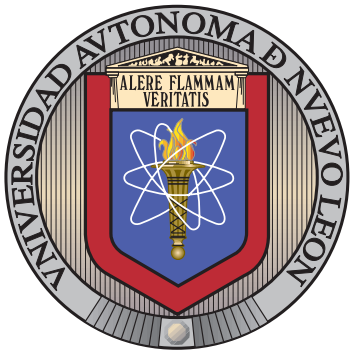
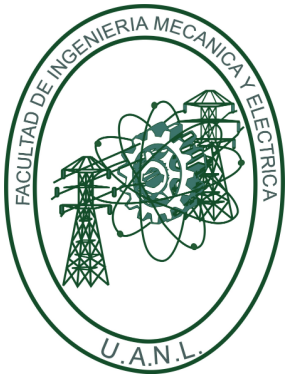
Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Posgrado en Ingeniería de Sistemas

Doctorado





Portafolio de Evidencias

de

Gerardo Palafox Castillo

1649275

para el curso de Modelos Probabilistas Aplicados,

con la profesora Satu Elisa Schaeffer.

Semestre Agosto 2020 - Enero 2021.

https://github.com/palafox794/AppliedProbabilityModels

Impact of COVID-19 pandemic in rapid transit ridership in Monterrey, Mexico.

G. Palafox

September 6, 2020

Abstract

In the following report, we make a comparison of passengers transported by Monterrey's rapid transit system (Metrorrey) between the years 2018 and 2020. We show the decrease of ridership expected from the movement restrictions imposed as a response to the current coronavirus pandemic.

Introduction

The metropolitan area of Monterrey is the third-largest in the country [2]. It is located in north eastern Mexico, south from Texas. Means of transportation in the city include a system of public buses, a rapid transit system, taxis, ride-sharing apps and personal vehicles. The rapid transit system, named Metrorrey, consists of two lines: line one is elevated and line two has an elevated and a subway component [3]. Metrorrey transported an average of 15.3 million passengers per month in the past two years. As many places around the world, Monterrey has imposed restrictions in the movement of their residents to stop the spread of the coronavirus epidemic. In the next section, we show a brief analysis of Metrorrey's ridership data, and use our results to observe how impactful the movement restrictions have been on rapid transit ridership.

Data Analysis

We obtained Metrorrey's ridership data corresponding to the January 2018 - June 2020 period from [1]. To perform a month-to-month comparison between different years, we only considered data from the first six months of each year. That is, we restricted ourselves to the January-June period for 2018, 2019, and 2020 (see table 1). Data analysis was performed with R Version 4.0.0 [6] on a Jupyterlab Notebook [5].

We processed the data found at [1] directly in R [6]. In order to extract information about the months we were interested in, and to create table 1, we used the following script[1]:

```
df <- read.csv(file = 'Tabulado-metrorrey.csv')
passng = df$Pasajeros.transportados..... Miles.de.pasajeros.
passng = passng[!is.na(passng)]
passng2018 = passng[1:6]
passng2019 = passng[13:18]
passng2020 = passng[25:30]
table <- data.frame(c("jan", "feb", "mar", "apr", "may", "jun"), passng2018, passng2019,
 passng2020)
names(table) <- c("month", "2018", "2019", "2020")
library("xtable")
xtable(table)
```

	Month	2018	2019	2020
1	jan	13529.98	14534.62	15220.38
2	feb	14404.83	14511.12	15548.33
3	mar	15102.95	14659.78	12554.89
4	apr	14991.67	14826.92	5653.46
5	may	15999.97	16515.39	4933.48
6	jun	14314.63	14750.24	7208.06

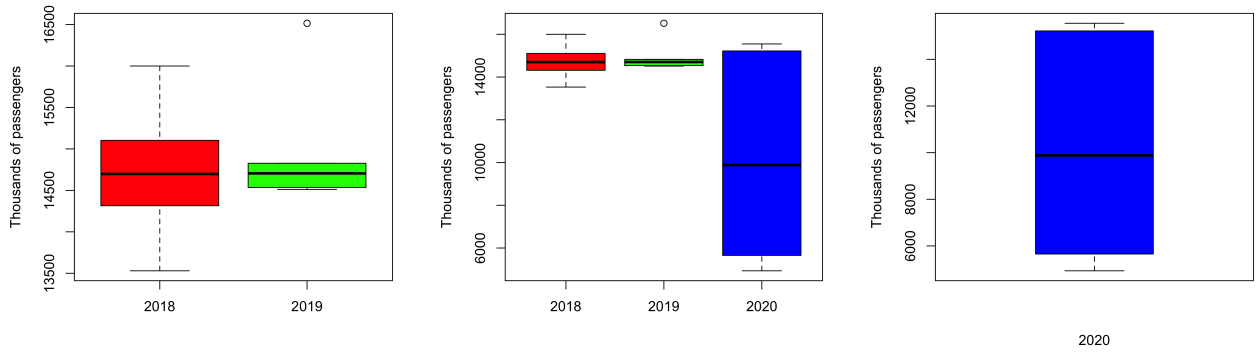
Table 1: Ridership data (in thousands) to be analysed

[1]The script and a Jupyter [5] notebook showing how we performed the data analysis and created the graphics in this report can be found at `https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment1`

A summary of the data is shown in table 2, where it can be seen that the smallest values (the minimum, and the first quartile) for ridership in 2020 differ drastically from those of previous years. This can be further shown with the boxplots in figure 1, the violin plots[2] in figure 2 and the bar plot in figure 3.

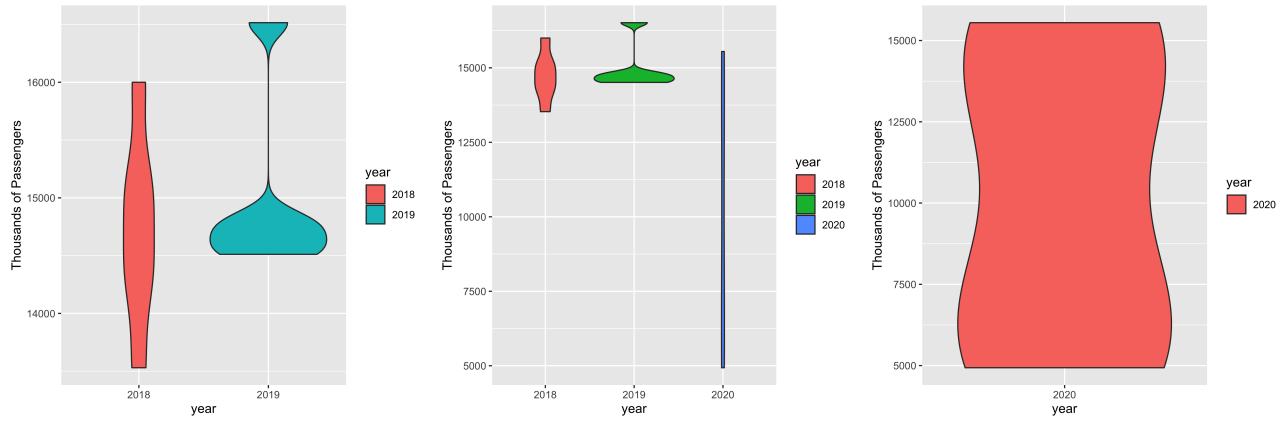
	Year (first half)	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
1	2018	13530	14337	14698	14724	15075	16000
2	2019	14511	14566	14705	14966	14808	16515
3	2020	4933	6042	9881	10186	14554	15548

Table 2: Ridership data summary. Passengers in thousands.



(a) (b) (c)

Figure 1: Boxplots of January-June ridership in different years.



(a) (b) (c)

Figure 2: Violin plots of January-June ridership in different years.

Conclusion

We observe that government measures to reduce mobility in Monterrey Metropolitan area had a noticeable impact, at least with respect to Metrorrey's passengers ridership. For instance, close to 5 million people rode Metrorrey on May 2020, an 11.5 million decrease when compared to May 2019. Further analysis is needed to establish if this reduction in mobility was similar in other ways of public transport. It can also be of interest to study whether this reduced mobility stopped the spread of covid-19 in a significant way.

[2]To read more about violin plots, see [4].

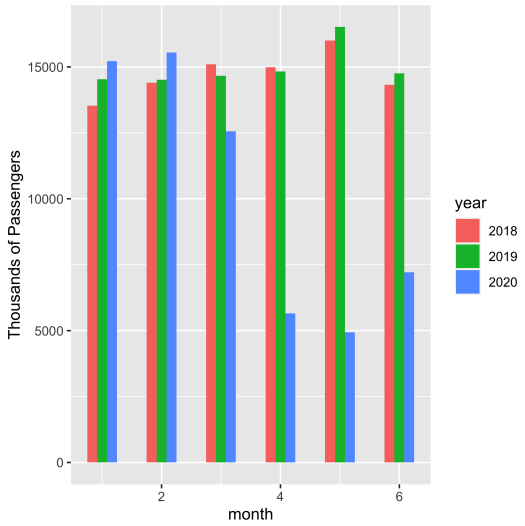


Figure 3: Bar plot of the data analysed.

References

[1] Instituto Nacional de Estadística y Geografía. Transporte Urbano de Pasajeros. Principales características del Sistema de Transporte Colectivo Metrorrey. `https://www.inegi.org.mx/app/tabulados/?nc=100100049`.

[2] Instituto Nacional de Estadística y Geografía. Delimitación de las Zonas Metropolitanas de México 2015. *Publicaciones*, 2018. `https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825006792`.

[3] Gobierno de Nuevo León. Registro Estatal de Trámites y Servicios. Metro. `http://retys.nl.gob.mx/servicios/metro`.

[4] Jerry L. Hintze and Ray D. Nelson. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998.

[5] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, et al. Jupyter notebooks—a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing*, page 87. IOS Press, 2016.

