List 2

Kamil Zaborniak

Contents

Task 1

a)

Let $\underline{X_n} = (X_1, \ldots, X_n)$ (where $n > 0$) be the sample from the Poisson distribution with expected value λ. Let consinder the test of:

$$H_0 : \lambda = 5,$$

$$H_1 : \lambda > 5.$$

At the significance level α. The test statistic is:

$$T\left(\underline{X_n}\right) = \sum_{i=1}^{n} X_i \sim Poiss\left(n\lambda\right).$$

Suppose $T \sim Poiss\left(n\lambda\right)$, then p-value:

$$p = P_{H_0}\left(T > T\left(\underline{X_n}\right)\right) = 1 - F_{Poiss(n\lambda)}\left(T\left(\underline{X_n}\right)\right)$$

The function determining the p-value in R is as follows:

```
p_for_Poiss<- function(sample, lambda=5){
  n<- length(sample)
  p<- ppois(sum(sample),
            lambda = lambda*n,
            lower.tail=F)
  return(p)
}
```

b)

Let's generate 1000 samples of size 100 from the $Poiss(5)$ distribution. Let's calculate the p-value for each of them. Based on the obtained vector of p-values: $\underline{p} = (p_1, \ldots, p_{1000})$, we will check the distribution of p-values when the sample comes from a discrete distribution.

Let's first draw a histogram of the obtained p-values, and then a qqplot.

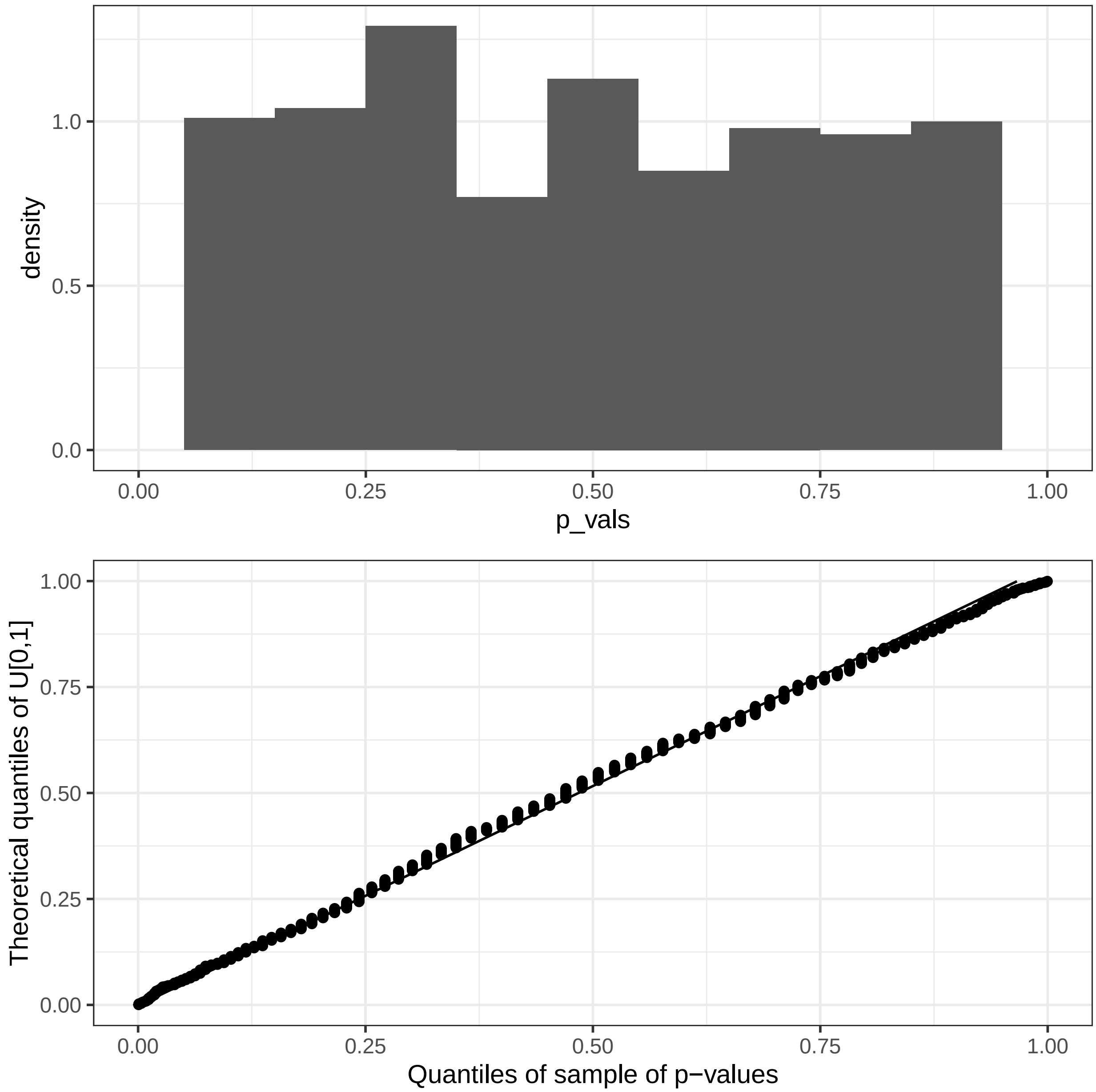

Figure 1: Histogram and qqplot of p-value.

