Theoretical Foundations of Large Data Sets

List 2

Paulina Podgórska

Global testing for the expected value of the Poisson distribution (1)

Let $X_1, ..., X_n$ be the sample from the Poisson distribution. Let us consider the test for the hypothesis

$$H_0 : E(X_i) = 5 \quad vs \quad H_1 : E(X_i) > 5,$$

which rejects the null hypothesis for large values of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.Therefore out T-statistics $T = \bar{X}$.

P-value (a)

We know that if H_0 is true, $\sum_{i=1}^{n} X_i \sim Pois(n\lambda)$. With that, let us derive the formula for the p-value:

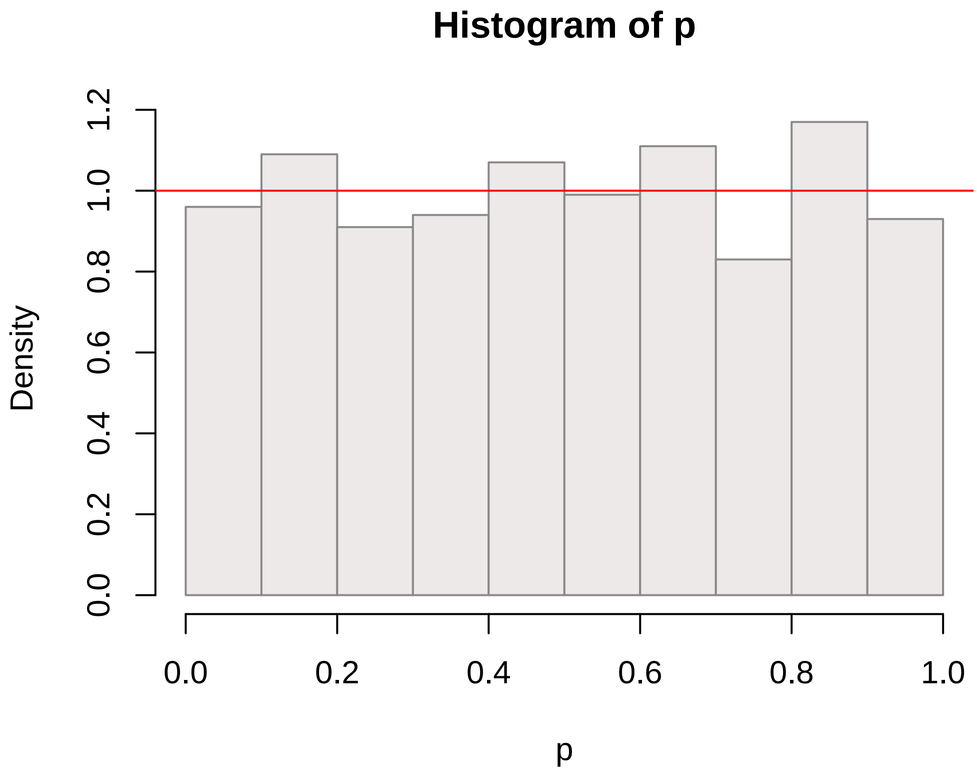
$$p = P_{H_0}\left(\frac{1}{n}\sum_{i=1}^{n} X_i > T\right) = 1 - P_{H_0}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq T\right) = 1 - P_{H_0}\left(\sum_{i=1}^{n} X_i \leq nT\right) = 1 - \Phi_{Pois(5n)}(nT).$$

We can see that nT is the observed sum $\sum_{i=1}^{n} X_i$. Therefore the function in R for calculating the p-value is as follows:

```
p_vals <- function(x){
  return( 1-ppois(sum(x), 5*100))
}
```

Calculating p-values from simulation (b)

Let us consider 1000 of the same hypothesis for $n = 100$. From that simulation we will draw histogram of p-values and discuss their distribution.



From the above histogram we can conclude that the distribution of the p-values is approximately uniform $U[0,1]$.

Bonferroni and Fisher tests of the global hypothesis (c)

Now, let us consider the meta problem of testing the global hypothesis $H_0 = \bigcap_{j=1}^{1000} H_{0j}$ and use the simulations to estimate the probability of type I error for the Bonferroni and Fisher tests at the significance level $\alpha = 0.05$.

