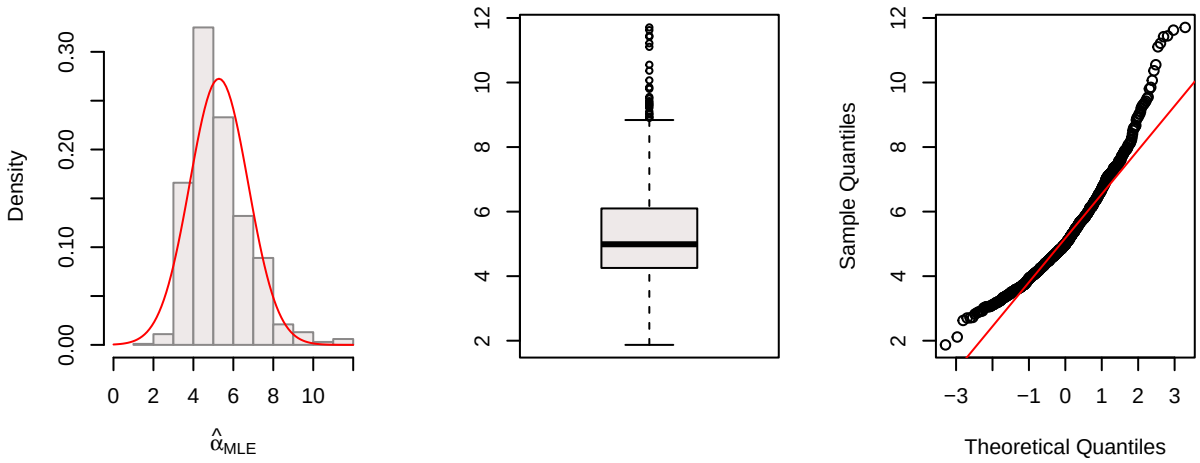
$\hat{\alpha}_{MLE}$	$\hat{\alpha}_{mom}$	$\alpha - \hat{\alpha}_{MLE}$	$\alpha - \hat{\alpha}_{mom}$	$(\alpha - \hat{\alpha}_{MLE})^2$	$(\alpha - \hat{\alpha}_{mom})^2$
4.249	4.166	0.751	0.834	0.565	0.696

We can notice, that although the difference is not significant – MLE estimator is more accurate.

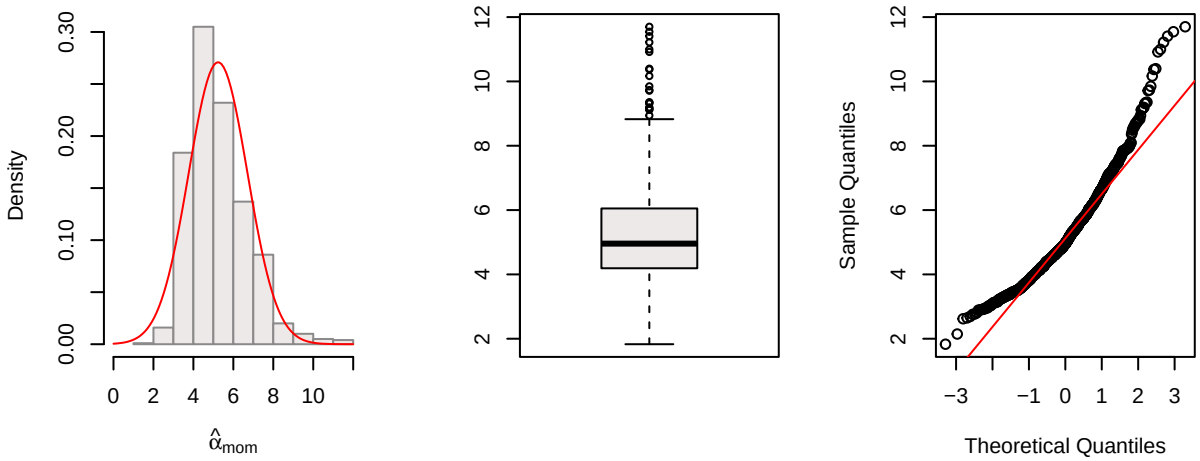
1000 samples, n = 20

To better verify above assumption let us generate 1000 samples of the size $n = 20$. Firstly we will compare a few plots for both estimators

Plots for the maximum likelihood estimator $\hat{\alpha}_{MLE}$



Plots for the moment estimator $\hat{\alpha}_{mom}$



The plots for both estimators are nearly identical. From the histograms and box plots, we can observe that their distributions closely resemble each other. Most frequently, the values we obtained for both estimators lie between 4 and 5. The median for both is around 5:

```
median(alpha_mle)
```

```
## [1] 4.988128
```

```
median(alpha_mom)
```

```
## [1] 4.956697
```

It seems that, with $\alpha = 5$ and $n = 20$ both estimation methods perform very well and we cannot determine a significant difference between them.

