Theoretical Foundations of the Analysis of Large Data Sets

Report 2

Magdalena Potok

Prepared on:
November 22, 2024

Excercise 1

Let $X_1, \ldots, X_n$ be a sample from the Poisson distribution. We consider a test for the hypothesis

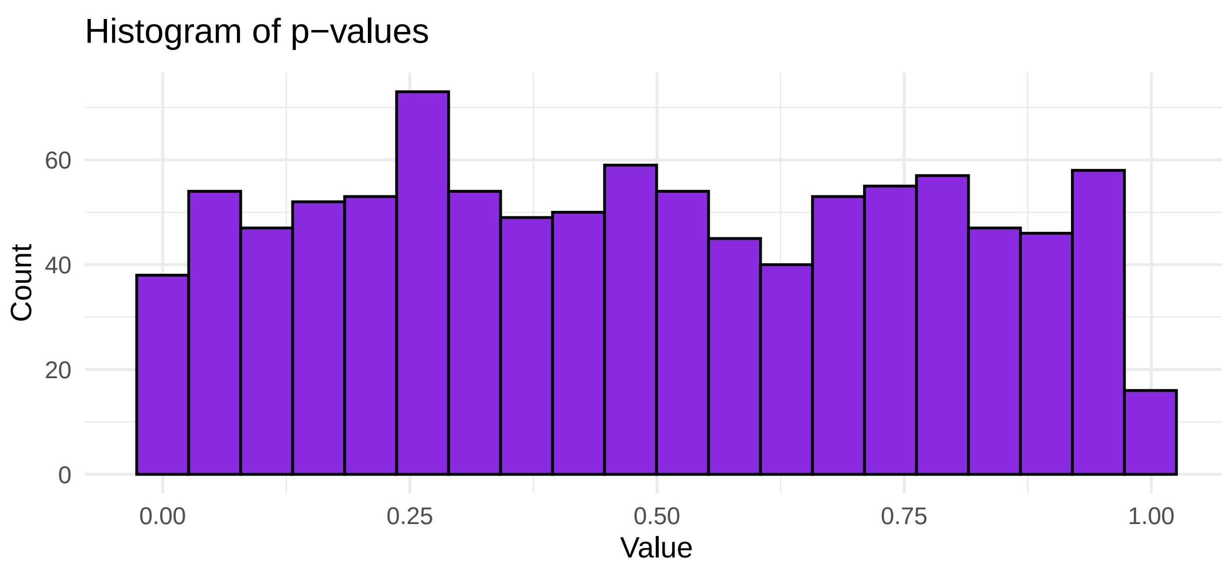
$$H_0 : \mathbb{E}(X_i) = 5, \quad vs \quad H_1 : \mathbb{E}(X_i) > 5,$$

which rejects the null hypothesis for large values of $\hat{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The p-value of this test can be calculated using the formula:

$$\mathbb{P}(T > \hat{X}) = \mathbb{P}(T > \frac{1}{n}\sum_{i=1}^{n} X_i) = \mathbb{P}(nT > \sum_{i=1}^{n} X_i) = \mathbb{P}(Y > \sum_{i=1}^{n} X_i), \text{ where } Y \sim Poiss(5n).$$

```
cal_pval = function(n){
  x = rpois(n, 5)
  p_val = 1 - ppois(sum(x), n*5)
  return(p_val)
}
```

Setting $n = 100$ for this function we calculate the p-value as 0.435.
Next, we consider 1000 repetitions of the same hypothesis test with $n = 100$ and calculate the p-values. The results are presented in a histogram.



As shown, the distribution of p-values does not follow a uniform distribution, as is typically expected under the null hypothesis. However, in the case of the Poisson distribution, p-values from discrete distributions exhibit uniform behavior only asymptotically. This characteristic is reflected in the histogram.
We then address the meta-problem of testing $H_0 = \cap_{j=1}^{1000} H_{0j}$ using simulations to estimate the type I error probability for Bonferroni and Fisher tests at a significance level of $\alpha = 0.05$.