Fisher's Combination Test rejects the null hypothesis when $T = -\sum_{i=1}^{n} 2log(p_i)$ is greater than $\chi^2_{2n}(1-\alpha)$.

Bonferroni's method rejects the null hypothesis when $\underset{1\leq i\leq n}{min}\, p_i \leq \frac{\alpha}{n}$.

Test	P(Type I Error)
Bonferroni	0.054
Fisher	0.166

We observe that the probability of type I error for the Bonferroni test aligns closely with the specified significance level $\alpha = 0.05$. On the other hand the probability of type I error in Fisher's test is higher than α. The reason for this difference might be that Fisher's test assumes that the test statistic follows a chi-squared distribution, a condition predicated on the p-values being uniformly distributed. As we saw in the previous task, our p-values only approximately follow a uniform distribution, which explains why the observed probability differs from the expected α.

Power of our test for needle in the haystack and many small effect (d)

Now let us use simulations to compare the power of the Bonferroni and Fisher tests for two alternatives:

- Needle in the haystack

$$E(X_1) = 7 \text{ and } E(X_j) = 5 \text{ for } j \in \{2, ..., 1000\}$$

Test	Power
Bonferroni	1
Fisher	0.77

In the needle in the haystack problem, Bonferroni's method outperforms Fisher's test with a power of 1, excelling at detecting a single strong signal. This showcases Bonferroni's ability in situations with one pronounced effect.

- Many small effects

$$E(X_j) = 5.2 \text{ for } j \in \{1, ..., 100\} \text{ and } \quad E(X_j) = 5 \text{ for } j \in \{101, ..., 1000\}.$$

Test	Power
Bonferroni	0.202
Fisher	0.99

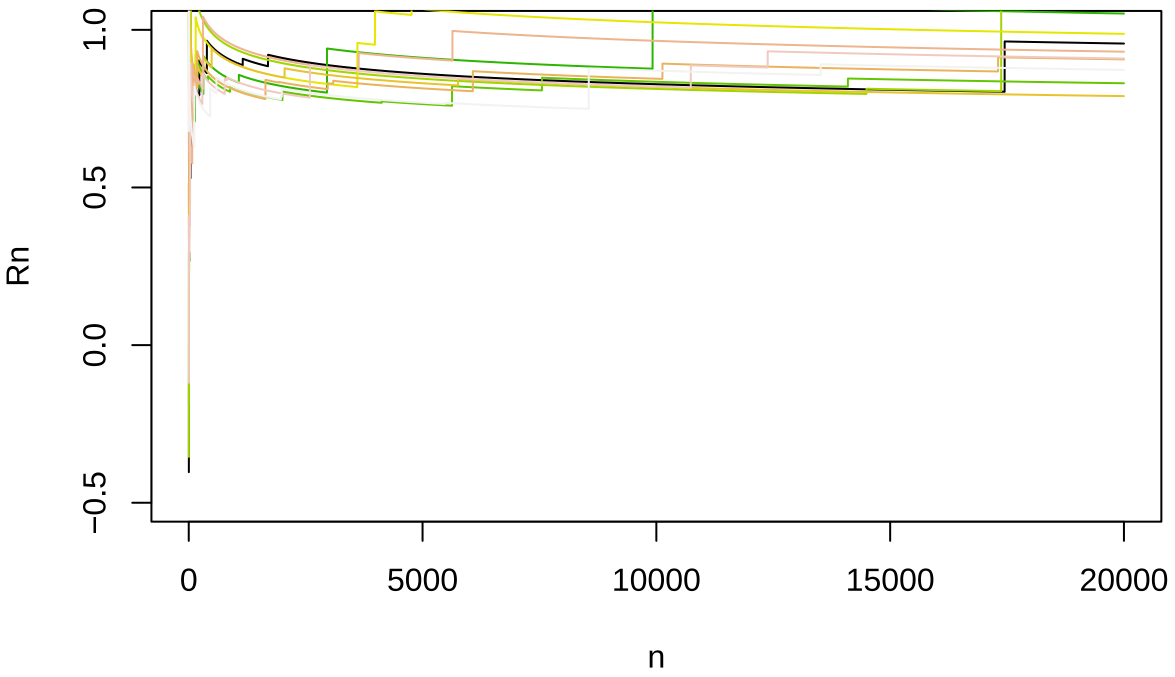
In the case of many small effects, Fisher's test achieved almost the maximum power. Bonferroni is likely too conservative in this situation, which leads to a failure in detecting true positives.

