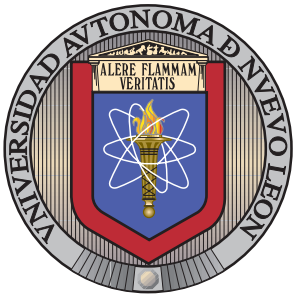
Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Posgrado en Ingeniería de Sistemas

Doctorado



Portafolio de Evidencias

de

Oscar Alejandro Hernández López

1985273

para el curso de Modelos Probabilistas Aplicados,

con la profesora Dra. Elisa Schaeffer.

Semestre Agosto 2020 - Enero 2021.

https://github.com/oscaralejandro1907/probability-in-R

Homework Assignment 1 Corrections

In this homework corrections in punctuantions of equations were made, added after the feedback.

Homework Assignment 1: Applied Probabilistic Models

Data Analysis of Growth Domestic Product category

5273

Introduction

For this work, data is collected on the official website of Instituto Nacional de Estadística y Geografía (INEGI) [2]. The chosen section is Growth Domestic Product, and within this section, a national macroeconomic indicator is selected: Global Index of Economic Activity (IGAE) for the data analysis. Data obtained from INEGI website are in `csv` format, edited in order to work with the general values of the three main representative activities each month, which the aforementioned indicator is based on. The objective is to evaluate the behavior of the group of activities and which has a major impact on the IGAE.
For the analysis it is used the R software in its version 4.0.2 [4] and the code used is available on the GitHub repository of [3]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

Data

Global Index of Economic Activity

The IGAE is an index that aproximates the calculation of the generated wealth in the country monthly. It is considered a trend index and marks the path that the national economic activity is reporting in the given month [5]. It is important to emphasize that the IGAE collects monthly figures since January 1993.

Calculation Methods

The method used to calculate the IGAE consists of monthly indexes of the physical volume of production for each of the selected classes, with a fixed base in 2013 [1].
The calculation of the physical volume index of industrial activities consists of preparing monthly indexes of the volume of real production for each of the classes that have information on quantities produced, production values, and prices at the product level.
In general, an index number is the percentage relationship that measures the change from one time to another in price, quantity, value, or some other element of interest, and can be constructed for different periods, ranging from high frequency (daily, weekly, monthly, quarterly, etc.) to annual.
In the case of IGAE physical volume of production indexes are elaborated, of Laspeyres-type (an index that systematically overvalues inflation), as well as simple indexes based on a related indicator, expressed base 2013 = 100.

Physical Volume Index of production, Laspeyres-type

$$Q_{0,n} = \frac{\sum P_0 Q_n}{\sum P_0 Q_0 / 12} * 100,$$

where:
$Q_{0,n}$ = Physical volume index of the production of period n in relation to period 0,
0 = base year,
n = period of reference,
Q_n = Quantity of a produced good during period n (period of study),
Q_0 = Quantity of a produced good during period 0 (base),
P_0 = price of a good corresponding to the base period.

Real value indexes (income or expenses)

$$IV = \frac{I_n}{I_0} * 100,$$

where:
IV = Value index,
I_n = Real income/expenses of study period,
I_0 = Real income/expenses of base period.
In the calculation of the IGAE, three fundamental areas or activities are involved: primary, secondary, and tertiary, which are detailed below.

Primary Activities

These activities include agriculture, breeding, and exploitation of animals.

Secondary Activities

These activities cover mining, generation, transmission, and distribution of electricity, water supply, and gas through pipelines to the final consumer, construction, and manufacturing industries.

Tertiary Activities

This group of activities contains wholesale and retail; transport, post, and storage; media information; services such as financial and insurance, real state, rental and intangible property, professional, scientific, educational and technical services, business support and waste management, health and social assistance, recreational, cultural and sports services, temporary accommodation; legislative, governmental, justice administration and international activities.

Representation and Analysis

For the data representation, it is selected data of 330 months starting from January 1993 to June 2020. A boxplot of the index values for each group of activities is represented in Figure 1. Here the horizontal axis represents each of the group of activities and the vertical one represents the corresponding values of the index.

```r
dat<-read.csv('/Users/oscarhernandezlopez/Dropbox/R/T1-Probabilidades/IGAE_1.csv')

dat_trimmed <- dat[,2:(ncol(dat))]
dat_transpose <- t(dat_trimmed)

png('boxplot_activities.png',width = 2000, height = 1600, res = 300)
boxplot(dat_transpose)
dev.off()
```

t1.R

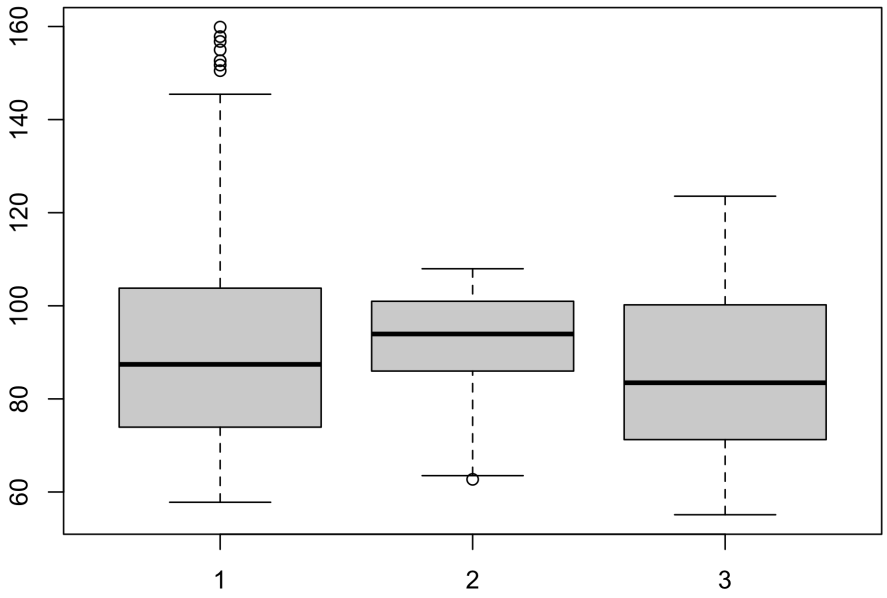


Figure 1: Boxplot of the three activities of the IGAE

After looking at the values of the three groups of activities it can be noticed a wider range of values in the primary activities. This primary activity also has higher values of the sample, with several outliers above the third quartile. The secondary activities are the least scattered so it the one that shows an overall performance over time. On the other hand, its higher values are not as great as primary and tertiary activities.

Primary activities can be analyzed further. In Figure 2 it can be seen last two months of the year are months with higher action and August and September are the month with lower activity. Also in Figure 3 it can be seen a growing trend with lows in the 19 and 26 boxes, as a consequence of the economic crisis of 2008 and COVID-19 respectively.

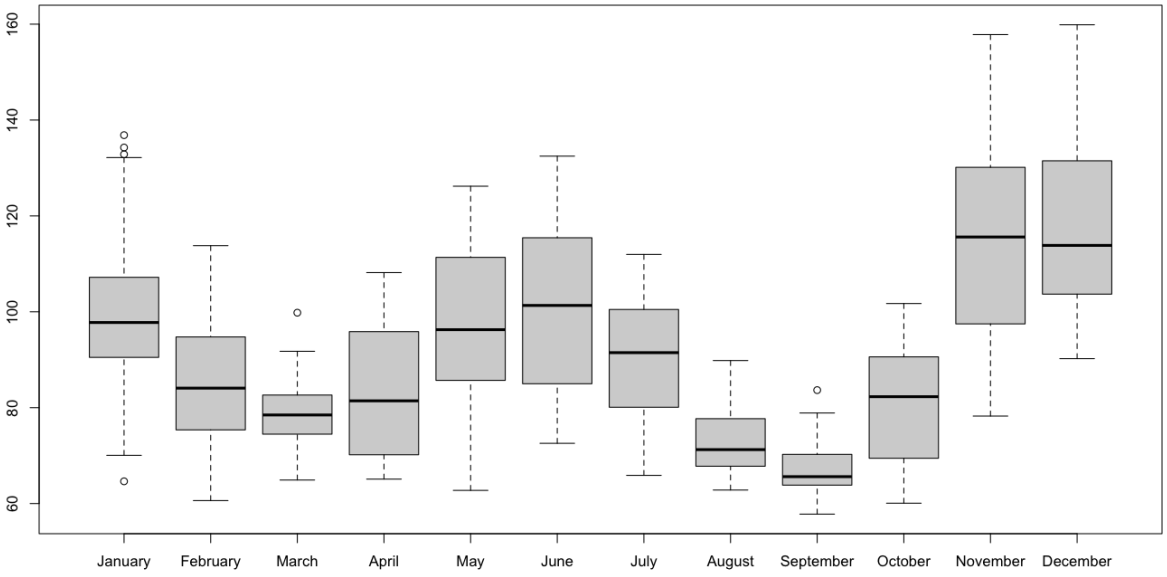


Figure 2: Boxplot of primary activities by months

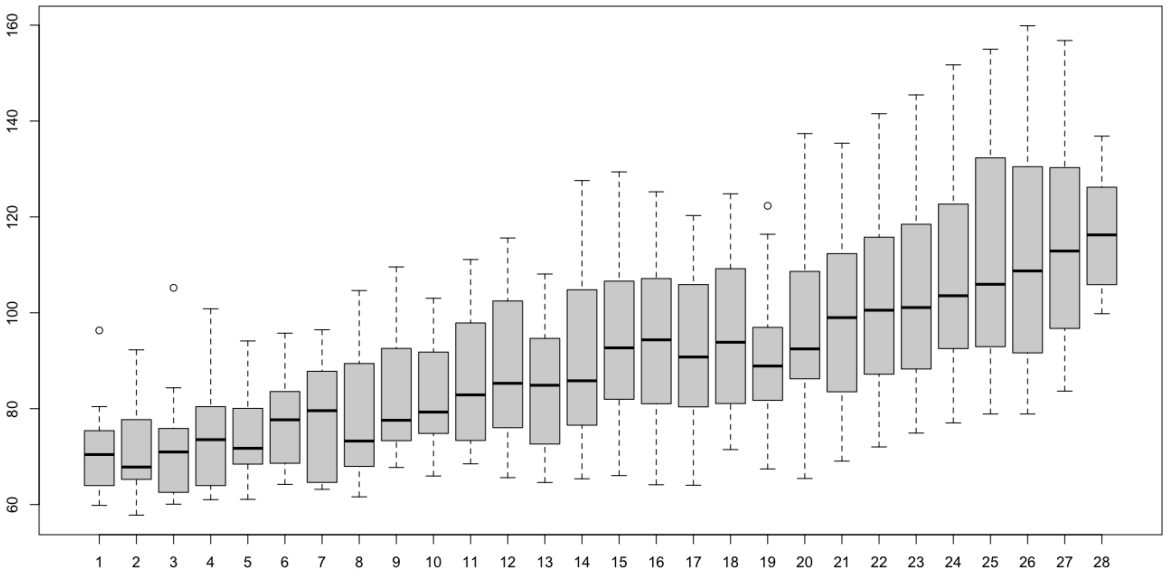


Figure 3: Boxplot of primary activites by year

References

[1] Instituto Nacional de Estadiística y Geografía. Sistema de cuentas nacionales de México. Fuentes y metodologías. año base 2013. Indicador Global de la Actividad Económica, 2017. https://www.inegi.org.mx/contenidos/programas/igae/2013/metodologias/SCNM_Metodo_IGAE_B2013.pdf, Last accessed on 2020-09-05.

[2] INEGI. Datos, 2020. https://www.inegi.org.mx/temas/igae/, Last accessed on 2020-09-05.

[3] Oscar Alejandro Hernadez Lopez. Probability in R. https://github.com/oscaralejandro1907/probability-in-R/blob/master/assignment1/t1.R, 2020.

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

[5] Eduardo Torreblanca. ¿En qué consiste el IGAE?, 2016. https://mvsnoticias.com/noticias/economia/opinion-en-que-consiste-el-igae-610/, Last accessed on 2020-09-05.

Homework Assignment 2: Applied Probabilistic Models

Analysis of the Book Structure

5273

1 Introduction

For this work, data is collected on the free eBooks library Project Gutenberg [1]. The chosen book for the analysis is: "The Autobiography of Benjamin Franklin". Data obtained from the Project Gutenberg are in `txt` format.

For the analysis, it is used the R software in its version 4.0.2 [3], and the code used is available on the GitHub repository [2]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Data

The book is downloaded directly from the web and in order to develop the analysis, the following code is used.

```r
require(gutenbergr) #Download books from online library
require(tidytext) #Clean text
require(dplyr)  #Data Manipulation

library(textshape)

#Load the book: "The Autobiography of Benjamin Franklin"
book<-gutenberg_download(c(148))
```

a2.R

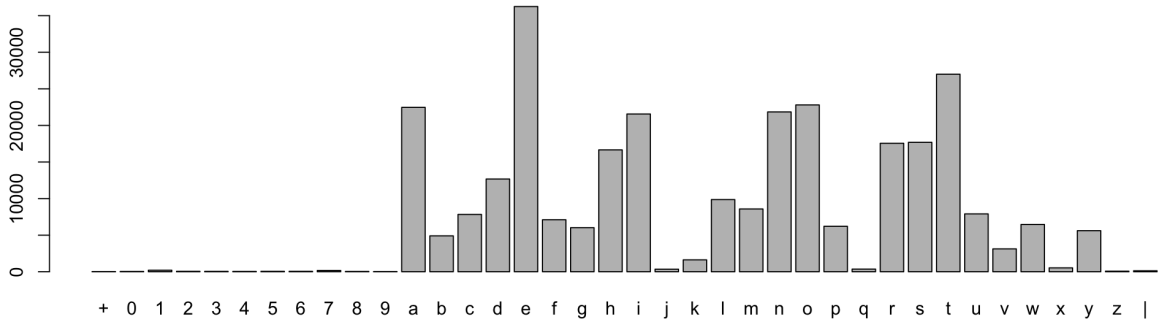
The book has a total of 294 003 characters (letters) and 66 520 words. This data is used to a further analysis of what the most important letters and words are according to its frequency. The next part of the code is set to this objective.

```r
#Variables used:
letters <- book %>% unnest_tokens(chars, text, "characters") #contains letters
words <- book %>% unnest_tokens(word, text, "words") #contains words

#Work with Letters:
png('barplot_letters.png',width = 2100, height = 768,res = 180)
barplot(table(letters$chars)) #Barplot of letters
dev.off()

freq_l <- as.data.frame(table(letters$chars))
names(freq_l) <- c('Letter', 'FrequencyL')

rl <- freq_l[freq_l$FrequencyL>500,] #Filter relevant letters
png('barplot_relevant_letters.png',width = 1366, height = 768,res = 150)
barplot(rl$FrequencyL, names.arg = rl$Letter)
dev.off()

rlo <- rl[order(rl$FrequencyL, decreasing=TRUE),]
png('barplot_relevant_letters_ordered.png',width = 1366, height = 768,res = 150)
barplot(rlo$FrequencyL, names.arg = rlo$Letter)
dev.off()

#Work with Words:
png('barplot_words.png',width = 1366, height = 768,res = 150)
barplot(sort(table(words$word), decreasing = TRUE)) #Barplot of words
dev.off()

freq_w <- as.data.frame(table(words$word))
names(freq_w) <- c('Word', 'FrequencyW')

muw <- freq_w[freq_w$FrequencyW > 100,]
png('barplot_most_used_words.png',width = 2100, height = 768,res = 180)
barplot(muw$FrequencyW, names.arg = muw$Word)
dev.off()

rw <- muw[muw$FrequencyW < 200,]
png('barplot_relevant_words.png',width = 2048, height = 768,res = 150)
barplot(rw$FrequencyW, names.arg = rw$Word)
dev.off()

rw_o <- rw[order(rw$FrequencyW, decreasing=TRUE),]
png('barplot_relevant_words_ordered.png',width = 2166, height = 768,res = 180)
barplot(rw_o$FrequencyW, names.arg = rw_o$Word)
barplot(rw_o$FrequencyW, names.arg = rw_o$Word, log = 'y')
dev.off()

places<-as.data.frame(table(grep("york|london|boston|newport|philadelphia|paris",
    words$word,value=TRUE)))
names(places) <- c('Place', 'FrequencyP')
rp <- places[places$FrequencyP>2,] #Filter relevant places
png('barplot_places.png',width = 1366, height = 768,res = 150)
barplot(rp$FrequencyP, names.arg = rp$Place)
dev.off()
```

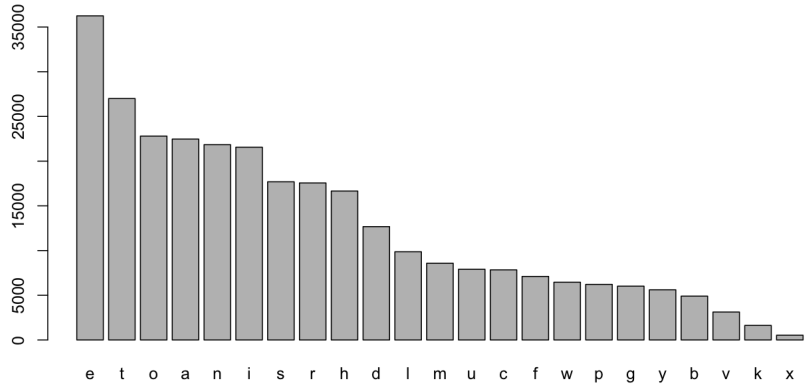
a2.R

2.1 Letters

For the analysis of the letters, barplots are generated. Here the horizontal axis represents all the used letters and the vertical one the corresponding frequencies. Figure 1 shows in (a) all the letters present in the document, sorted in decreasing order and by the same token in (b) it can be seen the most used



(a) Barplot of all letters present in the document



(b) Barplot of most used letters

Figure 1: Barplots of letters present in the document

letters. In this latter case frequency greater than 500 is the criteria to fall in this category. Frequencies of the six most used letters are shown in Table 1.

Table 1: Frequency of most used letters

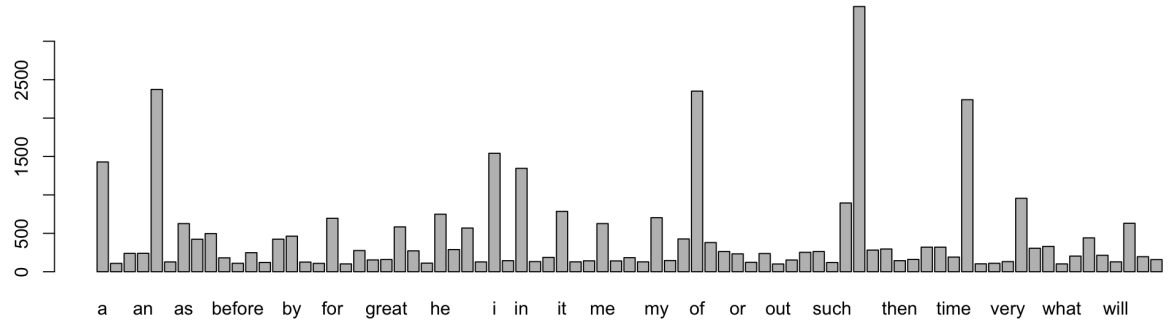
Letter	Frequency
e	36 252
t	27 001
o	22 803
a	22 472
n	21 845
i	21 561

2.2 Words

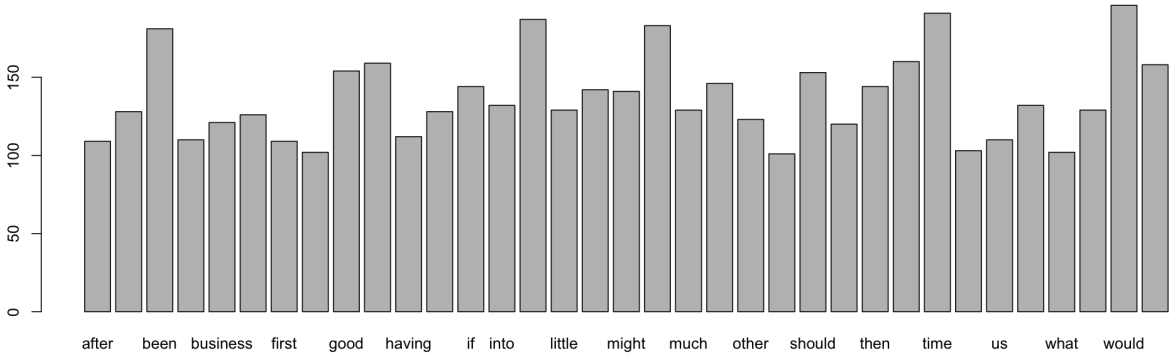
In a similar fashion, barplots are generated for the analysis of words. Here the horizontal axis represents all the used words in the document and the vertical one corresponds to the respective frequencies.

In the first attempt, it is hard to see a difference among words, that is why the following barplots are generated. Figure 2 shows the words filtered. In (a) it can be seen the frequency of the most used words. For this category, it is set as a requirement a frequency greater than 100. Furthermore, in (b), for this category, it is discarded the words with a frequency greater than 250 because articles and prepositions are included and these words are not very descriptive. Finally, those relevant words are sorted in decreasing order in Figure 3.

Last, in Figure 4 it is generated a barplot, which shows an analysis of several places mentioned in the autobiography, and it can be seen as the most relevant ones in the life of this character. From this barplot most of the events related to Franklin took place in the city of Philadelphia.



(a) Barplot of most used words present in the document



(b) Barplot of relevant words

Figure 2: Barplots of words present in the document

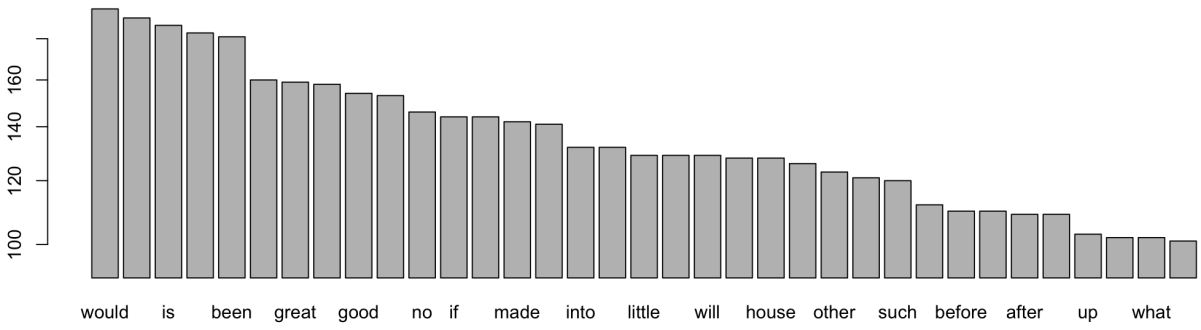


Figure 3: Barplot of relevant words present in the document (sorted)

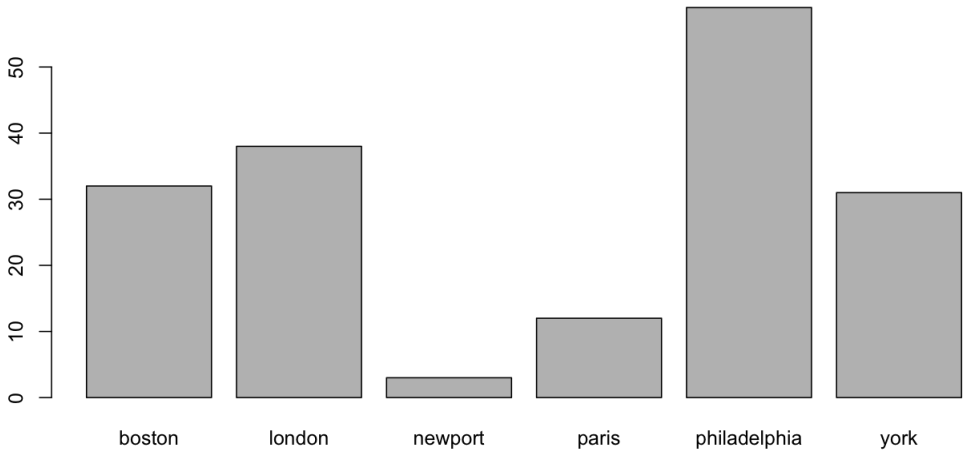


Figure 4: Barplot of relevant places mentioned in the document

References

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

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

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

Homework Assignment 3: Applied Probabilistic Models

Word Distributions

5273

1 Introduction

For this work, data is collected on the free eBooks library Project Gutenberg [5]. The chosen book for the analysis is: "The Autobiography of Benjamin Franklin" [4]. Data obtained from the Project Gutenberg are in `txt` format.