[6] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

A text analysis of Nietzsche's *Antichrist*

G. Palafox

September 15, 2020

Abstract

In the following report, we show the results of employing basic text analysis techniques on Friedrich Nietzsche's book *The Antichrist*. Graphics are shown to aid with the exposition of our findings.

1 Introduction

A text analysis of Nietzsche's *The Antichrist* [3] was done. We were able to find the most common words and characters in the text, omitting numbers and really common words (e.g., "the"). We used these words and characters to create various barplots and a word cloud visualization. Furthermore, we determined the most common pairs of words occurring together in the text, and show them with a network representation of our book.

2 Text Analysis

The text extraction and analysis was performed in a Jupyter[2] notebook running R[4] version 4.0.0[1]. We downloaded the book directly from Project Gutenberg's site using R's `gutenbergr` library. The book downloaded starts with an introduction by the translator, which we ommitted from the analysis, as the intention was to study the author's words. Our first step into analysing the text was to get the most frequent characters and words. For this we ommitted numbers, punctuaction, and so-called stop-words. Table 1 shows the ten most used letters and ten most used words in the text. Additionally, frequency of characters and words are shown in the barplots of Figure 1. For illustration purposes, we show a word cloud of the most frequent words in Figure 2. In a word cloud, the size of each word is proportional to the number of times it appears in the text.

2.1 Network representation

Next, we created a network representation of our text. For readers unfamiliar with network (or graph) theory, basic definitions can be found in Appendix A. For our analysis, we considered words in the text as our vertices, joining a pair of words with an edge if they appeared together in the text (that is, if they form a bigram). Notice that since every edge joins exactly two words that appeared together in the text, there is a direct correspondence between our edges and the bigrams in the text. We made this a weighted network by assigning to each edge a weight equal to the number of times its corresponding bigram appears in the text. We restricted ourselves to those bigrams appearing more than once. The results can be seen in Figure 3. We also extracted the largest connected component of the network, as can be seen in Figure 4. Vertices with highest degree and strength can be seen in Table 2. Finally, we computed the degree distribution of both the whole network and of the largest connected component. The degree distribution gives us the relative frequency of n-th degree vertices, with $n = 0, 1, \ldots, \max_v\{\deg(v)\}$. You can observe these in Figure 5.

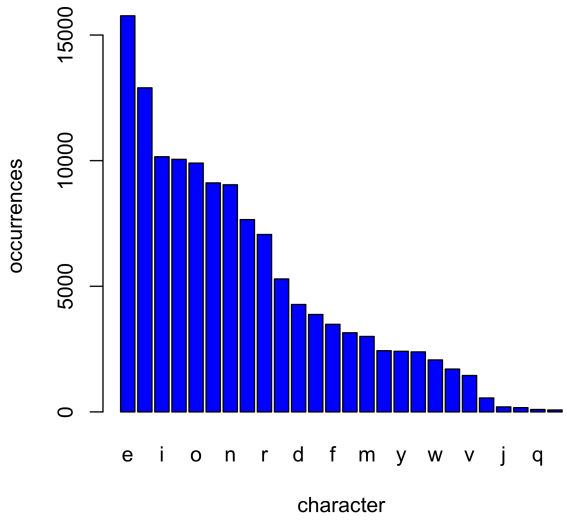
3 Conclusion

In accordance to what was expected given the theme of the book, the words *god*, *life*, *christian* and *christianity* appear the most in the text. The word *christian* is also the highest-degree word in our network. It is of interest to observe that while *instinct* is the 9th most appearing word, is the second word with highest degree (and strength) in our network. This means that, while overall is not the most used, it is still very "central"in joining words together. These are all very elementary findings. More sophisticated techniques can still be used to get a deeper study of Nietzsche's work, e.g., sentiment analysis. Further work can involve the comparison of different Nietzsche's books, or comparison of Nietzsche's books with works of other authors, in an attempt to characterize his writing style.

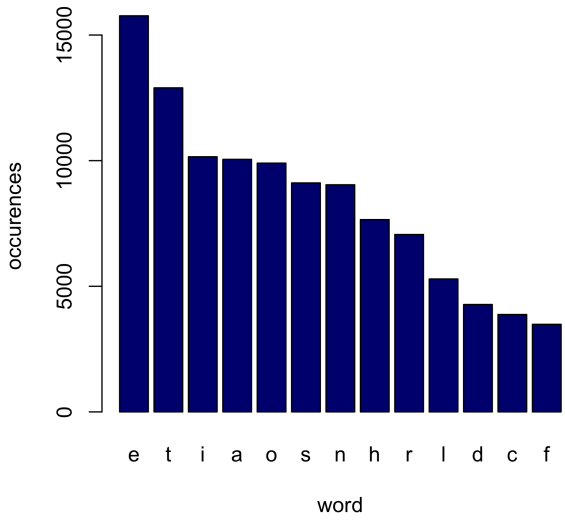
[1]The script and a Jupyter [2] notebook showing how we performed the data analysis and created the graphics in this report can be found at `https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment2`

Figure 1: Barplots of words and characters occurences.

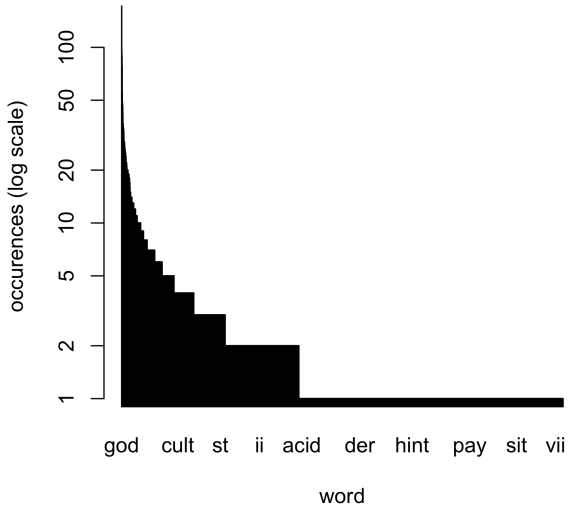
(a) Character occurrence.



(b) Most-frequent character occurrence.



(c) Word occurrence.



(d) Most-frequent word occurrence.

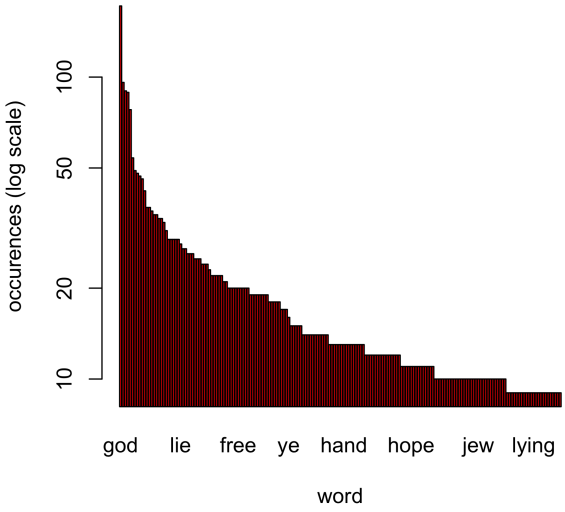


Figure 2: Word cloud of the book.

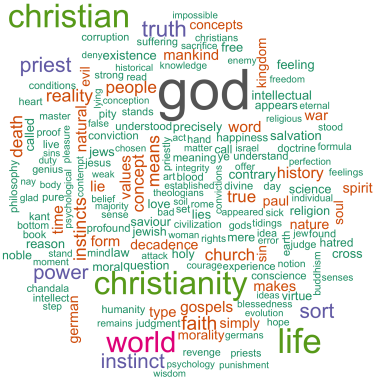


Figure 3: Book network representation.

(a) Network with words as vertices and edges joining words appearing as a bigram in the text.

(b) Same network, with vertex size and edge width proportional to their degree and weight.

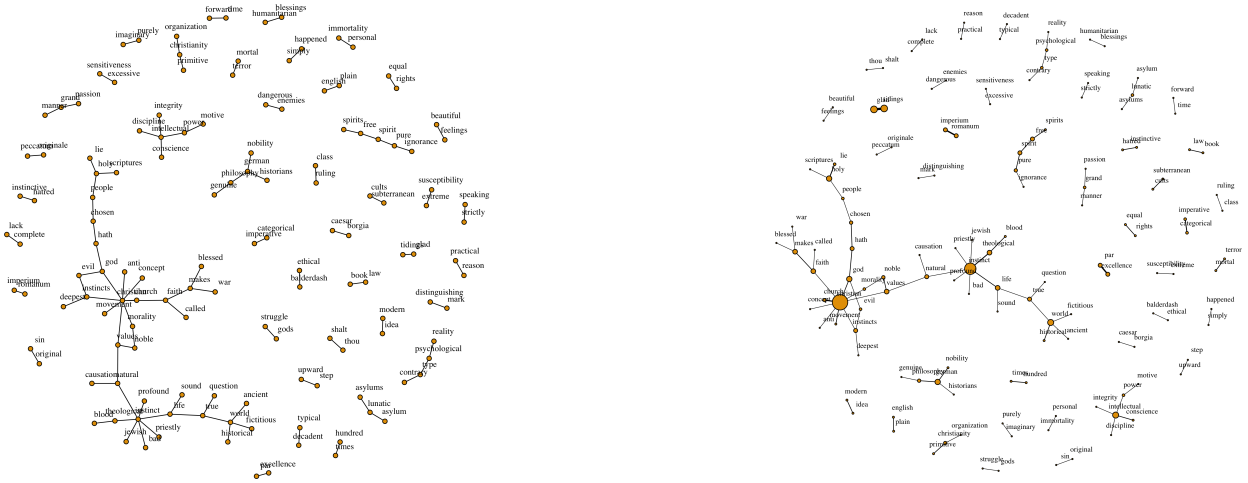


Figure 4: Largest component of network in Figure 3, with vertex sizes and edges width proportional to their degree and weight.