	Bonferroni	Fisher
Error rate	0.054	0.18

Both results are expected to be close to the specified significance level. As observed, Bonferroni's method more accurately approximates this value compared to Fisher's method. The Fisher statistic, $T = -\sum_{i=1}^{n} 2logp_i$, is designed for continuous distributions. Since our distribution is discrete and the p_i values (as we observed above) are not uniformly distributed, the distribution of the test statistic cannot be derived accurately, and the probability of type I error may deviate from expectations.

We will use simulations to compare the power of the Bonferroni and Fisher test for two alternatives:

- Needle in the haystack

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

- Many small effects

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

	Needle in the haystack	Many small effect
Bonferroni	1.000	0.186
Fisher	0.773	0.986

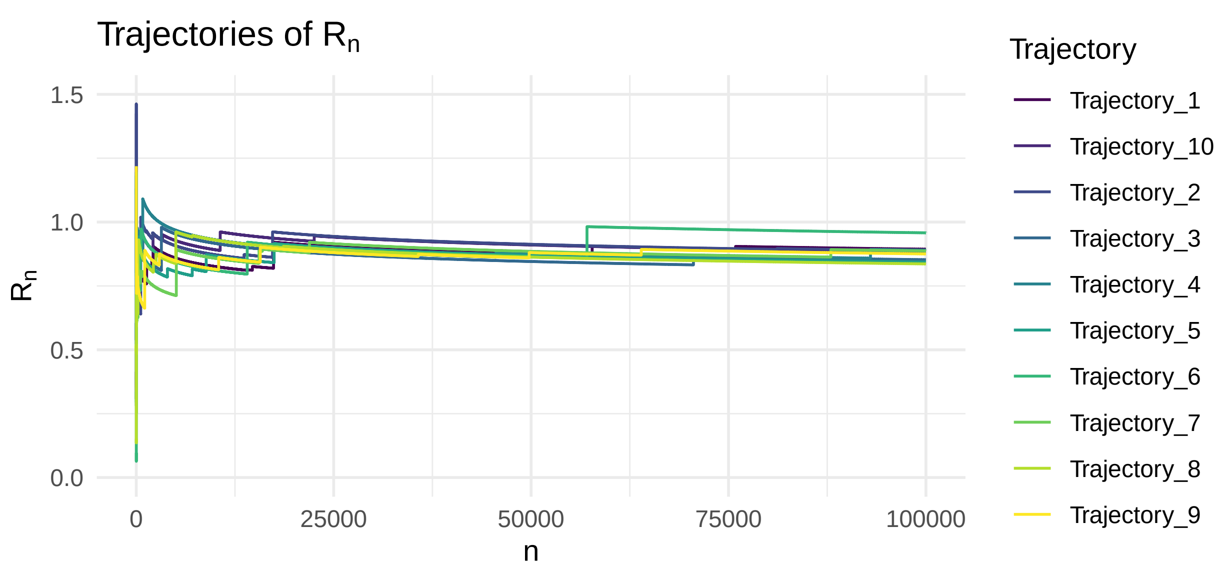
From the results, we observe that Bonferroni's method is more powerful in the "needle in the haystack" scenario. This test focuses on the smallest p-value, making it well-suited for detecting cases where at least one p-value is significant. However, it is less effective in detecting distributed small effects. Conversely, Fisher's method aggregates all p-values, which enhances its power in scenarios with many small effects but reduces its efficacy in cases like the needle in the haystack problem.

Exercise 2

Let $X_1, \ldots, X_{100000}$ be iid random variables from $N(0,1)$ For $n \in \{2, \ldots, 100000\}$ then we can calculate function

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

We will repeat the above experiment 10 times and plot the respective trajectories of R_n.



Bonferroni method rejects the global null when the smalles $p_i \leq \alpha/n$, we can equivalently check if $maxX_i > z_{\alpha/n}$. As n becomes large $z_{\alpha/n}$ behaves asymptotically as $z_{\alpha/n} \approx \sqrt{2logn}$. This means we reject when

$maxX_i > \sqrt{2logn}$, making the rejection threshold asymptotic to $\sqrt{2logn}$. In this plot, we divide $maxX_i$ by $\sqrt{2logn}$, resulting in the trajectories of R_n. As shown, for large n R_n is always below value 1

$$\frac{maxX_i}{\sqrt{2logn}} < 1 \Rightarrow maxX_i < \sqrt{2logn},$$

this means we do not reject the global null hypothesis H_0.
In the excercise 4, we assume that our needle equals a little bit more, than the rejection threshold $\sqrt{2logn}$. This means that as n will increase, the power of the test will increase too.

