Report 2

Klaudia Balcer

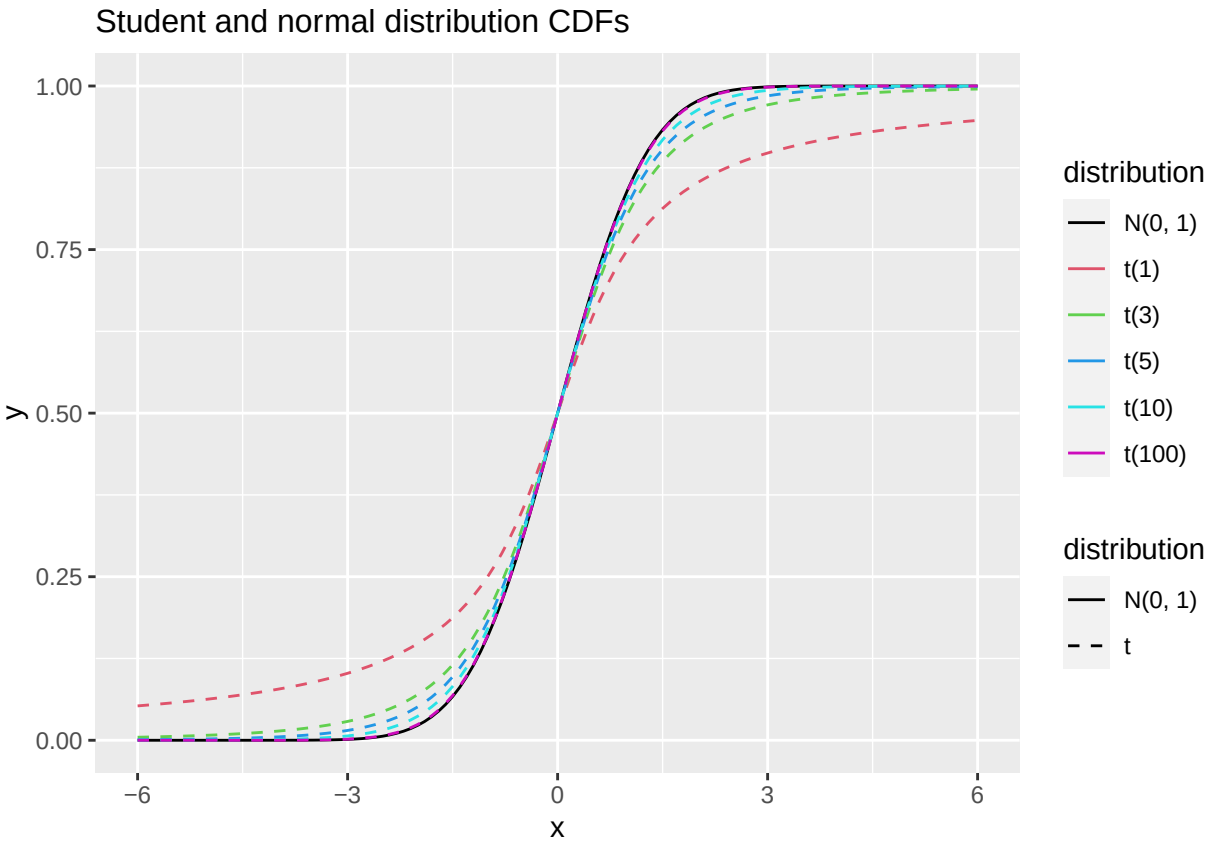
11/5/2021

Deadline: 02.12.2021

Contents

Task 1

In the first task, we will compare the **cumulative distribution functions** (CDFs) for the standard normal distribution and t-student distribution with certain degrees of freedom.



With the growing number of degrees of freedom, the student distribution converges to the standard normal distribution. When n is equal 100 the difference between the CDFs is invisible.

Task 2

In the first task, we will compare the **cumulative distribution functions** (CDFs) for the standard normal distribution and normalized χ^2 distribution with certain degrees of freedom.