Table 1: Word and character frequency

(a) Character frequency

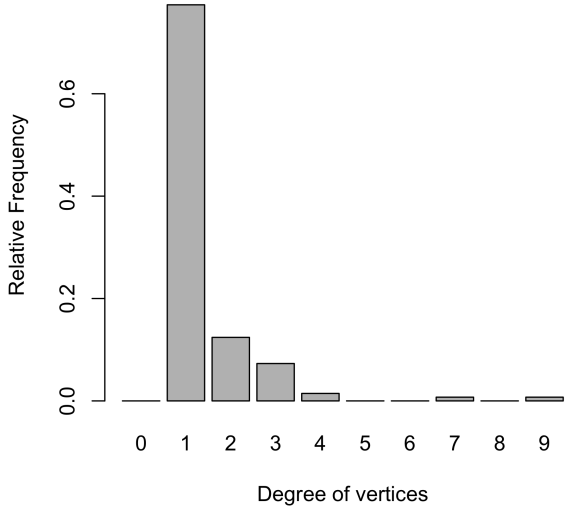
	Letter	Frequency
1	e	15765
2	t	12898
3	i	10155
4	a	10052
5	o	9902
6	s	9114
7	n	9040
8	h	7654
9	r	7061
10	l	5291

(b) Word frequency

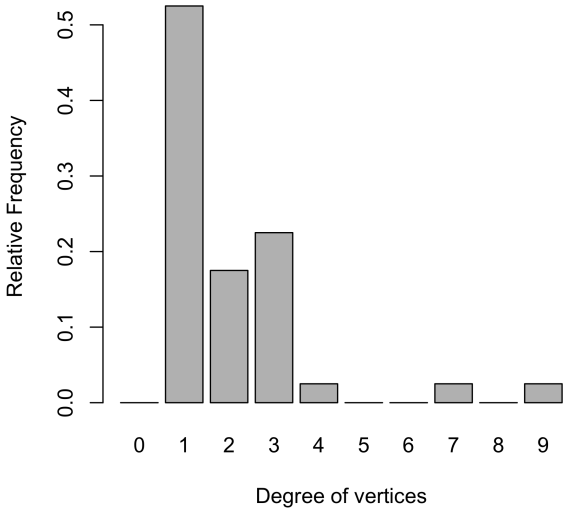
	Word	Frequency
1	god	172
2	life	96
3	christian	90
4	christianity	89
5	world	78
6	truth	54
7	priest	49
8	sort	48
9	instinct	47
10	power	46

Figure 5: Degree distributions.

(a) Relative frequency of nth degree vertices in our network.



(b) Relative frequency of nth degree vertices in the network's largest component.



A Network theory

The following theory, and more about networks, can be found at Jungnickel's book[1]. A *network* (or *graph* in the mathematics literature) is a pair $\mathcal{N} = (V, E)$ consisting of a non-empty, finite set V and a set E of two-element subsets of V[2]. An element $e = \{a, b\} \in E$ is called an *edge* with *end vertices* a and b. We say that a and b are *incident* with e and that a and b are *adjacent* or *neighbors* of each other, and write $e = ab$. The degree of a vertex v is defined as

$$\deg v := |\{u \in V : u \text{ is adjacent to v }\}|, \tag{1}$$

where $|A|$ denotes the cardinality of a set A. A network is *weighted* if there is a function $w : E \to \mathbb{R}$, and we say and edge e has weight $w(e)$. In a weighted network, we define the strength of a vertex v as the sum of the weights of the edges inciding on v. A sequence $(v_1, \ldots, v_k)$ of adjacent vertices is called a *walk* starting on v_1 and ending in v_k. Two vertices a and b are *connected* if there exists a walk starting in a and ending in b; we say the vertices are *disconnected* if no such walk exists. If all pairs of vertices of a network $\mathcal{N}$ are connected, we say $\mathcal{N}$ itself is connected. Given a network $\mathcal{N} = (V, E)$, and $V' \subseteq V$, we denote by $E_{V'}$ the set of all edges $e \in E$ which have both their end vertices in V'. The network $(V', E_{V'})$ is called the *induced subnetwork* on V'. Each network of the form (V', E') where $V' \subseteq V$ and $E' \subseteq E_{V'}$ is said to be a *subnetwork* of the network $\mathcal{N}$. A *connected component* of a network $\mathcal{N}$ is a connected subnetwork (V', E') such that any vertex in V' is disconnected from vertices not in V'.

[2]Some literature allows V to be infinite, but it will not be needed in our discussion. Also, for *directed* networks, E consists of ordered pairs of elements of V.

Table 2: Highest degree and strength vertices for largest connected component in the network.

(a) Sorted by degree.

vertex	degree	strength
christian	9.00	23.00
instinct	7.00	17.00
world	4.00	9.00
holy	3.00	8.00
faith	3.00	7.00
god	3.00	8.00
life	3.00	7.00
makes	3.00	7.00
natural	3.00	6.00
true	3.00	6.00

(b) Sorted by strength

vertex	degree	strength
christian	9.00	23.00
instinct	7.00	17.00
world	4.00	9.00
holy	3.00	8.00
god	3.00	8.00
theological	2.00	7.00
faith	3.00	7.00
life	3.00	7.00
makes	3.00	7.00
values	3.00	7.00

References

[1] D. Jungnickel. *Graphs, networks, and algorithms.* Number v. 5 in Algorithms and computation in mathematics. Springer, Berlin ; New York, 1999.

[2] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, et al. Jupyter notebooks—a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing*, page 87. IOS Press, 2016.

[3] Friedrich Nietzsche. *The Antichrist.* Project Gutenberg, September 2006. `http://www.gutenberg.org/files/19322/19322.txt`.

[4] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020.

Exploring the text distribution of Nietzsche's *Antichrist*

G. Palafox

September 22, 2020

Abstract

An elementary text analysis of Nietzsche's *The Antichrist* [3] is performed. The distribution of some of the obtained data is estimated.

1 Introduction

A text analysis of Nietzsche's *The Antichrist* [3] is performed. We find the frequency of common words, and explore the lengths of sentences in the text. We try to find the distribution of this data, or simulate it if we cannot find a fitting common distribution. Additionally, we compare a network representation of the book with a random network having the same degree distribution.

2 Text analysis and distributions

The text extraction and analysis is performed in a Jupyter [1] notebook running R [4] version 4.0.0[1]. The book is downloaded directly from Project Gutenberg's site using R's `gutenbergr` library. The book downloaded starts with an introduction by the translator, which we omitted from the analysis, as the intention was to study the author's words.

First we obtain two elements of the text: the length of its sentences, and the frequency of the words appearing in it. We discard so called stop-words, and keep only the most frequent words (those appearing over eight times). We create a histogram of the sentences' length, that can be seen in Figure 1a. From our knowledge of discrete distributions, we infer these data has a geometric distribution. We use R's function `fitdistr` from the `MASS` library to try and fit the data to a geometric distribution, obtaining a parameter $p = 0.36$. We generate a thousand pseudo-random numbers with geometric distribution and $p = 0.36$, and plot a histogram of these values, as seen in Figure 1b. Observing these two histograms side-by-side in Figure 1, we conclude the length of the sentences has a geometric distribution with $p = 0.36$.

In the case of the word frequency, which is plotted in Figure 2a, the function `fitdistr` does not give as good of a match. In this case, we pseudo-randomly select words from the text, aiming to obtain a distribution similar to the actual one. To do this, we do the following. We interpret the bar plot as a probability mass function (prior to normalizing the area to be one). Then, we partition the words in groups according to the frequency of which they occur. We calculate the area of each of these groups, relative to the total area, to partition get a partition $p_0 = 0, p_1, \ldots, p_8, p_9 = 1$ of the unit interval . Finally, we generate a pseudo-randomly generated number p, and select a word pseudo-randomly from group k if $p_{k-1} \leq p \leq p_k$. A bar plot showing the frequency of the randomly generated words is shown in Figure 2b. It can be compared to the actual frequency of words in the text in Figure 2. Additionally, Table 1 shows the most frequent words from both the text and our random selection, and Figure 3 shows a boxplot comparing these two data sets.

2.1 Network representation

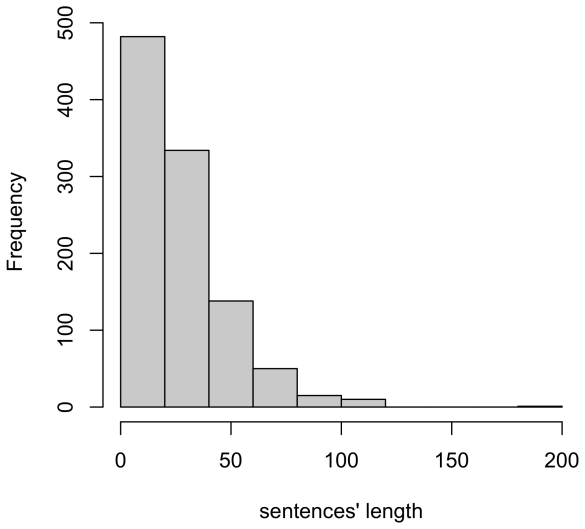
In a previous text analysis[2], a network representation of our book was made, as is seen in Figure 4a. Using R's library `igraph`, function `sample_degseq`, we create a random network which has the same degrees as our book network. This is done based on the configuration model [2], discarding graphs with multiple-edges or self-loops. We cannot say much about how these two compare, but decided to show them here nonetheless. The networks are shown side by side in Figure 4.