Function R_n (2)

Let $X_1, ..., X_{100000}$ be iid random variables from $N(0,1)$. For $n \in \{2, ..., 100000\}$ let us plot 10 graphs of the function

$$R_n = \frac{max\{X_i, i = 1, ..., n\}}{\sqrt{2logn}},$$

depending on the draw data.



The optimal Neyman - Pearson test (3)

Let $Y = (Y_1, ..., Y_n)$ be the random vector from $N(\mu, I)$ distribution. For the classical needle in the haystack problem: $H_0 : \mu = 0$ vs H_1 : one of the elements of μ is equal to γ, consider the statistics L of the optimal Neyman-Pearson test

$$L = \frac{1}{n}\sum_{i=1}^{n} e^{\gamma Y_i - \gamma^2/2}$$

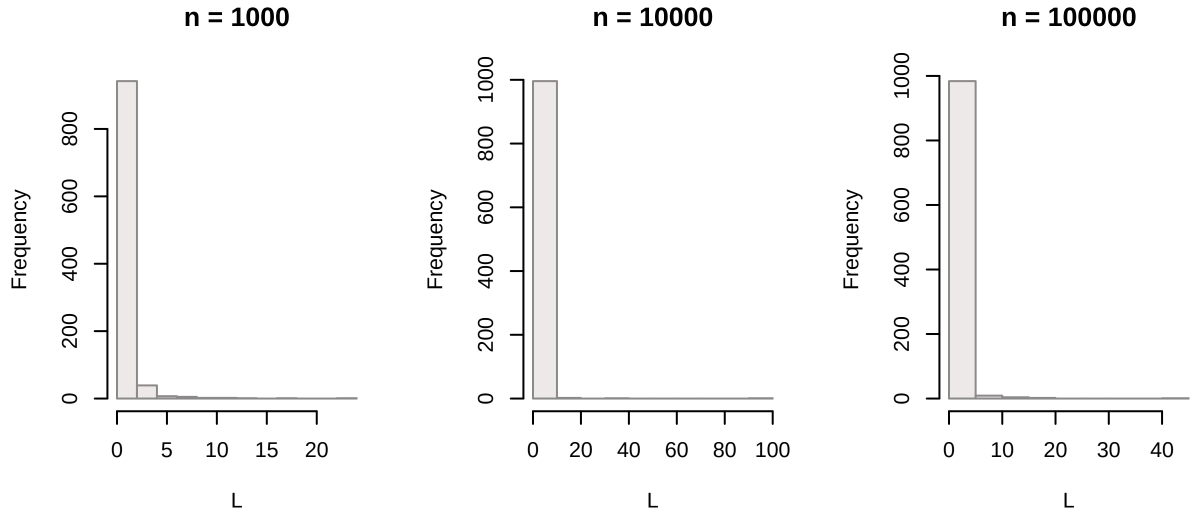
and its approximation

$$\tilde{L} = \frac{1}{n}\sum_{i=1}^{n}\left[e^{\gamma Y_i - \gamma^2}\mathbf{1}_{\{Y_i < \sqrt{2logn}\}}\right].$$

For $\gamma = (1-\epsilon)\sqrt{2logn}$ with $\epsilon = 0.1$ and $n \in \{1000, 10000, 100000\}$. We will use 1000 replicates.