The above graphs indicate that the p-value based on the sample coming from a discrete distribution also comes from a discrete distribution. However, the distribution of p-values is asymptotically uniform on [0,1]. Let us consider the proof shown in the example below.

Przykład 0.1 *Niech* $\underline{X} = (x_1, \ldots, x_n)$*, gdzie* $n \geq 1$*, będzie próbą niezależnych zmiennych losowych z rozkładu Poissona o parametrze* λ*. Rozważmy, na poziomie istotności* $\alpha \in [0, 1]$*, prawostronny test hipotez:*

$$\begin{aligned} H_0 :& \quad \lambda = \lambda_0, \\ H_1 :& \quad \lambda > \lambda_0, \end{aligned}$$

gdzie $\lambda_0 \in (0, \infty)$. *Wyznaczmy statystykę testową:*

$$T\left(\underline{X}\right) = \sum_{i=1}^{n} x_i.$$

Widzimy, że przy założeniu hipotezy zerowej, statystyka $T\left(\underline{X}\right)$ *przyjmuje rozkład Poissona o parametrze* $n\lambda_0$. *Załóżmy, że zmienna losowa* z *pochodzi z rozkładu Poissona o parametrze* $n\lambda_0$. *Wyznaczmy p-wartość:*

$$p = P\left(z > T\left(\underline{X}\right) \mid H_0\right) = 1 - \mathbb{F}\left(T\left(\underline{X}\right)\right),$$

gdzie $\mathbb{F}(\bullet)$ *jest dystrybuantą rozkładu Poissona o parametrze* $n\lambda_0$. *Niech* $x \in [0, 1]$. *Obliczmy prawdopodobieństwo, przy założeniu prawdziwości hipotezy zerowej, że* $p \leq x$:

$$\begin{aligned} P(p \leq x \mid H_0) &= P\left(1 - \mathbb{F}\left(T\left(\underline{X}\right)\right) \leq x \mid H_0\right) \\ &= P\left(\mathbb{F}\left(T\left(\underline{X}\right)\right) \geq 1 - x \mid H_0\right). \end{aligned} \tag{1}$$

Niech

$$k = \min\left\{ m: \quad \sum_{i=0}^{m} \frac{\left(n\lambda_0\right)^i e^{-n\lambda_0}}{i!} \geq 1 - x \right\}.$$

Z tego wynika, że:

$$\begin{aligned} P(p \leq x \mid H_0) &= P\left(T\left(\underline{X}\right) \geq k \mid H_0\right) \\ &= 1 - \mathbb{F}(k), \end{aligned} \tag{2}$$

gdzie:

$$1 - x \leq \mathbb{F}(k) \leq 1 - x + \frac{\left(n\lambda_0\right)^k e^{-n\lambda_0}}{k!}. \tag{3}$$

Oznaczmy $L = n\lambda_0$, $a_k = \frac{L^k e^{-L}}{k!}$. *Wykażmy, że przy* $n \to \infty$, $\mathbb{F}(k) \to 1 - x$. *Ponieważ* $L > 0$ *oraz* $k \geq 0$, *to* $\forall k : a_k > 0$. *Sprawdźmy monotoniczność ciągu* a_k. *Ciąg jest rosnący, kiedy:*

$$a_k < a_{k+1} \iff \frac{a_{k+1}}{a_k} > 1.$$

Wyznaczmy:

$$\frac{a_{k+1}}{a_k} = \frac{\frac{L^{k+1} e^{-L}}{(k+1)!}}{\frac{L^k e^{-L}}{k!}} = \frac{L}{k+1}.$$

Z tego wynika, że:

$$\frac{a_{k+1}}{a_k} > 1 \iff k < \lfloor L \rfloor,$$

gdzie $\lfloor r \rfloor$ *oznacza część całkowitą liczby rzeczywistej* r. *Natomiast ciąg* a_k *jest malejący, kiedy:*

$$a_k > a_{k+1} \iff \frac{a_{k+1}}{a_k} < 1 \iff k > \lfloor L \rfloor.$$

Z tego wynika, że:

$$\sup_{k \geq 0} \left(\frac{L^k e^{-L}}{k!} \right) = \frac{L^{\lfloor L \rfloor} e^{-L}}{\lfloor L \rfloor !}.$$

Wobec tego:

$$\begin{aligned} \lim_{n \to \infty} \left(\sup_{k \geq 0} \left(\frac{\left(n\lambda_0\right)^k e^{-n\lambda_0}}{k!} \right) \right) &= \lim_{L \to \infty} \left(\sup_{k \geq 0} \left(\frac{L^k e^{-L}}{k!} \right) \right) \\ &= \lim_{L \to \infty} \left(\frac{L^{\lfloor L \rfloor} e^{-L}}{\lfloor L \rfloor !} \right). \end{aligned}$$

Oszacujmy:

$$0 \leq \frac{L^{\lfloor L \rfloor} e^{-L}}{\lfloor L \rfloor!} \leq \frac{\left(\lfloor L \rfloor + 1\right)^{\lfloor L \rfloor} e^{-\lfloor L \rfloor}}{\lfloor L \rfloor!}. \tag{4}$$