3 Conclusion

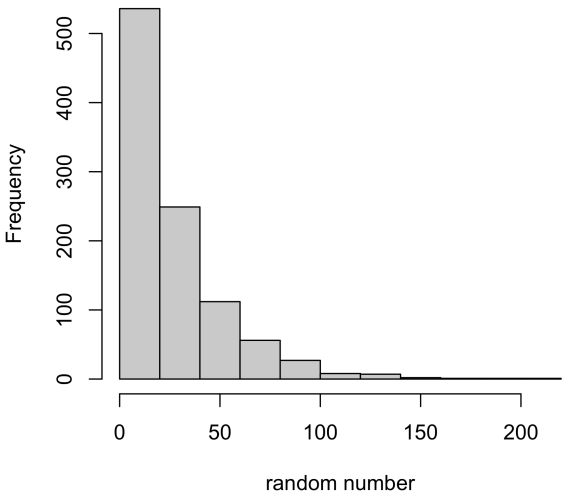
Overall, our exposition is very elementary. Further statistical analysis can be done to determine whether the geometric distribution really is a good fit for the sentences' length distribution. A deeper study of discrete probability distributions

[1] The script and a Jupyter [1] notebook detailing our analysis and graphics creation, can be found at `https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment3`

[2] It can be found at `https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment2`

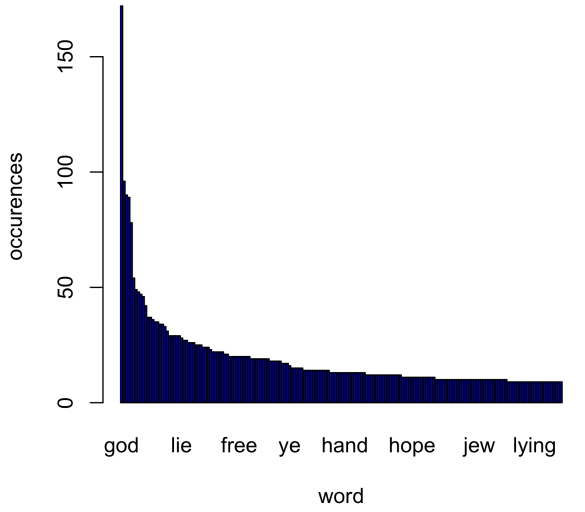


(a) Histogram of sentences' length.

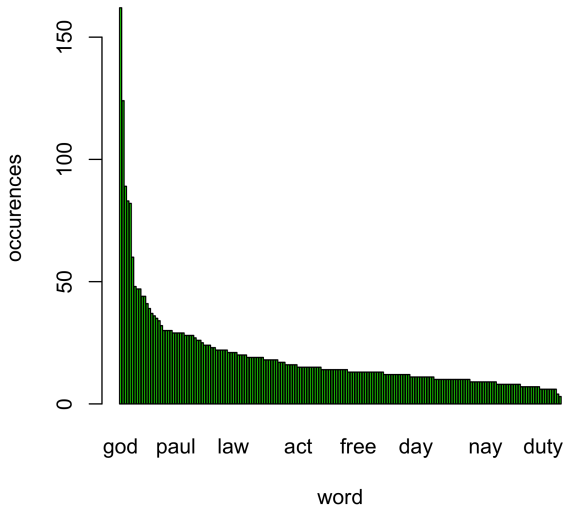


(b) Histogram of 1000 words with distribution Geo(0.36).

Figure 1: Sentences' length distribution.



(a) Word occurrence in the book.



(b) Word occurrence of words pseudo-randomly selected from the book.

Figure 2: Barplots of word occurrences.

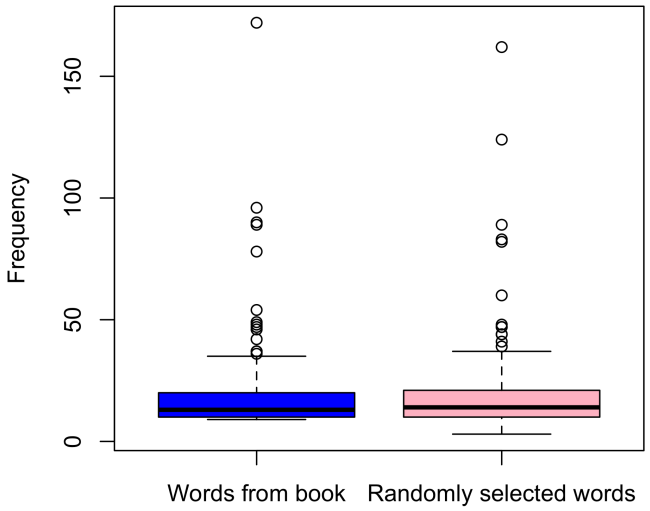


Figure 3: Comparison of word frequency.

(a) Book network. Words are vertices and edges joining words appear as a bigram in the text.

(b) Random network with same degree distribution as book network.

Figure 4: Network representation.

Table 1: Word frequency

(a) Frequency of words in the book

words	freq
god	172
life	96
christian	90
christianity	89
world	78
truth	54

(b) Frequency of pseudo-randomly selected words from book

rand_words	Freq
god	162
world	124
life	89
christian	83
christianity	82
sort	60

may help find a good fit for the distribution of how frequent words distribute in the text. More sophisticated network analysis techniques could be deployed, aiming to understand the differences and similarities of our book network and a random network with the same degree sequence.

References

[1] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., et al. Jupyter notebooks—a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing* (2016), IOS Press, p. 87.

[2] Newman, M. *Networks*, vol. 1. Oxford University Press, Oct 2018.

[3] Nietzsche, F. *The Antichrist*. Project Gutenberg, Sept. 2006. `http://www.gutenberg.org/files/19322/19322.txt`.

[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

Generation of Poisson distributed pseudo-random numbers

G. Palafox

September 29, 2020

Abstract

Two algorithms generating Poisson-distributed pseudo-random numbers are shown. A rigorous proof for one of them is provided, while the correctness of the other one is analyzed computationally. Lastly, an example of how the binomial distribution approaches the Poisson distribution is also given.

1 Introduction

In this work, two algorithms for generating Poisson distributed numbers are explained. The first one assumes access to an exponentially-distributed pseudo-random number generator, while the second one assumes access to a uniformly-distributed pseudo-random number generator. Graphics and elementary probability are used to support the validity of the algorithms. In the last section, an example of how the binomial distribution approaches the Poisson distribution is given. This study was performed with R version 4.0.0 [4] on a Jupyter [1] notebook[1].

2 Generating pseudo-random numbers

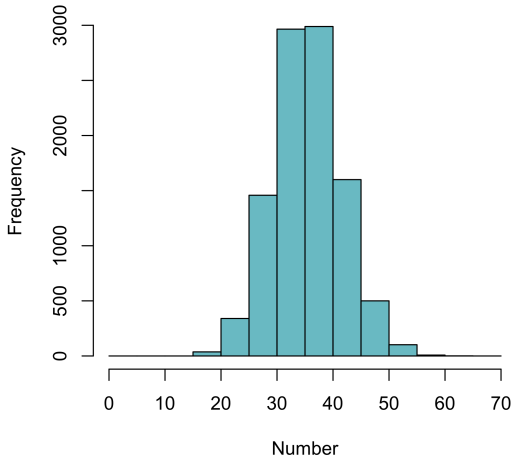
First, an algorithm which generates numbers with a Poisson distribution is shown in Algorithm 1. The idea is to generate numbers $x_i \sim \text{Exp}(\lambda)$ until $x_1 + \cdots + x_k$ first exceeds some number M. Then, we return $k-1$. This will generate numbers following a Poisson distribution with mean λM. A rigorous proof of this claim is not shown, however, the following intuitive explanation is given. In a Poisson process with rate λ, the number of events in an interval of length t is a Poisson distributed random variable with mean λt [6]. Additionally, the time between events in a Poisson process is distributed exponentially with mean $1/\lambda$. From these two facts we can see that the number of exponential variables generated with sum less than M must be the events occurring in a time interval of length M, so they must have distribution $\text{Pois}(\lambda M)$. A different, significantly more informal explanation, is that if each exponential random variable is $1/\lambda$ on average, $M\lambda$ of this variables will be needed on average for their sum to exceed M. Empirically, Figure 1 shows different comparisons between numbers generated by Algorithm 1 and numbers generated by R's `rpois` function with the mean we expected to see. Furthermore, R's function `MASS::fitdistr` supported our conclusions, as is seen in Table 1.

Algorithm 1 Poisson numbers from exponentials

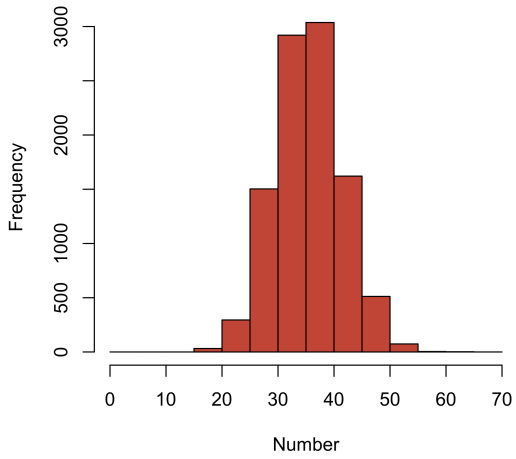
Input: Positive values `lambda, m`; positive integer `rep`.
Output: Array of `rep` numbers with Pois(`lambda` $*$ `m`) distribution.

