David Bieniakowski, Luke Halecki, and Tyler Pick

Dr. P. Brewer

Statistical Inference

December 7th, 2022

R-Assignment 5

Executive Summary and Introduction

Following from assignment 4 in which we computed a single p-value using the Neyman Pearson likelihood ratio for the given hypothesis test, we are now tasked with creating a distribution of many p-values and displaying the p-values in a histogram that appears left skewed. To do so, we simulate 100,000 values of R, the Neyman Pearson likelihood ratio, and follow by doing the same for 10,000 values of the given null hypothesis. From these 10,000 values we simulated, we estimated a p-value to create a histogram that appears left skewed with most values at p-value of .3 and just under .7. Then we redo this process with the null hypothesis and find similar results.

Data

The first thing we did was define several constants that we would use throughout the project, including:

- “lam_null” = .1, the null lambda value.

- “lam_alt” = 1, the alternative lambda value.

- “n” = 10, the number of samples.

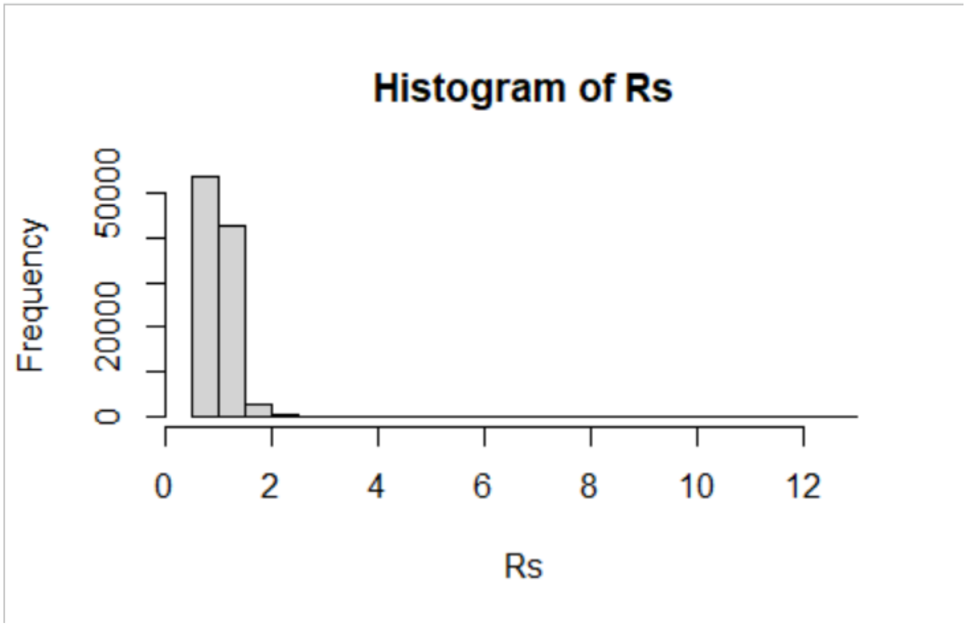
- “trials” = 10, the number of trials per sample for the binomial distribution.

- “num_R” = 100,000, the number of Neyman Pearson ratios, which we call R, we want to compute.

We used a for loop to create the many Neyman Pearson ratios ranging from 1 to “num_R”. In the loop, we simulated our own data in the form of a binomial distribution with the number of trials (trials) = 10 and probability of a success (lambda null) = .1 because we assumed the null hypothesis was true. From the data, we compute a Neyman Pearson ratio and append it to our empty matrix of R’s using the rbind() function.