Korzystając ze wzoru Stirlinga, dla $i \in \mathbb{N}_+$:

$$i! \approx \left(\frac{i}{e}\right)^i \sqrt{2\pi i}, \qquad \textit{gdzie} \lim_{i \to \infty} \frac{i!}{\left(\frac{i}{e}\right)^i \sqrt{2\pi i}} = 1, \tag{5}$$

obliczmy:

$$\begin{aligned}
\lim_{L \to \infty} \left(\frac{\left(\lfloor L \rfloor + 1\right)^{\lfloor L \rfloor} e^{-\lfloor L \rfloor}}{\lfloor L \rfloor!} \right) &= \lim_{L \to \infty} \left(\frac{\left(\lfloor L \rfloor + 1\right)^{\lfloor L \rfloor} e^{-\lfloor L \rfloor}}{\lfloor L \rfloor!} \cdot \frac{\lfloor L \rfloor + 1}{\lfloor L \rfloor + 1} \right) \\
&= \lim_{L \to \infty} \left(\frac{\left(\lfloor L \rfloor + 1\right)^{\lfloor L \rfloor + 1} e}{\left(\lfloor L \rfloor + 1\right)! \cdot e^{\lfloor L \rfloor + 1}} \right) \\
&\overset{(5)}{=} \lim_{L \to \infty} \left(\frac{\left(\lfloor L \rfloor + 1\right)^{\lfloor L \rfloor + 1} e}{\left(\frac{\lfloor L \rfloor + 1}{e}\right)^{\lfloor L \rfloor + 1} \sqrt{2\pi \left(\lfloor L \rfloor + 1\right)} \cdot e^{\lfloor L \rfloor + 1}} \right) \\
&= \lim_{L \to \infty} \left(\frac{e}{\sqrt{2\pi \left(\lfloor L \rfloor + 1\right)}} \right) \\
&= 0.
\end{aligned} \tag{6}$$

Z Twierdzenia o trzech ciągach, (4) *oraz* (6) *wynika, że:*

$$\lim_{n \to \infty} \left(\sup_{k \geq 0} \left(\frac{\left(n\lambda_0\right)^k e^{-n\lambda_0}}{k!} \right) \right) = 0. \tag{7}$$

Poprzez (3), (7) *pokazaliśmy, że:*

$$\mathbb{F}(k) \xrightarrow{n \to \infty} 1 - x. \tag{8}$$

Z (1), (2), (8) *wynika, że:*

$$P(p \leq x \mid H_0) \xrightarrow{n \to \infty} 1 - (1 - x) = x.$$

W ten sposób zareprezentowaliśmy, że rozkład p-wartości, dla tego przykładu, jest asymptotycznie jednostajny na przedziale $[0, 1]$.

c)

Now, let us consider testing the global null at the $\alpha = 0.05$ level of significance, using the Bonferroni's Method and Fisher's Combination Test.

Suppose, we have n samples $\underline{X}_i \sim Poiss(\lambda_i)$, where $i \in \{1, \ldots, n\}$. Each sample $\underline{X}_i$ is of size m_i and corresponds to a null hypothesis:

$$H_{0,i} : \lambda_i = 5.$$

and p-value: p_i. So, we have the global null:

$$H_0 = \bigcap_{i=1}^{n} H_{0,i} \Longleftrightarrow \forall_{i \in \{1,\ldots,n\}} \lambda_i = 5.$$

Bonferroni Test

We know, that test statistic has form:

$$T_B = \min_i \{p_i\},$$

And we reject global null when :

$$T_B \leq \frac{\alpha}{m},$$

where n is the number of null hypotheses.

Fisher's Combination Test

The test statistic for Fisher Test has form:

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

Note that the test statistic $T_F \sim \chi^2_{2n}$, then and only if $p_i \sim U[0,1]$ for all i. And then we reject the global null H_0, when:

$$T_F > \chi^2_{2n}(1-\alpha),$$

where $\chi^2_k(\beta)$ is the β quantile of a chi-square distribution with k degrees of freedom. We have p-values from a discrete distribution, so in our case T_F does not come from χ^2 distribution, but we will not transform the test. It's means that we should be prepared for the fact that in our case the Fisher's Combination Test will not control the type I error.

Simulation

Let's estimate the probability of the type I error for the Bonferroni and Fisher's tests at the significance level $\alpha = 0.05$. Let's repeat 1000 times the experiment of generating 100 random ten-element samples from the $Poiss(5)$. As a result of each experiment, we obtain 100 p-values and based on them we calculate the test statistics of the above tests, and then check whether they allow us to reject the global null . As a result of the entire simulation, we obtain the fraction of test statistics that rejected the global null hypothesis, which is the sought estimator of the probability of a type I error.

As a result of the simulation, we obtained that the estimator of the probability of committing a type I error is:
- for Bonferroni method: 0.048,
- for Fisher's Combination Test: 0.138.