Estimation of the bias, the variance and MSE of both estimators

Next, we will estimate the bias, the variance and mean-squared error of both estimators. We will also construct 95% confidence intervals for these parameters.

	$bias_{MLE}$	$bias_{mom}$	MSE_{MLE}	MSE_{mom}	var_{MLE}	var_{mom}
Value	0.28	0.24	2.22	2.22	2.14	2.17
CI	(0.189, 0.37)	(0.146, 0.328)	(2.129, 2.31)	(2.13, 2.313)	(1.967, 2.345)	(1.989, 2.371)

It is difficult to conclusively determine which estimator is better, as the statistics for both are highly similar. We can consider the differences to be minimal.

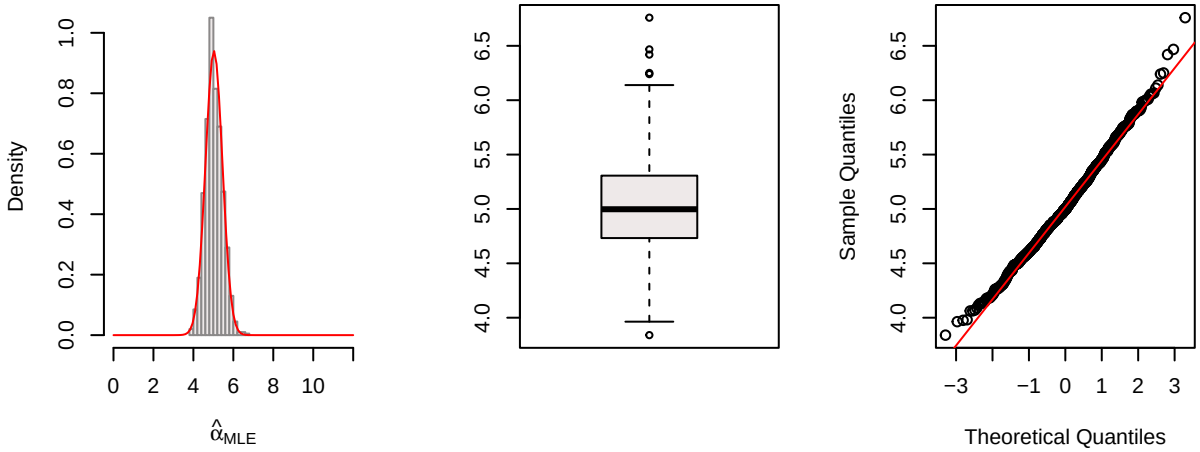
Theoretical values:

Table 3: Theoretical values of parameters for $\hat{\alpha}_{MLE}$

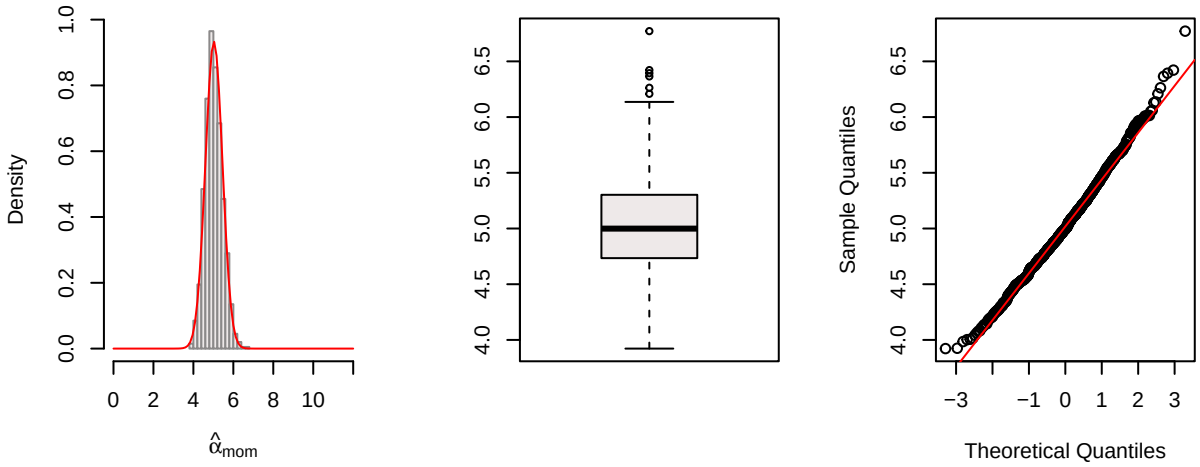
var	bias	MSE
1.8	0	1.8

1000 samples, n = 200

Plots for the maximum likelihood estimator $\hat{\alpha}_{MLE}$



Plots for the moment estimator $\hat{\alpha}_{mom}$



As we increase n to 200, we observe that both estimators improve, closely approximating the actual value of $\alpha = 5$. The number of outliers is reduced, as seen on the box plots and qq-plots. The majority of the values for the estimators fall between 4 and 6. Once again, the differences between $\hat{\alpha}_{MLE}$ and $\hat{\alpha}_{mom}$ are minimal. The histograms are considerably narrower, and the medians are closer to 5 compared to when $n = 20$:

```
median(alpha_mle)
```

```
## [1] 4.997204
```

```
median(alpha_mom)
```

```
## [1] 4.99901
```

Estimation of the bias, the variance and MSE of both estimators

	$bias_{MLE}$	$bias_{mom}$	MSE_{MLE}	MSE_{mom}	var_{MLE}	var_{mom}
Value	0.03	0.03	0.18	0.18	0.18	0.18
CI	(0.004, 0.056)	(0.002, 0.055)	(0.155, 0.207)	(0.157, 0.21)	(0.165, 0.197)	(0.168, 0.2)

Once again, the differences between the two estimation methods are minimal, reinforcing the observation that both approaches perform equivalently well. Specifically, with $n = 200$, there's an improvement in the performance of both estimators. The bias, MSE and variance for each estimator are significantly reduced. The confidence intervals associated with these three statistics have also narrowed, indicating greater precision in the estimations. We can conclude that with a larger sample size, both estimation methods deliver results with higher accuracy.

Theoretical values:

Table 5: Theoretical values of parameters for $\hat{\alpha}_{MLE}$

var	bias	MSE
0.18	0	0.18

Now, let us consider a simple random sample $X_1, ..., X_n$ from the exponential distribution $Exp(\lambda)$ with the density $f(x, \lambda) = \lambda e^{-\lambda x}$, for $x > 0$, $\lambda > 0$. Our task is to find the uniformly most powerful test at the level $\alpha = 0.05$ for testing the hypothesis $H_0 : \lambda = 5$ against $H_1 : \lambda = 3$.

Critical value for the test

Using Neyman - Pearson Theorem, we can provide the formula for the critical value of this test.

The likelihood function is given by:

$$L(x, \lambda) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}.$$

From this, we can determine the likelihood ratio as:

$$\frac{L(x, \lambda_1)}{L(x, \lambda_0)} = \frac{3^n e^{-3 \sum_{i=1}^{n} x_i}}{5^n e^{-5 \sum_{i=1}^{n} x_i}} = \left(\frac{3}{5}\right)^n e^{2 \sum_{i=1}^{n} x_i} > k$$

Consequently, this leads to:

$$\sum_{i=1}^{n} x_i > \frac{1}{2} log\left(\left(\frac{5}{3}\right)^n k\right) = k_1,$$

where k_1 is the critical value.

Given that $\alpha = 0.05$, we have:

$$\alpha = 0.05 = P_{H_0}\left(\sum_{i=1}^{n} x_i > k_1\right),$$

and under H_0, $X_i \sim Exp(5)$, so $\sum_{i=1}^{n} X_i \sim \Gamma(n, \frac{1}{5})$.