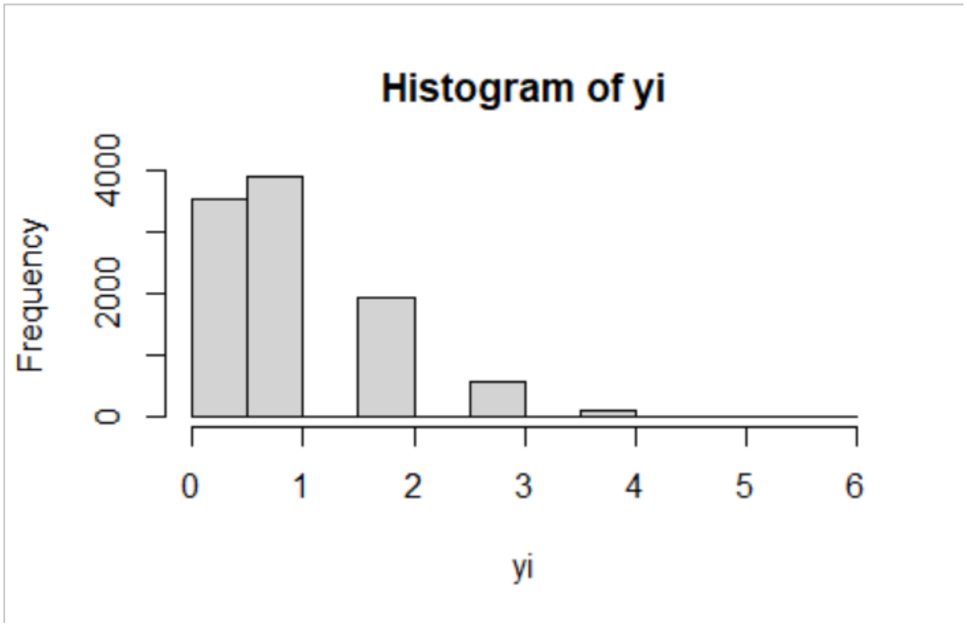
Next, we try to understand the distribution of R’s by printing the maximum of them and plotting a histogram. The values of the R’s change with each iteration on the code but the maximum we computed was about 12.8 and the histogram is approximately distributed like figure 1 below.

Figure 1:



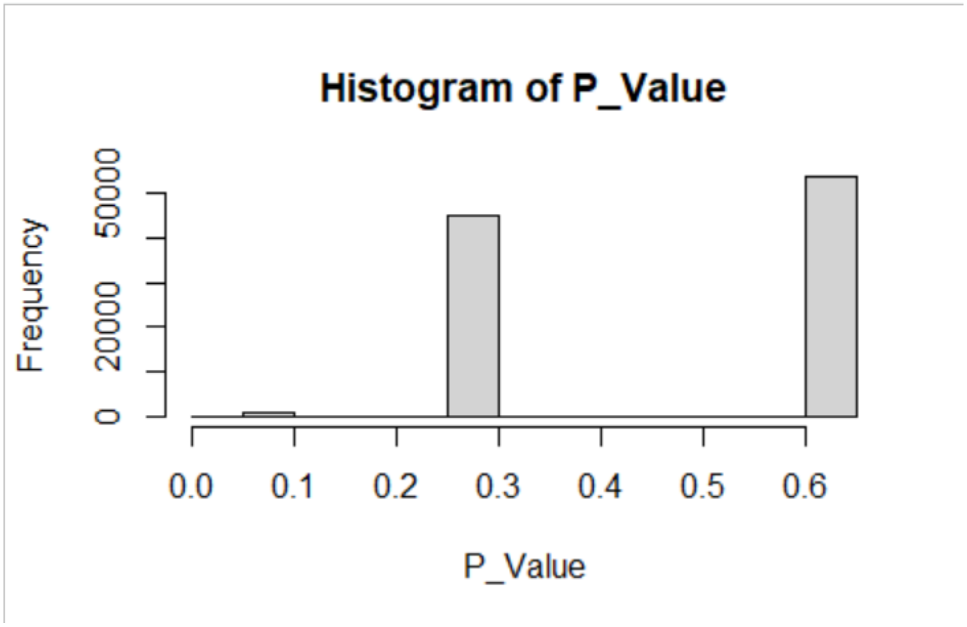
Next, we worked to compute the P-Values of the Rs in relation to the value 2.56 which was given in R-Assignment 4. Although it wasn't necessary, we found the probability that any given R from our distribution was over the 2.56 threshold to find about 0.28% were greater. We created a distribution based on the null hypothesis called "yi" with 10,000 samples, the distribution is shown in Figure 2 below.

Figure 2:



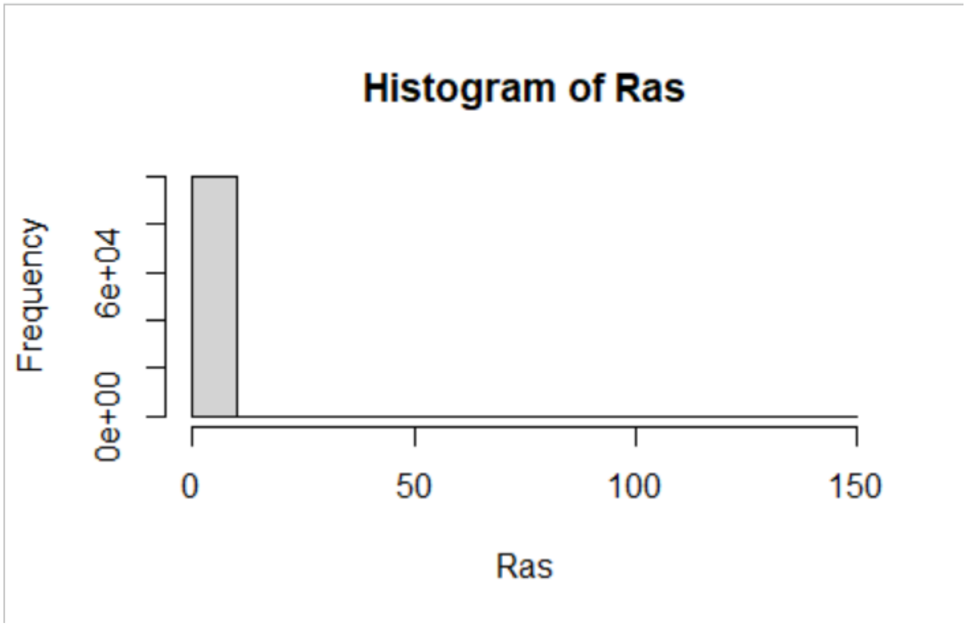
Like we did for the Rs, we create an empty matrix called P_Value and run a for loop to append values to it. To do so, we create a temporary variable in the loop names “l” which sums the yis from the binomial distribution that are greater than a given R and divide by the total samples from the binomial distribution. The distribution of the P-Values is displayed below in Figure 3. Most of the values are grouped together at 0.3 and 0.7 which proves our Neyman Pearson ratios are not accurate at determining the P-Value.

Figure 3:



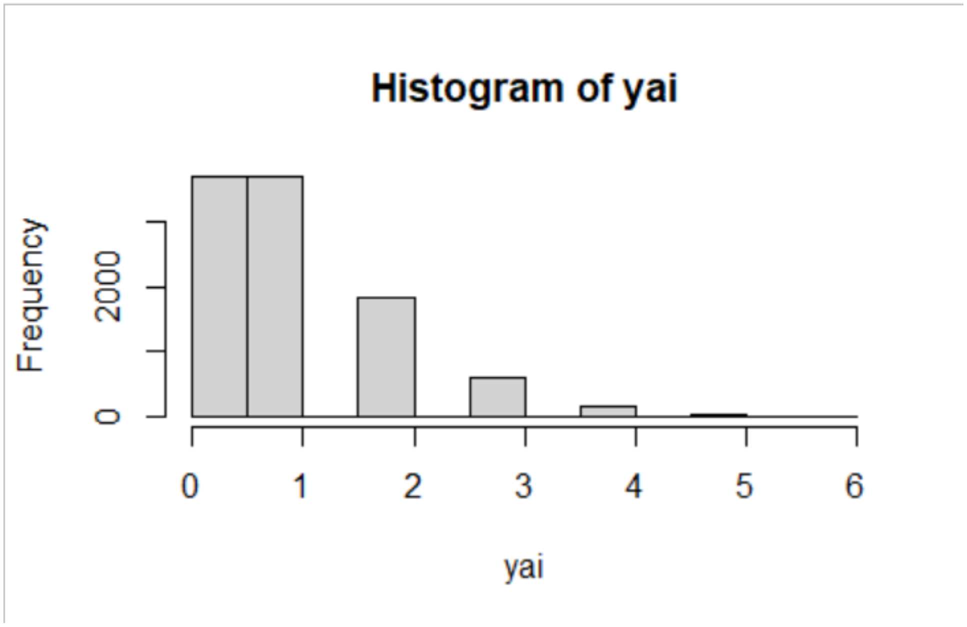
In the last part of the problem, we are asked to repeat the procedure but assuming the alternative hypothesis is true, so the distributions from which we derive the ratios and P-Values will be poisson instead of binomial. The code is the same but with slight differences in names, values, and graphs. We rename “Rs” to “Ras”, the binomial “xi” to the poisson “xai”, the binomial “yi” to the poisson “yai”, and “P_Value” to “P_Value_a”. Now, our maximum Ra is about 150, the probability of any given Ra being greater than 2.56 is 1.05% and the distribution of Ras is shown in Figure 4. The distribution of Ras looks like the distribution of Rs in Figure 1 but the x-axis reaches higher values.