We will use the following formula for standarizing the χ^2 distribution:

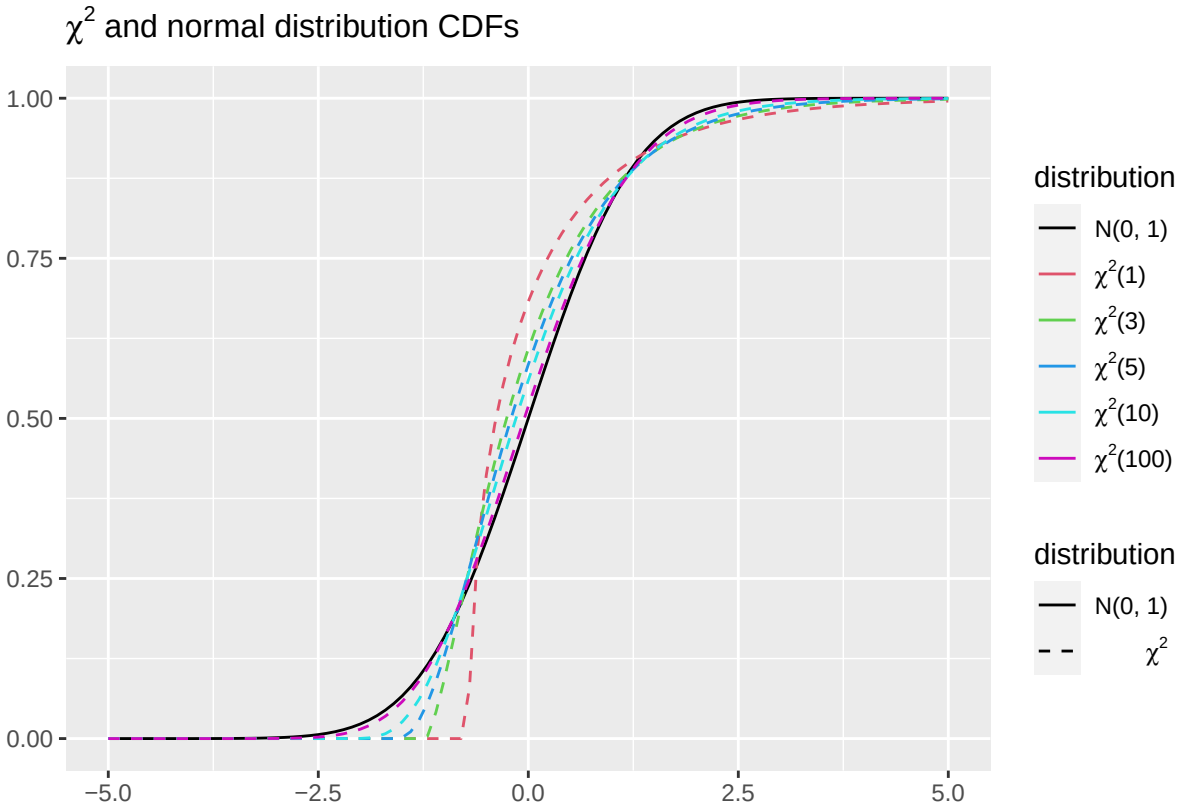
$$T = \frac{\chi^2_{df} - df}{\sqrt{2df}}$$

Let's take a look on the distribution of the standarized random variables:

$$P\Big(\frac{\chi^2 - df}{\sqrt{2df}} < k\Big) = P\Big(\chi^2 - df < k\sqrt{2df}\Big) = P\Big(\chi^2 < k\sqrt{2df} + df\Big)$$

Equivalently:

$$F_T(k) = F_{\chi^2_{df}}(k\sqrt{2df} + df)$$



With the growing number of degrees of freedom, the χ^2 distribution converges to the standard normal distribution. When n is equal 100 the difference between the CDFs is almost invisible. The convergence is slower than for student distribution.

Task 3

In this task, we will consider the global testing for the expected value of the Poisson distribution.

Problem Definition

The Poisson distribution with the rate parameter λ has the following **probability mass function** (PMF):

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Both mean and variance of a random variable from $Poi(\lambda)$ are equal λ.