For the analysis, it is used the R software in its version 4.0.2 [3], and the code used is available on the GitHub repository [6]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Data Distribution

The book is downloaded directly from the web and in order to develop the analysis, the following code is used.

```r
require(gutenbergr) #Download books from online library
require(tidytext) #Clean text
require(dplyr)  #Data Manipulation
require(textshape)
require(tokenizers)

library(fitdistrplus)

#Load the book: "The Autobiography of Benjamin Franklin"
book<-gutenberg_download(c(148))

#Variables used:
words <- book %>% unnest_tokens(word, text, "words")  #contains words
sentences <- book %>% unnest_tokens(sentence, text, "sentences")
paragraphs <- book %>% unnest_tokens(paragraph, text, "paragraphs")
```

a3.R

In this work, it is analyzed how frequencies of English words are distributed. The first discussed aspects are events that describe two known distributions. The second part corresponds to the other two events that can be worth finding which distribution best fits the data. For this last issue, it is recommended to see the paper of Delignette-Muller et al. [2].

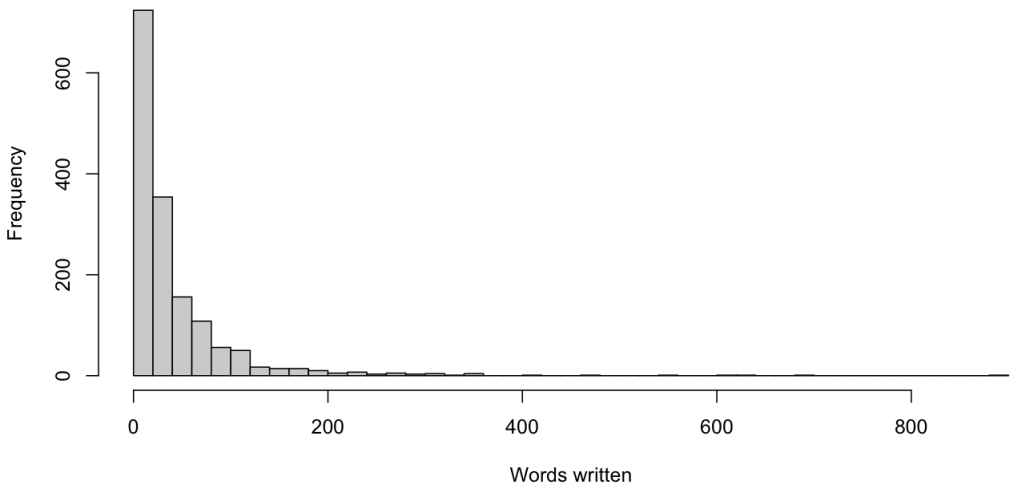


Figure 1: Histogram of the amount of words used before the pronoun "I"

2.1 Geometric Distribution

Since the book is an autobiography it is redacted in the first person, so it is expected the word "I" appears multiple times referring to facts corresponding to the author himself. Some statistics of this pronoun are studied. The first aspect to analyze is how many words have been used when the "I" appears. This may correspond to a geometric distribution, which can be described as the number of repetitions resulting in failures until the first success is achieved. In this case, all the words written until the pronoun is used can be considered as failures, and the use of "I" is the success. Figure 1 shows a histogram of how this aspect is represented in the book.

2.2 Binomial distribution

The number of times this pronoun is used in each sentence of the book is another event that could be analyzed. This case could be described as a binomial distribution, which describes the number of success obtained in n Bernoulli repetitions. For this example, those repetitions are the total number of sentences in the text whereas the number of success is given by the presence of the pronoun "I" in k times. This is shown in the histogram corresponding to Figure 2.

2.3 Other distributions

Length of English words and also its quantity when constructing paragraphs are the other considered aspects. Figure 3 shows a histogram of words length used throughout the book and Figure 4 shows a histogram of the amount of word used per paragraph in the document.

Last, Figure 5 show a skewness-kurtosis plot such as the one proposed by Cullen et al. [1] for the empirical distribution of both events. In this plot, values for common distributions are displayed in order to help the choice of distributions to fit data. The distribution is represented by a single point on the plot. According to Delignette-Muller et al. [2] skewness and kurtosis are known not to be

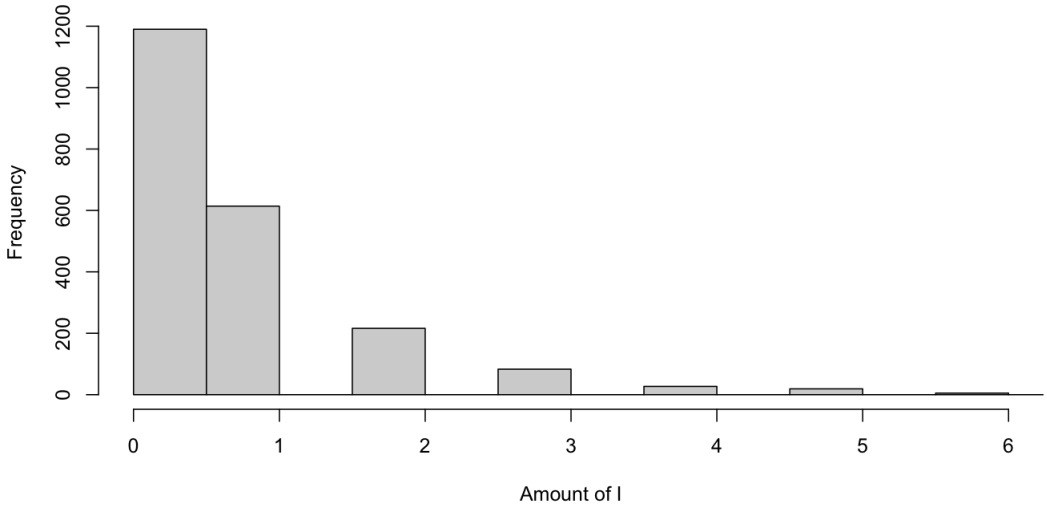


Figure 2: Histogram of the amount of times the pronoun "I" is mentioned in sentences

robust, thus the plot should then be regarded as indicative only. A non-zero skewness reveals a lack of symmetry of the empirical distribution. For words length and words per paragraph, the skewness has values of 0.996 and 3.128 respectively. The kurtosis value quantifies the weight of tails in comparison to the normal distribution for which the kurtosis equals 3. The values of kurtosis are 3.519 for word length and 27.017 for words per paragraph.

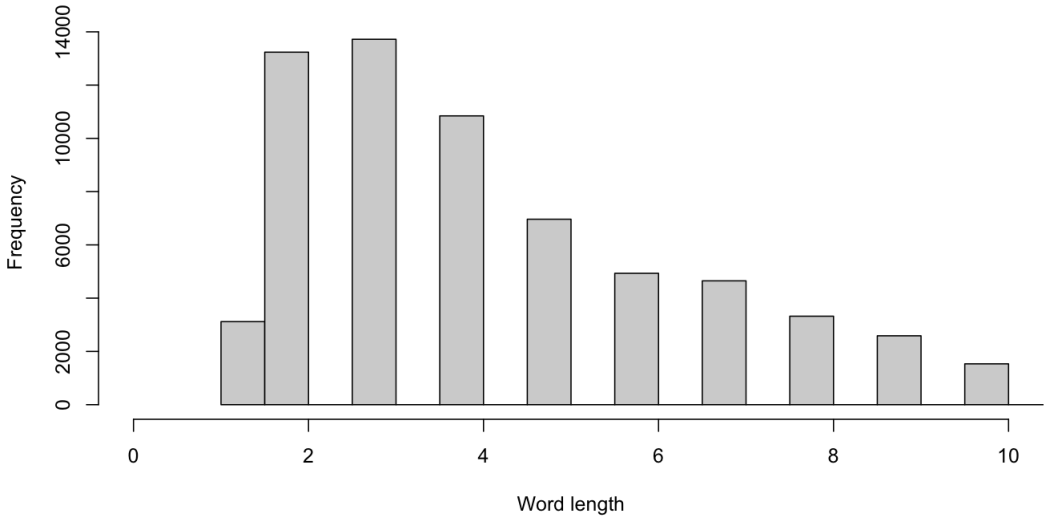


Figure 3: Histogram of lenght of words used in the book

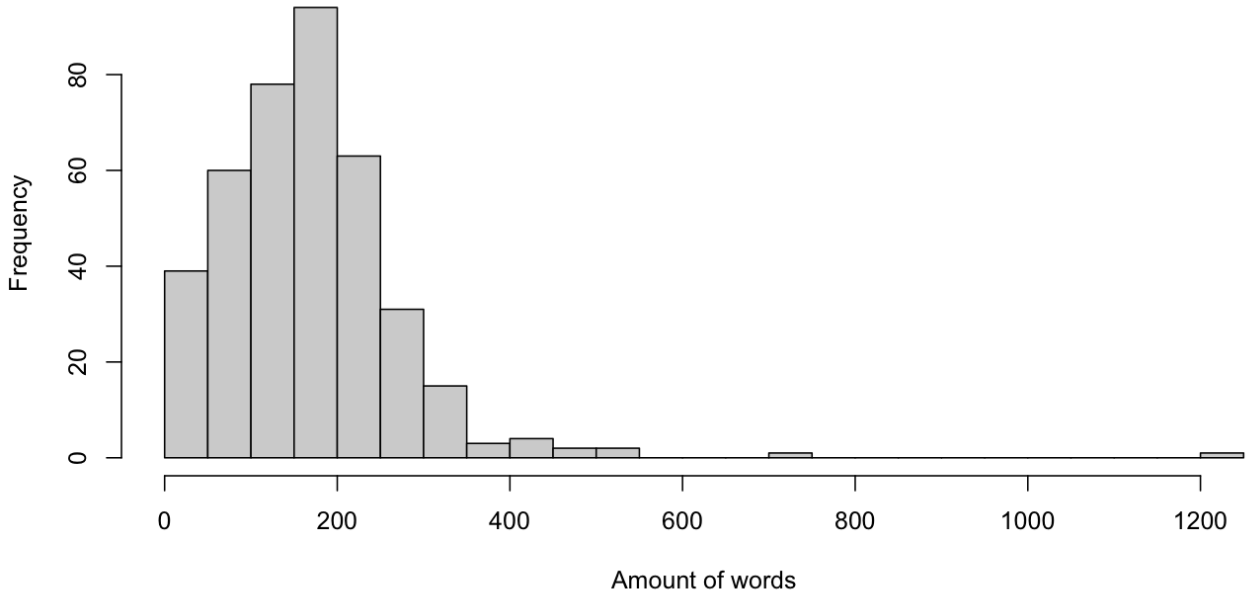
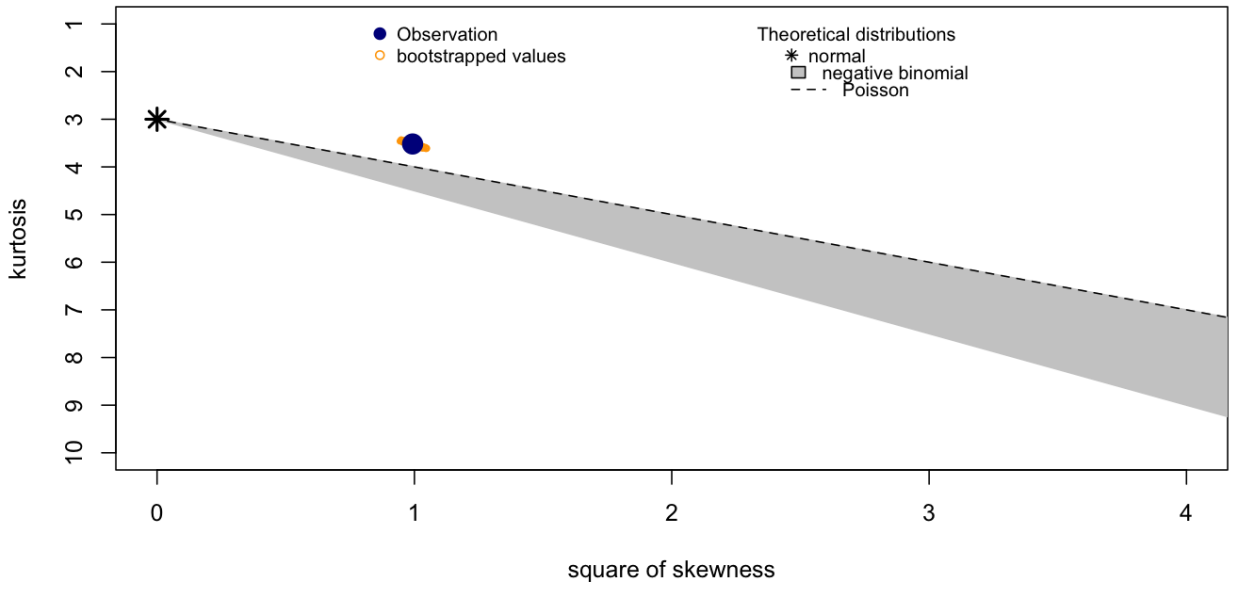
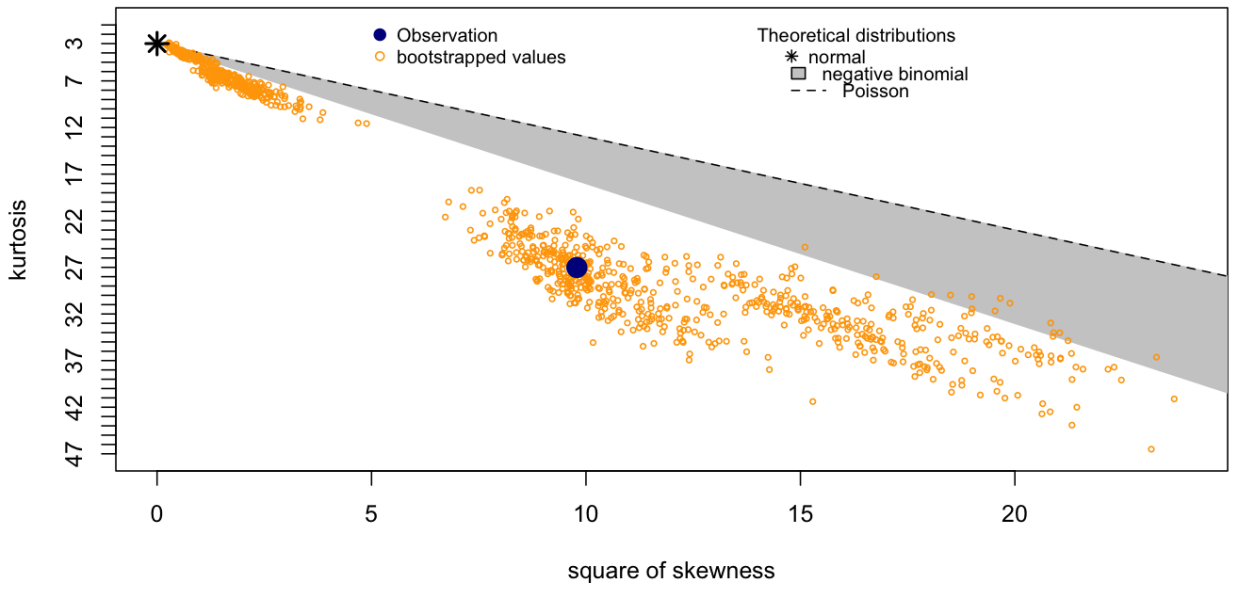


Figure 4: Histogram of the amount of words used per paragraph



(a) Cullen and Frey graph for words lenght present in the document



(b) Cullen and Frey graph for words per paragraphs

Figure 5: Cullen and Frey Graphs

References

[1] Alison C Cullen, H Christopher Frey, and Christopher H Frey. *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs*. Springer Science & Business Media, 1999.

[2] Marie Laure Delignette-Muller, Christophe Dutang, et al. fitdistrplus: An r package for fitting distributions. *Journal of statistical software*, 64(4):1–34, 2015.

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

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

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

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

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.

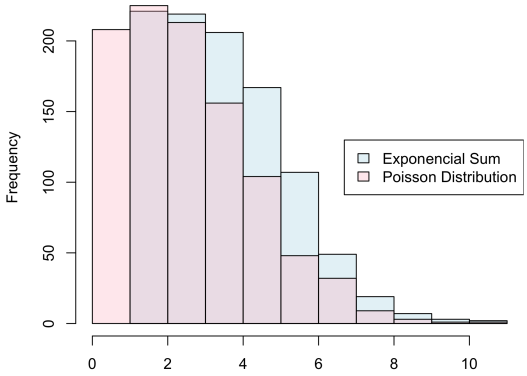
2.1 Relation with exponential distribution

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.

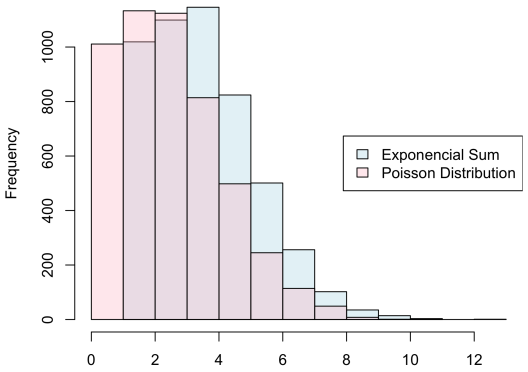
Figure 1 describe what happens when a Poisson distribution is generated, and a variation in the number of repetitions is executed. At this stage, the value of λ is fixed to 3, and the number of repetitions in where the exponential variable is sum are changed within 4 values: 1 000; 2 000; 10 000; and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

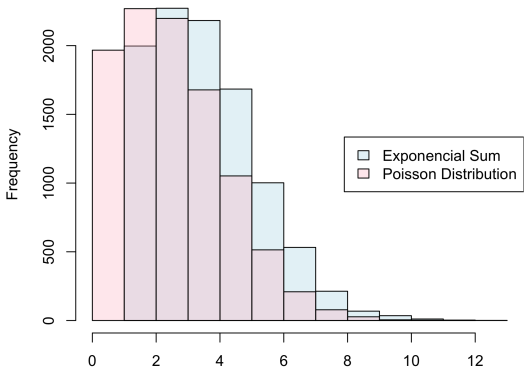
In conclusion, with this experiment, it can be seen that changing λ the exponential sum is closer to the generated pseudo-random Poisson distribution.



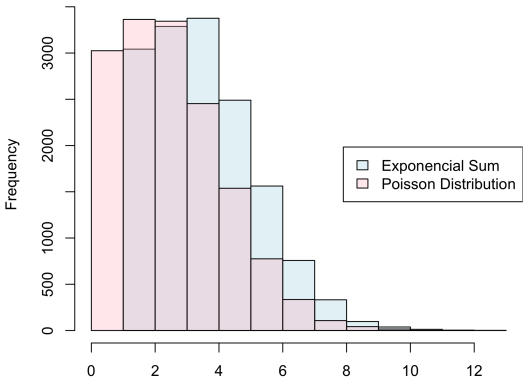
(a) Histogram with 1000 repetitions



(b) Histogram with 5000 repetitions

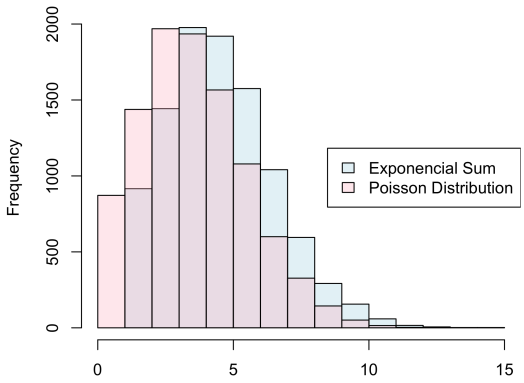


(c) Histogram of 10000 repetitions

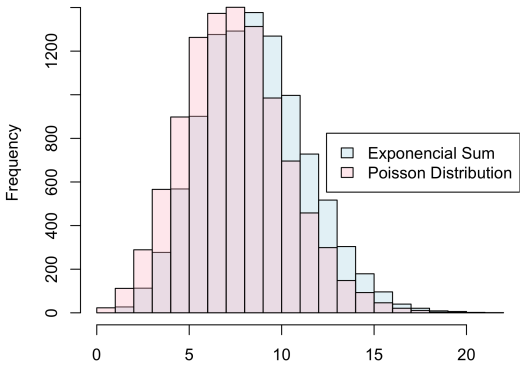


(d) Histogram of 15000 repetitions

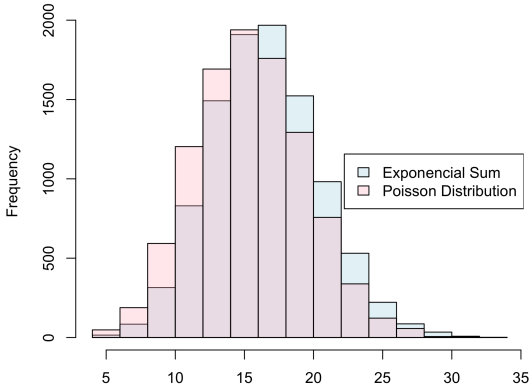
Figure 1: Histograms of the experiment changing the number of repetitions while $\lambda = 3$.



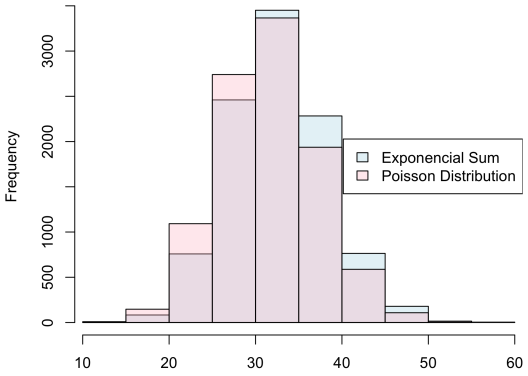
(a) Histogram with $\lambda = 4$



(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.