As we expected, the Fisher test does not control for type I error in our case.

d)

Suppose number of null hypothesis: $n = 1000$, size of each sample: $m_i = 100$, and samples, the global null from point c). To determine the power estimator for the considered tests, let's construct two alternative hypotheses (defining two different problems), opposite to our global null hypothesis.

$$H_{1,a} : \lambda_1 = 7, \forall_{i \in \{2,\ldots,n\}} \lambda_i = 5,$$

$$H_{1,b} : \forall_{j \in \{1,\ldots,100\}} \lambda_j = 5.2, \forall_{j \in \{100,\ldots,n\}} \lambda_j = 5.$$

Power

To determine the power estimator of the considered tests, we need to repeat the experiment 1000 times, consisting in generating 1000 hundred-element samples assuming an alternative hypothesis. For each experiment, we will calculate the test statistics of the discussed tests and check whether they allow us to reject the global null. As a result of 1000 repetitions of the experiment, we will obtain the fraction of rejections of the global null hypothesis of a given test, which is the sought power estimator. Since we are considering two alternatives, we need to repeat the simulations for each of them.

For the ''needle in a haystack" problem (hypothesis $H_{1,a}$), the estimated power of:
- Bonferroni method is equal to: 1,
- Fisher's Combination Test is equal to: 0.559.

For the problem with many small effects (hypothesis $H_{1,b}$), the estimated power of:
- Bonferroni method is equal to: 0.189,
- Fisher's Combination Test is equal to: 0.987.

The above results confirm that the Bonferroni method only works when there are very few strong effects. In particular, for the "needle in a haystack" problem, there is no better test than the Bonferroni method. When testing many small signals, one of the best tests is the Fisher's Test. It follows that there is no situation where the two tests under consideration perform equally well because they are opposite to each other.

Task 2

Let us consider the optimal detection threshold for the Bonferroni method in 'needle in a haystack' problems. We know from the lectures, that the considered threshold is equal to $\sqrt{2\log(n)}$, where n is a number of null hypothesis. We also know that the Bonferroni method rejects the global null when $\max_i |T_i| > \sqrt{2\log(n)}$, where T_i is the test statistic corresponding to the ith. null hypothesis.

Let us show by simulation that both these values are asymptotically equal. We will generate 100 000 independent random variables $\{X_1, \dots, X_n\}$ from a standard normal distribution. Then we will determine the trajectory of the:

$$R = \frac{\max\limits_{1<i\leq n} X_i}{\sqrt{2\log(100000)}}.$$

We will repeat the experiment 10 times.

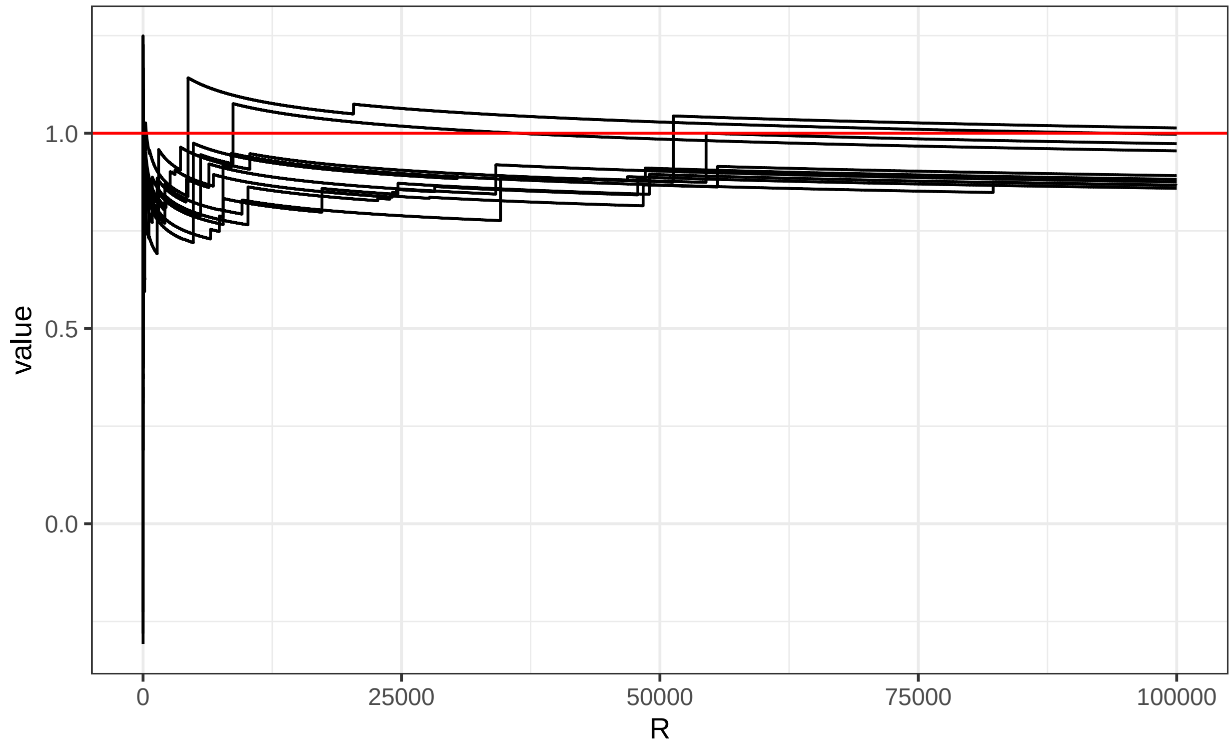

Figure 2: Plot of trajectories of R.