Exercise 3

Let $Y = (Y_1, \ldots, Y_n)$ be the random vector from $N(\mu, I)$ distribution. For the classical needle in haystack problem

$$H_0 : \mu = 0 \quad vs \quad H_1 : \text{one of the elements of } \mu \text{ is equal to } \gamma.$$

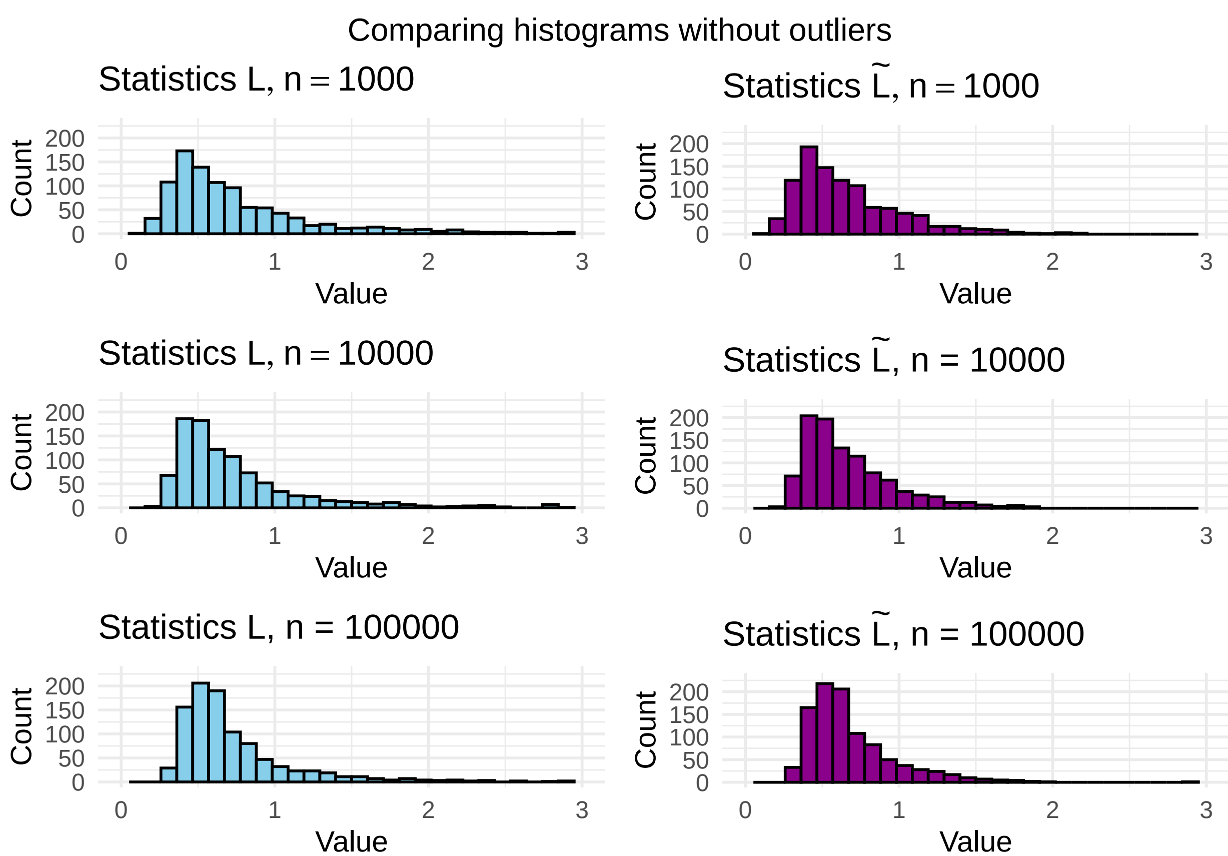
We consider the statistics L of the optimal Neyman-Pearson test