Make empty numeric vector `ce`
for `i` $= 1, 2, \ldots,$ `rep` **do**
    Make empty numeric vector `de`
    while `sum(de)` $<$ `m` **do**
        Generate pseudo-random `x` $\sim$ Exp(`lambda`)
        Append `x` to array `de`
    end while
    Append `length(de)`-1 to array `ce`
end for
return `ce`

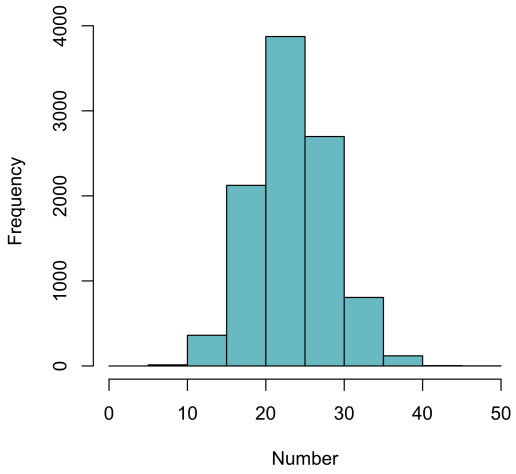
[1] The notebook with the code containing our analysis, as well as this report, can be found in the Github Repository: `https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment4`



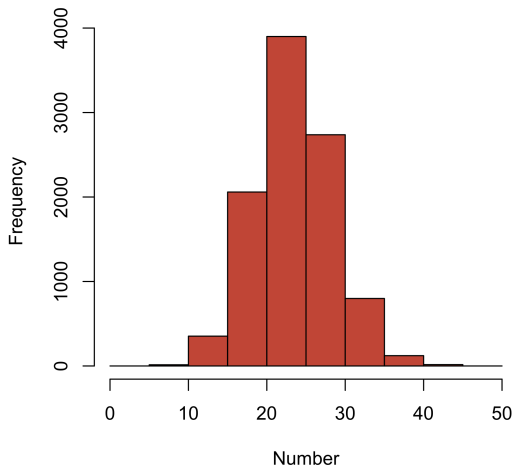
(a) Histogram of numbers generated by Algorithm 1 with `la` = 12, `me` = 3.



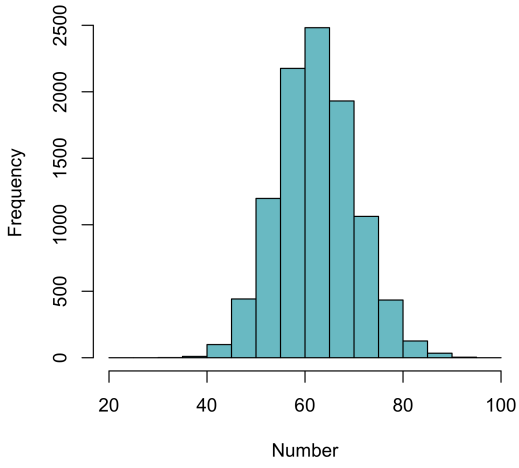
(b) Histogram of numbers generated by R with Pois(36) distribution.



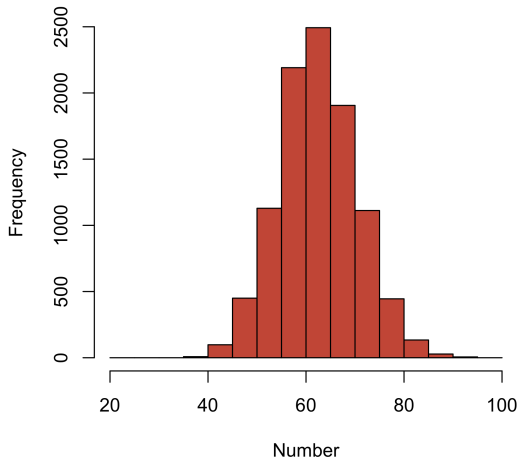
(c) Histogram of numbers generated by Algorithm 1 with `la` = 4, `me` = 6.



(d) Histogram of numbers generated by R with Pois(24) distribution.



(e) Histogram of numbers generated by Algorithm 1 with `la` = 9, `me` = 7.



(f) Histogram of numbers generated by R with Pois(63) distribution.

Figure 1: Comparison of Algorithm 1 vs R's generator.

Table 1: Proposed λ compared to λ given by `MASS::fitdistr`

Proposed λ	Estimated λ by `fitdistr`	Estimated standard error
36	35.954	0.059
24	23.895	0.048
63	62.853	0.079

2.1 Using uniform random variables

A different way of generating Poisson pseudo-random numbers is shown in Algorithm 2. The algorithm is attributed to Knuth [2], and it can be proven as a consequence of Algorithm 1.

Lemma 1. *If* $u \sim \text{Unif}(0,1)$ *and* $\lambda > 0$*, then* $\frac{-\log(u)}{\lambda} \sim \text{Exp}(\lambda)$.

Proof. First we find $\mathbb{P}\left(\frac{-\log(u)}{\lambda} \leq x\right)$. Observe that

$$\frac{-\log(u)}{\lambda} \leq x \Leftrightarrow -\log(u) \leq \lambda x \tag{1}$$

$$\Leftrightarrow \log(u) \geq -\lambda x \tag{2}$$

$$\Leftrightarrow u \geq \exp(-\lambda x) \tag{3}$$

Therefore,

$$\mathbb{P}\left(\frac{-\log(u)}{\lambda} \leq x\right) = \mathbb{P}\left(u \geq \exp(-\lambda x)\right) = 1 - \mathbb{P}\left(u \leq \exp(-\lambda x)\right) \tag{4}$$

From the cumulative distribution function for the uniform distribution, we see $\mathbb{P}\left(u \leq \exp(-\lambda x)\right) = \exp(-\lambda x)$, thus

$$\mathbb{P}\left(\frac{-\log(u)}{\lambda} \leq x\right) = 1 - \exp(-\lambda x) \tag{5}$$

This completes the proof. □

Proposition 1. *Knuth's Algorithm 2 generates pseudo-random numbers with a* $\text{Pois}(\lambda)$ *distribution when given a parameter* `lambda` $= \lambda$.

Proof. We know from Algorithm 1 that counting the number of exponential random variables with mean $1/\lambda$ such that $x_1 + \cdots + x_k < 1$ gives us a Poisson variable with mean λ. Using Lemma 1 we can rewrite this as $-\frac{\log(u_1)}{\lambda} - \cdots - \frac{\log(u_k)}{\lambda} < 1$, where $u_i \sim \text{Unif}(0,1)$. Then, observe that

$$-\frac{\log(u_1)}{\lambda} - \cdots - \frac{\log(u_k)}{\lambda} < 1 \Leftrightarrow -\log(u_1) - \cdots - \log(u_k) < \lambda \tag{6}$$

$$\Leftrightarrow -\log(u_1 u_2 \cdots u_k) < \lambda \tag{7}$$

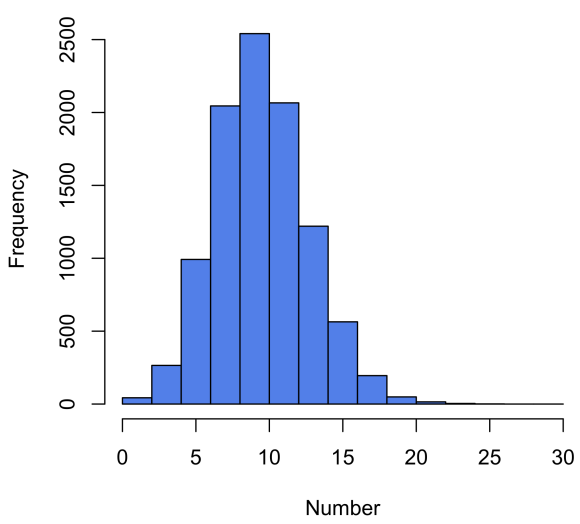
$$\Leftrightarrow u_1 u_2 \cdots u_k > \exp(-\lambda) \tag{8}$$

This shows that the number of exponential random variables generated before its sum exceeds one is the same as the number of uniform random variables generated before its product is less than $\exp(-\lambda)$ and so they must have the same distribution. □

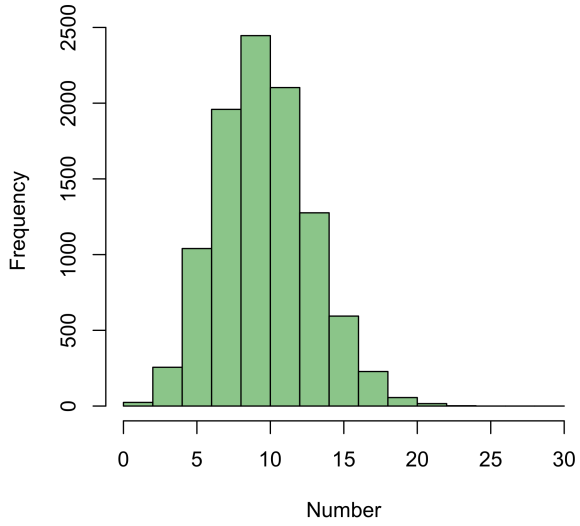
Supporting the formal proof, a comparison of numbers generated by Algorithm 2 and by R's `rpois` function is shown in Figure 2.

3 Binomial distribution tends to Poisson distribution

It is known [5] that for n large and small p, the binomial distribution $\text{Binom}(n,p)$ can be approximated with a Poisson distribution $\text{Pois}(np)$. This situation is exemplified with a random graph. An Erdős–Rényi random network [3], denoted as $G(n,p)$, is defined in the following way: n nodes are fixed, and an edge is placed between each distinct pair with independent probability p. It should be clear that the probability of any vertex having degree k is $\binom{n-1}{k}p^k(1-p)^{n-1-k}$. That is, the degree of vertices has a binomial distribution. A random network $G(10000, \frac{1}{10000})$ is created using R's library `igraph`. The degree distribution of this network is binomial, due to the justification given above. However, since n is sufficiently large and p is sufficiently small, the degree distribution can be approximated as a Poisson distribution. Figure 3a shows the degree distribution of the network, and Figure 3b shows an identical histogram for numbers having a Pois(1) distribution. Figure 3c shows a boxplot comparing these two sets of data.



(a) Histogram of numbers generated by Algorithm 2 with `la` = 10.