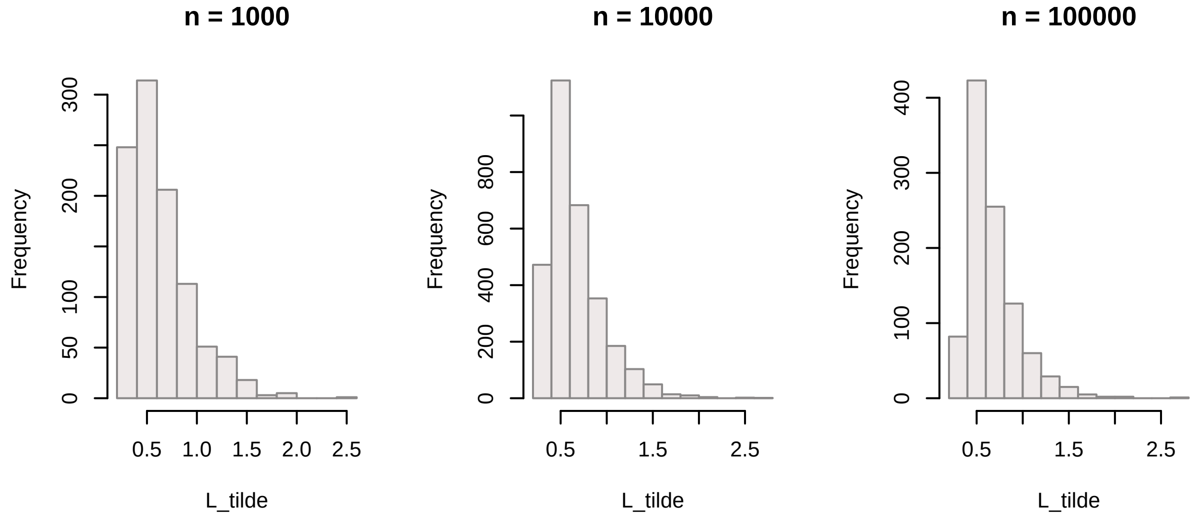
Histograms of L and $\tilde{L}$ (a)

First, let's look at the histograms of L depending on the value of n.



We do not observe significant differences between the histograms. It indicates that the value of n does not have a clear impact on the value of L. It is evident that the vast majority of L values are small.

Now, let's see histograms for $\tilde{L}$.



In this instance, we observe that the data is less spread. We can see that the values of $\tilde{L}$ range between 0 and 2. Again, we cannot determine a clear influence between n and the value of our statistic.

Variances of L and $\tilde{L}$ under null hypothesis (b)

Using our simulations lets calculate the variances of L $\tilde{L}$.

n	var(L)	var($\tilde{L}$)
1e+03	1.806	0.096
1e+04	10.227	0.090
1e+05	3.482	0.076

We observe that the variance of $\tilde{L}$ is significantly lower. We can also notice that with the increase of n the variance decreases. In case of L, the variances don't seem to be affected by the value of n – we record the highest variance for $n = 10000$.

Estimation of $P_{H_0}(L = \tilde{L})$ (c)

n	$P(L = \tilde{L})$
1e+03	0.914
1e+04	0.928
1e+05	0.923

We can compare the calculated probabilities to the theoretical ones. From the lecture we know, that the probability $P(L \neq \tilde{L})$ approaches 0 as $n \to \infty$:

$$P(L \neq \tilde{L}) \leq P(\underset{q \leq i \leq n}{max}\, y_i \geq T_n) \leq \sum_{i=1}^{n} P(y_I \geq T_n) \leq \sum_{i=1}^{n} \frac{1}{T_n} \frac{1}{2\pi} e^{-\frac{T_n^2}{2}} = \frac{1}{\sqrt{2\pi}} \frac{n \cdot \frac{1}{n}}{Tn} \underset{n \to \infty}{\to} 0.$$

From this, we can conclude that the probability $P(L = \tilde{L}) = 1 - 0 = 1$. It is evident that as n increases, our calculated probability gets closer to the theoretical one.

Critical value of the optimal Neyman-Pearson test (4)

In this section we will use simulations to find the critical value of the optimal Neyman-Pearson test and compare the power of this test and the Bonferroni test for the "needle in the haystack" problem with $n \in \{500, 5000, 50000\}$ and the needle $\gamma = (1 + \epsilon)\sqrt{2logn}$ with $\epsilon \in \{0.05, 0.2\}$.