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

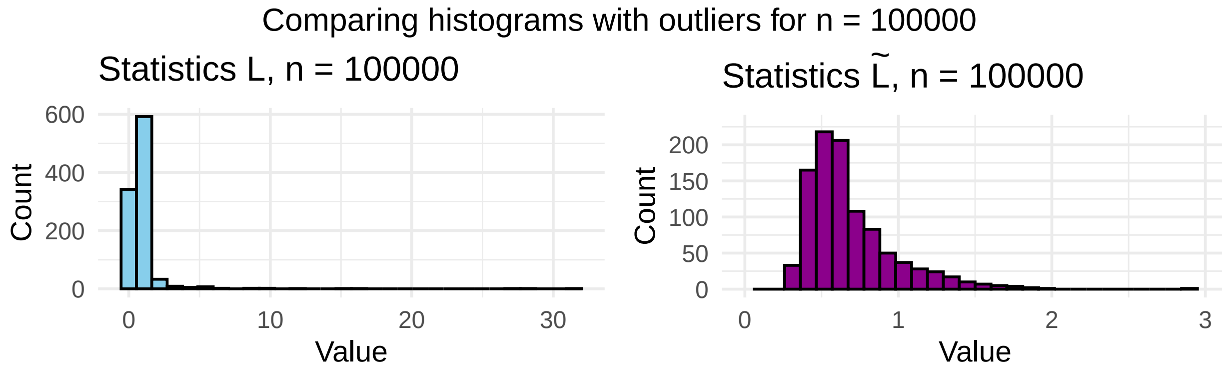
and its approximation

$$\tilde{L} = \frac{1}{n} \sum_{i=1}^{n} e^{\gamma Y_i - \gamma^2/2} \mathbb{1}_{(Y_i < \sqrt{2logn})}.$$

For $\gamma = (1 - \epsilon)\sqrt{2logn}$ with $\epsilon = 0.1$ and $n \in \{1000, 10000, 100000\}$ we will use 1000 replicates to study propeties of L and $\tilde{L}$ statistics.



For better comparison of histograms, the x-axis of L was restricted to a specific range to highlight the similarities between L and its approximation in their aggregation. The histogram of $\tilde{L}$ is showing a much more concentrated spread compared to L, especiallly when outliers are included. However, when outliers are excluded and under the null hypothesis, the histograms of L and $\tilde{L}$ appear similar.



When outliers are included, the histogram for L demonstrates a long tail, indicating a presence of extreme values. In contrast, $\tilde{L}$ shows a more concentrated spread. Comparision for $n = 1000, n = 10000$ and $n = 100000$ looks the same.
Properties of L and $\tilde{L}$ under the null hypothesis are shown in the table below. Theoretical probability is calculated from formula:

$$\mathbb{P}(\tilde{L} \neq L) \leq \mathbb{P}(max_j X_j > T_n) \leq \sum_{j=1}^{n} \frac{\phi(T_n)}{T_n} = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2logn}} \to 0, n \to \infty$$

$$\mathbb{P}(\tilde{L} = L) = 1 - \mathbb{P}(\tilde{L} \neq L) \geq 1 - \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2logn}} \to 0, n \to \infty$$

	$var(L)$	$var(\tilde{L})$	$\mathbb{P}_{H_0}(L = \tilde{L})$	Theoretical prob.
$n = 1000$	1.751	0.116	0.915	0.893
$n = 10000$	1.601	0.088	0.919	0.907
$n = 100000$	3.696	0.089	0.933	0.917

Variances of L are significantly higher than that of $\tilde{L}$ across all sample sizes n, as was evident in the histograms. Furthermore, the variance of $\tilde{L}$ decreases as n increases, indicating a stabilization effect under the null hypothesis. The computed probability $\mathbb{P}_{H_0}(L = \tilde{L})$ closely matches the theoretical bound, converging to 1 as $n \to \infty$.