We see that our simulation confirms the fact:

$$\frac{\max_{i\in\{1,\ldots,n\}} X_i}{\sqrt{2\log(n)}} \to 1,$$

when $n \to \infty$.

Task 3

In this task, we consider the "needle in a haystack" problem for a normal distribution, when the expected value of the ''needle" is less than the detection threshold.

Let $\underline{X} = \{X_1, \ldots, X_n\}$, such that $X_i \sim N(\mu_i, 1)$, for $n > 1$ and $i \in \{1, \ldots, n\}$. Let us consider the global null:

$$H_0 : \forall_{i\in\{1,\ldots,n\}} \mu_i = 0$$

and the alternative:

$$H_1 : \exists!_{i\sim U\{1,\ldots,n\}} \forall_{j\neq i} \mu_i = \gamma, \mu_j = 0.$$

Assume $\gamma = (1-\varepsilon)\sqrt{2\log(n)}$. From the Neyman-Pearson lemma we have:

$$L = \frac{1}{n}\sum_{i=1}^{n} \exp(\gamma X_i - \gamma^2/2),$$

and its approximation:

$$\tilde{L} = \frac{1}{n}\sum_{i=1}^{n} \exp(\gamma X_i - \gamma^2/2)\mathbb{I}_{\{X_i<\sqrt{2\log(n)}\}}.$$

Let's check the properties of L and $\tilde{L}$ using simulation. Let us consider cases where $n \in \{10^3, 10^4, 10^5\}$ and $\varepsilon = 0.1$. The simulation consists in repeating the experiment 1000 times, generating an n-element sample from the probability distribution under H_0. For each n, we will draw histograms L and $\tilde{L}$, calculate their variances and estimate the probability $P_{H_0}(L = \tilde{L})$.

a)

