Homework Assignment 4: Applied Probabilistic Models

Aspects of Poisson Distribution

5273

1 Introduction

For this work, data is collected on the free eBooks library Project Gutenberg [3]. The chosen book for the analysis is: "The Autobiography of Benjamin Franklin" [2]. Data obtained from the Project Gutenberg are in `txt` format. The book is downloaded directly from the web and in order to develop the analysis.

For the analysis, it is used the R software in its version 4.0.2 [1], and the code used is available on the GitHub repository [4]. Experiments are run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Data Distribution

An experiment of the Poisson distribution was made using the `rpois` function and comparing it with the sum of exponential variables. Data is generated, taking into account the number of repetitions and the λ value. Other parameters are fixed for aesthetics, such as the number of bins.

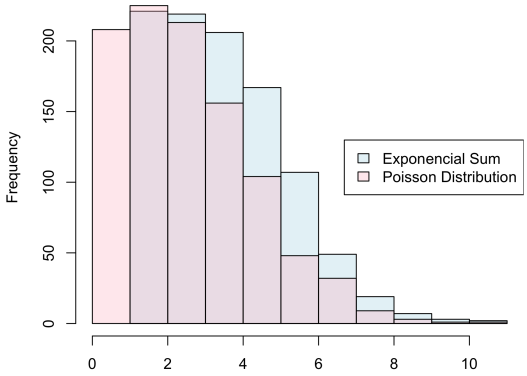
2.1 Relation with exponential distribution

As an experimentation strategy, it is performed a one factor at a time approach, where the number of repetitions changes as the λ value remains fixed. On the other hand, the opposite is executed, changing the λ values and fixing the number n of repetitions.

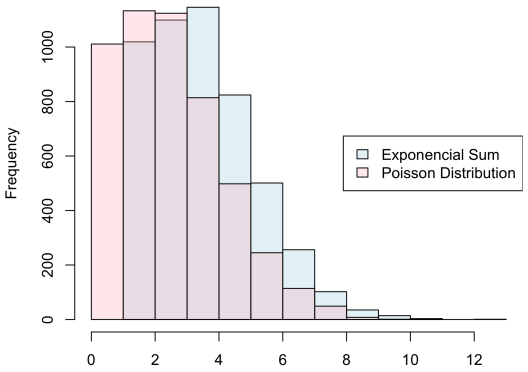
Figure 1 describe what happens when a Poisson distribution is generated, and a variation in the number of repetitions is executed. At this stage, the value of λ is fixed to 3, and the number of repetitions in where the exponential variable is sum are changed within 4 values: 1 000; 2 000; 10 000; and 15 000.

Alternatively, Figure 2 shows the experiment changing the values of λ to 4, 8, 16, and 32, and the number of repetitions is 10 000 for this case.

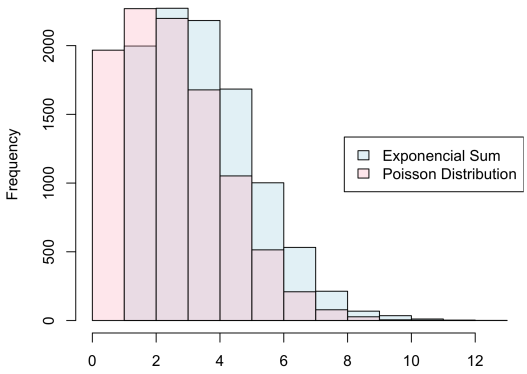
In conclusion, with this experiment, it can be seen that changing λ the exponential sum is closer to the generated pseudo-random Poisson distribution.