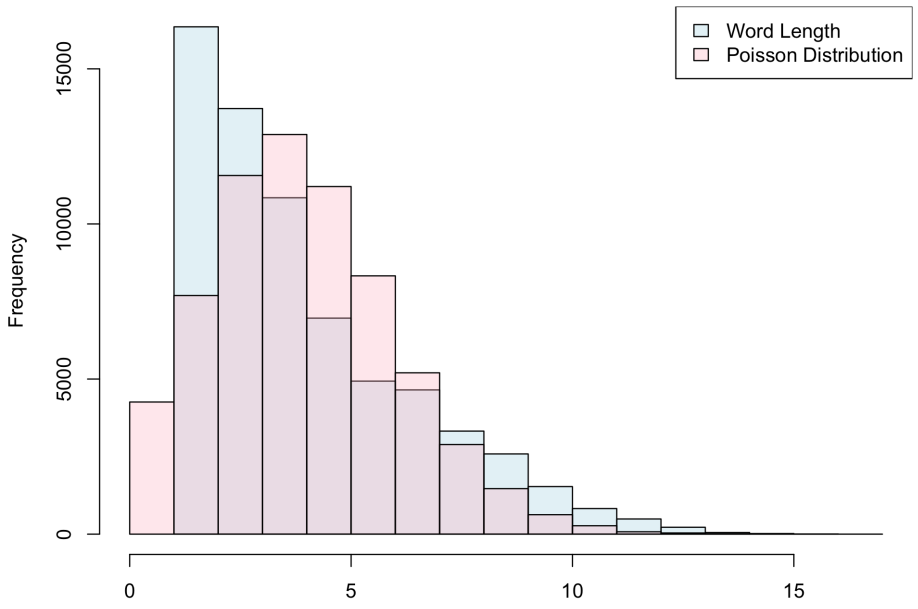


Figure 3: Histogram of Words Length and a Poisson Distribution

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
```

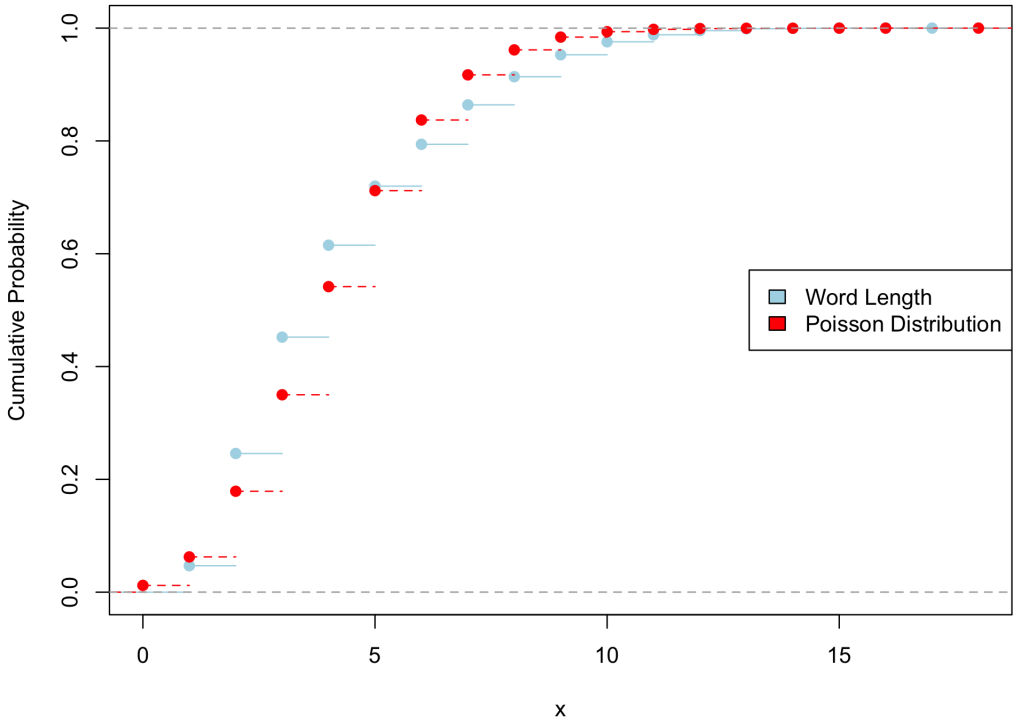


Figure 4: Kolmogorov–Smirnov test for Words Length and a Poisson Distribution

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.

Homework Assignment 5 Corrections

In this homework, data corresponding to smaller p-values were rewritten to avoid confusions when interpreting the scientific notation.

Homework Assignment 5: Applied Probabilistic Models

Aspects of Normal Distribution and Pseudo-random Numbers

5273

1 Introduction

In this work, it is studied the generation of pseudo-random numbers by different algorithms. Also, several experiments have been used to generate normal distributions from those generation methods. The algorithm's sensitivity is also tested changing parameters, and variables that define the different forms of generations are also part of the study.

For the analysis, it is used the R software in its version 4.0.2 [2], and the code used is available on the GitHub repository [3]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Algorithms

It is implemented the Box-Muller transform [1], which from two numbers uniformly distributed, generates normally distributed ones. Those numbers should be independent, but some tests are performed with several dependencies to see this algorithm's behavior. For this procedure, the following code is used, which creates a function that returns a pair of values resulting from applying the corresponding equations.

```r
gaussianNoise <- function (mu, sigma) { #Box-Muller Transform
  u <- runif(2)
  z0 <- sqrt(-2*log(u[1])) * cos(2*pi*u[2])
  z1 <- sqrt(-2*log(u[1])) * sin(2*pi*u[2])
  pair <- c(z0,z1)
  return (sigma * pair + mu) #Return a pair (zo,z1)
}
```

codes/a5.R

As can be seen, it takes two uniformly distributed random numbers using the function `runif` from R as a source, and it generates a pair of independent, normally distributed pseudo-random numbers.

On the other hand, it is analyzed the Linear Congruential Generator (LCG) [5], which is an algorithm that generates a sequence of pseudo-randomized numbers. This method is computed using the following code. This algorithm is defined by the seed or initial value (X_0), a modulus (m), a multiplier (a), and an increment (c) .

Table 1: Result of one experiment of the Shapiro test changing values of μ and σ.

	μ	σ	p-value
Test 1	400	20	0.548
Test 2	225	7	0.3549
Test 3	3	0.25	0.0860
Test 4	-23	3	0.5139
Test 5	-100	12	0.2405

```r
linearCongruentialGen <- function (n, seed) {
  a <- 11551
  c <- 27077
  m <- 39709
  x <- seed
  gen_data <- numeric()
  while (length(gen_data)<n){
    x <- (a * x + c) %% m
    gen_data <- c(gen_data,x)
  }
  return (gen_data/(m-1)) #Return a seed generation pseudo-random numbers
}
```

codes/a5.R

3 Experiments

Experiments for this work are based on the generation of Normal distributions from different methods. Several variations are studied to test the sensitivity of these algorithms as well. The parameters are the number of repetitions or sample size (n) with a value of 1000, then a mean (μ) of 10, a standard deviation (σ) of 2, and a seed of 27. Those values are arbitrarily selected at first and will be changed for further sensitivity analysis.

3.1 Box-Muller Transform

The first experiment is performed with the Box-Muller transform. This method generates a sequence of uniformly distributed values. This generation of pseudo-random numbers uses the elements generated by the variable Z_0. The same procedure is executed with the variable Z_1, and both distributions are compared with the generated using the `rnorm` function of R. This comparison is shown in Figure 1.

To perform a sensibility analysis for this algorithm, the parameters μ and σ were changed, and Shapiro Test is performed 30 times for each combination. An example of the results of one experiment is shown in Table 1. As a result of the test, it is important to highlight that it passed in all repetitions for every test, except for Test 3, in which the algorithm had a p-value less than 0.05 in two times.

In addition, it is generated a boxplot of these distributions (see Figure 2), and an analysis of variance (ANOVA) of one way is performed. In the ANOVA test, the null hypothesis (H_0) implies that there is not enough evidence to prove the mean of the group is different from another. And the alternative hypothesis (H_1) that at least, the mean of one group is different. In this case, the p-value is greater than 0.05, so there is no evidence to reject the null hypothesis, and therefore it is concluded that the means are identical.

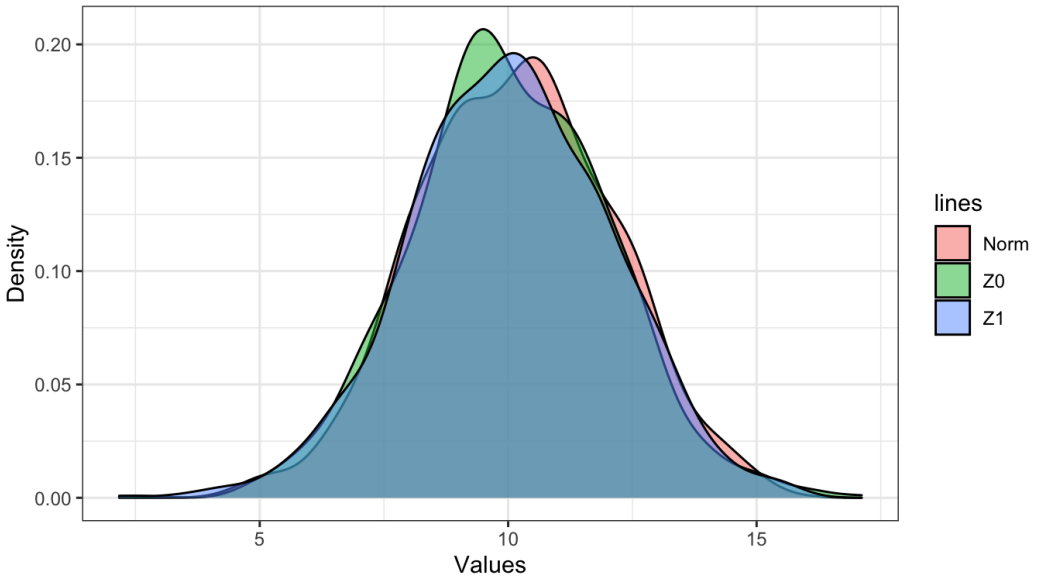


Figure 1: Density Plot of Normal Distributions

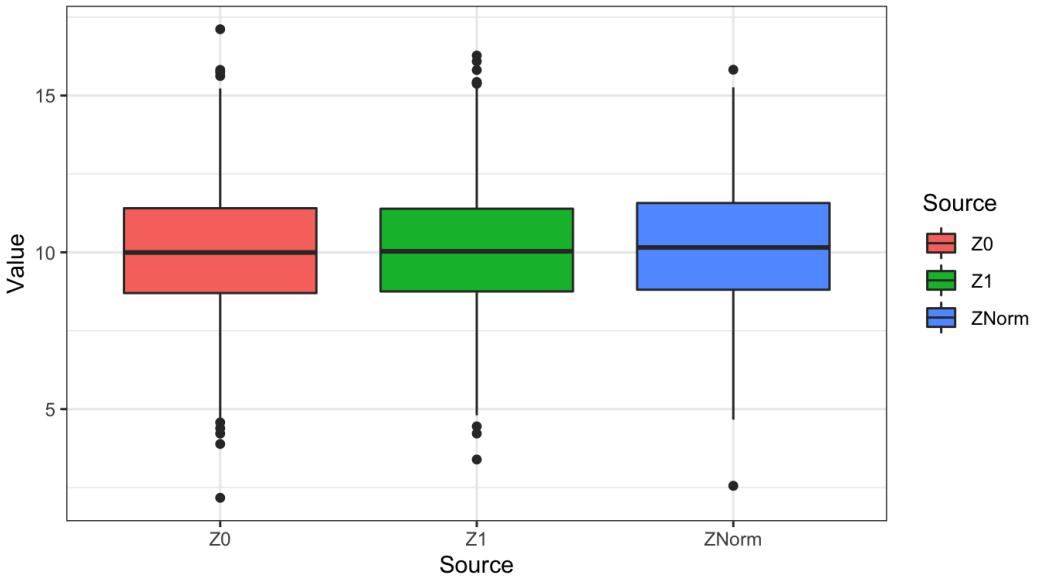


Figure 2: Box Plot of Normal Distributions

Table 2: Results of Shapiro Test with different relations.

Dependency	p-value
$u_2 > u_1$	2.67×10^{-13}
$u_2 = u_1$-`runif(1)`	0.3877
$u_2 =$ `runif(1,1,10)`-$2u_1$	0.6331
$u_2 = ($`runif(1)`-$u_1)/u_1$	0.0186
$u_2 = 2u_1$	4.5×10^{-16}
$u_2 = u_1$*`runif(1)`	2.35×10^{-13}

data.txt

```
             Df Sum Sq Mean Sq F value Pr>F
Source        2      8   3.875   0.984  0.374
Residuals  2997  11802   3.938
```

3.1.1 Testing Independency of the uniform pseudo-random generated numbers

A sensitivity analysis changing different relations between the uniformed pseudo-random numbers u_1 and u_2 from the Box-Muller transform is performed. Changes in which the algorithm is subject are defined by how the uniform pseudo-random numbers are generated. Then the transform is performed and repeated 1000 times, and a Shapiro-Test is performed to test the normality of these different generations. These changes are detailed in Table 2. In those cases, the algorithm only kept generating normal distributions if the relation is the difference between u_1 and u_2, and when it is generated a number u_2 between 1 and 10 (it can be considered as a bigger number), and it is subtracted with a multiple of u_1.

3.2 Linear Congruential Generator

Experiments with the LCG are performed based on the uniformed pseudo-random sequence generate with this algorithm. It is tested the distribution of the data with the Chi-squared test provided by the `uniform.test()` function over the histogram of the data. For this test, the p-value is greater than 0.05, so there is no evidence to reject the null hypothesis, and the data might be uniformly distributed.

data.txt

```
        Chi-squared test for given probabilities

data:  hist.output$counts
X-squared = 14.48, df = 9, p-value = 0.1062
```

3.2.1 Using LCG in the Box-Muller Transform

After concluding that data generated are uniformly distributed, it could be used to test these values in the Box-Muller transform as the values u_1 and u_2. The results of the histogram of this experiment are

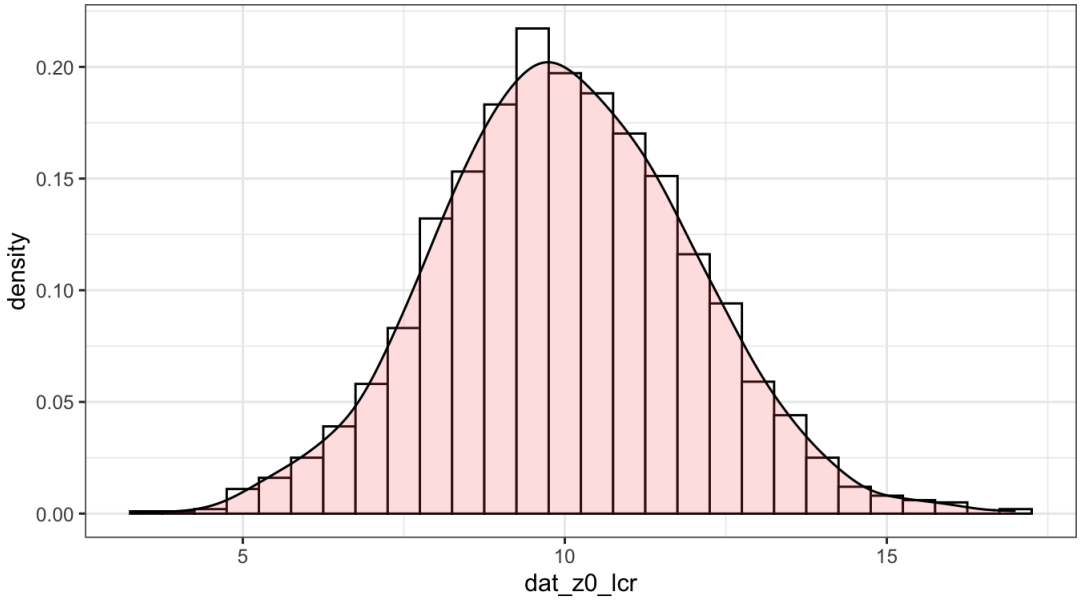


Figure 3: Histogram of the Box-Muller Transform using the LCG

Table 3: Results of the Shapiro Test changing the values a, c, and m.

	a	c	m	p-value
Test 1	3067	4751	7919	0.2793
Test 2	1993	2293	3001	0.2514
Test 3	1049	1459	1931	0.9003
Test 4	463	701	1033	0.0091
Test 5	229	347	461	2.2×10^{-16}

shown in Figure 3. It can be seen with a higher probability the normality of the data using the Box-Muller transform with those values. Shapiro Test throws a p-value of 0.4087, so it can be concluded data may proceed from a normal distribution.

Other experiments are performed, for example, changing the values of a, c, and m, resulting in the loss of normality when those values decrease with the same seed. This tests are shown in Table 3.

Table 4 also show experiments of different values of a, c, and m, but this time making them as non-prime values. Results show that with these numbers, it is susceptible to applying them into the gaussian algorithm because p-values show in multiple tests that the generated sequence is not normally distributed.

3.3 A nonlinear Generator

Non-linear generators can be defined for example by using a nonlinear transition function f, or a nonlinear output function g, or by combining two or more linear random linear generation, etc. In Knuth [4] it is analyzed a generator based on a quadratic recurrence of the form:

$$x_i = (ax_{i-1}^2 + bx_{i-1} + c) \bmod m$$

Table 4: Results of the Shapiro test with non prime values of a, c, and m.

	a	c	m	p-value
Test 1	11000	27070	39720	0.06912
Test 2	11343	20505	35100	2.2×10^{-16}
Test 3	15250	35400	65104	4.97×10^{-5}
Test 4	5640	10240	15800	1.83×10^{-9}
Test 5	2505	5005	15145	4.13×10^{-16}

For this generator the author gave conditions for a maximal period of m. If $m = 2^e$, then the period is maximal if and only if a is even, $(b - a - 1) \bmod 4 = 0$, and c is odd. The code used for this method is shown below and is adapted to return values in the interval (0,1). It is also tested with the `uniform.test` function or R, with $p - value$ of 0.1459, evidencing uniformity in data.

```r
nonlinearGen <- function (n, seed) {
  a <- 20  #an even number
  b <- 41
  c <- 15 #an odd number
  m <- 2^exp(1)
  x <- seed
  gen_data <- numeric()
  while (length(gen_data)<n){
    x <- (a * x^2 + b * x + c) %% m
    gen_data <- c(gen_data,x)
  }
  return (gen_data/(m-1))
}
```

codes/knuth.R

data.txt

```
        Chi-squared test for given probabilities

data:  hist.output$counts
X-squared = 15.872, df = 11, p-value = 0.1459
```

References

[1] George EP Box. A note on the generation of random normal deviates. *Ann. Math. Stat.*, 29: 610–611, 1958.

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

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

[4] Donald E Knuth. *Art of computer programming, volume 2: Seminumerical algorithms*. Addison-Wesley Professional, 2014.

[5] Pierre l'Ecuyer. History of uniform random number generation. In *Winter Simulation Conference (WSC)*, pages 202–230. IEEE, 2017.

Homework Assignment 6: Applied Probabilistic Models

Statistical Tests

5273

1 Introduction

For this work, data is collected on the official website of Instituto Nacional de Estadística y Geografía (INEGI) [3]. The chosen section is Growth Domestic Product, and within this section, a national macroeconomic indicator is selected: Global Index of Economic Activity (IGAE) for the data analysis. Data obtained from INEGI website are in `csv` format, edited in order to work with the general values of the three main representative activities each month, which the aforementioned indicator is based on.

For the analysis it is used the R software in its version 4.0.2 [7] and the code used is available on the GitHub repository of [4]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Theoretical Aspects

This section covers several aspects of statistical and tests and when it is appropriate to use them and their interpretation. Characteristics of hypothesis tests are treated as well.

2.1 Hypothesis and statistical tests

A statistical test is a form of evaluating the evidence that data provide. The objective is to prove a hypothesis which is denominated as the null hypothesis (H_0). Typically, H_0 establishes equality (between means, variances, or correlation coefficient to cite some examples). H_0 usually opposed to a hypothesis denominate alternative (H_1) and implies difference (between means, for example).

2.2 Rejecting the null hypothesis

If data do not provide sufficient evidence against H_0, it is not rejected. If, contrary, show strong evidence against H_0 it is rejected, and H_1 it is considered as true with a quantified risk (low) of being wrong.

2.3 Interpretation of a statistical test

The statistical test produces a number denominated p-value, which limits are 0 and 1. The p-value is the probability of obtaining data or extreme data below the null hypothesis.

In more práctical terms, p-value should be compared with alpha (see subsection 2.4) and it can be interpreted follow:

- If $p <$ alpha, H_0 is rejected, and H_1 is accepted with a proportional risk of the p-value of being wrong.
- If $p >$ alpha, H_0 is not rejected, but this does not imply it is necessary to be accepted. It means H_0 is true, but the experiment and the statistical test are not "strong" enough to produce a p-value less than alpha.

2.4 Selecting alpha

When a study is designed, a risk threshold must be specified, above which H_0 should not be rejected. This threshold is known as significance level, also denoted as alpha or α, and should be between 0 and 1. Low alpha values are more conservative [9].

The selection of alpha should depend on how dangerous is reject H_0 in case it is true. For example, in a study to prove the benefits of medical treatment, alpha should be low. On the other hand, when revising effects of many attributes in the appreciation of a product, alpha could be moderate. In most cases, alpha is set at 0.05, 0.01, or 0.001.

2.5 Common mistakes when interpreting p-value

A mistake that can be made when interpreting p-value is if a study is realized two times and $p \leq 0.05$ in one case and $p > 0.05$ in another, its results are in conflict. This is a mistake to assume because most of the experimental series should throw some null result. Another one is when $p \leq 0.05$ assures that 95% of effects will be true. The true probability that a significative effect reflects the existence of a real effect is given by the Positive Predictive Value (PPV), and it depends on the statistical power, plausibility of alternative hypotheses, and other questionable research practices. Another one is that a $p > 0.05$ implies there is no effect. This may be, but the truth is that there is no power enough to detect it. The other one is that $p \leq 0.05$ implies a found effect. The valid use of p-value is to control the rate of error type 1. The$p$-value does not say anything about concrete cases (current experiment) it only provides a general rule of behaviour [5].

2.6 Statistical power

When various statistical tests are available to be used, it is critical to choose the most convenient and powerful for the experimental design to be used. For that reason, it is necessary a selection criterion [2]. One of them is the statistical power which is the probability of rejecting the null hypothesis when it is false. Therefore, it can be considered a statistical test as appropriate when the probability of rejecting the null hypothesis (H_0) is small when is true; and big enough the probability of rejecting it when is false.

2.7 Parametric and non-parametric statistical tests

Parametric tests are for numerical data and, in general, are based on the normal or gaussian distribution. Possible tests to apply are t-student, the Pearson correlation coefficient, linear regression,

Table 1: Map to select the appropriate test

		Data type			
		Numeric (gaussian)	**Ordinal or numeric (non gaussian)**	**Numeric (outliers)**	**Nominal (binary)**
Objective	**Compare 2 independent groups**	t-student test for 2 independent samples	Mann-Whitney Test	Yuen test for independent samples	Fisher or Chi-Squared (for bigger samples)
	Compare 2 related groups	t-student test for 2 related samples	Wilcoxon Test for related samples	Yuen Test for related samples	McNemar Test
	Compare 3 or more independent groups	ANOVA one-way for independent samples	Kruskall-Wallis Test	Robust ANOVA one-way for independent smples	Chi-Squared Test
	Compare 3 or more related groups	ANOVA one-way for related samples	Friedman Test	Robust ANOVA one-way for related samples	Q of Cochrane Test
	Associate 2 variables	Pearson correlation	Spearman or Kendall correlation	Robust correlation	Cramer V coefficient

one-way Analysis of variance (One-way ANOVA), factorial analysis of variance (ANOVA), and Analysis of covariance (ANCOVA). Descriptive statistics such as standard deviation, mode, median and mean are used.

On the other hand, non-parametric tests are used with nominal and ordinal variables. They do not assume a particular distribution and the requirements regarding sample size are not as many as the parametric ones. Most used tests are Chi-Squared, Correlation and independency coefficients for cross-tabulation and Spearman's and Kendall's rank correlation coefficient.

2.8 Guide to find the needed statistical test

To begin with, the objective of the analysis needs to be clear. It could be about association or comparison. Both seek to establish relations (similarities or differences) between elements, but tests of comparison, in contrast with a test of association, evaluate relations between one or various groups. The type of variables has to be noticed (numeric or nominal) as well. Then, it needs to be distinguished if samples are independent or related. The next step corresponds to the fulfilment of the classical assumptions (Normality, homogeneity of variance, independency), this allows to choose between parametric, non-parametric and robust tests. Then, it can be chosen the appropriate test with the help of Table 1. Once the test is selected, the next step is to perform hypothesis testing, and finally it corresponds to the interpretation and if possible graph the results.

2.9 Assumptions to apply parametric techniques

To apply parametric techniques the following assumptions are needed:

- Observations are independent of each other.
- Populations should be normally distributed.
- These populations must have the same variance.
- Variables must be measured on at least an interval scale so that arithmetic operations can be used.

3 Application of Statistical Tests

In this section, several of the most used statistical tests are performed. These tests are applied considering data of the IGAE obtained from the INEGI website. The IGAE is an index that approximates the calculation of the generated wealth in the country monthly. It is considered a trend index and marks the path that the national economic activity is reporting in the given month [8]. The method used to calculate the IGAE consists of monthly indexes of the physical volume of production for each of the selected classes, with a fixed base in 2013 [1]. In the calculation of the IGAE, three fundamental areas or activities are involved: primary, secondary, and tertiary which are the ones analyzed in this work.

3.1 One Sample t-Test

This parametric test is used to test if the mean of a sample from a normal distribution could reasonably be a specific value [6]. In this case, data belong to the primary activities. The p-value is 0.4594, so the null hypothesis that the mean=90 cannot be rejected.

data.txt