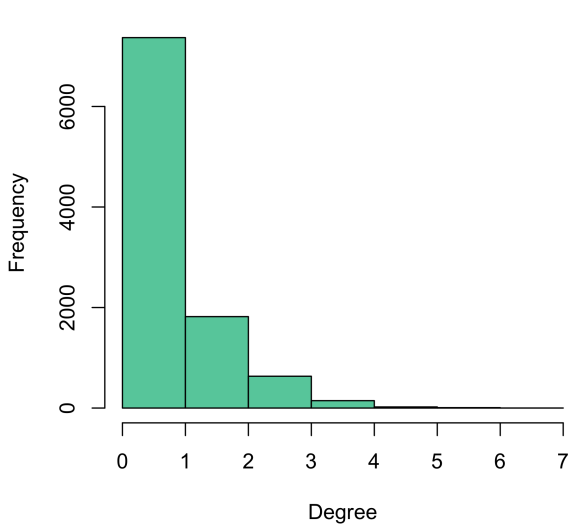
(b) Histogram of numbers generated by R with Pois(10) distribution.

Figure 2: Comparison of Algorithm 2 vs R's generator.

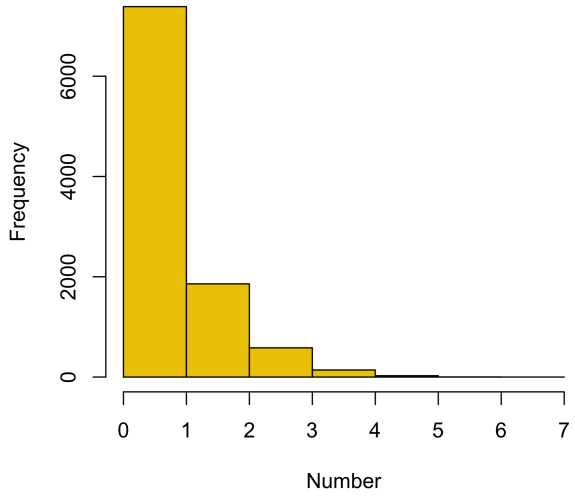
Algorithm 2 Knuth's algorithm

Input: Positive value `lambda`; positive integer `rep`.
Output: Array of `rep` numbers with Pois(`lambda`) distribution.

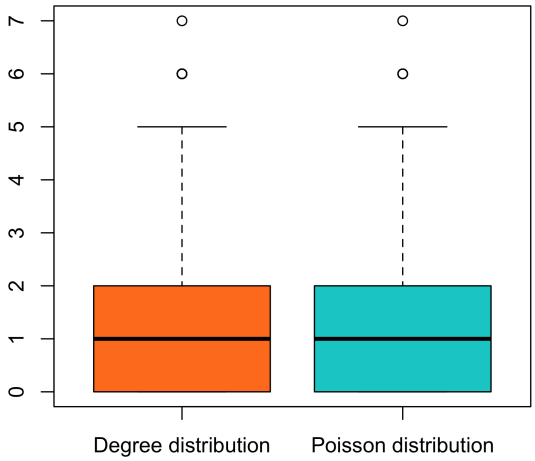
Make empty numeric vector `cu`
for `i` $= 1, 2, \ldots,$ `rep` **do**
    Make a numeric vector `du` containing only value 1
    while `prod(du)` $>$ `exp(-lambda)` **do**
        Generate pseudo-random `u` $\sim$ `Unif(0,1)`
        Append `u` to array `du`
    end while
    Append `length(du)-2` to array `cu`
end for
return `cu`



(a) Histogram of degrees in $G(10000, \frac{1}{10000})$ network.



(b) Histogram of 10 000 numbers generated by R with Pois(1) distribution.



(c) Boxplot comparing the data of Figures 3a and 3b.

Figure 3: Degree distribution of random network and Poisson-distributed numbers.

4 Conclusion

This work showed how some distributions (uniform, exponential, binomial) relate to the Poisson distribution. Further work may include the relation between normal and Poisson distributions. Additionally, it may be of interest to study how to use uniform values to generate pseudo-random numbers following distributions other than Poisson.

5 Acknowledgments

We thank Professor Elisa Schaeffer for providing an R code with Algorithms 1 and 2.

References

[1] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al., *Jupyter notebooks—a publishing format for reproducible computational workflows*, in Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing, IOS Press, 2016, p. 87.

[2] D. E. Knuth, *The art of computer programming*, vol. 2, Addison-Wesley, 3rd ed., 1997.

[3] M. Newman, *Networks*, vol. 1, Oxford University Press, Oct 2018.

[4] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[5] S. M. Ross, *A first course in probability*, Macmillan, 1976.

[6] ——, *Introduction to probability models*, Harcourt/Academic Press, 7th ed., 2000.

Pseudo-random number generators

G. Palafox

October 6, 2020

Abstract

Various pseudo-random number generators are exhibited and compared. Statistical tests are run to determine their performance. Variations are studied to analyze the sensitivity of the algorithms.

1 Introduction

In the present work, different pseudo-random number generators are studied. In particular, methods for generating uniform distributed pseudo-random numbers and normal distributed pseudo-random numbers are studied. Their efficacy is statistically tested with Frosini's uniformity test and Shapiro's Normality test, respectively.Additionally, the sensitivity of some of the methods is tested by varying their input parameters and measuring possible changes in the quality of the output. A time comparison is also carried out to compare the performance of some of the methods. This study was performed with R version 4.0.0 [4] on a Jupyter [3] notebook[1].

2 Methods for the uniform distribution

First, the simplest of the methods is studied: the linear congruential generator (LGC). Parameters are varied to look for differences in its output. Then, the Additive Congruential Random Number Generator is looked at. Frosini's test for uniformity [1] is performed on data outputted by both algorithms.

2.1 Linear Congruential Method

This method outputs a sequence of n numbers uniformly distributed on $(0, 1)$. Pseudo-code for this method is shown in Algorithm 1. It is based on the recurrence relation

$$X_{n+1} = (aX_n + c) \mod m, \tag{1}$$

where X_0, a, c and m are integers chosen by the user. The algorithm outputs at most m distinct numbers, but the possibility of patterns of length shorter than m occurring exist. Necessary and sufficient conditions for achieving an m-length sequence are known [2]: a, m must be coprime, $a-1$ must be divisible by all prime factors of m, and $a-1$ need be divisible by 4 if m is divisible by 4. Figure 1a shows an histogram of the output of LGC with a seed of $X_0 = 101$, and $a = 1151, c = 27077, m = 2^{32}$. The data in this histogram gives us a p-value of $0.38 > 0.05$ when Frosini's Uniformity test [1] is performed on it. To further study this method, a hundred sequences of length one-thousand were generated, and for each, the p-value associated to Frosini's test was computed and stored. This was done twice: first, with parameters $a = 2^{10} + 1, c = 2^{16} + 1, m = 2^{32}$ and then with parameters $a = 2^{10}, c = 2^{16} + 1, m = 2^{32}$. Note the first set of parameters satisfy the aforementioned conditions for maximum length while the second set does not. In particular, $a = 2^{10}, m = 2^{32}$ are not coprime. In the first scenario, the p-value was greater than 0.05 in 93 out of 100 times. A boxplot of this p-values is shown in Figure 1b. For the second scenario, all hundred p-values where significantly less than 0.05.

2.2 Additive Congruential Random Number Generator

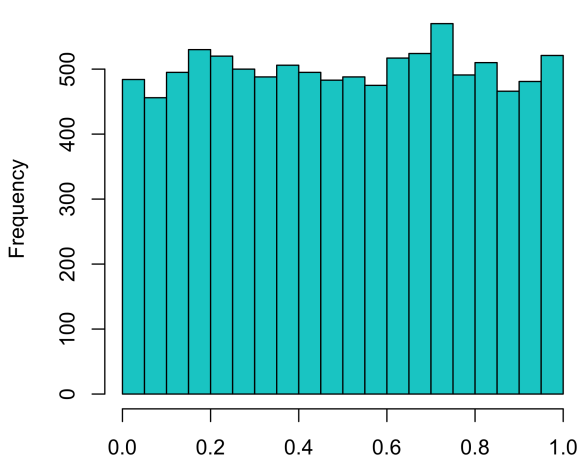
Similar as the LGC, the Additive Congruential Random Number Generator (ACORN) [7] is a method based on modular arithmetic and generates n uniformly distributed pseudo-random numbers. It starts with k initial values $Y_0^m, m = 1, 2, \ldots, k$ all being less than a modulus M. With these, the following are defined:

$$Y_n^0 = Y_{n-1}^0, n \geq 1, \tag{2}$$

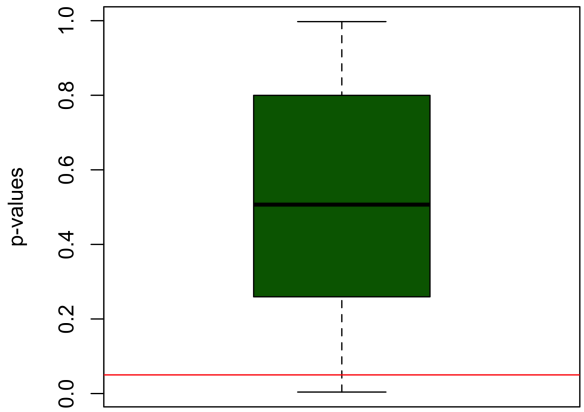
$$Y_n^m = (Y_n^{m-1} + Y_{n-1}^m) \mod M, n \geq 1, m = 1, 2, \ldots, k, \tag{3}$$

$$X_n^k = Y_n^k / M,\ n \geq 1. \tag{4}$$

[1] The notebook with the code containing our analysis, as well as this report, can be found in the Github Repository: `https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment5`



(a) Histogram of a LCG sample of size ten-thousand, with parameters $a = 1151, c = 27077, m = 2^{32}$



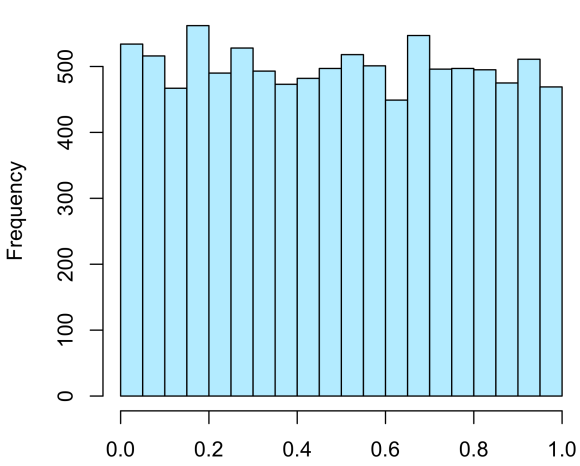
(b) Boxplot of the p-values of a hundred LCG samples, with $a = 2^{10} + 1, c = 2^{16} + 1, m = 2^{32}$. Red line is at 0.05

Figure 1: Linear congruential generator.