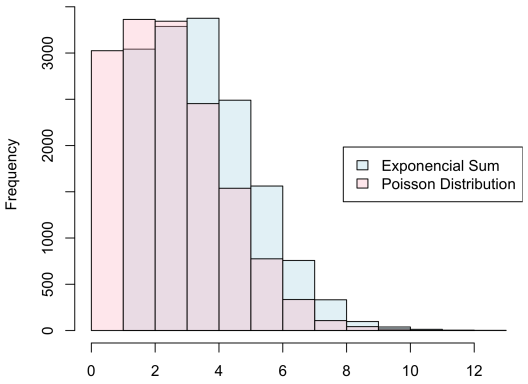
(a) Histogram with 1000 repetitions



(b) Histogram with 5000 repetitions

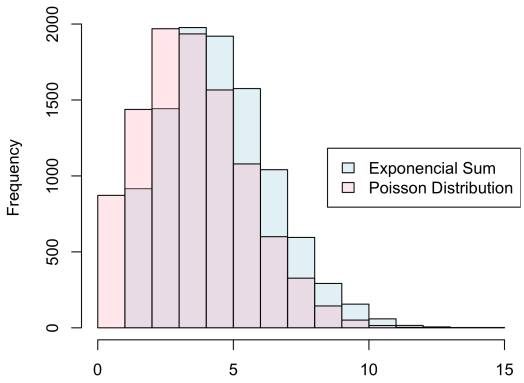


(c) Histogram of 10000 repetitions

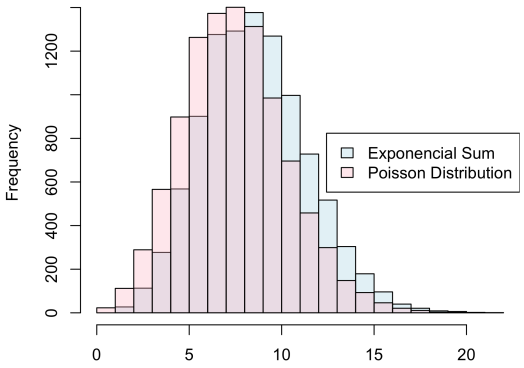


(d) Histogram of 15000 repetitions

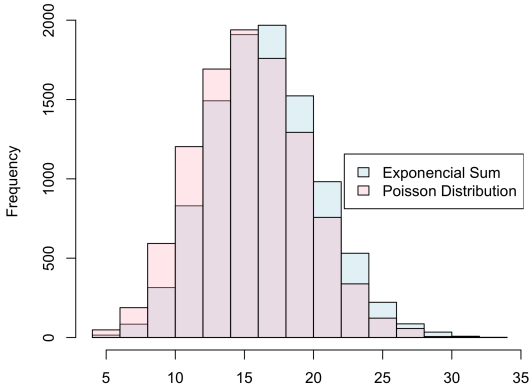
Figure 1: Histograms of the experiment changing the number of repetitions while $\lambda = 3$.



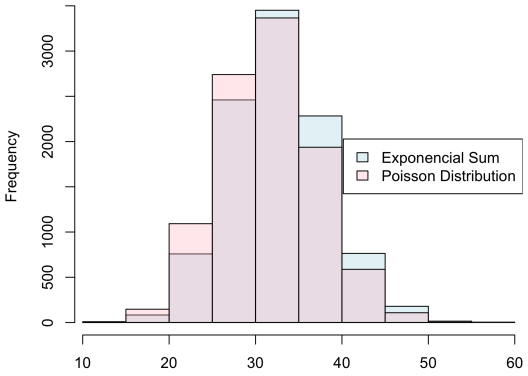
(a) Histogram with $\lambda = 4$



(b) Histogram with $\lambda = 8$



(c) Histogram with $\lambda = 16$



(d) Histogram with $\lambda = 32$

Figure 2: Histograms of the experiment changing λ while the number of repetitions is fixed to 10 000.