```
        One Sample t-test

data:  pri
t = 0.74072, df = 329, p-value = 0.4594
alternative hypothesis: true mean is not equal to 90
95 percent confidence interval:
 88.54626 93.20969
sample estimates:
mean of x
 90.87798
```

3.2 Wilcoxon Signed Rank Test

This test can be considered as an alternative to t-Test, especially when the data sample is not assumed to follow a normal distribution and it is a non-parametric method [6]. For this test, data belong to secondary activities. The p-value is less than the significance level of 0.05, so the null hypothesis is rejected and accept the alternative that true mean is not equal to 90.

data.txt

```
        Wilcoxon signed rank test with continuity correction

data:  as.numericsec
V = 33775, p-value = 0.0001926
alternative hypothesis: true location is not equal to 90
95 percent confidence interval:
 91.41908 94.14020
sample estimates:
pseudomedian
      92.88731
```

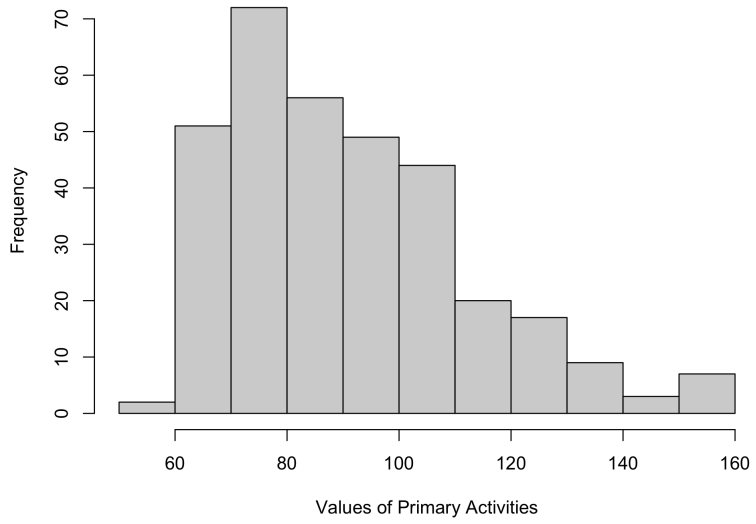


Figure 1: Histogram of the values of Primary Activiites

3.3 Two Sample t-Test and Wilcoxon Rank Sum Test

Here both t-Test and Wilcoxon rank test are used to compare the mean of 2 samples, passing the numeric vectors of primary and secondary activities. The p-value is 0.9976, so the null hypothesis cannot be rejected.

data.txt

```
        Wilcoxon rank sum test with continuity correction

data:  as.numericpri and as.numericsec
W = 47533, p-value = 0.9976
alternative hypothesis: true location shift is greater than 0
```

3.4 Shapiro Test

This is used to test if a sample follows a normal distribution. Data belong to primary activities. In this case, p-value is 3.29×10^{-10}, so the null hypothesis is rejected, which means data are not normally distributed. Figure 1 shows a histogram of the values reached by primary activities.

data.txt

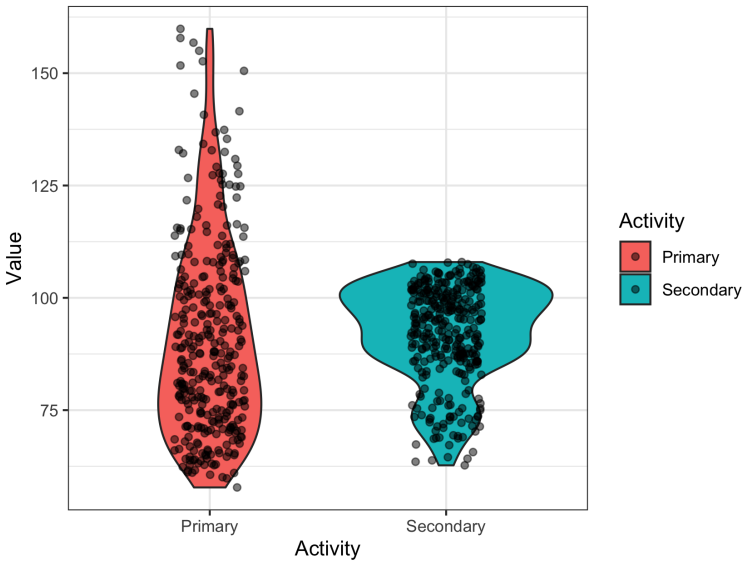


Figure 2: Violinplot of Primary and Secondary Activities

```
        Shapiro-Wilk normality test

data:  as.numericpri
W = 0.94082, p-value = 3.294e-10
```

3.5 Kolmogorov-Smirnov Test

This test is used to check whether two samples follow the same distribution. In this case, data belong to primary and secondary activities; the p-value is 3.70×10^{-10}, so the null hypothesis is rejected, which means data do not come from the same distribution. A violin plot is generated in Figure 2 to show how both groups are distributed.

`data.txt`

```
        Two-sample Kolmogorov-Smirnov test

data:  as.numericpri and as.numericsec
D = 0.26061, p-value = 3.695e-10
alternative hypothesis: two-sided
```

3.6 Fisher's F-Test

Fisher's F-test can be used to check if two samples have the same variance [6]. For this case, data of secondary and tertiary activities are taken as samples. The p-value is 2.20×10^{-16}, so the null hypothesis is rejected, which means the two samples do not have the same variance.

data.txt

```
        F test to compare two variances

data:  as.numericsec and as.numericter
F = 0.341, num df = 329, denom df = 329, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2746199 0.4234194
sample estimates:
ratio of variances
         0.3409977
```

3.7 Chi-Squared Test

It can be used to test if two categorical variables are dependent, by means of a contingency table [6]. In this case, the experiment is performed to evaluate if the results of the level of activities that conform to the IGAE is independent to the months in the year 1993.

For this test, there are two ways to tell if the variables are independent. The first one if by looking at the p-value, which in this case is 0.9931, as is greater than 0.05, the null hypothesis that the two variables are independent can not be rejected. Also, if look at the calculated Chi-Squared value, it is 9.0381, which is less than 33.9244 (critical value), so it also indicates the null hypothesis can not be rejected. In conclusion, as the two approaches lead to the same conclusion, it can be said there is no evidence to reject that the two variables are independent.

data.txt

```
        Pearson's Chi-squared test

data:  data_trimmed
X-squared = 9.0381, df = 22, p-value = 0.9931

Critical Value:  33.92444
```

3.8 Correlation

Correlation is used to test the linear relationship of two continuous variables. In this case, data come from primary and secondary activities. The p-value is 2.20×10^{-16}, which indicates the null hypothesis that there is a true correlation between the two variables is rejected. Figure 3 a scatter plot is generated in order to show how data is distributed.

data.txt

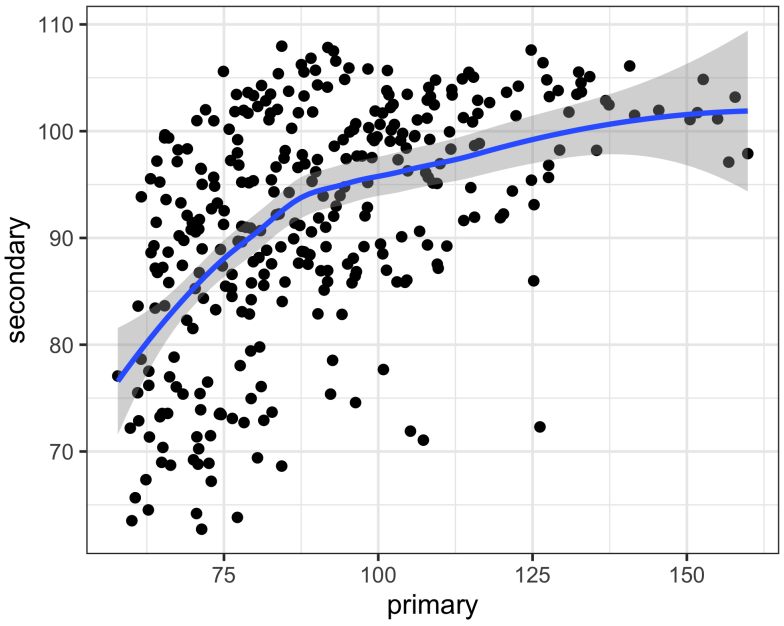


Figure 3: Scatterplot of Primary and Secondary Activities

```
        Pearson's product-moment correlation

data:  as.numericpri and as.numericsec
t = 10.528, df = 328, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4172422 0.5791088
sample estimates:
      cor
0.5025669
```

References

[1] Instituto Nacional de Estadiística y Geografía. Sistema de cuentas nacionales de México. Fuentes y metodologías. año base 2013. Indicador Global de la Actividad Económica, 2017. https://www.inegi.org.mx/contenidos/programas/igae/2013/metodologias/SCNM_Metodo_IGAE_B2013.pdf, Last accessed on 2020-09-05.

[2] EcuRed. Pruebas estadísticas, 2019. https://www.ecured.cu/Pruebas_estad%C3%ADsticas, Last accessed on 2020-10-07.

[3] INEGI. Datos, 2020. https://www.inegi.org.mx/temas/igae/, Last accessed on 2020-09-05.

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

[5] José C. Perales. Errores comunes en la interpretación del valor p, 2018. https://masalladelpvalor.wordpress.com/2018/12/04/unidad-3-errores-comunes-en-la-interpretacion-del-valor-p/, Last accessed on 2020-10-08.

[6] Selva Prabhakaran. Statistical tests, 2016. http://r-statistics.co/Statistical-Tests-in-R.html, Last accessed on 2020-10-11.

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

[8] Eduardo Torreblanca. ¿En qué consiste el IGAE?, 2016. https://mvsnoticias.com/noticias/economia/opinion-en-que-consiste-el-igae-610/, Last accessed on 2020-09-05.

[9] XLSTAT. ¿Qué es una prueba estadística?, 2019. https://help.xlstat.com/s/article/que-es-una-prueba-estadistica?language=es#:~:text=Una%20prueba%20estad%C3%ADstica%20es%20una,nula%2C%20y%20suele%20denominarse%20H0.&text=H0%20normalmente%20se%20opone%20a,alternativa%2C%20denominada%20H1%20o%20Ha, Last accessed on 2020-10-07.

Homework Assignment 7 Corrections

In this homework, grammatical mistakes were corrected as well as some typos.

Homework Assignment 7: Applied Probabilistic Models

Curve Fitting

5273

1 Introduction

Fitting a distribution from a dataset consists of finding the parameters' value, which, with greater probability, that distribution could have generated the observed data [7]. For example, the normal distribution has two parameters (mean and variance); once these two parameters are known, the entire distribution is known.

For the analysis, the R software is used in its version 4.0.2 [8], and the code used is available on the GitHub repository of [5]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Data

For this work, four functions have been created. Independent variables x_1, x_2 are the result of generated values using the R function `runif()`, x_3 is generated by the function `rchisq()` with two degrees of freedom, and x_4, using `rnorm` with its defuault values of mean 0 and standard deviation of 1.

$$y = 4x_1 + 5x_2, \tag{1}$$

$$y = (x_1)^2 + \texttt{rexp(1)}, \tag{2}$$

$$y = e^2(x_1)^2(x_2)^3, \tag{3}$$

$$y = (4x_1)^3(20x_2) + \frac{x_3}{5} + \log(x_4)^2. \tag{4}$$

In real case scenarios, the relation of the variables is unknown most of the time, meaning that a model that best describes this relation has to be found.

3 Experiments

The experiments for this work are based on the generated functions described in the previous section. The parameter used is the number of repetitions or sample size (n), with a value of 100. The previous functions are used in the transformations described in this report and analyzed one of them in each section of the experiments.

3.1 Multiple Linear Regression

For the Function 1, and assuming that the relationship between variables is unknown, it is considered to find the parameters or coefficient that best described this relation. In this case, it could be useful to perform a regression analysis.

data.txt

```
Call:
lmformula = y ~ x1 + x2, data = df1

Residuals:
       Min         1Q     Median         3Q        Max
-9.265e-15 -7.650e-17  9.510e-17  2.450e-16  6.130e-15

Coefficients:
             Estimate Std. Error   t value Pr>t
Intercept 1.421e-15  3.546e-16 4.008e+00  0.00012 ***

x1          4.000e+00  4.560e-16 8.771e+15  < 2e-16 ***

x2          5.000e+00  4.619e-16 1.083e+16  < 2e-16 ***

---

Residual standard error: 1.305e-15 on 97 degrees of freedom
Multiple R-squared:      1,        Adjusted R-squared:      1
F-statistic: 8.979e+31 on 2 and 97 DF,  p-value: < 2.2e-16
```

Regression in R software can be performed with the `lm()` function, which can be obtained the coefficients that best described a relation given by the form $Y = \beta_1 + \beta_2 X + \epsilon$ [6]. In this case, lets assume that it is known that function y depends on x_1 and x_2. For this reason, a multiple linear regression analysis is needed. This is an extension of simple linear regression used to predict an outcome variable (y) based on multiple distinct predictor variables (x). With two predictor variables (x), the prediction of y is expressed by the equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ [2]. Once the experiment is performed it throws in the column "Estimate" of the R output these β coefficients. In this case $\beta_1 = 4$ and $\beta_2 = 5$ which corresponds to the values previously declared when generating the function. The intercept value correspond with an error when estimate this relation.

3.2 Box-Cox Transformation

In some cases, if assumptions of the simplicity of structure for $E(y)$), the constancy of error variance and normality of distributions are not satisfied in terms of the original observations, a non-linear transformation of y may improve the analysis [1]. Box-Cox transformation is given by the equation:

$$y^{\lambda} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & if \quad \lambda \neq 0, \\ \\ \log y & if \quad \lambda = 0. \end{cases} \tag{5}$$

Figure 1 show plots of the diagnosis of function 2, fitting a model without doing a transformation first. The "Residual vs Fitted" plot shows a more concentrate residual points between 0 and -1 lines.

The "Quantile-Quantile" plot shows the normality of the errors, and it is not great specially at the upper tail, which goes off from the straight line. The "Scale-Location" plot, which shows there is no homoscedasticity and the "Residuals vs. Leverage" plot which shows if there are outliers in the residuals.

Then, Figure 2 shows a Maximum Likelihood plot for the values of λ with a 95% confidence interval. This plot gives a threshold with bounds of the recommended λ values for the Box-Cox transformation. Then the best value can be extracted, which represents the value when the curve is higher. This threshold can be seen by simple inspection; it moves from 0.1 to 0.5 approximately. Finally, the calculated value of λ is 0.263.

Figure 2 represents the same model diagnosis, but after performing the Box-Cox transformation with the previously calculated parameter, λ=0.263. Here main changes can be appreciated in the "Quantile-Quantile" Plot, which shows a better approximation to the normal line.

3.3 Log Transformation

Log transformations are another method that can be valuable for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics [4]. This log transformation it is used in Function 3, which depends of the variables x_1 and x_2.

data.txt

```
Intercept     logx1     logx2
  1.4739959   0.3864703   0.4155402
```

In the R output show for this section, the logarithm's coefficient for each independent variable and the intercept's value can be seen.

3.4 Tukey's Ladder of Power

Tukey [9] describes an orderly way of re-expressing variables using a power transformation suggesting exploring simple relationships such as $y^{\lambda} = b_0 + b_1 X$ where λ is a parameter chosen to make the relationship as close to a straight line as possible. Table 1 shows examples of the Tukey's ladder of transformations.

data.txt

```
    lambda      W Shapiro.p.value
428  0.675 0.9584        0.003112

if lambda >  0{TRANS = x ^ lambda}
if lambda == 0{TRANS = logx}
if lambda <  0{TRANS = -1 * x ^ lambda}
```

In the above output corresponding to the `transformTukey()` function of R, applied to Function 2 it shows the value of λ which data could be transformed, which is $\lambda = 0.675$ and the transformation should be x^{λ}. Figure 4 shows a histogram of the transformation.

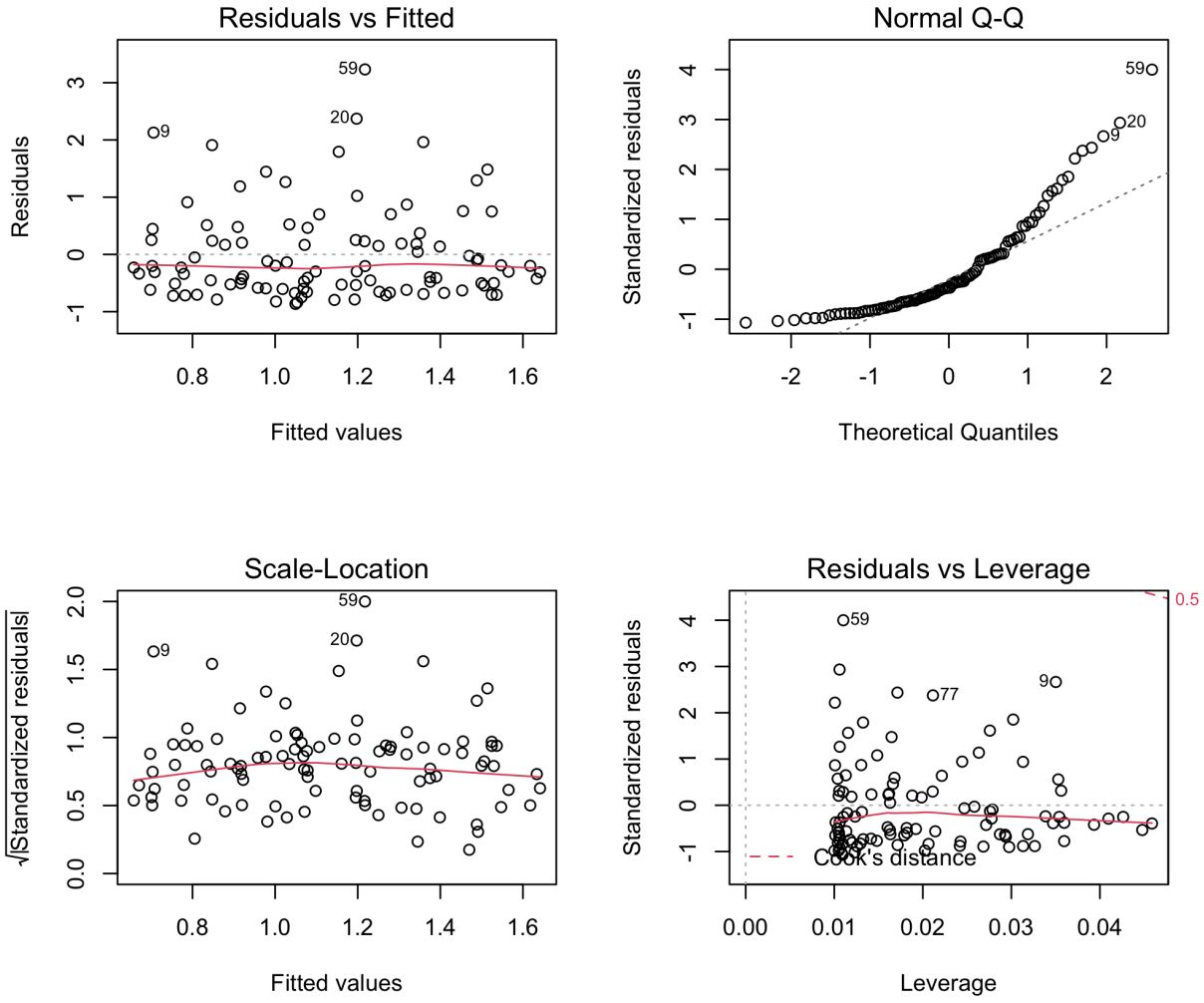


Figure 1: Plots of the Function 2 without transformations

Table 1: Tukey's Ladder of Transformations

λ	-2	-1	$-\frac{1}{2}$	0	$\frac{1}{2}$	1	2
y	$\frac{1}{x^2}$	$\frac{1}{x}$	$\frac{1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	x	x^2

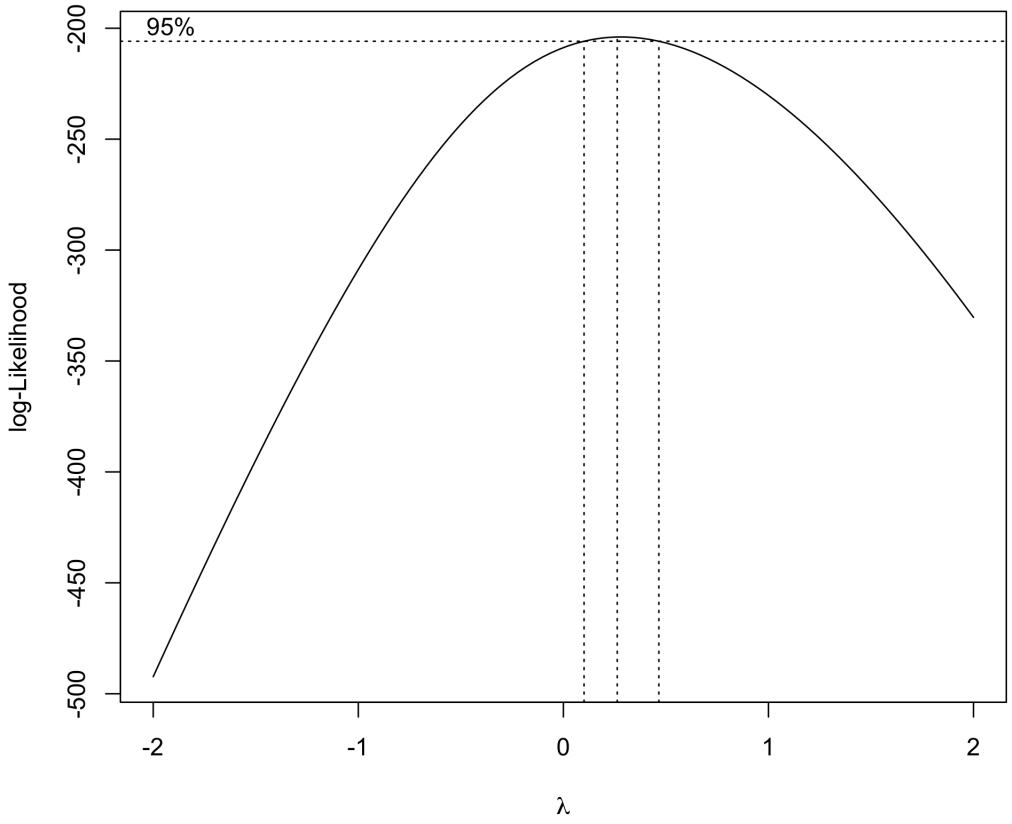


Figure 2: Max-Likelihood plot to determine the best λ for the Box-Cox Transformation

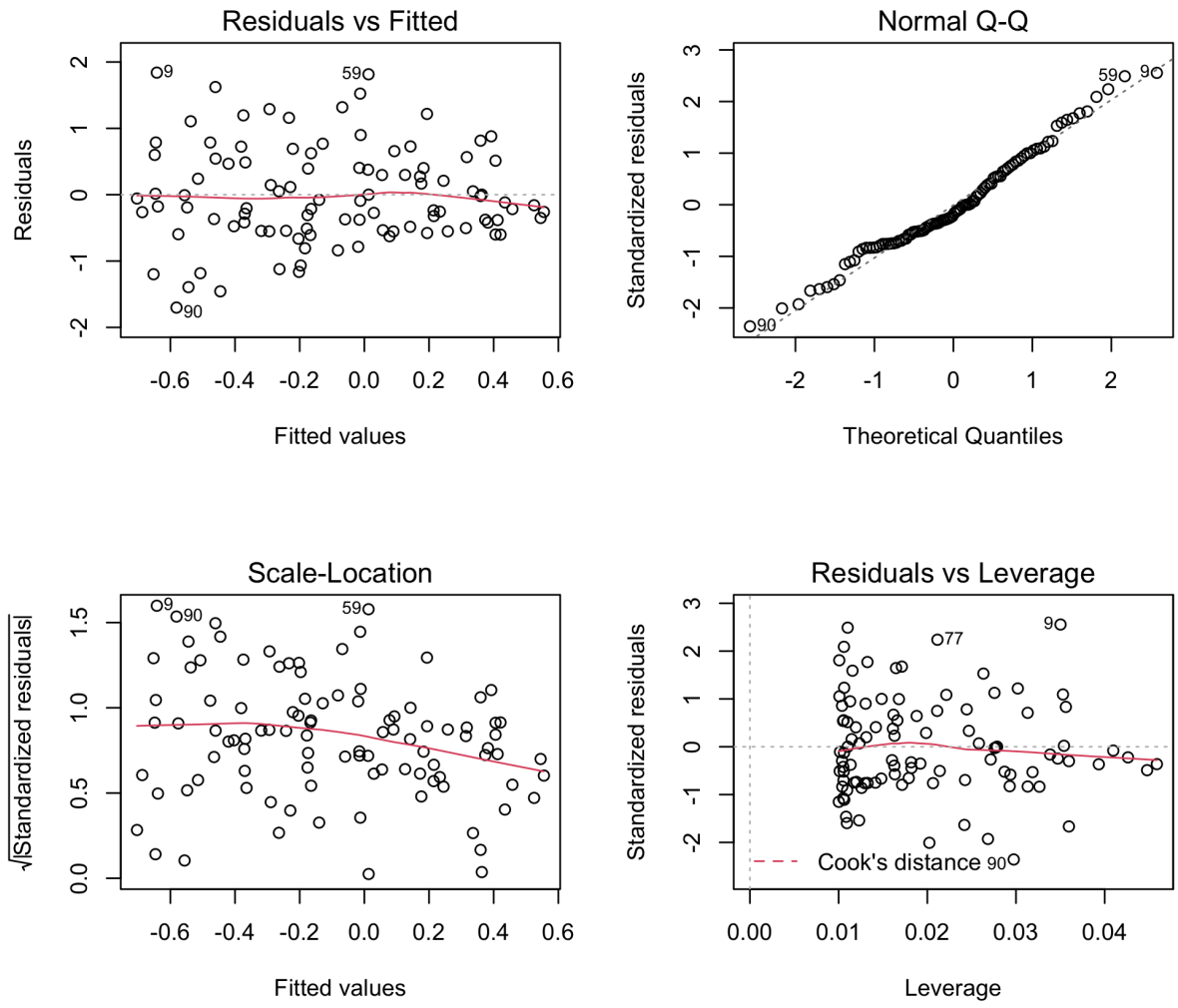


Figure 3: Plots of the Function 2 using the Box-Cox Transformation

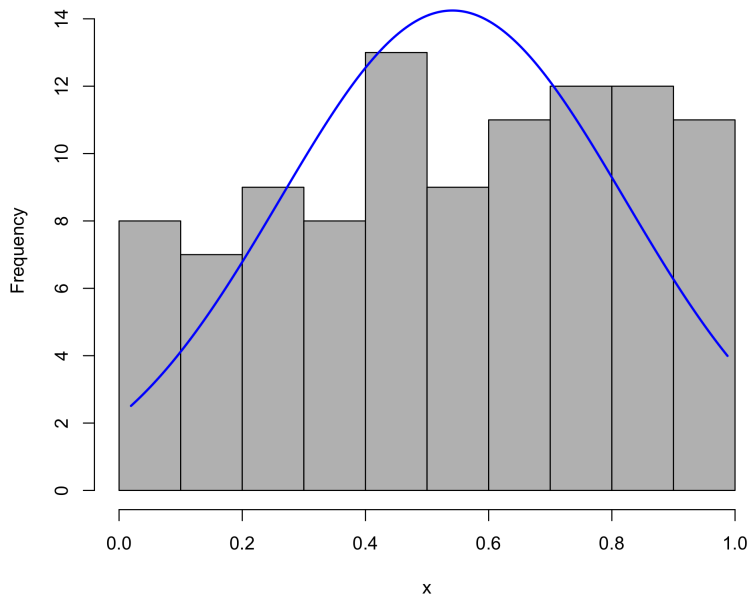


Figure 4: Histogram of Function 2 using the Tukey Ladder of Power Transformation

3.5 Stepwise Regression

The stepwise regression (or stepwise selection) consists of iteratively adding and removing predictors in the predictive model to find the subset of variables in the data set, resulting in the best performing model, that is, a model that lowers prediction error [3].

Here a comparison of the resulting fit model of multiple linear regression with the stepwise regression is performed. The R output showed below shows the results of fitting data applying multiple linear regression, indicating the intercept and values of the coefficient of all x variables.

data.txt