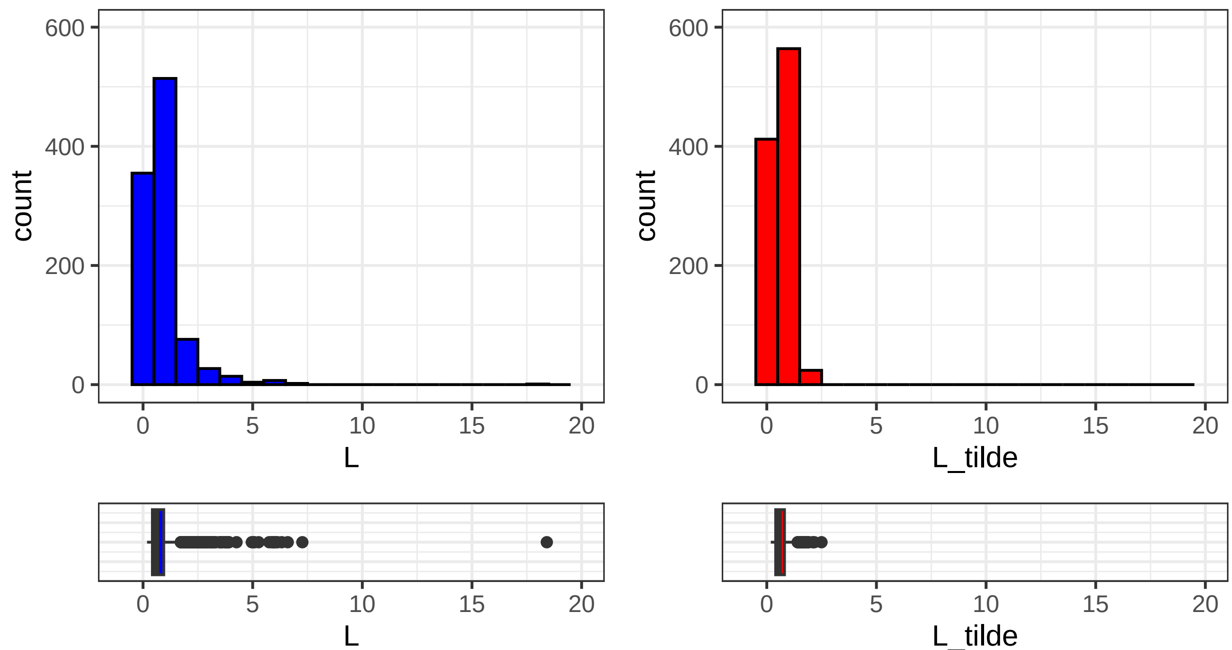

Figure 3: Histograms and boxplots of L and $\tilde{L}$ for $n = 10^3$

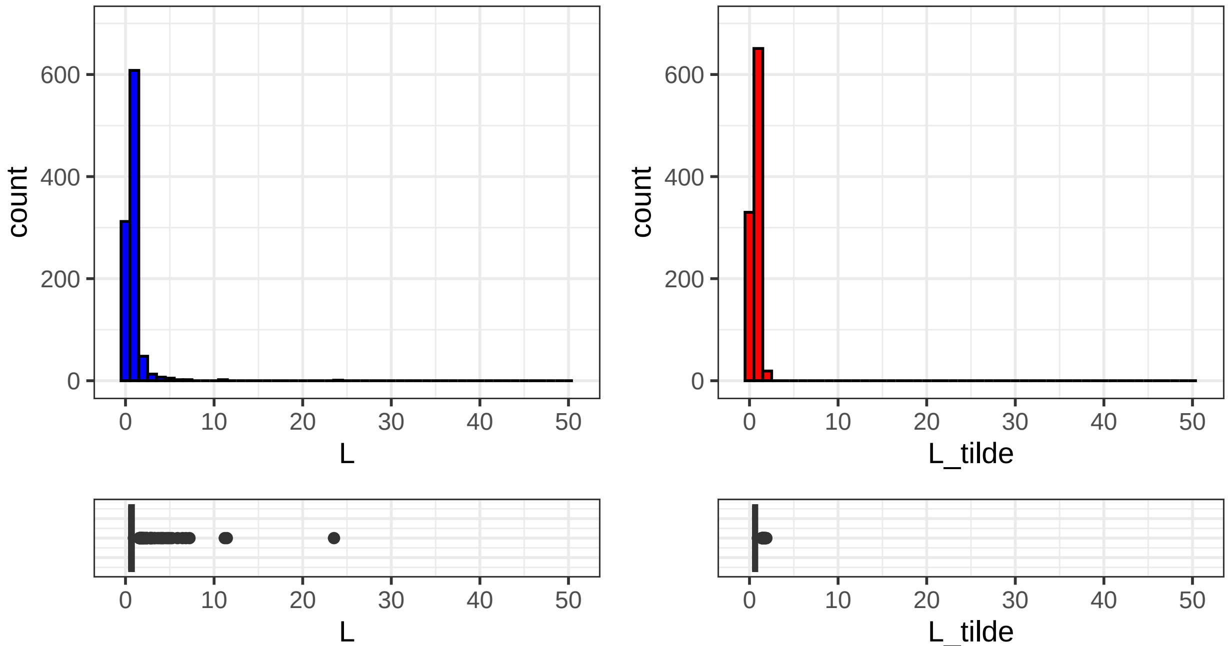

Figure 4: Histograms and boxplots of L and $\tilde{L}$ for $n = 10^4$

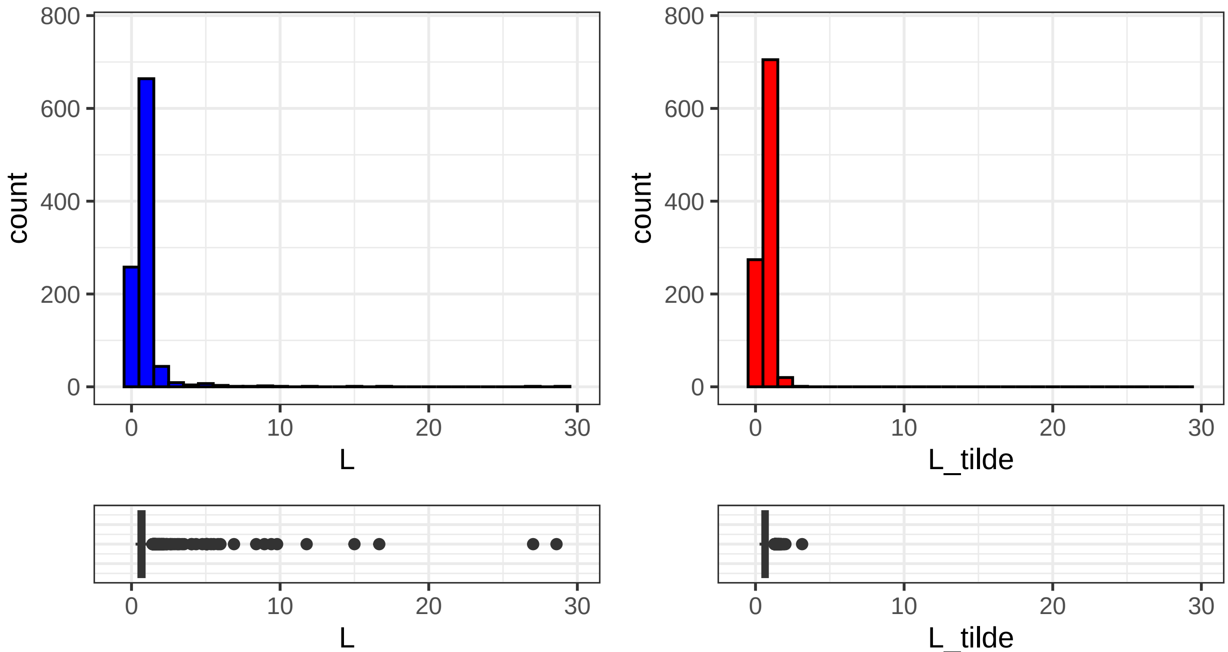

Figure 5: Histograms and boxplots of L and $\tilde{L}$ for $n = 10^5$

We can see that L and $\tilde{L}$ are similarly distributed. L compared to $\tilde{L}$ can assume an extremely high value (the indicator used in the formula for L' avoids this). However, as n increases, the distribution of both variables tends to a cumulative distribution of 1.

b)

		$\mathbb{E}(L)$	$Var(L)$	$\mathbb{E}(\tilde{L})$	$Var(\tilde{L})$	$\hat{P}_{H_0}(L=\tilde{L})$
n	10^3	1.028	13.306	0.638	0.102	0.898
	10^4	0.941	1.999	0.664	0.098	0.903
	10^5	1.115	30.748	0.692	0.077	0.912

Table 1: Table of variances, expected values and probabilities.