Ultimately our critical value is expressed as:

$$1 - \alpha = 0.95 = \Phi_{\Gamma(n,\frac{1}{5})}(k_1)$$
$$k_1 = \Phi^{-1}_{\Gamma(n,\frac{1}{5})}(0.95).$$

Power of the test

$$P_{H_1}\left(\sum_{i=1}^{n} x_i > k_1\right) = 1 - P_{H_1}\left(\sum_{i=1}^{n} x_i \leqslant k_1\right) = 1 - \Phi_{\Gamma(n,\frac{1}{3})}(\Phi^{-1}_{\Gamma(n,\frac{1}{5})}(0.95)).$$

P-value

The formula for the p-value for a given random sample:

$$P_{H_0}\left(\sum_{i=1}^{n} X_i \geqslant d\right) = 1 - \Phi_{\Gamma(n,\frac{1}{5})}(d),$$

where $d = \sum_{i=1}^{n} x_i$.

Next, for $n = 20$, we will generate one random sample from H_0 and one random sample from H_1 and find the respective p-values.

```
p1 # for H0
```

```
## [1] 0.9309067
```

```
p2 # for H1
```

```
## [1] 4.731284e-06
```

The observed p-value for the sample drawn from H_0 is notably higher, surpassing our significance level of $\alpha = 0.05$ – the null hypothesis would not be rejected. On the other hand, the sample from H_1 is close to zero. In this situation we would reject the null hypothesis, indicating that the $\lambda = 3$.

Distribution of the p-value when data comes from H_0

$$P_{H_0}(p \leqslant x) =^{x\in(0,1)} P_{H_0}\left(1 - \Phi_{\Gamma(n,\frac{1}{5})}\left(\sum_{i=1}^{n} X_i \leqslant x\right)\right) = 1 - P_{H_0}\left(\sum_{i=1}^{n} X_i \leqslant \Phi^{-1}_{\Gamma(n,\frac{1}{5})}(1-x)\right) =$$

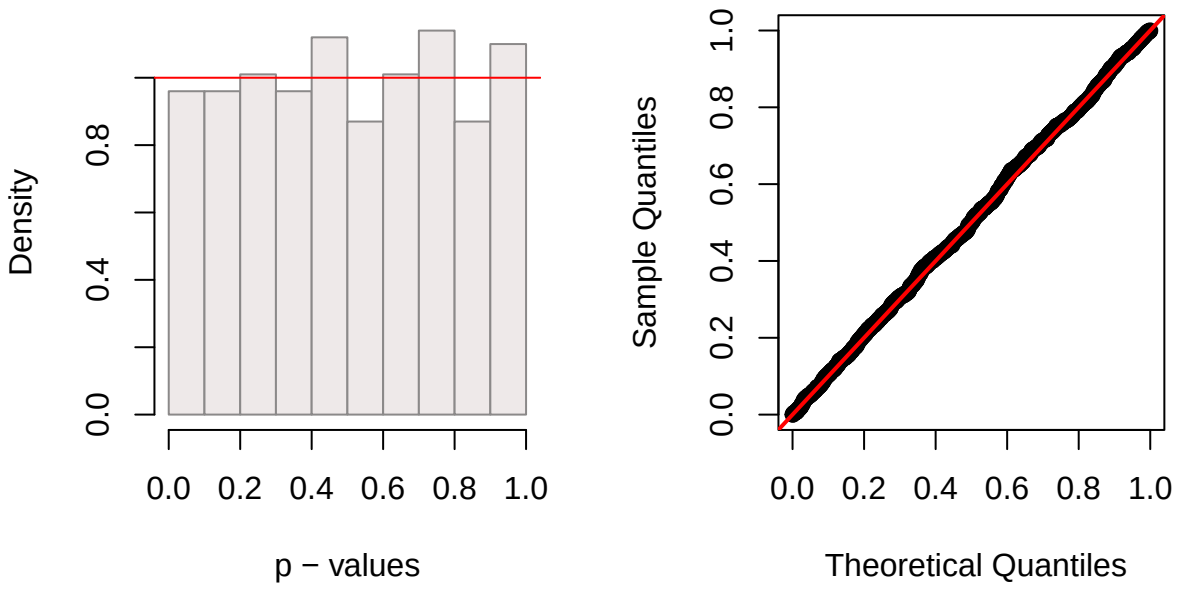
$$= 1 - \Phi_{\Gamma(n,\frac{1}{5})}\left(\Phi^{-1}_{\Gamma(n,\frac{1}{5})}(1-x)\right) = 1 - (1-x) = x$$

When data comes from H_0, the distribution of the p-value is uniform $U(0, 1)$.

1000 samples with $n = 20$

In this step we will generate 1000 samples of the size $n = 20$ from H_0 and calculate respective p-values.

Plots



Distribution of out p-values is very close to the theoretical one.