```
Call:
lmformula = y ~ ., data = df5

Residuals:
    Min      1Q  Median      3Q     Max
-15.162  -5.996  -2.152   3.966  34.388

Coefficients:
            Estimate Std. Error t value Pr>t
Intercept -16.9595     2.2949  -7.390 5.67e-11 ***

x1           31.8100     2.8789  11.049  < 2e-16 ***
x2           20.0134     2.8394   7.049 2.87e-10 ***
x3           -0.2517     0.4446  -0.566    0.573
x4            0.3228     0.8442   0.382    0.703
---
```

Table 2: Resulting Models from the two methods for Function 4

Multiple Linear Regression Model	**Stepwise Regression Model**
$31.81x_1 + 20.02x_2 - 0.25x_3 + 0.32x_4 - 16.96$	$43.49x_1 + 26.48x_2 - 27.09$

```
Residual standard error: 8.5 on 95 degrees of freedom
Multiple R-squared:  0.6305,        Adjusted R-squared:  0.6149
F-statistic: 40.52 on 4 and 95 DF,  p-value: < 2.2e-16
```

This last R output shows the final model results by performing the stepwise regression, where it can be seen it removes variables x_3 and x_4 from the initial model to predict y, in contrast with the multiple linear regression. Finally, it can be expressed the final models resulting form this two methods in Table 2.

data.txt