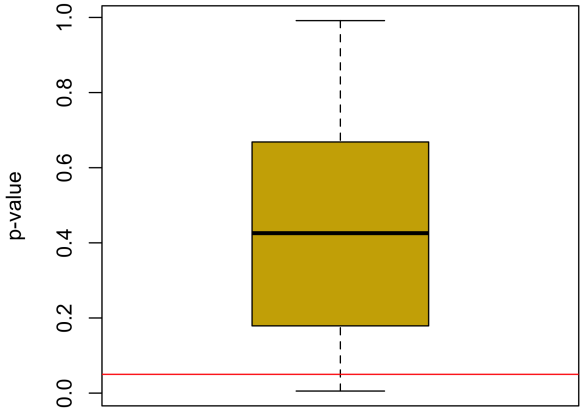
Algorithm 1 Linear Congruential Generator

Input: Positive integers `n, a, c, m, seed`.
Output: Array of `n` numbers with Unif(0,1) distribution.

1: Make empty numeric vector `numbers`
2: Make `x` = `seed`
3: **while** length(`numbers`) < `n` **do**
4: Make `x = (a*x + c) mod m`
5: Append `x/(m-1)` to `numbers`
6: **end while**
7: **return** `numbers`



(a) Histogram of one ACORN sample.



(b) Boxplot of the p-values of a hundred ACORN samples.

Figure 2: ACORN generated numbers.

The sequence X_n^k is the sequence of ACORN generated pseudo-random numbers. Pseudo-code for this method is found in Algorithm 2. In a similar fashion as with LGC, a hundred samples of ACORN outputs were tested with Frosini's test, and the respective p-values were stored. Here, 95 out of a 100 were greater than 0.05. These are displayed in Figure 2b. An histogram of ACORN generated numbers is found in Figure 2a.

Algorithm 2 Additive Congruential Random Number Generator

Input: Positive integers `M, n, y11, y12, ..., y1k`.
Output: Array of `n` numbers with Unif(0,1) distribution.

```
Create n × k matrix A with first row equal to y11, y12, ..., y1k
for row = 2, 3, ..., n do
    for col = 1, 2, ... k do
        if col == 1 then
            A[row][col] = A[row-1][col]
        else
            x = A[row][col-1] + A[row-1][col]
            A[row][col] = x mod M
        end if
    end for
end for
Make vector numbers equal to the k-th column of A
return numbers / M
```

3 Methods for the normal distribution

Methods for generating normal-distributed pseudo random numbers are shown next. First, the Box-Muller method and its polar form variation are presented. The sensitivity of Box-Muller method is explored. Finally, an algorithm based on the rejection method [6] is given.

3.1 Box-Muller method

Algorithm 3 presents the Box-muller method for sampling a pair z_0, z_1 of normal distributed pseudo-random numbers. A proof of why the algorithm works is given by Ross [5]. The algorithm was used to generate one-hundred samples of

Table 1: Number of p-values smaller and larger than 0.05 for different normal samples

Method	Less than 0.05	Greater than 0.05
Box-Muller, keeping only z_0	3	97
Box-Muller, keeping only z_1	6	94
Box-Muller, keeping both z_0, z_1	7	93
Using `rnorm`	4	96

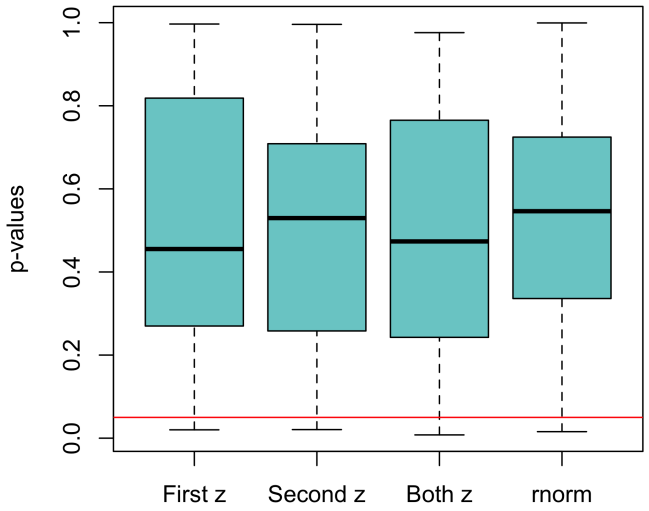


Figure 3: p-values obtained with different methods. Red line represents 0.05

five-thousand normally distributed values, where the p-values resulting of performing the Shapiro Test on each sample was stored. This was compared to the p-values of R's `rnorm` method, and to variations of the Box-Muller method were only the first (z_0) or second (z_1) values sampled are kept. The comparison of these p-values is shown in Figure 3. A summary is also shown in Table 1.

Algorithm 3 Box-Muller

Input: Real numbers `mu, sigma`.
Output: Number with Normal(`mu, sigma`) distribution.

1: Generate numbers `u1, u2` $\sim$ Unif$(0, 1)$
2: Make `z0 = sqrt(-2 * log(u1)) * cos (2 * pi * u2)`
3: Make `z1 = sqrt(-2 * log(u1)) * sin (2 * pi * u2)`
4: **return** `sigma * z0 + mu, sigma * z1 + mu`

Next, to study the sensitivity of Algorithm 3, some variations were performed: first, instead of using R's `runif` to generate the uniform numbers, Algorithm 1 (LGC) was used instead. The other two variations consisted of using two *non-independent* uniform distributed numbers. The first one uses $u_1 \sim \text{Unif}(0, 1)$ and $u_2 = \frac{u_1}{2}$, while the second one uses $u_1 \sim \text{Unif}(0, 1)$ and $u_2 = \frac{u_1+1}{2}$. As before, we generated a hundred samples and stored the p-values from Shapiro's Test. The results are summarized in Table 2.

It is observed that using a different method for generating the uniform pseudo-random numbers did not have much effect on the output. Using non-independent uniform numbers, however, made caused all samples to fail Shapiro's test.

3.1.1 Box-Muller polar form

A known variation of the Box-Muller method [5] is given in Algorithm 4. It works with the same principle, but avoids the computation of trigonometric functions. The performance of both Box-Muller forms was compared, concluding that

Table 2: Number of p-values smaller and larger than 0.05 for different Normal samples

Method	Less than 0.05	Greater than 0.05
Box-Muller, using LGC	8	92
Box-Muller, non-independent uniform values ($u_1 \sim \text{Unif}(0,1)$ and $u_2 = \frac{u_1}{2}$)	100	0
Box-Muller, non-independent uniform values ($u_1 \sim \text{Unif}(0,1)$ and $u_2 = \frac{u_1+1}{2}$)	100	0

Algorithm 4 performs slower than Algorithm 3 77% of the time, with an average difference of 0.001 seconds.

Algorithm 4 Box-Muller Transform. Polar version

Input: Real numbers `mu, sigma`.
Output: Number with Normal(`mu, sigma`) distribution.

repeat
    Generate numbers `u1, u2` $\sim$ Unif$(0,1)$
    Make `V1 = 2 * u1 - 1`
    Make `V2 = 2 * u2 - 1`
    Make `S = V1`^{2} `+ V2`^{2}
until `S` ≤ 1
Make `Z1 = sqrt( (-2 * log(S) ) / S ) * V1`
Make `Z2 = sqrt( (-2 * log(S) ) / S ) * V2`
return `sigma * Z1 + mu, sigma * Z2 + mu`

3.2 Rejection method

Algorithm 5 is based on the Rejection Method, and is explained in Ross' Simulation book [6]. As in previous sections, a hundred samples were generated using Algorithm 5 and the p-values of this samples under Shapiro's test were stored, as can be seen in Figure 4c. A histogram of a five-thousand sample generated with Algorithm 5 is shown in Figure 4a, next to a histogram of five-thousand numbers generated by R's `rnorm` in Figure 4b.

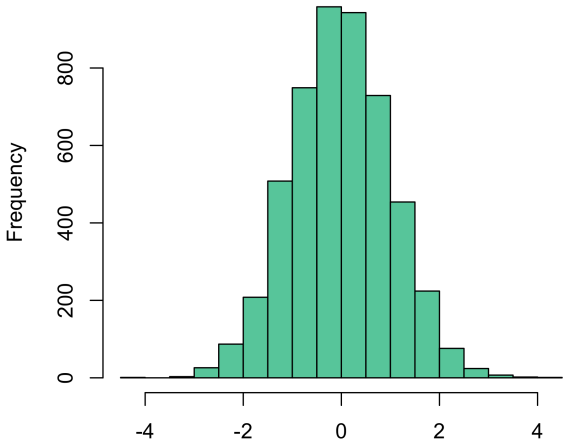
Algorithm 5 Rejection method. Normal distribution

Input: Real numbers `mu, sigma`.
Output: Number with Normal(`mu, sigma`) distribution.