Simple Hypothesis Test

Let's define the simple hypothesis test problem:

$$H_{0,i} : \lambda = 5 \Leftrightarrow \mathbb{E}X = 5$$
$$H_{1,i} : \lambda > 5 \Leftrightarrow \mathbb{E}X > 5$$

In this problem, the test statistics is as follows:

$$T(X) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

The test statistics has the following distribution:

$$nT(X) \sim Poi(n\lambda)$$

Let's calculate the critical value of the right-sided critical region:

$$\mathbb{P}(T(X) > c) = \mathbb{P}(nT(X) > nc) = 1 - F_{Poi(n\lambda)}(nc) = \alpha = 0.05$$

$$c = \frac{1}{n} F^{-1}_{Poi(n\lambda)}(0.95)$$

Note: as the distribution is discrete, the significance level and the size of the test may differ.

P-value:

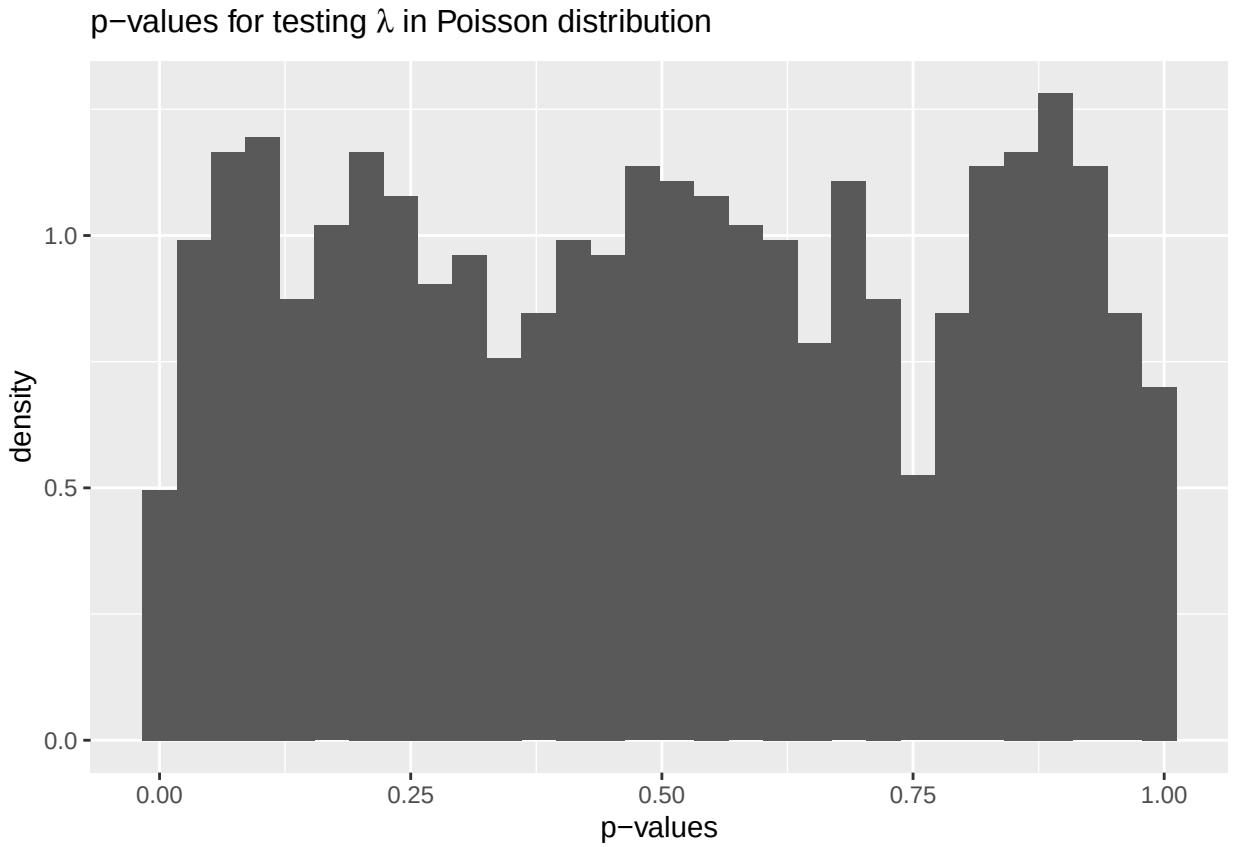
Definition of p-value:

$$p = \mathbb{P}_0(T > T(X)) = \mathbb{P}_0(nT > nT(X)) = 1 - F_{Poi(n\lambda)}(nT(X))$$