From the table we see that the variance and mean of L do not depend on n. As we expected, we can see that the variance of $\tilde{L}$ tends to 0 and its mean tends to 1 as n increases. This confirms the fact from the lecture that, under H_0:

$$\tilde{L} = \Phi\left(\varepsilon\sqrt{2\log(n)}\right) + o(1),$$

where:

$$\mathbb{E}_{H_0}(\tilde{L}) = \Phi\left(\varepsilon\sqrt{2\log(n)}\right),$$

$$Var_{H_0}(\tilde{L}) = o(1).$$

We also see that the probability $P\left(L=\tilde{L}\right)$ tends to 1 as n increases. It confirms the following fact:

$$P_{H_0}\left(L=\tilde{L}\right) = 1 - P_{H_0}\left(L \neq \tilde{L}\right) \geq 1,$$

where

$$P_{H_0}\left(L \neq \tilde{L}\right) \leq P_{H_0}\left(\max(X_i) \geq \sqrt{2\log(n)}\right) \longrightarrow 0.$$

Task 4

Based on the testing problem from the previous task and the significance level $\alpha = 0.05$, let us use simulation to determine the critical value of the Neyman-Pearson test and its power, and the power of the Bonferroni method, when $n \in \{500, 5000, 50000\}$, $\gamma = (1+\varepsilon)\sqrt{2\log(n)}$, $\varepsilon \in \{0.05, 0.2\}$.

We know that the Neyman-Pearson test rejects H_0 when:

$$L > k,$$

which is equivalent to:

$$\log(L) > \log(k) = k^*.$$

For the given cases, let's determine the critical value k. Let's generate a sample of 10000 L under H_0, for each n and ε. Then let's logarithmize the sample and determine its $1-\alpha$ quantile. We use this procedure because we don't know the distribution of the L and $\log(L)$. The resulting quantiles are the critical values of the considered test for individual cases.

		ε	
		0.05	0.2
n	500	0.9784	0.7910
	5000	0.7716	0.4310
	50000	0.7009	0.2732

Table 2: Critical values of the Neyman-Pearson Test.

With the critical values, we can determine the power of the Neyman-Pearson test and compare it with the power of the Bonferroni method. We will do it similarly to task 2.

		epsilon			
		0.05		0.2	
		Bonferroni Method	Neyman-Pearson's test	Bonferroni Method	Neyman-Pearson's test
n	500	0.524	0.528	0.738	0.739
	5000	0.550	0.559	0.779	0.783
	50000	0.588	0.595	0.814	0.816

Table 3: Powers of Bonferroni and Neyman-Pearson tests depending on n and ε.

From the above results, we see that the power of each test increases with n. This confirms the fact from the lecture that as n approaches to infinity, the power of the Bonferroni test and the Neyman-Pearson test tends to 1 when the signal's mean is greater than the detection threshold. Moreover, it was also confirmed that the larger the ε, the greater the test power. Whatever the case, the differences between the two tests are slight, but the Neyman-Pearson test is better in either case because it is the most powerful test. Moreover, we see again that Bonferroni method works very well in the "needle in a haystack" problem.

Task 5

a)