The calculated critical values c:

- $\epsilon = 0.05$

n	c
500	0.791
5000	0.696
50000	0.592

- $\epsilon = 0.05$

n	c
500	0.589
5000	0.446
50000	0.411

We observe that for both values of ϵ, the critical values decrease as n increases. This indicates that with larger sample sizes, our tests become more strict in rejecting the null hypothesis which makes it less likely to incorrectly reject the null hypothesis. We can also observe a relation between the critical value c and the value of ϵ – with $\epsilon = 0.2$ we achieve significantly lower critical value.

Now, let us compare the ower of the Bonferroni test ad the optimal Neyman-Pearson test.

- $\epsilon = 0.05$

n	Power Bonf.	Power N-P
500	0.437	0.488
5000	0.482	0.520
50000	0.500	0.541

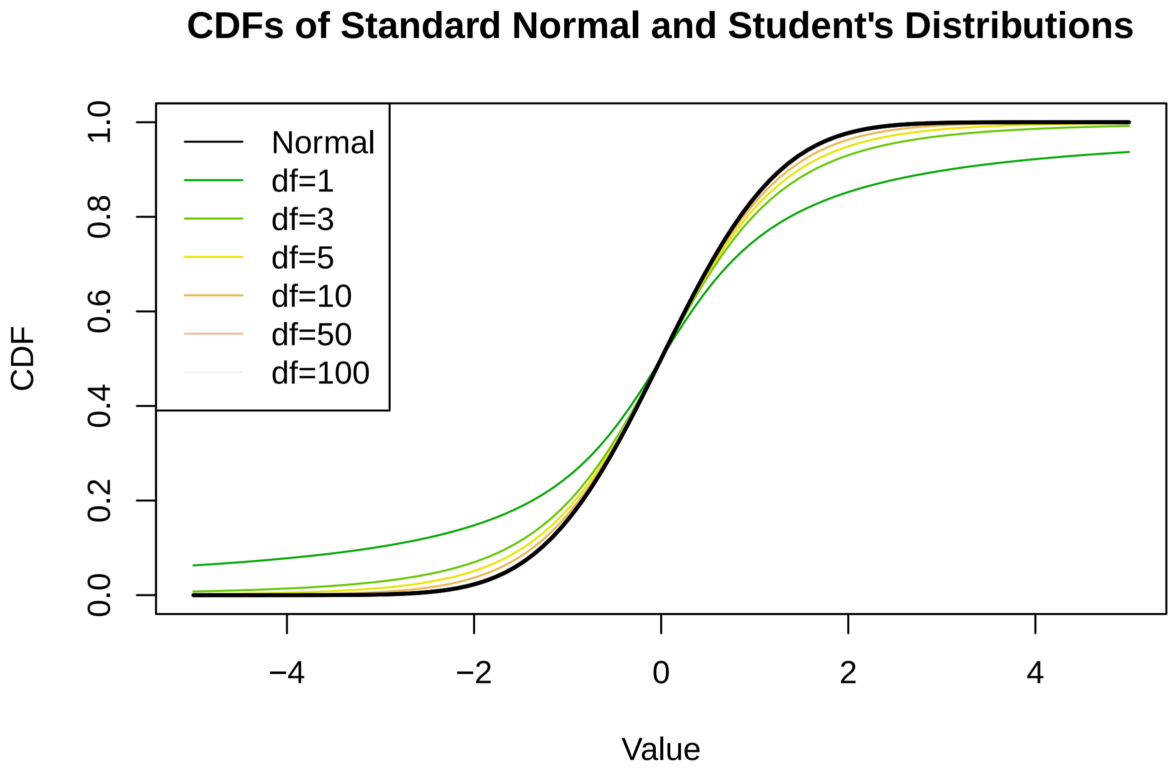
- $\epsilon = 0.2$

n	Power Bonf.	Power N-P
500	0.626	0.688
5000	0.699	0.751
50000	0.762	0.794

In every examined scenario, the Neyman-Pearson test consistently achieves the highest power, with the observed difference being approximately 0.05. Across both values ofϵ, we notice the that the power of tests increases with the growth of n. As the size of the needle increases (bigger ϵ), it's easier to find it. Correspondingly, as expected, a larger sample size leads to improved outcomes, even achieving a power of 0.8.