```
Subset selection object
4 Variables  and intercept
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
1 subsets of each size up to 2
Selection Algorithm: backward
         x1  x2  x3  x4
1   1  "*" " " " " " "
2   1  "*" "*" " " " "
Intercept          x1          x2
  -27.09467    43.49111    26.48317
```

References

[1] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[2] Alboukadel Kassambara. Multiple linear regression in R, 2018. `http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/`, Last accessed on 2020-10-17.

[3] Alboukadel Kassambara. Stepwise regression essentials in R, 2018. `http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/`, Last accessed on 2020-10-17.

[4] David Lane, David Scott, Mikki Hebl, Rudy Guerra, Dan Osherson, and Heidi Zimmer. *Introduction to statistics*. David Lane, 2003.

[5] Oscar Alejandro Hernadez Lopez. Probability in R. `https://github.com/oscaralejandro1907/probability-in-R/blob/master/assignment1/t1.R`, 2020.

[6] Selva Prabhakaran. Linear regression, 2016. `http://r-statistics.co/Linear-Regression.html`, Last accessed on 2020-10-17.

[7] Joaquín Amat Rodrigo. Ajuste de distribuciones con R, 2020. `https://www.cienciadedatos.net/documentos/55_ajuste_distribuciones_con_r.html`, Last accessed on 2020-10-15.

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

[9] John W Tukey. *Exploratory data analysis*, volume 2. Reading, MA, 1977.

Homework Assignment 8: Applied Probabilistic Models

Bayes' Theorem

5273

1 Introduction

In this work, are analyzed some applications of Bayes' theorem. Several documents discuss the interpretation of Bayes' theorem in COVID-19 test results and its true accuracy and how data can be interpreted in subjects which have been tested to find out if have the disease. Several of these documents prove how applying Bayes' theorem may lead to counterintuitive results.

For the analysis, the R software is used in its version 4.0.2 [11], and the code used is available on the GitHub repository of [8]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Document discussion

To begin with, some basic concepts treated in the documents need to be reviewed, adapted with COVID-19 tests.

1. True positive: A person with COVID-19 tests positive for COVID-19.
2. False positive: A person without COVID-19 tests positive for COVID-19.
3. False negative: A person with COVID-19 tests negative for COVID-19.
4. True negative: A person without COVID-19 tests negative for COVID-19.

The term sensitivity is the probability that a person tests positive, given that they have the disease. The specificity is the probability that a person tests negative, given that they do not have the disease; also, the terms accuracy and precision can be resumed in Table 1.

Equation 1 refers to the Bayes's theorem. In this equation, $P(A)$ is sometimes called the base rate. $P(B)$ can be expressed in Equation 2, where the term $notA$ means "not the case". The conditional probability $P(A \mid B)$is what we want to find out.

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}, \tag{1}$$

$$P(B) = P(A) * P(B \mid A) + P(notA) * P(B \mid notA). \tag{2}$$

Table 1: Concepts about tests

Concept	**Interpretation**
Accuracy	$\frac{\text{true positives + true negatives}}{\text{all results}}$
Precision	$\frac{\text{true positives}}{\text{true positive + false positive}}$
Sensitivity	$\frac{\text{true positives}}{\text{true positive + false negative}}$
Specificity	$\frac{\text{true negatives}}{\text{true negative + false positive}}$

Ranjan [9] starts saying that no test is 100% accurate to detect the coronavirus. However, it is common to hear tests that are 98.5% accurate in detecting COVID infections, but it is important to know what this accuracy means.

In Lewis [6] shows how the probability of having COVID-19 given a positive test result depends on numbers, about which there is some uncertainty. The author compares three scenarios based on the number of COVID-19 cases in the United States (US) on April 6^{th}. By that time, the US had 336 000 confirmed cases and a population of about 329.4 million. That gives a probability of having COVID-19, let say $P(A) = 0.001$ and consequently a 0.999 value of $P(notA)$. For illustrative purposes, let assume a test with a sensitivity of 99% is owned, which gives a $P(B \mid A) = 0.99$ and 1% of those who do not have the disease test positive for it (false positives), this gives a $P(B \mid notA) = 0.999$. This example constitutes the first scenario; the second one is that this confirmed number of cases is underestimated by a factor of 10, as suggested by Dr Dean Blumberg of UC Davis Children's Hospital [4], and the third scenario is a hypothetical one, where it is assumed that the factor is underestimated by a factor of 100. Applying the Bayes' Theorem to those situations, results are shown in Table 2.

The first scenario shows that only about 9 of every 100 people who test positive would actually be Covid-19 cases, which implies a lot of false positives. In the second one the base rate increase about 1% and the probability that someone has COVID-19 given that they test positive for it is about 50%. In the third scenario, the rate increase by about 10% and the calculated probability is about 92%. With these experiments, it can be seen a pattern that even when using a very sensitive test, of 99%, the lower the base rate of the disease the more likely it is to obtain false positives.

A similar example is given in Ranjan [9], where there is a case in which a random person from a population is picked up and tested. He tested positive, and what we know is the probability that given a person who has the disease, the test will be positive. Again, assuming a high sensitivity of the test (99%), the interest is to find the probability that given a person tests positive, he actually has the virus. That probability is less than 0.5%. If there is an area where chances of catching the virus have increased 10 fold, results will not differ much from the above. With that being said, the author explains why test random people for COVID-19 would not be a wise idea. In contrast, Bello [1] shares a different opinion, supporting the idea that testing is one of the most important tools to slow and reduce the spread and impact of a virus.

Good et al. [3], applied Bayesian analysis to interpret negative and positive COVID-19 polymerase chain reaction (PCR) assay results for two clinical scenarios. The first one estimated with a high pre-test probability of infection at 90% and the second one the opposite with an estimate of up to 10% of infection. Results shows for the first scenario, a post-test probability of a false negative test ranged from 47 to 73%; on the other hand, the second scenario this probability ranged from 0.5 to 3.2%.

Table 2: Probability that a person has Covid-19 given that they have tested positive for it.

	Scenario 1	**Scenario 2**	**Scenario 3**
$P(A \mid B)$	0.09	0.5	0.92

With PCR testing, false negative tests are concerning, potentially leading to an inappropriate sense of security. Screening tests are performed in Chan [2], where can also be concluded that a negative test result, in this paradigm, is never absolutely negative. Rather it adjusts the pre-test probability of having disease lower.

3 Experiments

Data for the analysis were collected from Mario Romero [7], which were transcribed from a database on the Serendipia website [10]. Data are updated at the time of writing (October 25^{th}, 2020), and it shows a cumulative of the confirmed cased by states of Mexico, daily updated. The objective of the experiment is to apply the Bayes' theorem to calculate the conditional probability that a person has COVID-19, given that he tested positive.

For this calculation event, A is a subject who has COVID-19 and event B is a test with a positive result. It is assumed a test with a result of 99% of sensitivity ($P(B \mid A) = 0.99$) and 1% of false positive results (those who do not have the disease but test positives), which constitute $P(B \mid A) = 0.01$. The base rate is calculated from the total confirmed cases reported in the database divided by the Mexican population and, with this number, conversely it is obtained the probability of not having the virus. With all these values, Bayes' theorem can be applied and calculate the desired probability:

$$P(A \mid B) = \frac{(0.007)(0.99)}{(0.007)(0.99) + (0.993)(0.01)} = 0.4132.$$

Therefore, if a random person is tested positive, there is a chance of 41,32% that he actually is infected.

Figure 1 shows a graph were are functios that calculate the sensitivity, specificity and predictive values and prevalence of a test can be seen. The positive predictive value (ppv) is defined as the percent of predicted positives that are actually positive while the negative predictive value (npv) is defined as the percent of negative positives that are actually negative [5].

Continuing with the example in Ranjan [9] of pool testing, and let the probability that a person living in Mexico has COVID-19 is 0.007 (as calculated before) to 0.012. If it is pooled x samples and probability that a person has the disease is p, then the probability that at least 1 person will have covid is given by:

$$P = 1 - (1 - p)^x. \tag{3}$$

If it is an interest to know the probability that at least 1 person has COVID-19, from a pool of x samples tested positive, to be more than 99% or might be less; Bayes' Theorem can also be applied. This result can be shown in Figure 2, which can be used to choose the number of samples for pool testing.

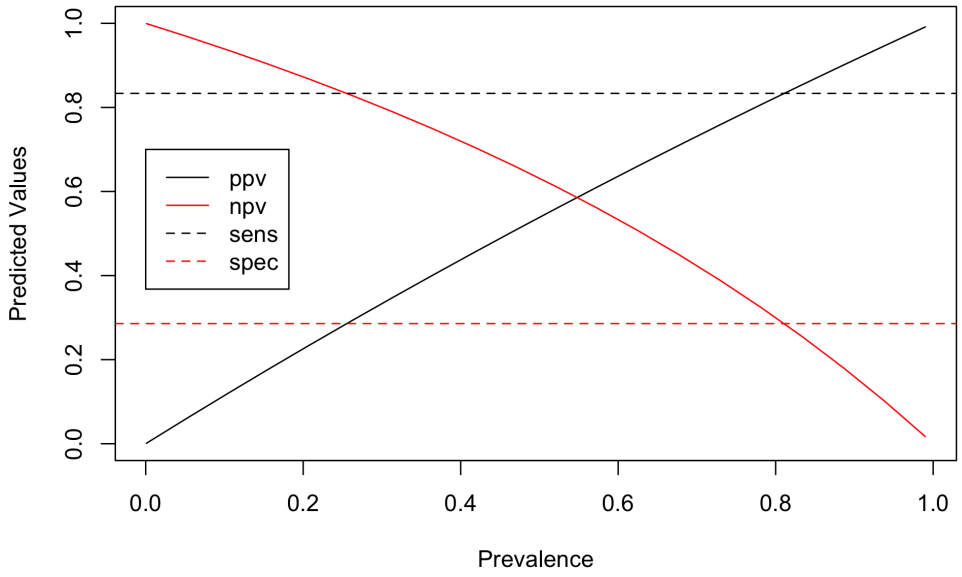


Figure 1: Plots of the probability values for sample size

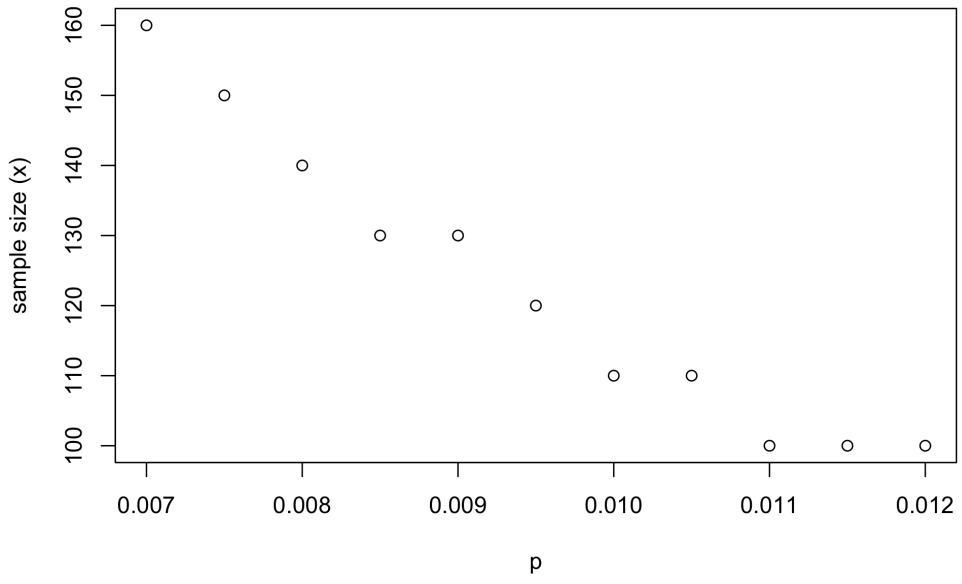


Figure 2: Plots of the probability values for sample size

References

[1] Miriam Bello. The Accuracy of COVID-19 Tests, 2020. https://mexicobusiness.news/health/news/accuracy-covid-19-tests, Last accessed on 2020-10-26.

[2] Gar Ming Chan. Bayes' theorem, covid19, and screening tests, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7315940/, Last accessed on 2020-10-24.

[3] Chester B. Good, Inmaculada Hernandez, and Kenneth Smith. Interpreting covid-19 test results: a bayesian approach, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7269418/, Last accessed on 2020-10-24.

[4] Nicole Karlis. US may have already failed to contain COVID-19, experts say, 2020. http://web.archive.org/web/20201005025223/https://www.salon.com/2020/03/12/us-may-have-already-failed-to-contain-covid-19-outbreak-experts-say/, Last accessed on 2020-10-25.

[5] Max Kuhn. Calculate sensitivity, specificity and predictive values, 2020. https://rdrr.io/cran/caret/man/sensitivity.html, Last accessed on 2020-10-26.

[6] Michael A. Lewis. Bayes' theorem and covid-19 testing, 2020. http://web.archive.org/web/20201005040306/https://www.significancemagazine.com/science/660-bayes-theorem-and-covid-19-testing, Last accessed on 2020-10-24.

[7] Mario Romero. COVID-19 Time Series. https://github.com/mariorz/covid19-mx-time-series/tree/master/data, 2020.

[8] Oscar Alejandro Hernadez Lopez. Probability in R. https://github.com/oscaralejandro1907/probability-in-R/blob/master/assignment1/t1.R, 2020.

[9] Archit Ranjan. COVID-19, Bayes' theorem and taking probabilistic decisions, 2020. https://towardsdatascience.com/covid-19-bayes-theorem-and-taking-data-driven-decisions-part-1-b61e2c2b3bea, Last accessed on 2020-10-24.

[10] Serendipia. Periodismo de datos. Datos abiertos sobre casos de Coronavirus COVID-19 en México. https://serendipia.digital/2020/03/datos-abiertos-sobre-casos-de-coronavirus-covid-19-en-mexico/, 2020.

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

Homework Assignment 9: Applied Probabilistic Models

Expected Value and Variance

5273

1 Exercises

Exercises solved in this work are provided on the book Grinstead and Snell [1].

1.1 Expected Value of Discrete Random Variables

For this exercises the expected value and the variance of discrete random variables are discussed.

Exercise 1 page 247

As the deck consists of cards between 2 through 10, it will have 36 cards. Let X be the number of the selected card. The player will win a dollar if the number of the card is odd and loses one dollar if the number is even, so the expected value of his winnings will be:

$$E(X) = -1\left(\frac{4}{36}\right)+1\left(\frac{4}{36}\right)-1\left(\frac{4}{36}\right)+1\left(\frac{4}{36}\right)-1\left(\frac{4}{36}\right)+1\left(\frac{4}{36}\right)-1\left(\frac{4}{36}\right)+1\left(\frac{4}{36}\right)-1\left(\frac{4}{36}\right) = -\frac{1}{9}.$$

Exercise 15 page 249

In this exercise, the game stops whenever it is one dollar profit or run out of gold balls. For one gold ball, a player wins one dollar and loses one dollar if a silver ball is drawn as the box contains two gold balls and three silver balls and let X be the results of the draws until the game is finished. There are seven possible outcomes, which are detailed below with its corresponding probability. Therefore, the expected value is:

$$E(X) = 1\left(\frac{2}{5}\right) + 1\left(\frac{1}{10}\right) + 0\left(\frac{1}{10}\right) + 0\left(\frac{1}{10}\right) - 1\left(\frac{1}{10}\right) - 1\left(\frac{1}{10}\right) - 1\left(\frac{1}{10}\right) = \frac{2}{10} = \frac{1}{5}.$$

Because $E(X) > 0$ it can be said this is a favorable game.

Exercise 18 page 249

Six similar keys are given, and let X be the number of tried keys before the success of opening the door.

$$E(X) = 0\left(\frac{1}{6}\right) + 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) = \frac{5}{2}.$$

Table 1: Possible outcomes for Exercise 15 page 249.

Outcome	**Probability**	**Profit**
G	$\left(\frac{2}{5}\right)$	1
SGG	$\left(\frac{3}{5}\right)\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)$	1
SGSG	$\left(\frac{3}{5}\right)\left(\frac{1}{2}\right)\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)$	0
SSGG	$\left(\frac{3}{5}\right)\left(\frac{1}{2}\right)\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)$	0
SSSGG	$\left(\frac{3}{5}\right)\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)(1)(1)$	-1
SGSSG	$\left(\frac{3}{5}\right)\left(\frac{1}{2}\right)\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)$	-1
SSGSG	$\left(\frac{3}{5}\right)\left(\frac{1}{2}\right)\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)$	-1

Exercise 19 page 249

For every correct answer, the student gets three points and for every incorrect one loses one point. The problem has four possible answers. Let X be the result if the answer is correct or incorrect. Therefore, the probability of choosing the correct answer just guessing is 0.25, whereas the probability of choosing the incorrect one is 0.75. The expected value is:

$$E(X) = 3(0.25) - 1(0.75) = 0.$$

Exercise 31 page 254

(a) For a pooled sample of k people, if the test is positive, it means that each person has a positive result on the test independently with probability p. Therefore, the probability that each person has a negative result is $1-p$. The probability that all of the k subjects have negative result is $(1-p)^k$. Consequently:

$$\begin{aligned} P(\text{sample is positive}) &= 1 - P(\text{all } k \text{ subjects have negative results}), \\ &= 1 - (1-p)^k. \end{aligned}$$

(b) Let X be the number of blood tests necessary under the plan (2). There are $\dfrac{N}{k}$ groups of k individuals. For each of these groups, if someone is positive, a $k+1$ tests are needed, and otherwise, only one test. The expected value for the number of tests for one group is:

$$(k+1)P(Positive) + 1 * P(Negative) = (k+1)(1-(1-p)^k) + 1(1-p)^k,$$

For the whole group of N subjects is:

$$\begin{aligned} E(X) &= \frac{N}{k}\left[(k+1)(1-(1-p)^k) + 1(1-p)^k\right] \\ &= \frac{N}{k}\left[(k+1) - (k+1)(1-p)^k + (1-p)^k\right] \\ &= \frac{N}{k}\left[k+1-k(1-p)^k - (1-p)^k + (1-p)^k\right] \\ &= \frac{N}{k}\left[k+1-k(1-p)^k\right]. \end{aligned}$$

(c) To minimize the expected value, it can be calculated the derivate of with respect to k and set equal to 0.

$$E(k) = \frac{N}{k}\left[k + 1 - k(1-p)^k\right],$$

$$\frac{\mathrm{d}E}{\mathrm{d}k} = 0$$
$$\frac{-n\ln(1-p)(1-p)^k k^2 + N}{k^2} = 0,$$

If in the expression above, a sufficient small p is considered approximately 0, the equality would be:

$$\frac{-n\ln(1-\cancelto{0}{p})(1-p)^k k^2 + N}{k^2} = 0$$
$$\frac{N}{k^2} = 0,$$

When substituting the value of k to $\frac{1}{\sqrt{p}}$:

$$\frac{N}{k^2} = 0$$
$$\frac{N}{\left(\frac{1}{\sqrt{p}}\right)^2} = 0$$
$$\frac{N}{\left(\frac{1}{p}\right)} = 0$$
$$Np = 0.$$

At this point, if the value of p is sufficient small (approximately 0), the equality is fullfilled for a value of $k = \frac{1}{\sqrt{p}}$, where the expected number of test is minimum.

Exercise 1 page 263

If $\{S = -1, 0, 1\}$:

- The expected value would be:

$$E(X) = \frac{\sum(x)}{N} = -1\left(\frac{1}{3}\right) + 0\left(\frac{1}{3}\right) + 1\left(\frac{1}{3}\right) = 0.$$

- To calculate the variance it can be used Table 2. Then, the variance would be:

$$\sigma^2 = \frac{\sum(x-\bar{x})^2}{N} = \frac{(-1-0)^2 + (0-0)^2 + (1-0)^2}{3} = \frac{2}{3}.$$

- The standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{2}{3}} \approx 0.816$.

Table 2: Variance in Exercise 1 page 263.

x	$x-\bar{x}$	$(x-\bar{x})^2$
-1	-1	1
0	0	0
1	1	1
$\sum x_i = 0$		$\sum (x_i-\bar{x})^2 = 2$

1.2 Expected Value of Continuous Random Variables

This section corresponds to exercises of expected value and variance of continuous random variables.

Exercise 3 page 278

The expected value of a continuous random variable is given by

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx,$$

Therefore the expected lifetime of the light bulb would be:

$$\begin{aligned}
E(T) &= \int_0^{\infty} t(\lambda)^2 te^{-\lambda t}dt \\
\lambda^2 \int t^2 e^{-\lambda t}dt &= \lambda^2 \left(\frac{-t^2e^{-\lambda t}}{\lambda} - \int \frac{-2te^{-\lambda t}}{\lambda}dt \right) \\
&= \lambda^2 \left[\frac{-t^2e^{-\lambda t}}{\lambda} + \frac{2}{\lambda}\left(-\frac{te^{-\lambda t}}{\lambda} - \int \frac{-e^{-\lambda t}}{\lambda}dt \right) \right] \\
&= \lambda^2 \left[\frac{-t^2e^{-\lambda t}}{\lambda} + \frac{2}{\lambda}\left(-\frac{te^{-\lambda t}}{\lambda} + \frac{1}{\lambda}\left[\frac{-e^{-\lambda t}}{\lambda} \right] \right) \right] \\
&= \left[\frac{\left(-t^2\lambda^2 - 2t\lambda - 2\right)e^{-\lambda t}}{\lambda} \right]_0^{\infty} \\
&= \frac{2}{\lambda} \\
&= 40.
\end{aligned}$$

The variance would be:

$$\begin{aligned}
\lambda^2 \int t^3 e^{-\lambda t}dt &= \lambda^2 \left(\frac{-t^3e^{-\lambda t}}{\lambda} - \int \frac{-3t^2e^{-\lambda t}}{\lambda}dt \right) \\
&= \lambda^2 \left[\frac{-t^3e^{-\lambda t}}{\lambda} + \frac{3}{\lambda}\left(-\frac{t^2e^{-\lambda t}}{\lambda} - \frac{2te^{-\lambda t}}{\lambda^2} - \frac{2e^{\lambda t}}{\lambda^3} \right) \right] \\
&= \left[\frac{\left(-t^3\lambda^3 - 3t^2\lambda^2 - 6t\lambda - 6\right)e^{-\lambda t}}{\lambda^2} \right]_0^{\infty} \\
&= \frac{6}{\lambda^2} \\
&= 2400,
\end{aligned}$$

$$\begin{aligned} V(T) &= E(X^2 - E(X)^2) \\ &= 2400 - 1600 \\ &= 800. \end{aligned}$$

Exercise 12 page 280

The variables X and Y are independent, and both are uniformly distributed on $[0, 1]$. The expected value of both variables is given by:

$$\begin{aligned} E(X^Y) &= \int_0^1 \int_0^1 x^y f(x) f(y) dx dy \\ &= \int_0^1 \left[\frac{x^{y+1}}{y+1} \right]_0^1 dy \\ &= \int_0^1 \frac{1}{y+1} dy \\ &= [\ln(y+1)]_0^1 \\ &= \ln 2 \\ &\approx 0.6931. \end{aligned}$$

Results of the simulation are shown in Figure 1. It is performed in R software in its version 4.0.2 [3], and the code used is available on the GitHub repository of [2].

Exercise 28 page 284

The length of the needle is L (much bigger than 1). If it is dropped on a grid with a horizontal line, it will form an angle θ with intersections is $L \cos\theta$ and with a vertical line with intersections is $L \sin\theta$. The uniform probability density function of θ between 0 and $\frac{\pi}{2}$ is:

$$\begin{cases} \frac{2}{\pi} & \text{if } 0 \leq \theta \leq \frac{\pi}{2}, \\ 0 & \text{elsewhere.} \end{cases}$$

Therefore the average number of lines crossed approximately is:

$$\begin{aligned} a &= \frac{2}{\pi} \int_0^{\frac{\pi}{2}} L \left[cos\theta + sin\theta \right] d\theta \\ &= \frac{2L}{\pi} \left[sin\theta - cos\theta \right]_0^{\frac{\pi}{2}} \\ &= \frac{4L}{\pi}. \end{aligned}$$

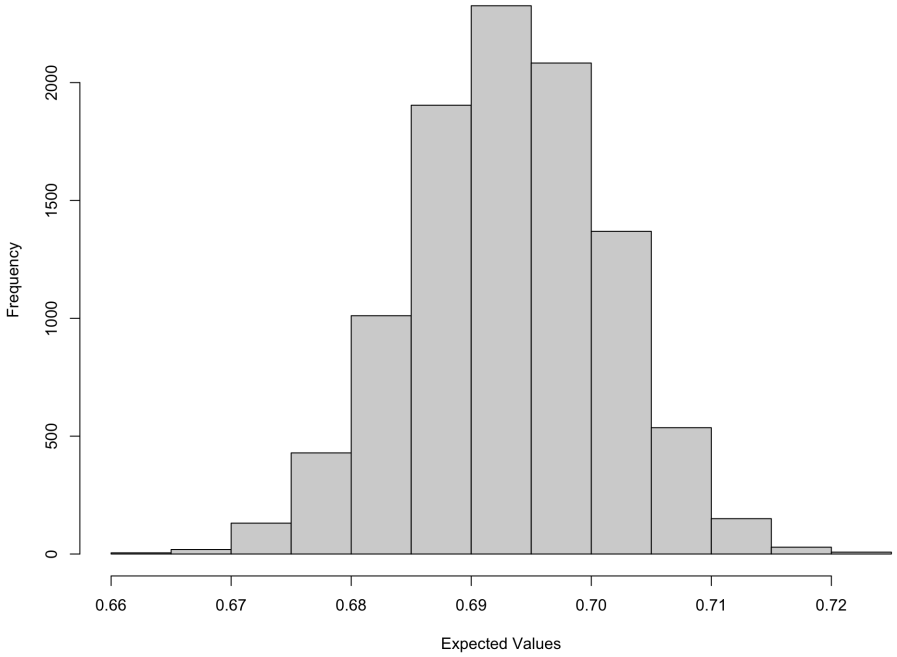


Figure 1: Histogram of the simulated expected values in Exercise 12 page 280

References

[1] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 2012.

[2] Oscar Alejandro Hernadez Lopez. Probability in R. `https://github.com/oscaralejandro1907/probability-in-R/blob/master/assignment1/t1.R`, 2020.

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

Homework Assignment 10 Corrections

In this homework, Exercise 18 page 249 is corrected, adjusting the number of iterations when Monte Carlo simulation is performed, eliminating pointless iterations due to the size of the sample space. Code is modified as well.

Homework Assignment 10: Applied Probabilistic Models

Monte Carlo Simulations

5273

1 Introduction

In this work, Monte Carlo simulations are employed to estimate expected value in the exercises. To perform simulations, sampling distributions are considered to compare the expected values with the theoretical values obtained in the previous assignment.

The expected value of a random variable can be considered as the long-run average value of its outcomes when the number of repeated trials is large [2], and Monte Carlo simulations play an important role for this kind of experiments. An example of Monte Carlo simulation is repeating an experiment a large enough number of times to make the results practically equivalent to doing it over and over forever [1].

For the analysis, the R software is used in its version 4.0.2 [5], and the code used is available on the GitHub repository [4]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Exercises

For this work exercises that have been simulated are provided in the book Grinstead and Snell [3].

2.1 Exercise 1 page 247

For this exercise, a deck is given which consists of cards between 2 through 10. The player will win a dollar if the number of the card is odd and loses one dollar otherwise. The calculated expected value $E(X) = -\frac{1}{9}$. Then it is executed by simulations (10 000 repetitions), which yields very close values with respect to the previously calculated. In turn, other repetitions of this experiment are executed, represented in Figure 1, which gives a threshold of all the expected values around the theoretical one.

2.2 Exercise 6 page 247

In this situation, let X denote the sum of two numbers that turn up when a die is rolled twice, and Y the difference of the numbers. This exercise requires demonstrate that $E(XY) = E(X)E(Y)$. As states before, simulations are executed to validate conclusions. The experiments are executed 10 000 times and then means are calculated the same amount of times also, Figure 2 shows very similar results.

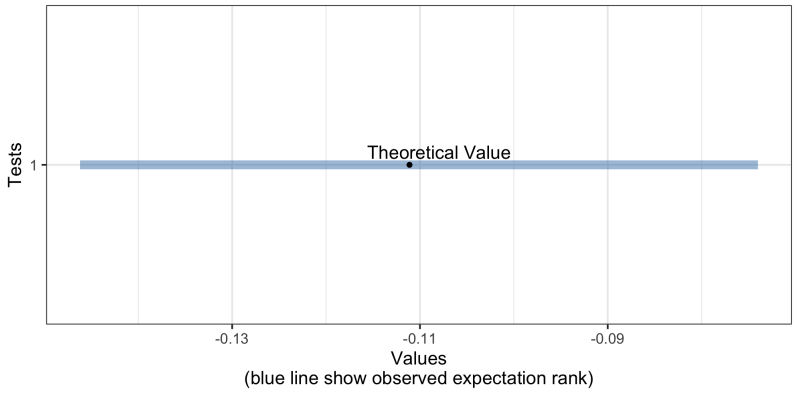


Figure 1: Expected value threshold in experiments in Exercise 1 page 247

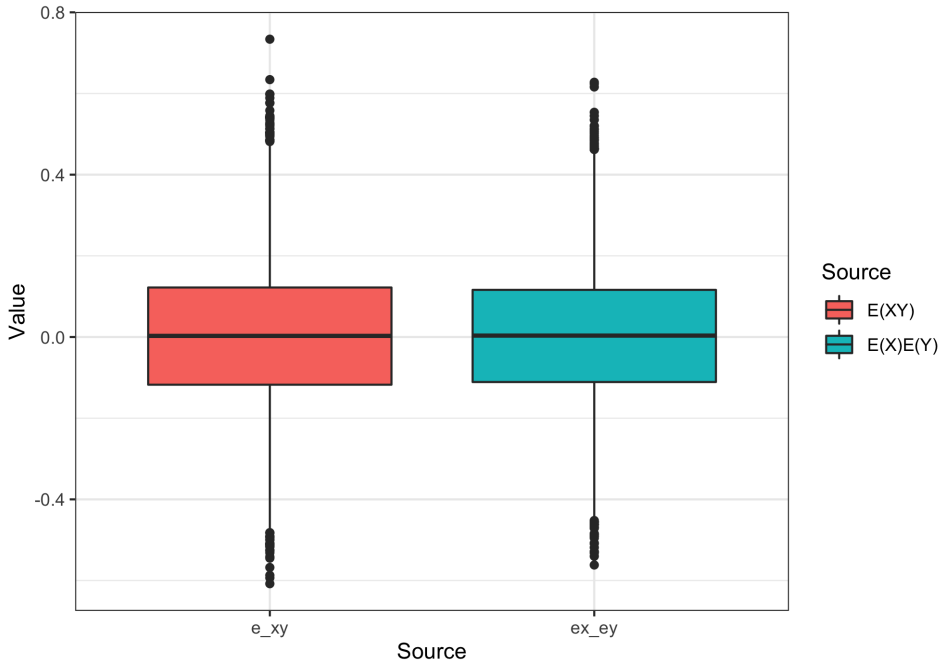


Figure 2: Boxplot of the expected values for both situations in Exercise 6 page 247

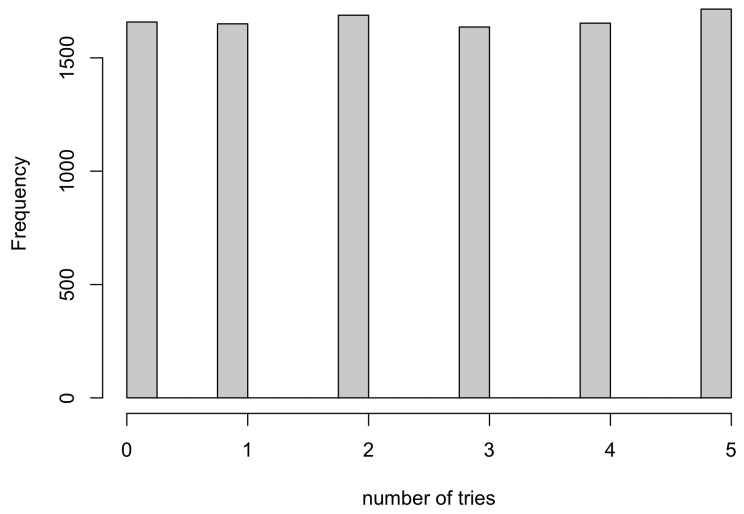


Figure 3: Histogram of the number of tries before opening the door in Exercise 18 page 249

2.3 Exercise 18 page 249

In this exercise, six similar keys are given and let X be the number of tried keys before the success of opening the door. This situation gives an expected value $E(X) = 2.5$. After computing, the sample means of 120 tries, which are the total of combinations, the experiment gives a value of 2.4818, which seems to be fairly close to the expected value. Figure 3 shows a histogram of the number of tries throughout the 120 repetitions of this Monte Carlo simulation. In addition, Figure 4 represents a plot that shows after the value of 120 repetitions the expected value begins to stabilize, which is an approach to determine how many Monte Carlo experiments are enough.

2.4 Exercise 3 page 278

For this exercise, the density function of the variable lifetime (in hours) of the ACME super light bulb is given. The calculated expected value $E(T) = 40$. The plot of the density function is generated and shown in Figure 5, which shows the highest area of the curve around the calculated expected value of 40.

2.5 Exercise 12 page 280

The variables X and Y are independent, and both are uniformly distributed on $[0, 1]$. The calculated expected value $E(X^Y) = \ln 2$; this value correspond approximately to 0.6931. Simulations are shown in the histogram of Figure 6 and Figure 7, the latter shows how the expected value begin to stabilize as the number of repetition grows around the previously calculated value.

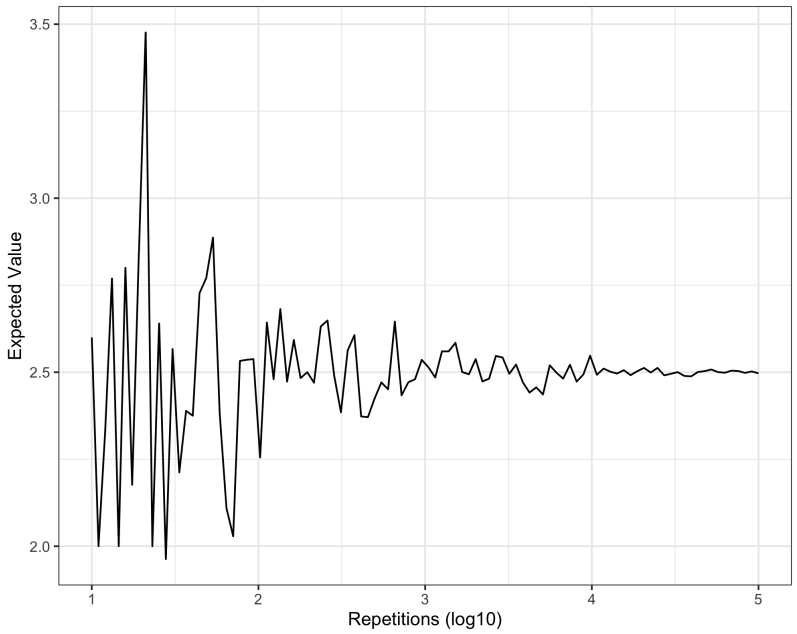


Figure 4: Stabilization of the expected value in the number of tries before opening the door in Exercise 18 page 249

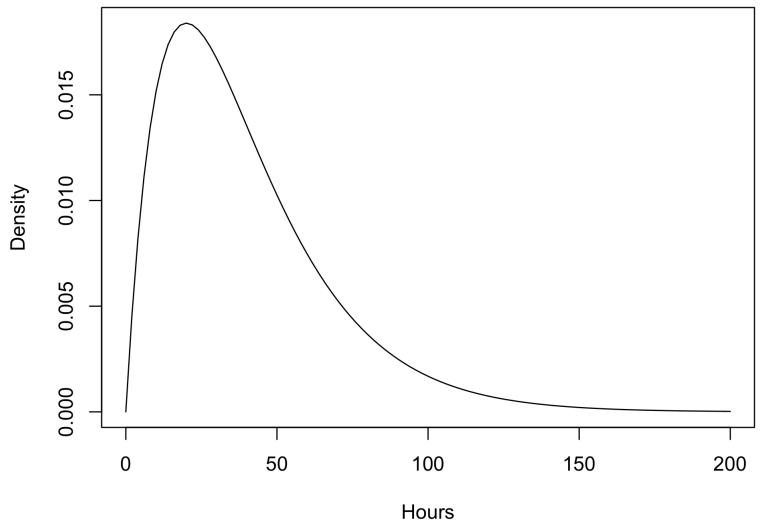


Figure 5: Density curve of the lifetime of the light bulb in Exercise 3 page 278

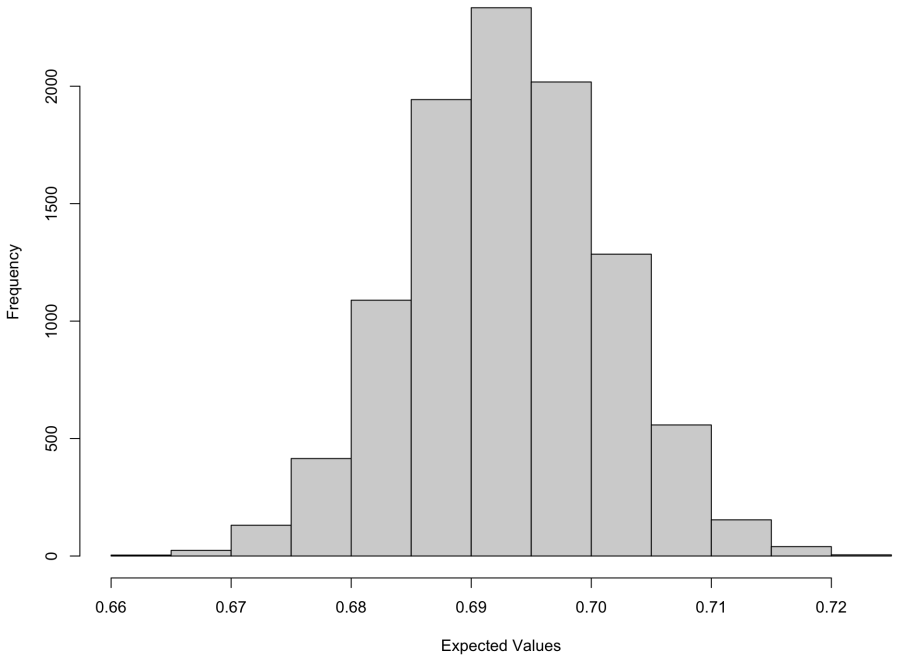


Figure 6: Histogram of the simulated expected values in Exercise 12 page 280

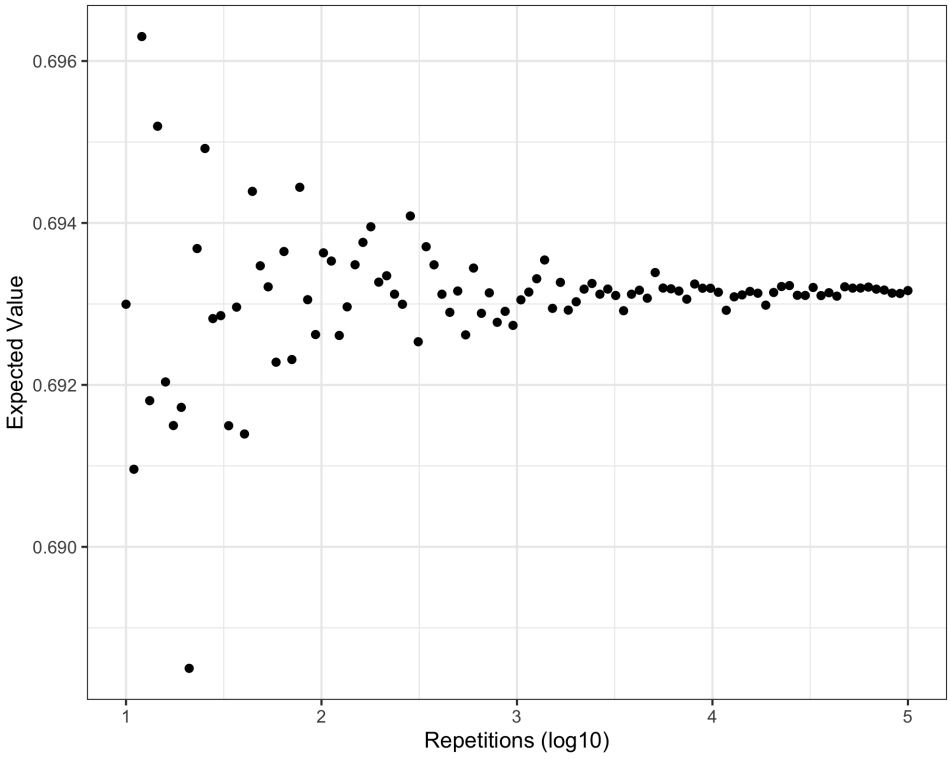


Figure 7: Stabilization of expected values in Exercise 12 page 280

References

[1] Faisal Akbar. Data science: Probability, 2019. https://rstudio-pubs-static.s3.amazonaws.com/487679_4087cbd1e9f74f8b8a4ddcd7eb0de0e6.html, Last accessed on 2020-11-8.

[2] Alexander Gerber Christoph Hanck, Martin Arnold and Martin Schmelzer. Introduction to econometrics with r, 2020. https://www.econometrics-with-r.org/index.html, Last accessed on 2020-11-5.

[3] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 2012.

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

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

Homework Assignment 11 Corrections

In this homework, is added the demonstrations of this properties:

- $\text{Cov}[aX + b, cY + d] = ac\text{Cov}[X, Y]$,
- $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$.

Homework Assignment 11: Applied Probabilistic Models

Multivariate distributions and probability densities

5273

1 Introduction

For this work aspects of multivariate probability densities and distribution are discussed. To begin with, a convolution of distribution example is performed, in the area of finance, specifically in stock prices of two companies. As a second point, a Chi-Squared test is performed for two categorical variables in the analysis of the results of my thesis work [4]. Finally, a numerical covariance analysis is performed, in order to prove two properties of this statistical value, also with prices on the stock market between two other enterprises.

For the analysis, the R software is used in its version 4.0.2 [6], and the code used is available on the GitHub repository [3]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Convolution of Distributions

This section is about the convolution of probability distributions. The convolution can be considered as the operation of forming linear combinations of random variables. Then, the probability density function (PDF) of a sum of random variables is the convolution of their corresponding PDF. This sum, let be $Z = X + Y$, for example, of two random variables, can be expressed as mentioned in Equation 1 for discrete variables and Equation 2 for continuous ones.

$$P(Z = z) = \sum_{k=-\infty}^{\infty} P(X = k)P(Y = z - k) \tag{1}$$

$$h(z) = (f * g)(z) = \int_{-\infty}^{\infty} f(z - t)g(t)dt = \int_{-\infty}^{\infty} f(t)g(z - t)dt \tag{2}$$

For the convolution example data are downloaded from Yahoo! Finance [1], corresponding to the values of the stock prices of Amazon and Tesla Motors enterprises, which correspond to the timeframe of September 2019 - September 2020. The data used for the analysis is the adjusted close price of the stock, which is the closing price after adjustments for all applicable splits and dividend distributions. It includes trading days only, which means that Saturdays, Sundays, and national holidays are not quoted as the stock market is not open on those days. Figure 1 and Figure 2 show a boxplot and density plot respectively created to represent differences of the data. Then, Figure 3 shows a histogram of the convolution of the distributions.

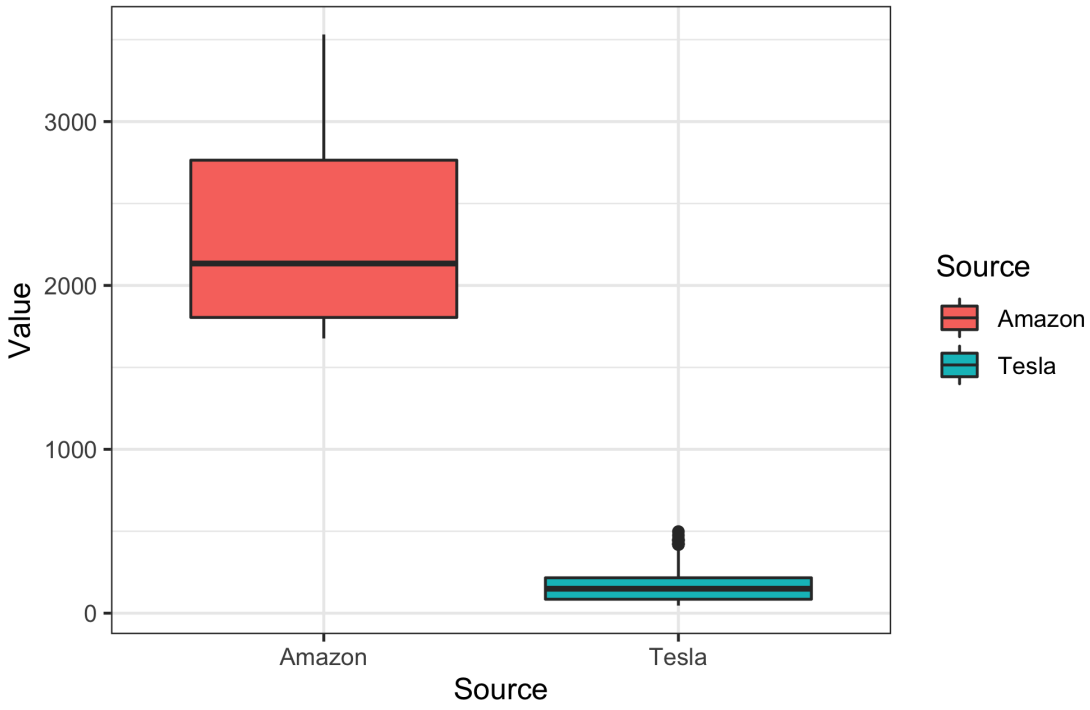


Figure 1: Boxplots of the Adjusted Close Price of Amazon and Tesla

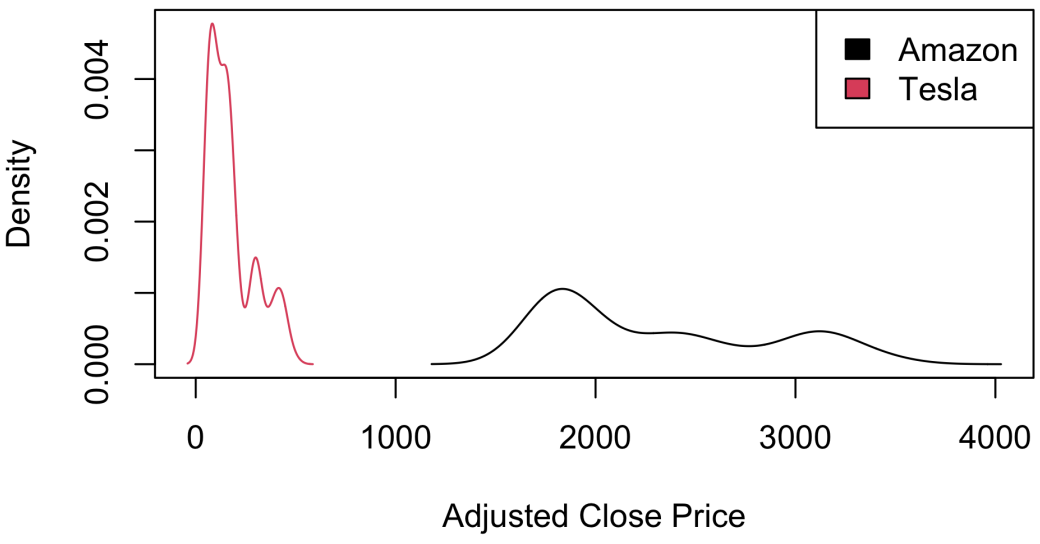


Figure 2: Densplot of the Adjusted Close Price of Amazon and Tesla

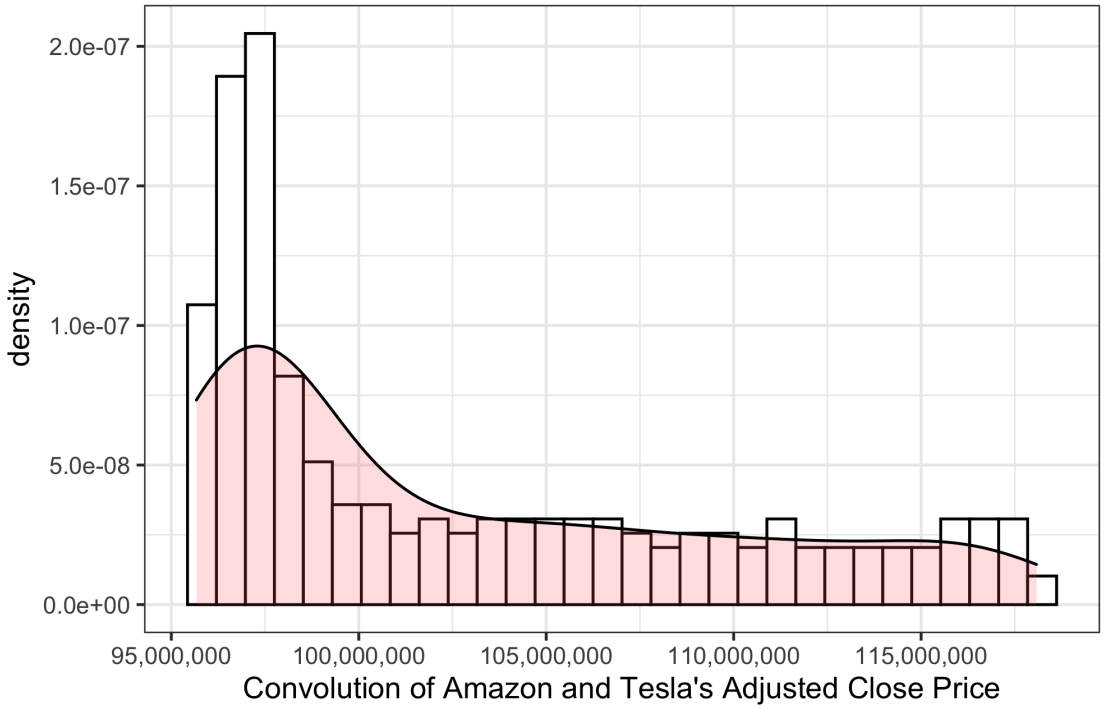


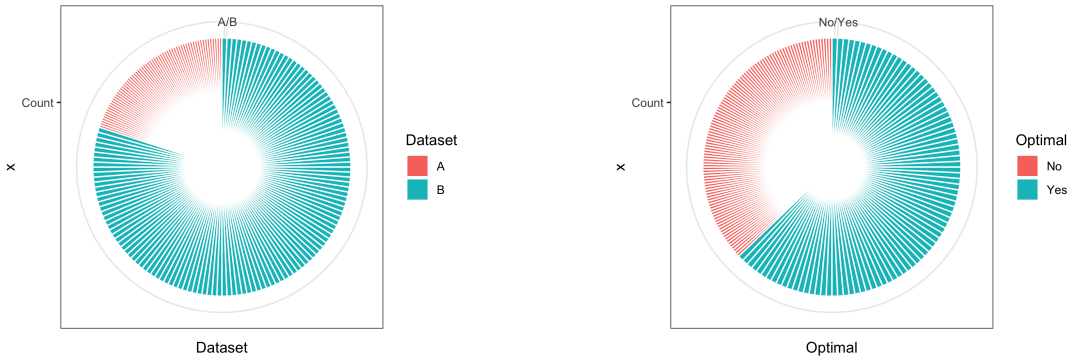
Figure 3: Histogram of the Convolution of the Adjusted Close Price of Amazon and Tesla

3 Chi-Squared Test

The Chi-Square test of independence tests whether there is a relationship between two or more categorical variables [5]. Categorical or nominal data refers that one uses labels instead of numbers; for example race and gender are categorical variables. The central tendency of categorical variables is given by its mode, since median and mean can only be computed on numerical data. Therefore, it does not follow a normal bell-curve distribution, and cannot be analyzed with tests based on a normal distribution such as the t-test or ANOVA. The hypotheses are:

- H_0: The variables are independent there is no relationship between the two categorical variables. Knowing the value of one variable does not help to predict the value of the other variable.
- H_1: The variables are dependent, there is a relationship between the two categorical variables. Knowing the value of one variable helps to predict the value of the other variable.

To perform the test, data are taken from the results of my thesis work. For this example,*Dataset* and *Optimal* are the two categorical variables which are the ones to be analyzed. Figure 4 shows how instances performed with these variables. Besides, a contingency table analysis is performed, which is shown below. The hypothesis to be tested is that *Dataset* and *Optimal* are not associated with one another. The p-value is 1.33×10^{-7}, which is less than the significance level of 0.05, so the null hypothesis is rejected, and the conclusion from this hypothesis test is that the *Dataset* and *Optimal* values are not independent, therefore are associated somehow. Finally, the last contingency table contains the expected values which would be true the null hypothesis.



(a) Groups of data of Instances

(b) Instances that reached optimal values

Figure 4: Pie Charts of Results

data.txt