Function in R calculating the p-value for testing a random vector `X` for parameter `lambda`:

```
pval <- function(X, lambda=5) {
  1 - ppois(sum(X), length(X) * lambda)
}
```

Histograms of p-values for simple hypothesis:



The above plot shows a histogram of 1000 p-values. Each of those p-values was calculated using a random vector of length 1000. The distribution of p-values does not seem to be uniform. And in fact, it is not as we consider a discrete distribution.

Global Hypothesis Test

Now, let's consider $m = 1000$ simple hypothesis. Instead of testing them separatly, we will perform a global test.

First, we need to define the global hypotheses:

$$H_0 : \bigcap_{j=1}^{n} H_{0,j}$$

$$H_1 : \bigcup_{j=1}^{n} H_{1,j}$$

To perform the global testing, we will use Bonferroni and Fisher methods. We will compare the properties of those approaches under H_0 and H_1.

Test Statitics

Let's denote p_i as the p-value in the simple testing problem $H_{0,i}$ vs $H_{1,i}$.

Bonferroni Test The Bonferronie test statistics has the formula as follows:

$$T_{bonf} = min\{p_i\}$$

We reject the null hypothesis for small values of the test statistics:

$$\varphi_{bonf} = T_{bonf} < \frac{\alpha}{m}$$

Fisher Test The Fisher test statistics has the formula as follows:

$$T_{fish} = -2\sum_{i=1}^{n} log(p_i)$$

We reject the null hypothesis for big values of the test statistics:

$$\varphi_{fish} = T_{fish} > \chi^2_{2n}(1-\alpha)$$

Note: The limiting distribution of the statistics T_{fish} can be prescribed only for testing in continuous distributions. For a discrete distribution, the p-values p_i in the simple tests are not uniformly distributed. Thus, the distribution of the test statistics cannot be derived. We use the χ^2_{2n} distribution as a simple satisfying approximation. However, **in the case of discrete distribution, the probability of Type I Error may not exactly meet our expectations.**

Type I Error

Probability of Type I Error has an upper bound of the significance level of the test. For test statistics with continuous distribution, the probability of Type I Error is equal α. For discrete distributions, in most cases the probability of Type I Error is lower than α.

For Bonferroni test, the probability of Type I error is equal to the probability of Type I Error for simple hypothesis for independent p_is and lower for dependent p_is.

For Fisher test, the probability of Type I Error converges to α for continuous distributed random variables. In other cases, we have no exact formula for the distribution of the test statistics. Thus, the probability of Type I Error is expected to be a value around α (without more specific expectations).

Estimation of Type I Errors:

P(Type I Error | Bonferroni) = 0.045

P(Type I Error | Fisher) = 0.072

Power

We will provide simulation under two alternatives:

- one strong effect (needle in the haystack problem),
- many small effects.

As Bonferroni takes only the lowest p_i in account, it is more likeli that it will discover the single strong signal. Fisher looks on all the signals and summarises all effects. Thus, this test is likeli to discover a very strong (in respect to the number of hypotheses) signal or many small signals.