Figure 4:



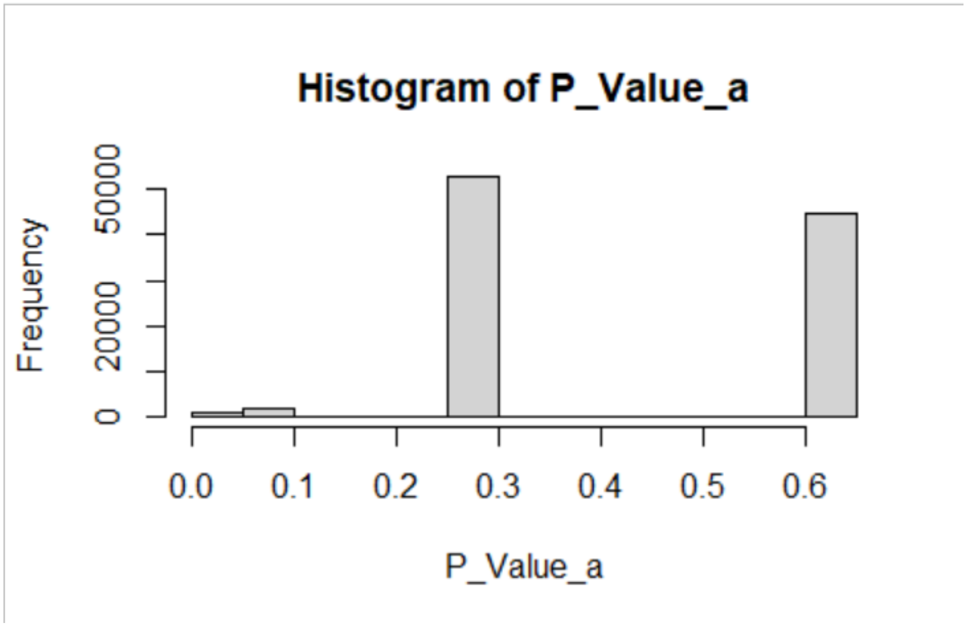
Then the distribution of "yai" of Figure 5 is close to a poisson distribution because that is what it is meant to represent.

Figure 5:



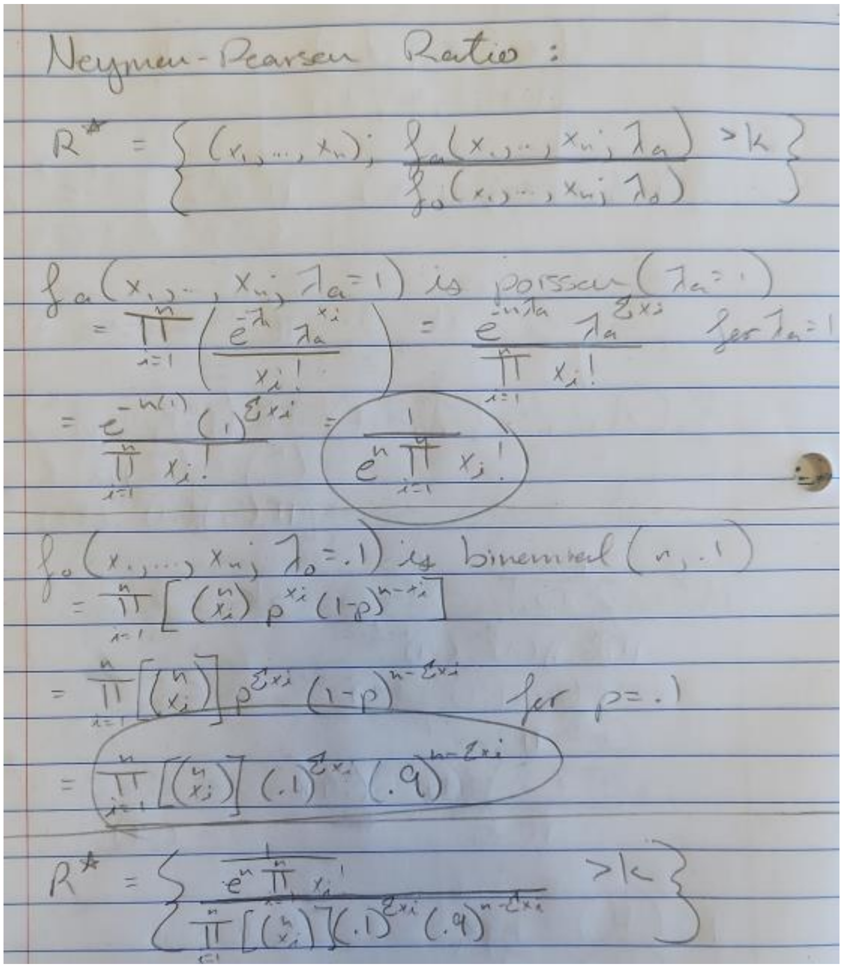
Finally, the distribution of P-Values for the alternative hypothesis is like the distribution of P-Values for the null hypothesis but with values gathered at 0.3 and 0.6. The distribution is shown in Figure 6 below.