```
       Contingency Table of Dataset and Optimal Solutions
         A     B
  No  0.163 0.837
  Yes 0.534 0.466
```

data.txt

```
       Pearson's Chi-squared test with Yates' continuity correction

data:  tbl
X-squared = 27.821, df = 1, p-value = 1.331e-07
```

data.txt

```
       Contingency Table of Values which H0 would be true
            A        B
  No  34.66667 69.33333
  Yes 29.33333 58.66667
```

4 Covariance

Covariance provides a measure of the strength of correlation between two variables or more sets of variables. In the covariance matrix de C_{ij} element corresponds to the covariance of x_i and x_j, whereas the element C_{ii} is the variance of x_i. The following properties can also be recognized:

- If $\text{COV}(x_i, x_j) = 0$ then variables are uncorrelated,
- If $\text{COV}(x_i, x_j) > 0$ then variables are positively correlated,
- If $\text{COV}(x_i, x_j) < 0$ then variables are negativetily correlated.

For the experiments, data of the adjusted close price of the stock market are downloaded from the Yahoo Finance website, in the timeframe from January 2007 to March 2017, the chosen companies are Procter & Gamble and its german peer Beiersdorf.

The first proof is that $\text{Cov}[aX + b, cY + d] = ac\text{Cov}[X, Y]$. For this experiment coefficients a, b, c, d of different distributions are generated 30 times each, in order to have a certain grade of diversity. In all the iterations results were the same. The used code is shown below:

```r
#1. Proof: Cov[aX+b, cY+d] = acCov[X,Y]
LHS <- numeric()
RHS <- numeric()

for (i in 1:30) {
  a<-runif(1,0,100)
  b<-rnorm(1,0,1)
  c<-rpois(1,5)
  d<-rbinom(1,5,0.4)

  lhs1 <- cov(a * df$PG + b, c * df$BEI.DE + d)
  LHS <- c(LHS,lhs1)
  rhs1 <- a * c * cov(df$PG, df$BEI.DE)
  RHS <- c(RHS,rhs1)
}
```

code/covariance.R

The second proof is that $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$. For this experiment, the mentioned variances of the adjusted close prices are calculated, and then the values of both sides of the equation are compared. The result is confirmed with a value of 968.83. The used code is shown below:

```r
#2. Proof: Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]
lhs2 <- var(df$PG + df$BEI.DE)
rhs2 <- var(df$PG) + var(df$BEI.DE) + 2 * cov(df$PG,df$BEI.DE)
```

code/covariance.R

To prove that $\text{Cov}[aX + b, cY + d] = ac\text{Cov}[X, Y]$, by definition [2] and with parameters a, b, c, and d holding constants:

$$\begin{aligned}
\text{Cov}[aX + b, cY + d] &= \mathbb{E}[(aX + b)(cY + d)] - \mathbb{E}[aX + b]\mathbb{E}[cY + d] \\
&= \mathbb{E}[acXY + adX + bcY + bd] - (a\mathbb{E}[X] + b)(c\mathbb{E}[Y] + d) \\
&= ac\mathbb{E}[XY] + ad\mathbb{E}[X] + bc\mathbb{E}[Y] + bd - ac\mathbb{E}[X]\mathbb{E}[Y] - ad\mathbb{E}[X] - bc\mathbb{E}[Y] - bd \\
&= ac(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) \\
&= ac\text{Cov}[X, Y].
\end{aligned}$$

To prove that $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$:

$$\begin{aligned}
\mathrm{Var}[X+Y] &= \mathbb{E}(X+Y)^2 - \mathbb{E}[X+Y]^2 \\
&= \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\
&= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2(\mathbb{E}XY - \mathbb{E}X\mathbb{E}[Y]) \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y].
\end{aligned}$$

References

[1] Yahoo! Finance. Stock Market Live, Quotes, and Business, 2020. https://finance.yahoo.com, Last accessed on 2020-11-14.

[2] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 2012.

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

[4] Oscar Alejandro Hernández López. Study of Mixed Integer Programming Models for the Concrete Delivery Problem, 2020.

[5] Antoine Soetewey. Chi-square test of independence in r, 2020. https://www.statsandr.com/blog/chi-square-test-of-independence-in-r/, Last accessed on 2020-11-13.

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

Homework Assignment 12: Applied Probabilistic Models

Generating Functions

5273

1 Exercises

Exercises solved in this work are provided in the book Grinstead and Snell [1].

1.1 Generating Functions for Discrete Densities

For these exercises generating functions for discrete densities are discussed. These exercises are related to branching processes

1.1.1 Exercise 1 page 392

The exercise describes a branching process. Let $h(z)$ be the ordinary generating function for the p_i:

$$h(z) = p_0 + p_1 z + p_2 z^2 + \cdots. \tag{1}$$

By Theorem 10.2 in Grinstead and Snell [1], if the mean number m of offspring produced by a single parent is ≤ 1, then $d = 1$ and the process dies out with probability 1. If $m > 1$ then $d < 1$ and the process dies out with probability d.

(a) For this case $h'(z)|_{z=1} = m = \frac{1}{4} + \frac{1}{2}(1) = \frac{3}{4}$. Since $m < 1$, then $d = 1$.

(b) For this case $h'(z)|_{z=1} = m = \frac{1}{3} + \frac{2}{3}(1) = 1$. Since $m = 1$, then $d = 1$.

(c) For this case $h'(z)|_{z=1} = m = \frac{4}{3}$. Since $m > 1$, then the process dies out with probability d. To find this value, the exercise states that at most two offspring can be produced, therefore the condition $z = h(z)$ yields the equation

$$d = p_0 + p_1 d + p_2 d^2, \tag{2}$$

which is satified by $d = 1$ and $d = p_0/p_2$. Thus, in addition to the root $d = 1$, this second root $d = \frac{1}{2}$, and represents the probability that the process will die out.

(d) For this case $h'(z)|_{z=1} = m = \sum_{j=0}^{\infty} \left(\frac{n}{2^{n+1}}\right)$. In looks like $m \to 1$ as it is added up to ∞ offspring, then $d = 1$.

(e) For this case $h'(z)|_{z=1} = m = \sum_{j=0}^{\infty} \frac{j}{3}\left(\frac{2}{3}\right)^j$. This sumation is greater than 1 then, $m > 1$. To calculate d if notice this geometric series has the form $\frac{1}{3-2z}$, then the condition $z = h(z)$ yields the equation

$$(3-2z)z = 1,$$
$$2z^2 - 3z + 1 = 0.$$

which is satified by $d = 1$ and $d = \frac{1}{2}$.

(f) To estimate d numerically it is used R software [2], with the following code:

```
pj <- function (j){
  return (exp(-2)*2^j/factorial(j)) #Probability pj of j offspring
}

d <- pj(0)