Exercise 4

Using simulations, we determine the critical value of the optimal Neyman-Pearson test and compare its power to that of the Bonferroni test for the needle in the haystack problem. The analysis is conducted for $n \in \{500, 5000, 50000\}$ and the needle $\gamma = (1 + \epsilon)\sqrt{2logn}$ with $\epsilon \in \{0.05, 0.2\}$.
In the Neyman-Pearson test, we reject H_0 when $L = \frac{1}{n} \sum_{i=1}^{n} e^{\gamma Y_i - \gamma^2/2} > c$, where c is the critical value the critical value chosen such that the probability of rejecting H_0 under the null hypothesis does not exceed the significance level of the test. Equivalently, when the statistic L is too complicated, we reject H_0 when $log(L) > log(c) = c'$. Since the likelihood ratio and the log-likelihood ratio do not follow a standard distribution, simulations are used to determine the critical value of the test. For each n and ϵ we generate 1000 log-likelihood ratios and coompute the $1 - \alpha$ to get the critical value. The results are summarized in the table below.

	$n = 500$	$n = 5000$	$n = 50000$
$\epsilon = 0.05$	0.9871388	0.8525299	0.6613813
$\epsilon = 0.2$	0.7133681	0.5443617	0.3014625

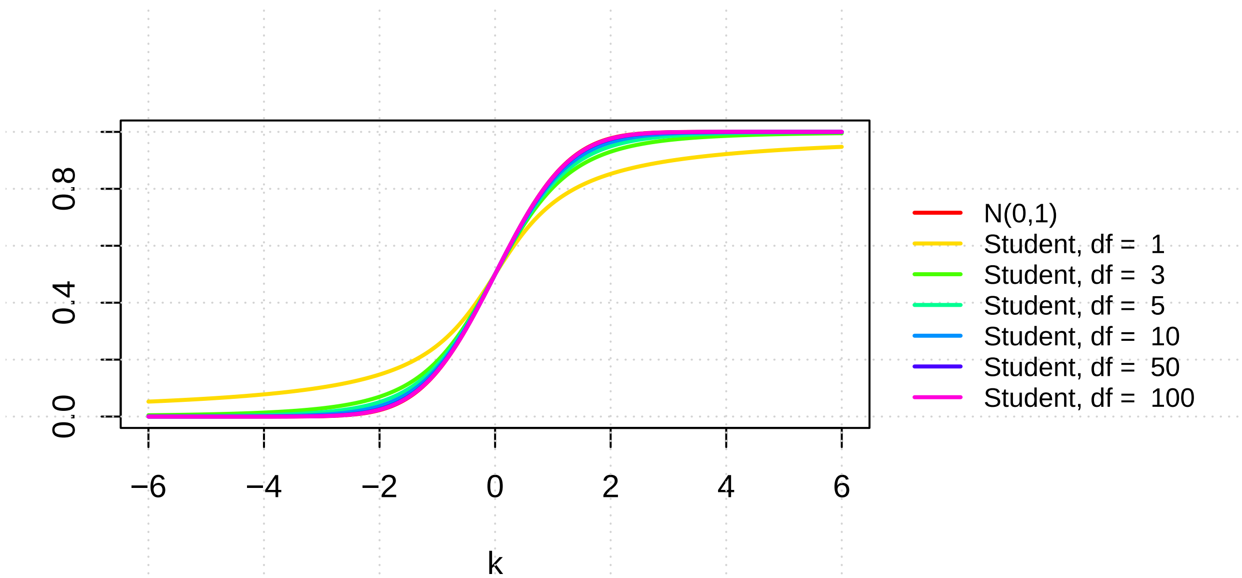
For each combination n and ϵ we perform a random sample test to compare the power of the Neyman-Pearson test against the Bonferroni test.

	Power of Neyman-Pearson test			Power of Bonferroni test		
	$n = 500$	$n = 5000$	$n = 50000$	$n = 500$	$n = 5000$	$n = 50000$
$\epsilon = 0.05$	0.6895	0.7100	0.7457	0.4536	0.4935	0.5174
$\epsilon = 0.2$	0.8219	0.8593	0.9022	0.6433	0.7156	0.7653

The results demonstrate that the power of the Neyman-Pearson test increases with the number of observations n. Additionally, larger values of ϵ, which correspond to a more significant γ, also enhance the power of the test. The Bonferroni test consistently exhibits lower power compared to the Neyman-Pearson test for all values of n and ϵ. However, its power also increases with n and ϵ, following a trend similar to the Neyman-Pearson test. This result is reasonable, as we mentioned that in excercise 2.