95% confidence interval

Let us construct the 95% confidence interval for the type I error in the following way:

$$\hat{p} = \frac{1}{k} \sum_{i=1}^{k} \mathbb{I}_{p_i < \alpha}$$

With that, the confidence interval can be calculated with:

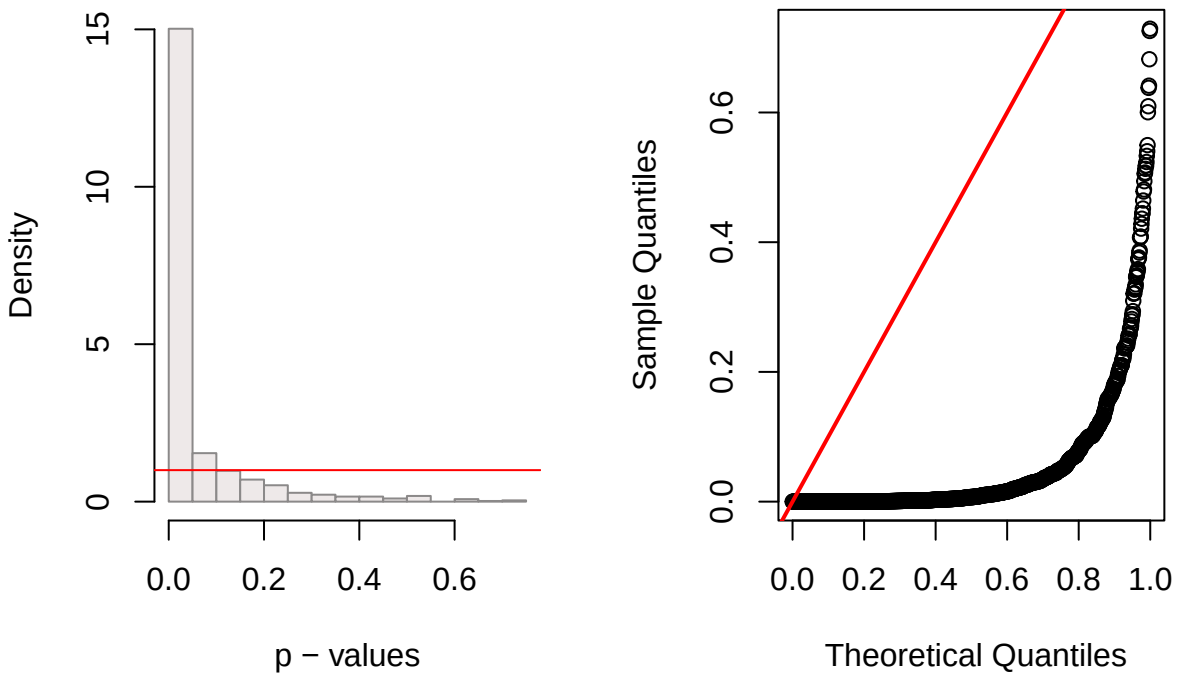
$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{k}}.$$

For our data, the confidence interval is: (0.034, 0.06).

1000 samples of size $n = 20$ for H_1

Let us generate 1000 samples of the size $n = 20$ from H_1 and calculate respective p-values. Then we will compare the distribution of p-values under H_0 and under H_1.

Plots



From plots, it's evident that the p-values don't originate from a uniform distribution, unlike in the situation when the data comes from H_0. The histogram suggests that most of p-values are below approximately 0.05.