Alternative with single strong signal Power of Bonferroni: 1

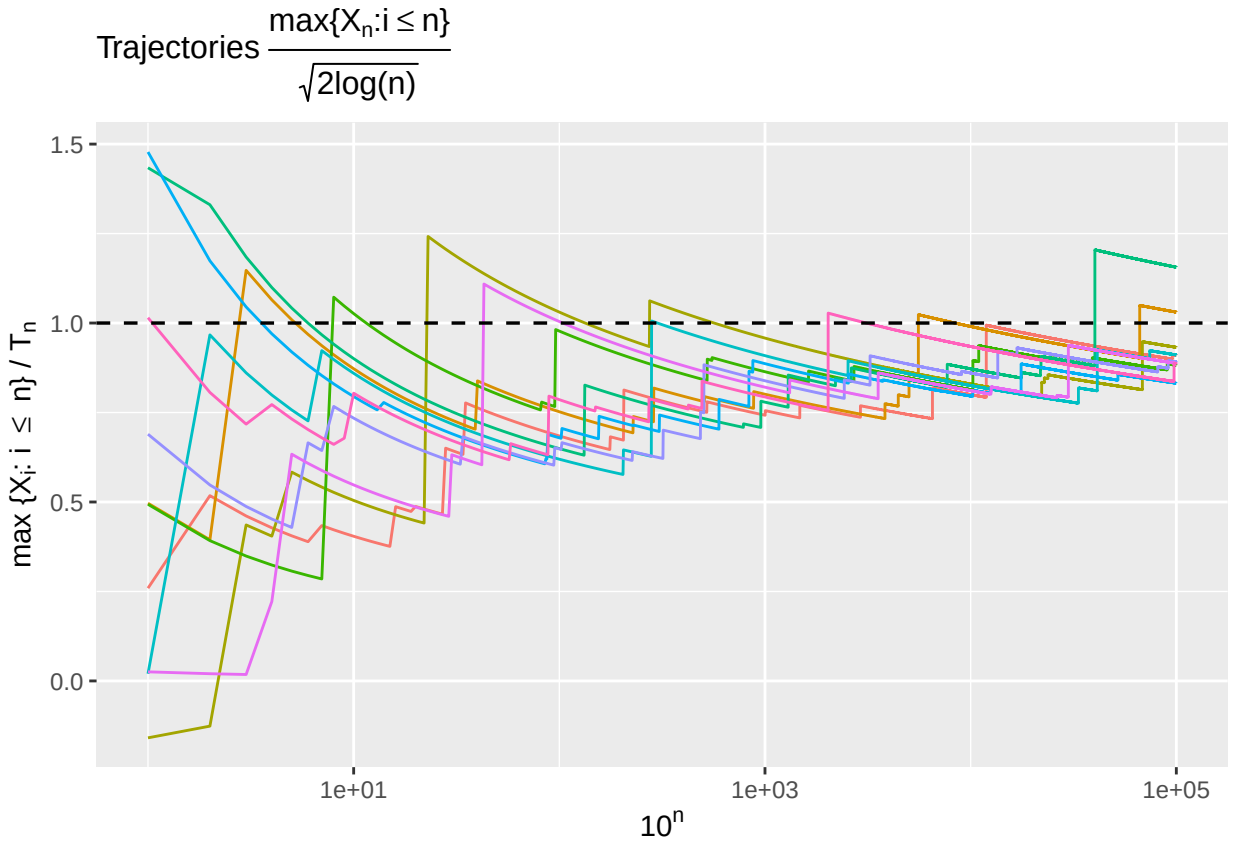
Power of Fisher: 0.81 #### Alternative with many small signal

Power of Bonferroni: 0.25

Power of Fisher: 1 The powers calculated in simulations have met the theoretical expectations.

Task 4

In the next tasks, we are going to show the optimality of detection treshold for Bonferroni in needle in Haystack problem. In this task, we will show an approximation of maximum of a random vector from a standard normal distribution.



At the above plot, we can see that the maximum of n independent random variables from the standard normal distribution converges to $\sqrt{2log(n)}$.

Task 5

In this task, we will study the properties of the likelihood function and its approximation in needle in the haystack problem.

Problem Definition

Let's have a random vector $Y = (Y_1, Y_2, \ldots, Y_n) \sim N(\vec{\mu}, I)$ and the testing problem: $H_0: $ all μ_j are equal 0 against $H_1 : \mu_i = \gamma = (1 - \epsilon)\sqrt{2log(n)}$ for $\epsilon \in (0, \frac{1}{2})$, the rest of μ_j are equal 0.

The alternative hypothesis is not simple. We could translate it as: $H_1 : \vec{\mu} \sim \pi$. Than $P(\mu_i \neq 0) = \frac{1}{n}$ for all $i = 1, \ldots, n$.