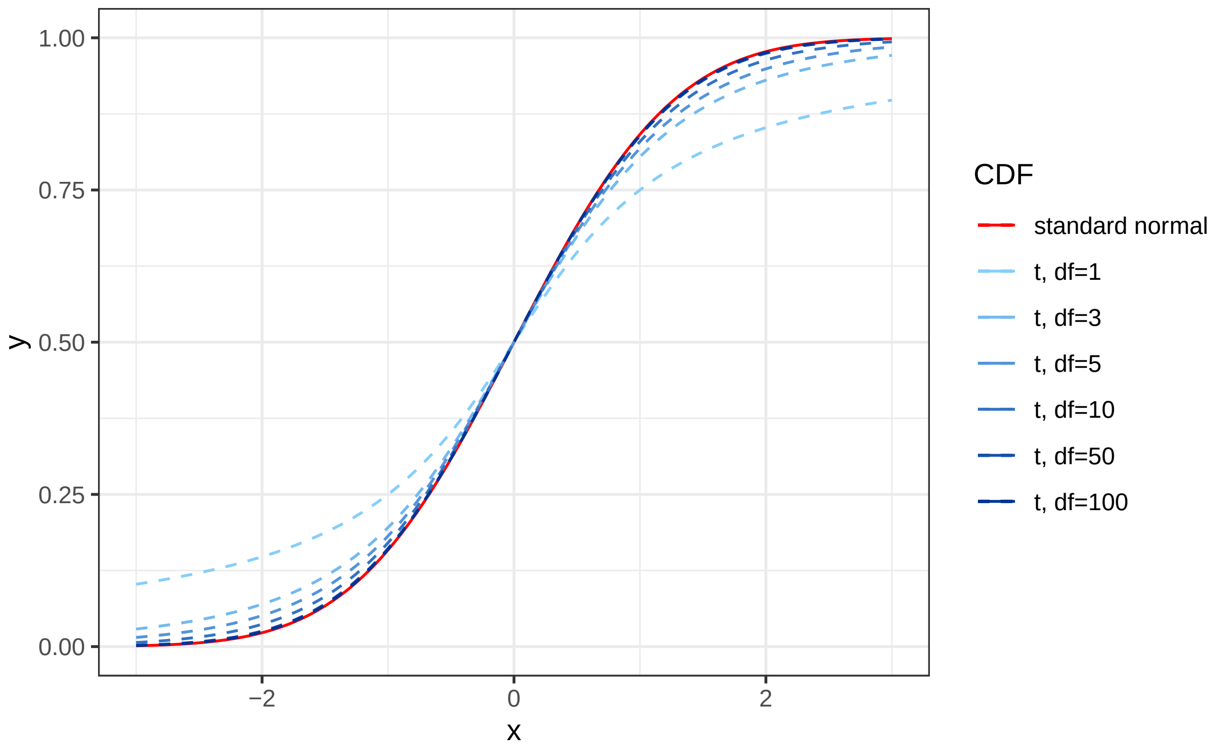

Figure 6: CDF's of standard normal distribution and Student's t-distributions.

By comparing the standard normal distribution with the student's t-distribution, we can see that the Student's t-distribution converges to the standard normal distribution as the degrees of freedom increase. With $df = 100$, the difference between the Student's t-distributions' CDF's and the CDF of standard normal distribution is unnoticeable on the plot.

b)

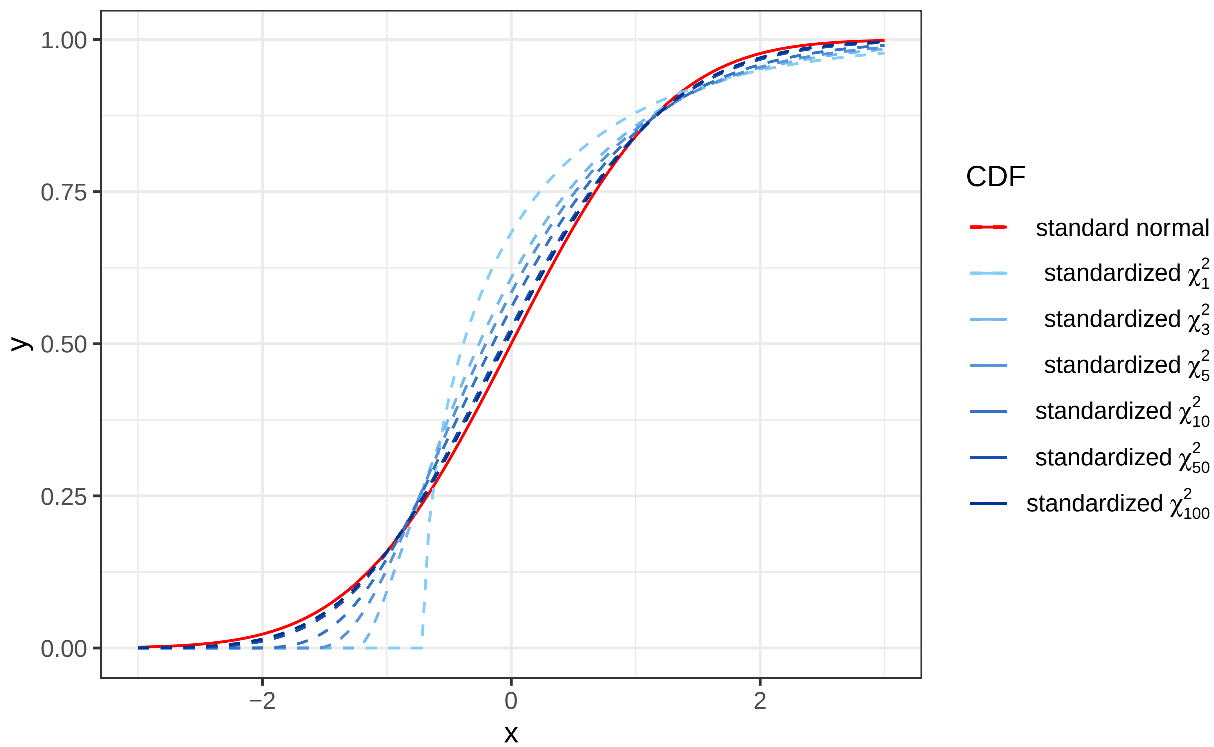

Figure 7: CDF's of standard normal distribution and standardized χ^2 distribution.

In the case of the standardized chi square distribution, we also see that as the degrees of freedom increase, it converges to the standard normal distribution. However, the convergence of the χ^2 distribution is slower compared to the Student's t-distribution.