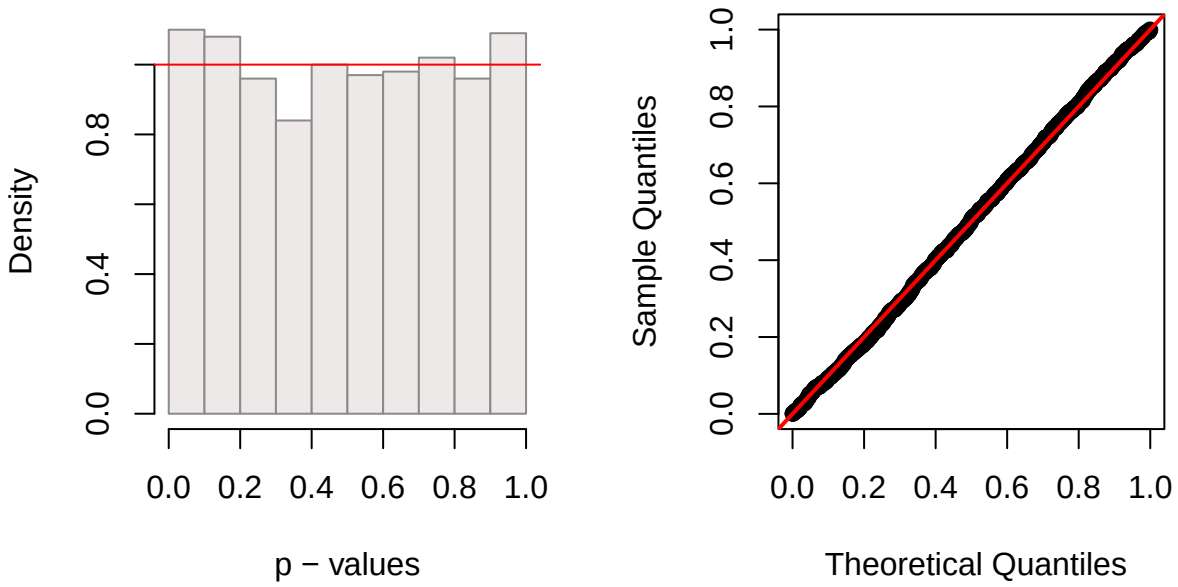
95% confidence interval for the power of the test

Based on our simulations, the 95% confidence interval for the test's power is (0.724, 0.778). The theoretical power for this test is 0.758. The confidence interval is narrow and the theoretical value falls within it - our estimation is precise.

Comparison with $n = 200$

Samples from H_0 – p - values

Plots



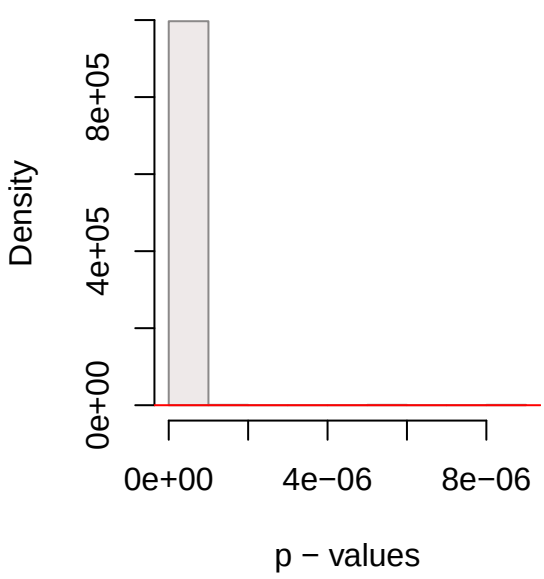
95% confidence interval

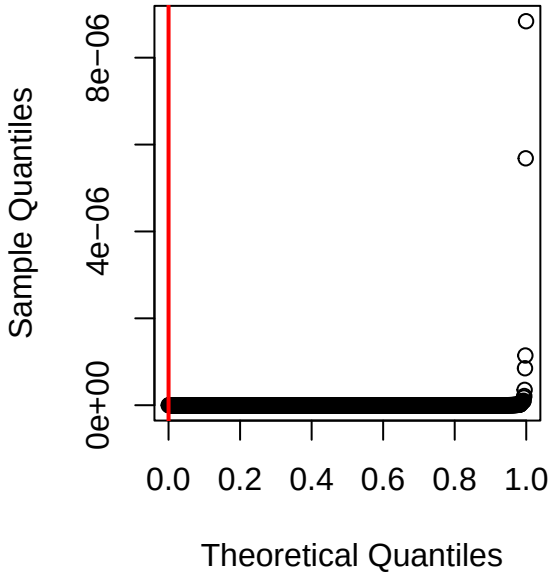
For our data, the confidence interval is: (0.033, 0.059).

There isn't a significant difference in the plots or confidence intervals between $n = 20$ and $n = 200$. We can conclude that increasing n doesn't substantially impact the p-values when data is drawn from H_0.

1000 samples of size $n = 200$ for H_1

Plots





95% confidence interval for the power of the test

For our data, the confidence interval is: (1, 1). Theoretical value of power: 1.

With the increase in sample size from 20 to 200, we observe that the power of the test reaches 1, indicating perfect sensitivity. Additionally, a vast majority of p-values drawn from H_1 are near zero.