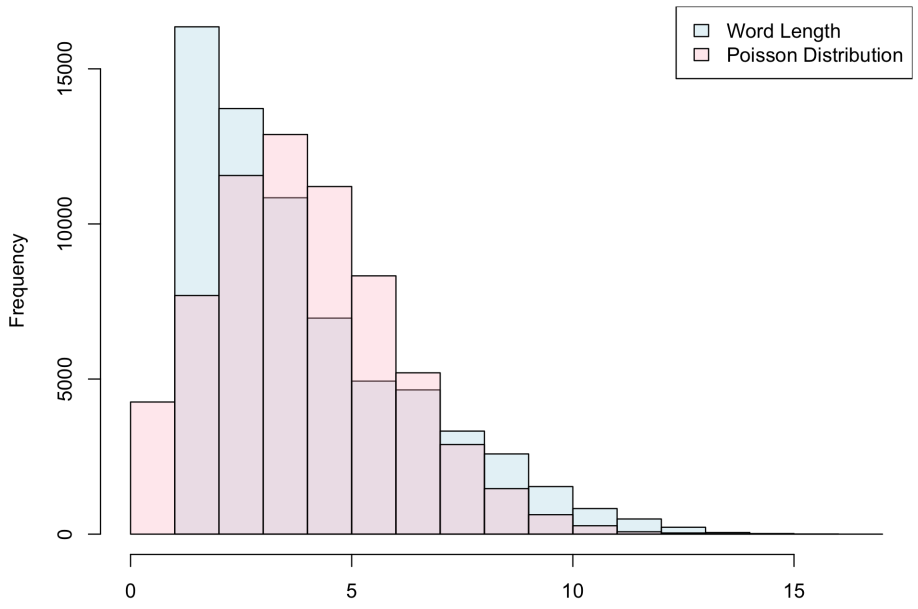


Figure 3: Histogram of Words Length and a Poisson Distribution

2.2 Application in the selected book

A comparison of the distribution of words length in the book and a similar Poisson distribution is made. For this process, it is assumed that word length is a variable that possibly has a Poisson distribution. It can be defined as X: Number of characters in a word. In this book, there are 66 520 words, so that would be our sample n, and the mean in word length would be our λ. With that fixed parameter, it is proceeded to generate the corresponding histogram (see Figure 3).

To determine if the two samples are significantly different, a Kolmogorov–Smirnov test is considered a very efficient way to do so. The Kolmogorov –Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples [5]. Figure 4 shows a representation of this test. After the test, as the p-value is less than 0.05, it is rejected the null hypothesis, meaning there are variations between the two data samples.

data.txt

```
        Two-sample Kolmogorov-Smirnov test

data:  lchar and poi
D^- = 0.05436, p-value < 2.2e-16
alternative hypothesis: the CDF of x lies below that of y
```

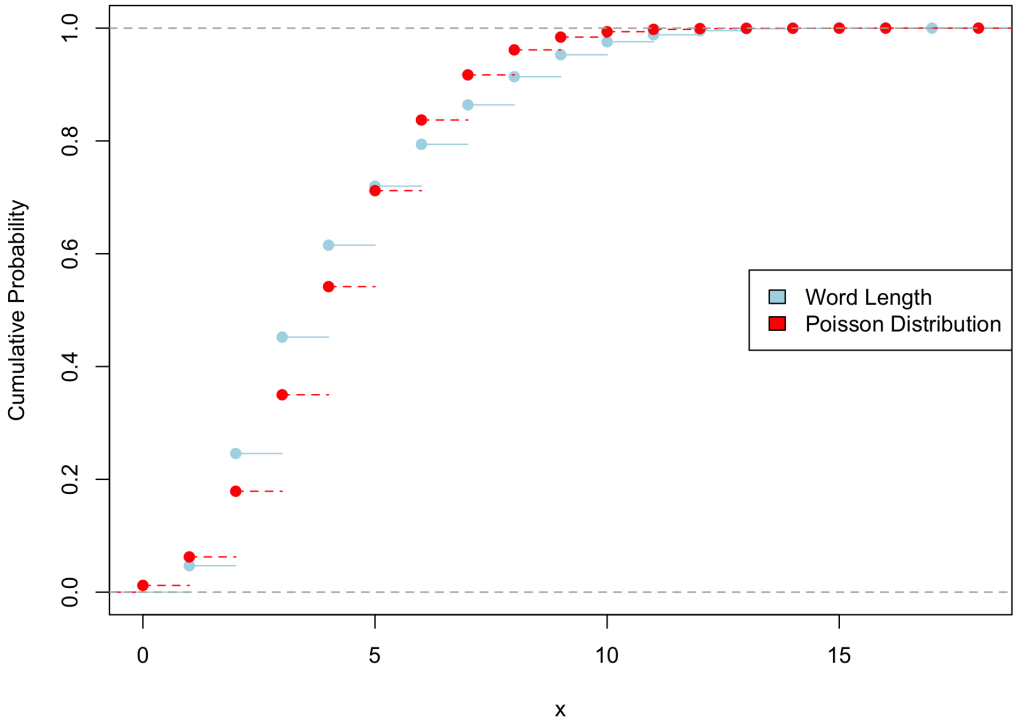


Figure 4: Kolmogorov–Smirnov test for Words Length and a Poisson Distribution

References

[1] The R Foundation. The R Project for Statistical Computing. https://www.r-project.org/, 2020.

[2] Benjamin Franklin. *The Autobiography of Benjamin Franklin: 1706-1757*, volume 1. 2007.

[3] Michael Hart. Project Gutenberg, 1971. http://www.gutenberg.org/ebooks/, Last accessed on 2020-09-09.

[4] Oscar Hernandez. Probability in R. https://github.com/oscaralejandro1907/probability-in-R/blob/master/assignment1/t1.R, 2020.

[5] Frank J Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.