Comparison of the distributions (5)

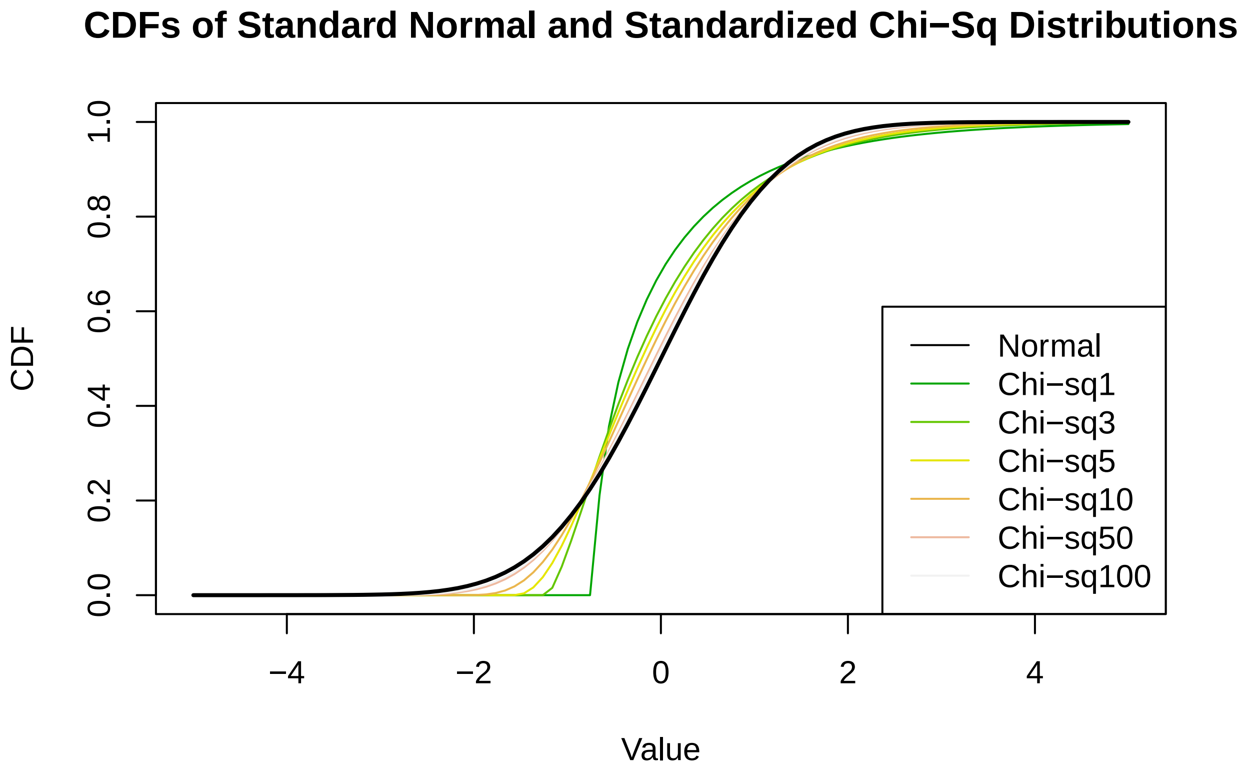
Let us draw one graph with cdfs of the standard normal distribution and the Student's distribution with degrees of freedom $df \in \{1, 3, 5, 10, 50, 100\}$.



We observe that with increasing degrees of freedom, Student's distributions converge towards the standard normal distribution. Especially for $df = 50$ and $df = 100$, we notice that Student's distributions closely approximate the distribution of $N(0,1)$. In contrast, the graph for $df = 1$ shows the greatest difference from the normal distribution – which is easily observed in the tails.

Now, let's draw a graph with cdfs of the standard normal distribution and the standardized chi-square distribution with the same degrees of freedom. The standardization is of the form

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



With increasing degrees of freedom, the standardized chi-square distribution more closely approximates the standard normal distribution, especially evident in the tails, contrasting with the Student's distribution which aligns more in the center. The chi-square distribution with 100 degrees of freedom nearly coincides with the normal distribution.