Exercise 5

We compare the cumulative distribution functions (CDFs) for the standard normal distribution and Student's t-distribution for degrees of freedom $df \in \{1, 2, 5, 10, 50, 100\}$.

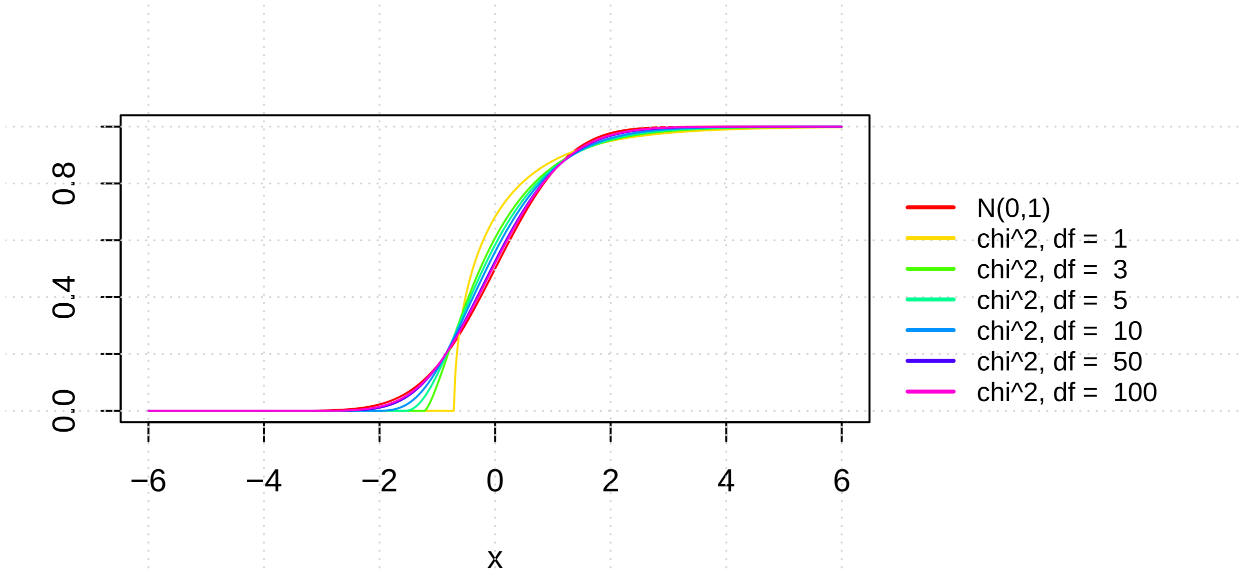


As the degrees of freedom increase, the Student's t-distribution converges to the standard normal distribution. When $df = 100$ the difference between CDFs is negligible. This result aligns with the theoretical fact that as $df \to \infty, \quad t_{df} \to N(0, 1)$.

Next, we compare the CDFs of the standard normal distribution with those of the standardized chi-squared distribution for $df = \{1, 3, 5, 10, 50, 100\}$. The standarization is defined as $T = \frac{\chi^2_{df} - df}{\sqrt{2df}}$. Since the distribution of T is not standard, we derive its CDF using the following modification:

$$\mathbb{P}\left(\frac{\chi^2 - df}{\sqrt{2df}} < k\right) = \mathbb{P}\left(\chi^2 < k\sqrt{2df} + df\right) = F_{\chi^2}(k\sqrt{2df} + df),$$

where F_{χ^2} is the CDFs of the chi-squared distribution.



As the degrees of freedom increase, the chi-squared distribution converges to the standard normal distribution. This convergence is a consequence of the central limit theorem. When $df = 100$ the difference is almost invisible, but the convergence is slower than for the student distribution.