for (i in 0:100){
  d_new <- 0
  for (j in 0:100){  #Sumation of d for j offspring
  d_new <- d_new + pj(j)*(d^j)
  }
  d <- d_new
}
```

branching.R

This experiment estimates a value of $d \approx 0.2032$.

■

1.1.2 Exercise 3 page 392

In the chain letter problem the expected number of letters we send is $m = p_1 + 2p_2$ and the expected payoff is equal to $-100 + 50(m + m^{12})$. The expected profit is asked to be calculated.

Case a:

If $p_0 = \frac{1}{2}, p_1 = 0,$ and $p_2 = \frac{1}{2}$; then $m = 0 + 2\left(\frac{1}{2}\right) = 1$. Therefore:

$$\mathbb{E}(\text{Profit}) = -100 + 50(1 + 1^{12}) = 0.$$

Case b:

If $p_0 = \frac{1}{6}, p_1 = \frac{1}{2},$ and $p_2 = \frac{1}{3}$; then $m = \frac{1}{2} + 2\left(\frac{1}{3}\right) = \frac{7}{6}$. Therefore:

$$\mathbb{E}(\text{Profit}) = -100 + 50\left[\frac{7}{6} + \left(\frac{7}{6}\right)^{12}\right] = 276.26.$$

If $p_0 > \frac{1}{2}$ we canot expect to make a profit because let say $p_0 = 0.51$, then p_1 and p_2 are force to take other two values, assume $p_1 = 0.16$, and $p_2 = 0.33$. Therefore:

$$\mathbb{E}(\text{Profit}) = -100 + 50\left[0.49 + (0.49)^{12}\right] = -75.49.$$

■

1.2 Generating Functions for Continuous Densities

For the continuous case, the moment generating function $g(t)$ for X is defined as:

$$g(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx. \tag{3}$$

1.2.1 Exercise 1 page 401

For this exercise let X be a continuous random variable with values in $[0, 2]$ and for each case there is a given density function f_X.

Case a:

$$\begin{aligned} g(t) &= \int_0^2 \frac{1}{2} e^{tx} dx \\ &= \frac{1}{2}\left[\frac{e^{tx}}{t}\right]_0^2 = \frac{1}{2}\left(\frac{e^{2t-1}}{t}\right) = \frac{e^{2t}-1}{2t}. \end{aligned}$$

Case b:

$$\begin{aligned} g(t) &= \int_0^2 \frac{1}{2} x e^{tx} dx \\ &= \frac{1}{2}\left[\frac{(tx-1)e^{tx}}{t^2}\right]_0^2 = \frac{1}{2}\left[\frac{(2t-1)e^{2t}+1}{t^2}\right] = \frac{(2t-1)e^{2t}+1}{2t^2}. \end{aligned}$$

Case c:

$$\begin{aligned} g(t) &= \int_0^2 \left(1 - \frac{x}{2}\right) e^{tx} dx \\ &= \left[\frac{e}{2t^2} - \frac{(x-2)e^{tx}}{2t}\right]_0^2 = \left[-\frac{(t(x-2)-1)e^{tx}}{2t^2}\right]_0^2 = \frac{e^{2t}}{2t^2} - \frac{2t+1}{2t^2} = \frac{e^{2t}-2t-1}{2t^2}. \end{aligned}$$

Case d:

$$\begin{aligned} g(t) &= \int_0^2 |1-x|\, e^{tx} dx \\ &= \left[\frac{(x-1)(tx-t-1)e^{tx}}{t^2|x-1|}\right]_0^2 = \left[\frac{(t-1)e^{2t}}{t^2} + \frac{2e^t}{t^2} - \frac{t+1}{t^2}\right] = \frac{(t-1)e^{2t}+2e^t-t-1}{t^2}. \end{aligned}$$

Case e:

$$\begin{aligned} g(t) &= \int_0^2 \frac{3}{8} x^2 e^{tx} dx \\ &= \left[\frac{3(t^2x^2-2tx+2)e^{tx}}{8t^3}\right]_0^2 = \frac{(6t^2-6t+3)e^{2t}-3}{4t^3}. \end{aligned}$$

■

1.2.2 Exercise 6 page 402

According to Grinstead and Snell [1], $k_X(\tau)$, called the characteristicfunction of X is the Fourier transform of f_X, and has an inverse given by

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\tau x} k_X(\tau) d\tau. \tag{4}$$

Therefore:

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\tau x}e^{-|\tau|}d\tau, \\ &= \frac{\frac{ix}{x^2+1}+\frac{1}{x^2+1}}{2\pi} + \frac{\frac{1}{x^2+1}-\frac{ix}{x^2+1}}{2\pi} \\ &= \frac{1}{\pi(x^2+1)} \end{aligned}$$

■

1.2.3 Exercise 10 page 403

(a)

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} x\frac{e^{-|x|}}{2}dx \\ &= \frac{1}{2}\int_{-\infty}^{\infty}\frac{x}{|x|}e^{-|x|}|x|dx \\ &= -\frac{e^{-|x|}|x|}{2} - \frac{e^{-|x|}}{2} \\ &= \left[\frac{e^{-|x|}(-|x|-1)}{2}\right]_{-\infty}^{\infty} \\ &= 0. \end{aligned}$$

$$\begin{aligned} \mathbb{V}(X) &= \int_{-\infty}^{\infty} x^2\frac{e^{-|x|}}{2}dx \\ &= \frac{1}{2}\int_{-\infty}^{\infty} x^2e^{-|x|}dx \\ &= \frac{1}{2}\left(\frac{x}{|x|}\right)\int_{-\infty}^{\infty} x^2e^{-|x|}\frac{|x|}{x}dx \\ &= \frac{1}{2}\left(\frac{x}{|x|}\right)\int_{-\infty}^{\infty} xe^{-|x|}|x| \\ &= \frac{1}{2}\left[\frac{x(-2e^{-|x|}|x| - x^2e^{-|x|} - 2e^{-|x|})}{|x|}\right] \\ &= -\left[\frac{xe^{-|x|}|x|(2|x| - x^2 + 2)}{2|x|}\right]_{-\infty}^{\infty} \\ &= 2. \end{aligned}$$

(b) Let X_1 be a trial process. The moment generating function would be

$$\begin{aligned} g(t) &= \mathbb{E}(e^{tx}) \\ &= \int_{-\infty}^{\infty} e^{tx}\left[\frac{e^{-|x|}}{2}\right]dx \\ &= \frac{1}{2}\left[\int_{-\infty}^{0} e^{xt+x}dx + \int_{0}^{\infty} e^{xt-x}dx\right] \\ &= \frac{1}{2}\left[\left[\frac{e^{x(t+1)}}{t+1}\right]_{-\infty}^{0} - \left[\frac{e^{x(1-t)}}{1-t}\right]_{0}^{\infty}\right] \\ &= \frac{1}{2}\left[\frac{1}{t+1} + \frac{1}{1-t}\right] \\ &= \frac{1}{2}\left[\frac{2}{1-t^2}\right] \\ &= \frac{1}{1-t^2}. \end{aligned}$$

If S_n is $X_1 + X_2 + ... + X_n$, then the moment generating function is given by

$$\begin{aligned} [g(t)]^n &= \left(\frac{1}{1-t^2}\right)^n \\ &= \frac{1}{(1-t^2)^n}. \end{aligned}$$

Then, if $S_n^* = \dfrac{(S_n - n\mu)}{\sqrt{n\sigma^2}}$, the moment generating function is given by

$$\begin{aligned} \left[g\left(\frac{t}{\sqrt{n}}\right)\right]^n &= \left(\frac{1}{1-t^2}\right)^n \\ &= \left[\frac{1}{\left[1-\left(\frac{t}{\sqrt{n}}\right)^2\right]}\right]^n \\ &= \frac{1}{\left[1-\left(\frac{t}{\sqrt{n}}\right)^2\right]^n} \end{aligned}$$

(c) According to the obtained expression of S_n^* as $n \to \infty$ then the expression may reduce to $\dfrac{1}{1^n}$. When the limit is calculated it is equal to 1.

■

References

[1] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 2012.

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

Homework Assignment 13: Applied Probabilistic Models

Law of Large Numbers

5273

1 Introduction

In this work, it is studied the law of large numbers in the context of probability and statistics. In general terms, this law states that as a sample size grows, its mean gets closer to the average of the whole population. In other words, if you repeat an experiment independently a large number of times and average the result, what is obtained will be close to the expected value. This is, having a sequence of random variables, ξ_1, ξ_2,... which have been drawn independent and identically distributed from some probability distribution P, the mean converges to the mean of the underlying distribution itself when the sample size goes to infinity [6]:

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \to \mathbb{E}(\xi) \quad \text{for } n \to \infty. \tag{1}$$

According to Grinstead and Snell [2] it can be expresed, for any $\epsilon > 0$ as

$$P\left(\left|\frac{Sn}{n} - \mu\right| \geq \epsilon\right) \to 0 \tag{2}$$

as n $\to \infty$. Equivalently,

$$P\left(\left|\frac{Sn}{n} - \mu\right| \geq \epsilon\right) \to 0 \tag{3}$$

as n $\to \infty$.

It is an important concept in statistics because it states that even random events with a large number of trials may return stable long-term results. The theorem deals only with a large number of trials since the average of the results of the experiment repeated a small number of times might be substantially different from the expected value. However, each additional trial increases the precision of the average result.

If one think of playing dice, each time we roll a fair die, the results are $\Omega = \{1, 2, 3, 4, 5, 6\}$. Each possible outcome can occur with the same probability $p = \frac{1}{6}$. Thus one can expect that the average of the resulting die values is 3.5. Figure 1 shows this experiment, where increasing the number of times one rolls the die (up to 50 000 in this case) gets closer and closer to the average value. With smaller sample sizes, the sample means are broadly spread around the population mean of 3.5. However, the more we go to the right extreme of the x-axis (and thus the larger the sample size), the narrower the sample means are spread around the population mean [3].

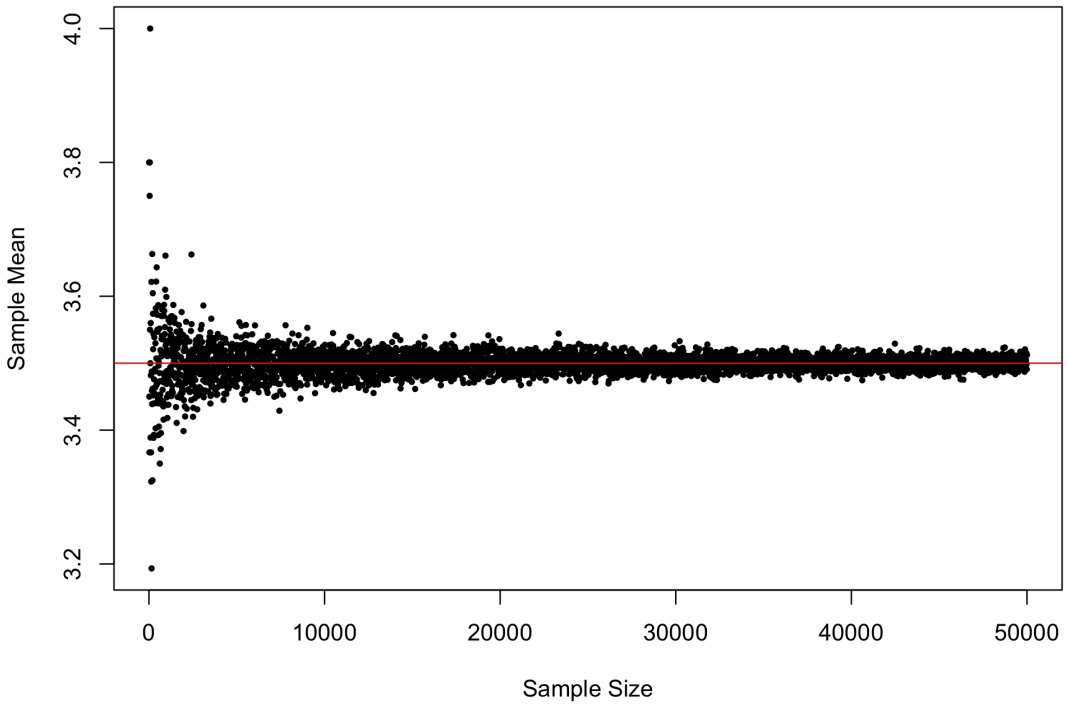


Figure 1: Plot of Expected Value as sample grows

2 Computing probabilities

The law of large numbers gives a way to compute probabilities. An example of this application is the hat check problem [4]. Let see as an example a class with $n = 20$ students and a professor who completely forgets the names and faces of all of them. Suppose he randomly hands back the midterm exams, let p_n be the probability that no one receives their ow test back. It is known that as $n \to \infty$ one have $p_n \to e^{-1}$.

It is generated a random permutation of the 20 exams, as a permutation of 20 numbers, and each random permutation is considered independent. The experiment is repeated 10 000 times as shown in the following code; it is used R software [5]. The result gives an approximation to a probability of 0.3644, which is quite good for even n=20 with respect to the expected value of e^{-1}.

```r
order <- seq(1,20)

give_hand <- function (){
  x <- sample(20,20)
  is.element(TRUE, x==order)
}

r <- replicate(10000,give_hand())
sum(r==FALSE)/10000
```

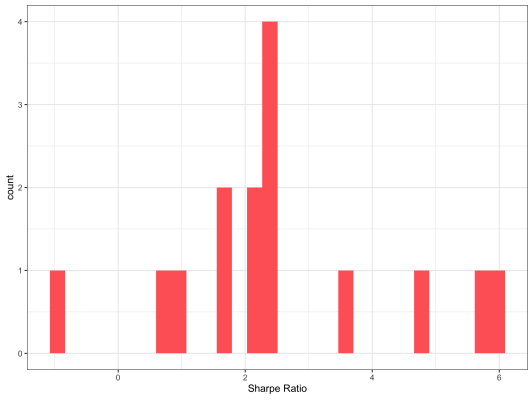
code/hat_check.R

3 Application in trading

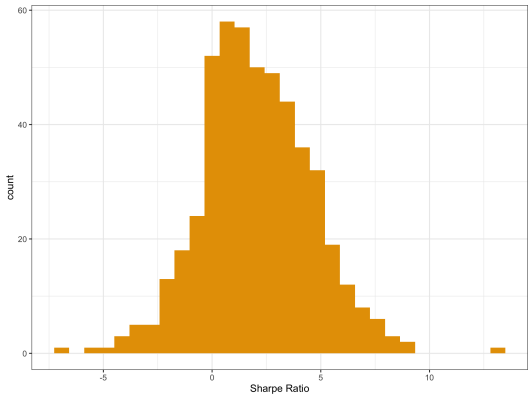
The law of large numbers can be seen in the world of trading too. A clear example of this is given by Bsl [1]. The author suggests that a large number of trades with a higher reward to risk ratio (Sharpe ratio) will tend to be more effective than a smaller number of trades.

Let assume a scenario with a relevant strategy where what matters is how much one will make when right and how much one lose when one is wrong when putting on a trade. In this situation, basis the law of large numbers, a trade can be wrong the majority of the time but still be profitable.

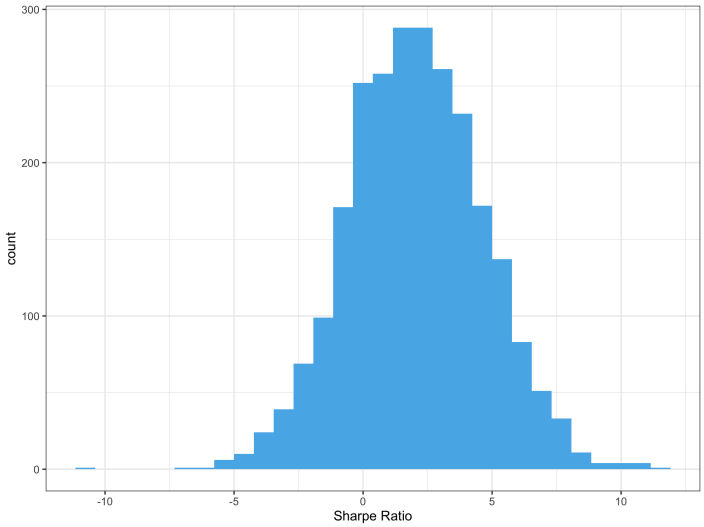
Backtesting on the longest period possible gets one closer to the mean annualized returns. Let assume a backtested strategy with a true Sharpe ratio of 2. Trying this over a 40 day trading period (broadly around two months), and assuming a normal distribution with mean returns of 0.1% one can see results in Figure 2, representing how the Sharpe ratios tend to be with an increasing number of trades. Histogram 2a shows with few trades that one would end up with a lesser Sharpe ratio than what was seen through backtesting and might observe more losses than wins. On the other hand, Histograms 2b and 2c one can see as the trials keep increasing, the Sharpe ratio is tending towards the backtested number (by tending towards a normal distribution).



(a) Histogram of 15 trades



(b) Histogram with 500 trades



(c) Histogram of 2500 trades

Figure 2: Histograms of the Sharpe ratios as the number of trades grows.

References

[1] Gayathri Bsl. Law of large numbers in finance, 2020. https://towardsdatascience.com/law-of-large-numbers-in-finance-using-python-86945eaee444, Last accessed on 2020-11-30.

[2] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 2012.

[3] Ulrich Matter. The Law of Large Numbers and The Central Limit Theorem, 2017. https://umatter.github.io/courses/berkstats/Berkstats_LLM_CLT.html, Last accessed on 2020-11-30.

[4] Richard Scoville. The hat-check problem. *The American Mathematical Monthly*, 73(3):262–265, 1966.

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

[6] Ulrike von Luxburg and Bernhard Schölkopf. Statistical learning theory: Models, concepts, and results. In Dov M. Gabbay, Stephan Hartmann, and John Woods, editors, *Inductive Logic*, volume 10 of *Handbook of the History of Logic*, pages 651 – 706. North-Holland, 2011. doi: https://doi.org/10.1016/B978-0-444-52936-7.50016-1. URL http://www.sciencedirect.com/science/article/pii/B9780444529367500161.

Homework Assignment 14: Applied Probabilistic Models

Central Limit Theorem

5273

1 Introduction

Central Limit Theorem (CLT) is an approximation one can use when the population to study is quite big (it would take a long time to gather data about each individual) and identifying its characteristics is desired. In statistical terms, one collects samples from a population and by combining the information from the samples, conclusions can be drawn about the population. In a nutshell, the approach of the CLT could be:

- Draw multiple samples sufficient in size.
- Calculate the individual mean of these samples.
- Calculate the mean of these sample means, and this value will give the approximate mean of the studied variable.
- Additionally, the histogram of the sample means will resemble a bell curve or normal distribution.

2 Applications

In this section, it is seen how the CLT can be used in real-world problems and how to apply it. It helps to solve problems where the population is not normal.

2.1 Manufacturing

Let assume a pipe manufacturing organization produces a different kind of pipes and the monthly data of the wall thickness of certain types of pipes are given. The organization wants to analyze the data by constructing confidence intervals to implement some strategies in the future and the challenge is that the distribution of the data is not normal. Data is simulated in R software [3]. Figure 1 shows a histogram of all the observations of the data. This graph denotes with a vertical red line the population mean, which is 12.802 and one can see that the population is not normal.

Therefore, to apply the CLT, it is needed to draw sufficient samples of different sizes and compute their means (known as sample means). Figure 2 shows this experiment, where sufficient samples are drawn, increasing its sizes. Means are calculated and are plotted in R. It is known that the minimum sample size taken should be 30, but even with samples of size 10, (see Figure 2a) nice bell-shaped curves are evidenced. The sampling distribution should approach a normal distribution as the sample sizes increase. Therefore, one can consider the sampling distributions as normal and the pipe manufacturing organization can use these distributions for further analysis.

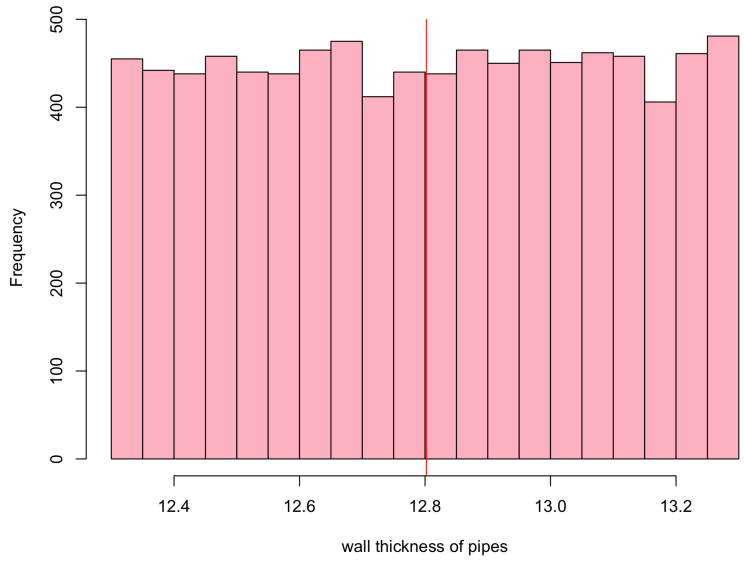


Figure 1: Histogram for Wall Thickness

2.2 Baseball

Another application can be found by Anderson and Bay [1], where CLT found an interesting application as hypothesis testing. This author answers the following question using the CLT: *Is there such thing as home-field advantage in Major League Baseball?*

Concerning this problem, the null hypothesis and the alternative hypothesis are:

$$H_0\text{: There is no home-field advantage,}$$
$$H_1\text{: There is a home-field advantage.}$$

To test this notion, in a Major League Baseball (MLB) season 2431 games are played. Let us take the 2013 MLB season, were 1308 of those games were won at home, therefore the observed value $\hat{p} = 0.5381$. To test the hypothesis, our null hypothesis will be 0.5, that is 50% of the MLB games are won at home and the other half on the road, hence, there is no home-field advantage. One method is using a confidence interval. For testing,

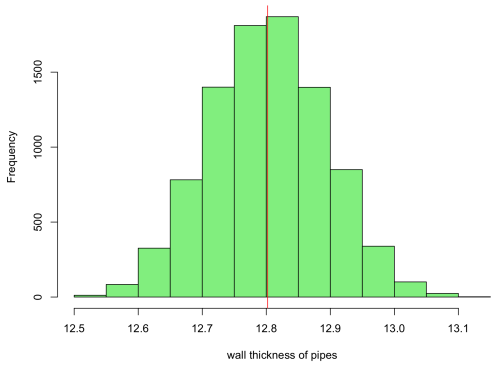
$$H_0\text{: } p = 0.50,$$
$$H_1\text{: } p > 0.50.$$

at the 0.05 level of significance a right-sided 95% confidence interval for p can be constructed. If our test statistic of $p = 0.5$ is in the interval, then we fail to reject H_0 at the 0.05 level of significance. If $p = 0.5$ is not in the interval, we reject H_0. The right-sided $100(1-\alpha)\%$ confidence interval for p for a large sample is given by

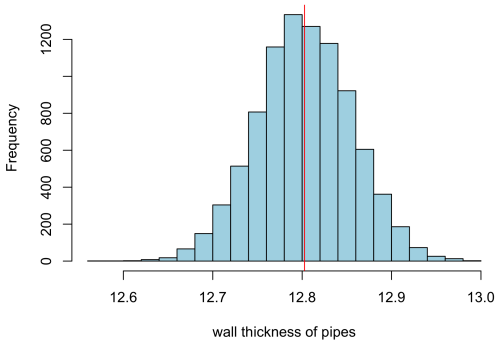
$$\hat{p} - z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p \leq 1,$$

where α is the level of significance.

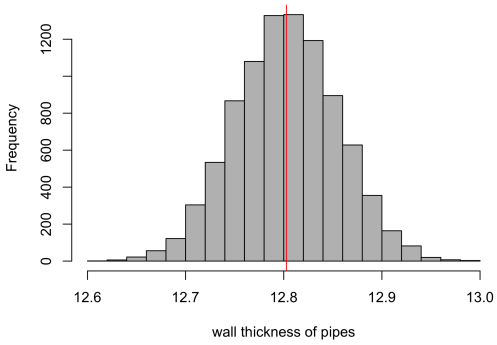
Since $n = 2431, \hat{p} = 0.5381, \alpha = 0.05$, and $z_{0.05} = 1.645$, obtained from the standard normal table [2], which is used to find the probability that a statistic is observed below, above, or between values on



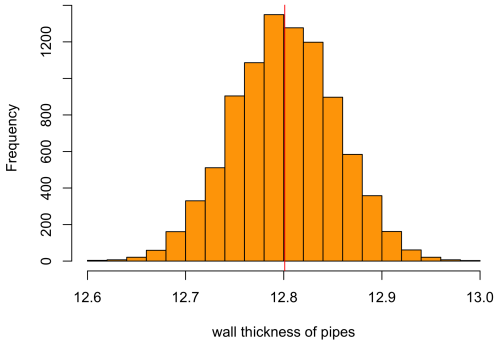
(a) Histogram of samples of size 10



(b) Histogram of samples of size 30



(c) Histogram of samples of size 50



(d) Histogram of samples of size 500

Figure 2: Histograms of the sample means

the standard normal distribution, and by extension, any normal distribution. Therefore a right-sided 95% confidence inerval for p is:

$$\begin{aligned} 0.5381 - 1.645\sqrt{\frac{(0.5381)(1-0.5381)}{2431}} < p \leq 1, \\ 0.5381 - 1.645(0.0101114) < p \leq 1, \\ 0.5215 < p \leq 1. \end{aligned}$$

Since $0.5 \notin (0.5215, 1]$, reject $H_0 : p = 0.5$ in favor of $H_1 : p > 0.5$ at the 0.05 level of significance, that is, there is enough evidence to support that there is a home-field advantage, and the home team wins more than 50% of the games played at home, hence there is such thing as home-field advantage in MLB.

Also, if one wants to see if there is a difference between the American League and the National League a similar analysis can be performed. For these separate leagues, a 99% confidence interval is used, and for the National League it is obtained that $0.5 \notin (0.5117, 1]$, therefore it is concluded that the National League has a home-field advantage. On the other hand, in the American League $0.50 \in (0.4978, 1]$, fail to reject $H_0 : p = 0.50$. That is, one does not have enough evidence to support that there is a home-field advantage in the American League based on the 2013 season.

References

[1] Nicole Anderson and Thunder Bay. Central limit theorem and its applications to baseball. 2014.

[2] F. James Rohlf, Robert R. Sokal, et al. *Statistical tables*. Macmillan, 1995.

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

Homework Assignment 15: Applied Probabilistic Models

Project Proposals

5273

Proposal 1

This proposal is about fitting probability models to frequency data from my thesis work about the delivery of concrete [1]. This data is generated from preliminary results pertaining to different solution methods, applying χ^2 Goodness-of-fit test. In addition, create a model that helps predict the probability of an instance to reach an optimal value given other characteristics, performing generalized linear models.

Proposal 2

The second proposal is the forecast for manufacturing operation through time series and Auto-Regressive Integrated Moving Average (ARIMA) model. This will help to forecast sales/demand for a period of time.

Proposal 3

The third proposal is an analysis of capacity and tolerance indices and Six Sigma metrics to measure if a manufacturing process has been fulfilling its specifications. Analyzing these capacity indices will allow one to know if the process is centered with respect to the specifications and therefore give recommendations to improve it. Also, with the design of tolerance limits can be defined the specifications of upper and lower values of to the nominal one that components of the product should have.

References

[1] Oscar Alejandro Hernández López. Study of Mixed Integer Programming Models for the Concrete Delivery Problem, 2020.

Homework Assignment 16: Applied Probabilistic Models

Feedback of proposals

5273

Alberto Martínez Noa

Simulación de soluciones para instancias grandes

Realizar un análisis de los resultados anteriormente mencionados, que nos permitan encontrar un modelo para simular soluciones de instancias grades (más de 300 elementos). Con esto se pretende saber cómo se podría comportar el modelo en instancias que no se han podido probar por la complejidad de este en su versión actual.

Erick Cervantes

Ingeniería de Calidad en la empresa

Se ve muy interesante el proyecto entiendo que te apoyarás de las cartas de control aunque no me queda claro qué medidas utilizarías para poder establecer los límites que la producción debía cumplir.

Johanna Bolaños

PIB en Colombia

Considero que convendría analizar un poco más de variables o más bien analizarlas por grupos en caso de ser posible, según la fuente de la que obtengas los datos, ya que el PIB es un indicador que depende de muchos factores además de los que mencionas y podría llevar a conclusiones equivocadas.