Let's look at the likelihood ratio:

$$L(X,\gamma) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\prod_{j=1,j\neq i}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x_j^2}{2\sigma^2}}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\gamma)^2}{2\sigma^2}}\right]}{\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x_j^2}{2\sigma^2}}}$$

$$L(X,\gamma) = \frac{(\frac{1}{\sqrt{2\pi}\sigma})^n}{(\frac{1}{\sqrt{2\pi}\sigma})^n}\frac{1}{n}\sum_{i=1}^{n}\frac{\prod_{j=1,j\neq i}^{n}e^{-\frac{x_j^2}{2\sigma^2}}\cdot e^{-\frac{(x_i-\gamma)^2}{2\sigma^2}}}{\prod_{j=1}^{n}e^{-\frac{x_j^2}{2\sigma^2}}}$$

$$L(X,\gamma) = \frac{1}{n}\sum_{i=1}^{n}\frac{e^{-\frac{(x_i-\gamma)^2}{2\sigma^2}}}{e^{-\frac{x_j^2}{2\sigma^2}}}$$

$$L(X,\gamma) = \frac{1}{n}\sum_{i=1}^{n}e^{\gamma x_i-\gamma^2/2}$$

Let's consider a approximation of the likelhood ratio:

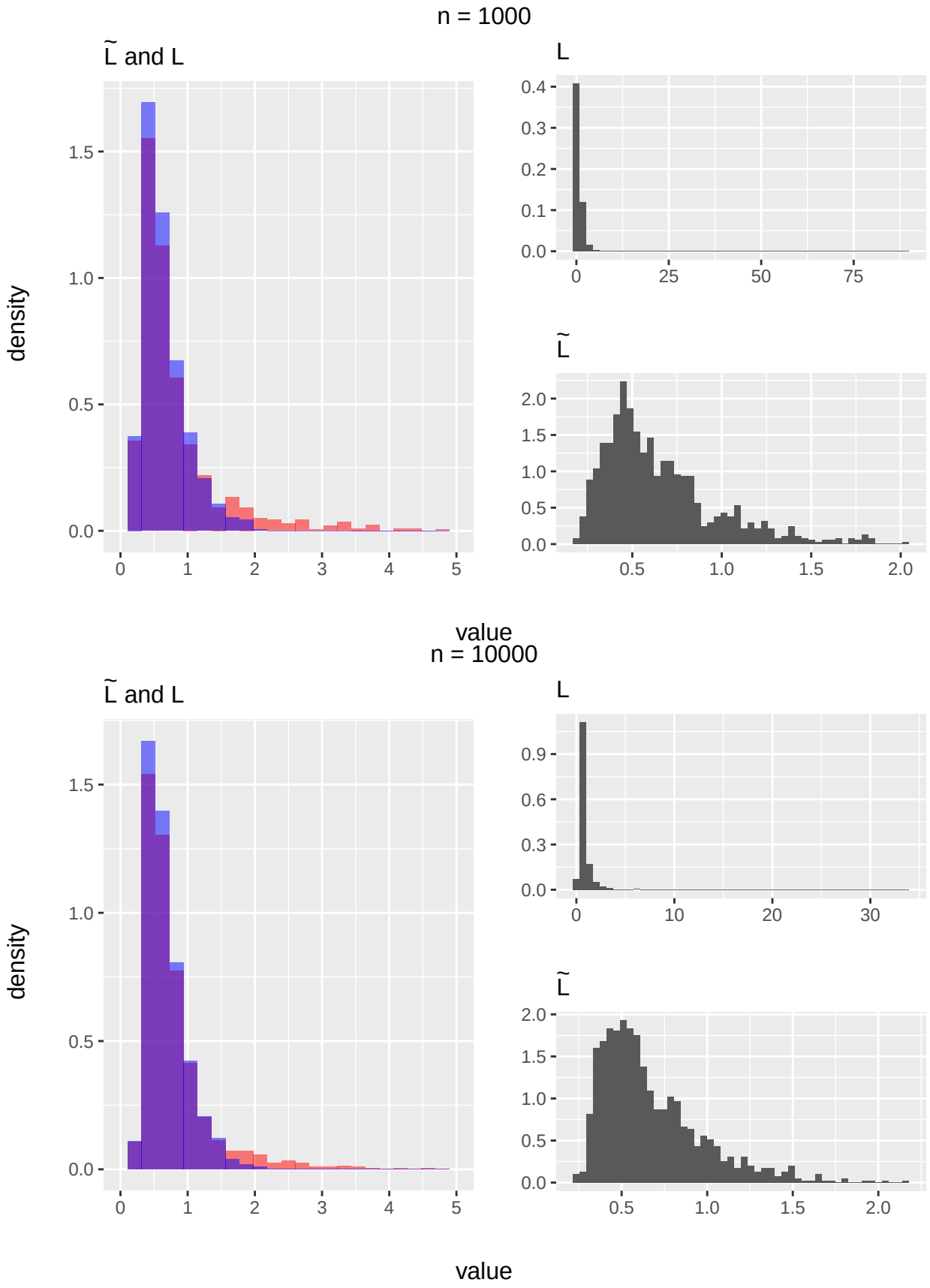
$$\tilde{L}(X,\gamma) = \frac{1}{n}\sum_{i=1}^{n}e^{\gamma x_i-\gamma^2/2}\cdot\mathbb{1}\{Y_i<\sqrt{2log(n)}\}$$