Figure 6:



Analysis

For this project, we just needed to reuse our Neyman Pearson formula for the previous assignment. The method of computing the Neyman Pearson ratio is shown in Figure 7. We needed to formulate the joint probability mass function (pmf) for both the null and alternative hypotheses given their respective parameters. The alternative hypothesis is the numerator of our ratio and is the product of n random variable from a poisson distribution with lambda=1.The null hypothesis and the denominator of our ratio, with p=.1. Then, if the data was closely related to the alternative hypothesis, the ratio would be large, so large values of k would reject the null hypothesis.



Neyman-Pearson Ratio :

$$R^* = \left\{ (x_1, \dots, x_n) ; \frac{f_a(x_1, \dots, x_n ; \lambda_a)}{f_0(x_1, \dots, x_n ; \lambda_0)} > k \right\}$$

$f_a(x_1, \dots, x_n ; \lambda_a = 1)$ is poisson $(\lambda_a = 1)$

$$= \prod_{i=1}^{n} \left(\frac{e^{-\lambda_a} \lambda_a^{x_i}}{x_i!} \right) = \frac{e^{-n\lambda_a} \lambda_a^{\Sigma x_i}}{\prod_{i=1}^{n} x_i!} \quad \text{for } \lambda_a = 1$$

$$= \frac{e^{-n(1)} (1)^{\Sigma x_i}}{\prod_{i=1}^{n} x_i!} = \boxed{\frac{1}{e^n \prod_{i=1}^{n} x_i!}}$$

$f_0(x_1, \dots, x_n ; \lambda_0 = .1)$ is binomial $(n, .1)$

$$= \prod_{i=1}^{n} \left[\binom{n}{x_i} p^{x_i} (1-p)^{n-x_i} \right]$$

$$= \prod_{i=1}^{n} \left[\binom{n}{x_i} \right] p^{\Sigma x_i} (1-p)^{n-\Sigma x_i} \quad \text{for } p = .1$$

$$= \boxed{\prod_{i=1}^{n} \left[\binom{n}{x_i} \right] (.1)^{\Sigma x_i} (.9)^{n-\Sigma x_i}}$$

$$R^* = \left\{ \frac{\frac{1}{e^n \prod_{i=1}^{n} x_i!}}{\prod_{i=1}^{n} \left[\binom{n}{x_i} \right] (.1)^{\Sigma x_i} (.9)^{n-\Sigma x_i}} > k \right\}$$

Figure 7, the work we did to compute our Neyman-Pearson Ratio.

Then, we were asked to compare our value to the actual likelihood ratio of 2.56, which we discussed in the data section previously. The probability that our R was greater than 2.56 from the null hypothesis was 0.28% and for the alternative hypothesis was 1.05%. Fortunately, this is what we expected because large values of the Rs agree with the alternative hypothesis.

Here is a table explaining the code we used:

Code	Explanation

-lam_null=.1 -lam_alt=1 -n=10 -trials=10 -num_R=100,000	These are all variables that we set and refer to in later code. This helps us easily alter the variable while testing and provides more description in future lines.
matrix(ncol=1,nrow=0)	Creates a matrix.
For (I in 1:num_R){}	Runs a loop num_R times.
binom(n,trials,lam_null)	Creates a vector of binomially distributed random variables assuming the null hypothesis.
- num=1/((exp(n*lam_alt))*prod(factorial(xi))) - den=prod(choose(n,xi))*(lam_null^sum(xi))*((1-lam_null)^(n-sum(xi))) - R=num/den	This is simple algebra which calculates the each Neyman-Pearson Ratio.
Rbind()	Append a value to a list as a new row.

(sum(Rs>2.56))/10,000	Calculates the probability of a R being greater than 2.56.
Rpois(n,lam_alt)	Creates a poisson distribution with parameter lam_alt.

Conclusion

To conclude, we were asked to simulate the distribution of the p-value for the hypothesis test which had the null hypothesis as binomially distributed with n=10 and lambda = 0.1 and the alternative hypothesis as Poisson distributed with lambda = 1. To solve this problem, we first simulated 100,000 values of R, the Neyman-Pearson likelihood ratio, to serve as an approximation for the distribution of R. Then we simulated 10,000 values of the null hypothesis, and for each value simulated, we estimated their p-value by using the empirical distribution we found in the first step. Lastly, we made a histogram of the 100,000 p-values. Based on the histogram, we think the distribution of the p-values is left-skewed. We then repeated all the steps again, but this time we used the alternative hypothesis instead of the null hypothesis. Based on the histogram from the alternative hypothesis, we think the distribution of the p-values is like that of the null hypothesis.