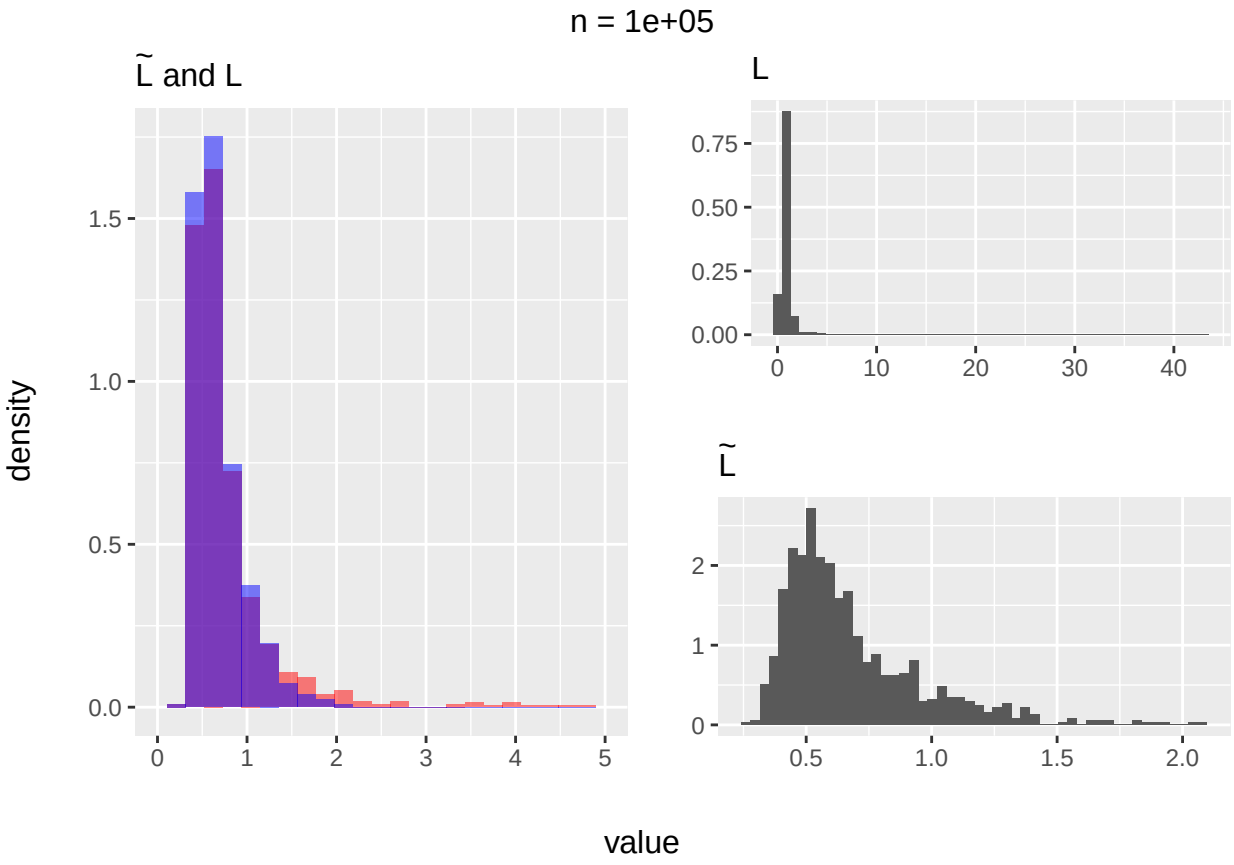
which sums only components restricted by $\sqrt{2log(n)}$. We know that $\mathbb{P}\Big(maxX_i \geq \sqrt{2log(n)}\Big) \to 0$ unger H_0, so we expect $\mathbb{P}(L \neq \tilde{L}) \to 0$.

In this task we will study the properites of L and $\tilde{L}$ under H_0.

Histograms of L and $\tilde{L}$

Please note: joint histograms have X-axis with limits [0, 5]; those plot do not include outliers. The full distributions can be seen on the individual histograms.


n = 1000
L̃ and L
L
L̃
density
value
n = 10000
L̃ and L
L
L̃
density
value



Properties of L and $\tilde{L}$

n	Var(L)	Var(L_approx)	P(L != L_approx)
1000	11.347338	0.1012157	0.892
10000	2.253234	0.0901697	0.928
1e+05	4.186953	0.0766553	0.939

As $\mathbb{P}\Big(maxX_i \geq \sqrt{2log(n)}\Big) \to 0$, the probability, that L and $\tilde{L}$ won't be equal also converges to 0 with the growth of n (**only for weak signals**). We can see it in the simulation results.

The main difference between L and $\tilde{L}$ is the variance. We can consider $\tilde{L}$ as L *without outliers*. We expect that $Var(\tilde{L}) = o(1)$ and we can see it in the pictures above. Variance of $\tilde{L}$ is relatively small anddoes not depend on n.

Task 6

In the last task we will study the needly in haystack problem again. This time: under alternative hypothesis.

Let's start with the likelihood ratio:

$$L(X, \gamma) = \frac{1}{n} \sum_{i=1}^{n} e^{\gamma x_i - \gamma^2/2}$$

In the the most powerfull test, we reject H_0 when $L(X, \gamma) > c$, where c is the critical value (a number such that under the null hypothesis, the probability of rejecting the null hypothesis has an upper bound of

significance level of the test). Equivalently, we reject H_0 when $l(X,\gamma) = log(L(X,\gamma)) > log(c) = c'$.

We will run study the problem for $n = 500, 5000, 50000$ and $\epsilon = 0.05, 0.1, 0.2$.

As the likelihood ratio and the loglikelihood ratio do not have a standard distribution, we will use simulations to state the critical value of the test. For each n we will provide $m = 10000$ log-likelihood ratios and get the quantile of $1 - \alpha$ to get the critical value.

Table 2: Critical values of the NP test based on the loglikelihood ratio

500	5000	50000
0.907	0.728	0.628

Then, for each pair of n and ϵ we will provide a random sample and test the hypothesis using:

- the most powerfull test
- Bofnerroni test
- Bonferroni test with corrected α (such that the probability of Type I Error for Bonferroni is α)

Please note: in this task we have a two-sided alternative.

Table 3: Power of Bonferroni test

	0.05	0.1	0.2
500	0.457	0.517	0.666
5000	0.502	0.556	0.711
50000	0.499	0.616	0.752

Table 4: Power of Bonferroni test (with corrected significance level for simple tests)

	0.05	0.1	0.2
500	0.458	0.522	0.666
5000	0.503	0.559	0.715
50000	0.500	0.620	0.753

Table 5: Power of Neyman-Person test

	0.05	0.1	0.2
500	0.523	0.601	0.709
5000	0.571	0.613	0.749
50000	0.549	0.668	0.771

As we can see, correcting α for Bonferroni makes only small changes. Bonferroni's power is very close to the maximal.