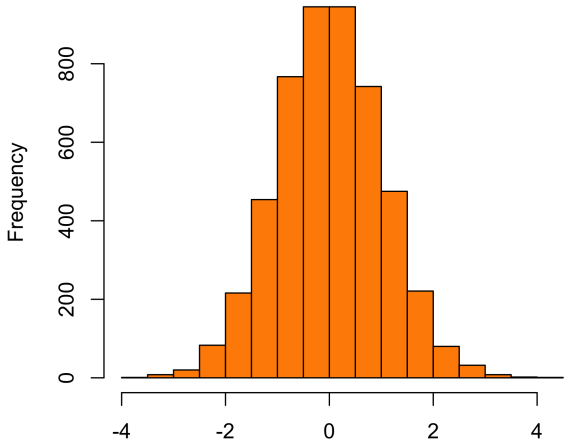
Generate `y1, y2` $\sim$ Exp(1)
while `y2 - (y1-1)`^{2} `/ 2` $<$ `0` **do**
    Generate `y1, y2` $\sim$ Exp(1)
end while
Generate `u` $\sim$ Unif$(0,1)$
if `u` $\leq 1/2$ **then**
    Make `z = y1`
else
    Make `z = - y1`
end if
return `sigma * z + mu`

4 Conclusions

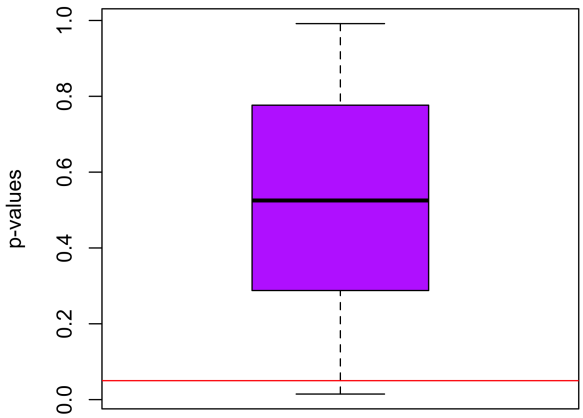
It is observed how methods for different generating non-uniform pseudo-random numbers commonly rely on the generation of uniform distributed pseudo-random numbers. The LGC and ACORN methods both seem to give good results, even when used with methods like Box-Muller's. Further study should include pseudo-random number generators for different distributions. Given how Frosini's and Shapiro's test are both distribution dependent (uniform, normal), the study of generators for new distributions should be accompanied with the study of new statistical tests.



(a) Histogram of one five-thousand number sample given by Algorithm 5.



(b) Histogram of one five-thousand number sample given by `rnorm`.



(c) Boxplot of the p-values of a hundred samples given by Algorithm 5.

Figure 4: Comparing Algorithm 5 to `rnorm`.

5 Acknowledgments

We thank Professor Elisa Schaeffer for providing code for the Box-Muller method in Algorithm 3 and the Linear Congruential Generator in Algorithm 1.

References

[1] P. Blinov and B. Lemeshko, *A review of the properties of tests for uniformity*, in 2014 12th International Conference on Actual Problems of Electronics Instrument Engineering (APEIE), IEEE, Oct 2014, p. 540–547.

[2] T. E. Hull and A. R. Dobell, *Random number generators*, SIAM Review, 4 (1962), pp. 230–254.

[3] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al., *Jupyter notebooks—a publishing format for reproducible computational workflows*, in Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing, IOS Press, 2016, p. 87.

[4] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[5] S. M. Ross, *Introduction to probability models*, Harcourt/Academic Press, 7th ed., 2000.

[6] ——, *Simulation*, Elsevier Academic Press, 4th ed ed., 2006.

[7] R. Wikramaratna, *ACORN—a new method for generating sequences of uniformly distributed pseudo-random numbers*, Journal of Computational Physics, 83 (1989), p. 16–31.

Statistical tests

G. Palafox

October 12, 2020

Abstract

Elementary facts of statistical tests are presented. Common statistical tests are performed on real data sets.

1 Introduction

In this report, basic facts about statistical tests are presented, along with examples of tests performed on real data sets of Mexico City subway system. The data was downloaded directly from INEGI's website [4], and includes kilometers traveled, passengers transported, and energy consumed by Mexico City's subway in 2018 and 2019. A fragment of this data is shown in Table 1. The work was coded in R version 4.0.0 [7] on a Jupyter [5] notebook[1].

2 Statistical hypothesis testing

The following is a brief summary of the theory regarding statistical hypothesis testing. It is not original work by the author. All presented here, and further topics, can be found in excellent books such as the one by Ross [8], Casella and Berger [2], or Navarro's online electronic book [6]. A statistical hypothesis is a statement about the nature of data, usually in terms of some statistical parameter [8]. To test the hypothesis, it must be decided whether a data sample appears to be consistent with said hypothesis For example, it may be the case where the real mean of some data is unknown, and a sample is tested to see if it is consistent with data having mean m_0. The hypothesis to be tested is called *null hypothesis*, denoted by H_0, and the alternative is called *alternative hypothesis*, denoted by H_1 or H_a. In the previous example, the test would be written as

$$H_0 : \mu = m_0, \quad H_1 : \mu \neq m_0, \tag{1}$$

where μ denotes the real mean of the population. The alternative hypothesis can also be written as $\mu > m_0$ or $\mu < m_0$. The null hypothesis can also be written as $\mu - m_0 = 0$, which would change the alternative hypothesis to $\mu - m_0 > 0$, and so on. The typical set up consists of specifying a small value α (called *significance level*, commonly set as 0.05) and then requiring the test to have the property that, whenever H_0 is true, its probability of being rejected is less than or equal to α. Statistical tests often output a p-value: this is the smallest significance level at which we would allow for the rejection of the null hypothesis for the given data [6, 8].

2.1 Practical considerations

When performing statistical tests, certain questions arise, such as choosing significance levels, or interpreting the output of the test. We intend to address some of these questions here, showing examples of statistical tests performed in R.

[1]The notebook with the code containing our analysis, as well as this report, can be found in the Github Repository: `https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment6`

Table 1: Fragment of Mexico City subway data.

Period	Km traveled (thousands)	Passengers Transported (millions)	Energy consumed (thousands of KWH)
2018			
January	3816.46	129.10	89340.53
February	3389.20	124.69	80122.84
March	3688.76	131.28	89370.53
April	3646.12	132.84	84142.34
May	3768	135.99	89340.53

Table 2: p-values from Shapiro-Wilk normality test on different data samples.

Data	p-value
Passengers	0.245
Passengers in 2018	0.346
Passengers in 2019	0.927
Kilometers traveled	0.495
Energy consumption	0.112

2.1.1 Interpreting the output of a statistical test

If the p-value is less than the significance level α, the null hypothesis is rejected and the alternative hypothesis is accepted. On the contrary, if $p \geq \alpha$, we do not reject H_0. Rejecting the null hypothesis should be taken as a strong indicative that it does not appear consistent with the observed data [8], while not rejecting H_0 is a weak indicative that H_0 is consistent with the data [8]. For example, many tests require an assumption of normality. Shapiro-Wilk normality test is a test whose null hypothesis is: *data comes from a normal distribution.* In R, it is easy to perform this test on a numeric array `passng` containing the (millions) of passengers transported by Mexico City subway each month. It is shown in Listing 1. This gives a p-value of 0.24. Taking a significance level $\alpha = 0.05$, the p-value is greater than α, so we do not reject the hypothesis that data comes from a normal distribution, and accept it instead. The same test was performed on different data samples, whose outputs are shown in Table 2.

Listing 1: Normality test

```
shapiro.test(passng)
#       Shapiro-Wilk normality test

#data:  passng
#W = 0.94795, p-value = 0.2445
```

On the other hand, suppose it is of interest to see whether the number of passengers and the kilometers traveled had the same variance. Fisher's F test can be performed, as seen in Listing 2, and it outputs a p-value smaller than 2.2×10^{-16}. In this case, the null hypothesis is rejected, and the ratio of the variance is assumed to be distinct to one.

Listing 2: Fisher's F test

```
var.test(passng, km)

F test to compare two variances

#data:  passng and km
#F = 0.0015576, num df = 23, denom df = 23, p-value < 2.2e-16
#alternative hypothesis: true ratio of variances is not equal to 1
#95 percent confidence interval:
#0.0006738075 0.0036006134
#sample estimates:
#ratio of variances
#0.001557601
```

2.1.2 Meaning of rejecting the null hypothesis

It should be noted that the goal of a statistical test is not to determine whether H_0 is true or not, but to determine whether it being true is consistent with the given data [8]. Rejecting the null hypothesis should mean that our observed data is very unlikely if H_0 is true [8]. Consider the case of performing a One Sample t-Test to see whether the mean of passengers transported is `mu`= 100. Carrying this out in R (see Listing 3) gives a p-value smaller than 2.2×10^{-16}, which leads to rejecting the null hypothesis of the sample having mean equal to a hundred. This means that, if the true mean was 100, it would be very unlikely to observe data as the observed in `passng`.

Listing 3: One sample t-test.

```
t.test(passng, mu = 100)
#       One Sample t-test

#data:  passng
#t = 26.513, df = 23, p-value < 2.2e-16
#alternative hypothesis: true mean is not equal to 100
#95 percent confidence interval:
#130.2191 135.3338
#sample estimates:
#mean of x
```

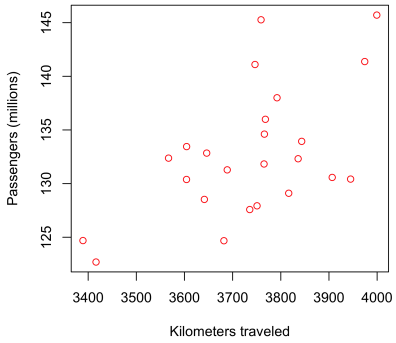


Figure 1: Plot of kilometers traveled against passengers transported.

```
#132.7765
```

A similar conclusion can be reached with Wilcoxon signed rank test, which unlike one sample t-test, does not assume a normal distribution for our data. The results are seen in Listing 4, which show a p-value of $1.192 \times 10^{-7} < 0.05$.

Listing 4: Wilcoxon signed rank test.

```
wilcox.test(passng, mu=100, conf.inf=TRUE)

#Wilcoxon signed rank exact test

#data:  passng
#V = 300, p-value = 1.192e-07
#alternative hypothesis: true location is not equal to 100
```

2.1.3 Parametric and non-parametric tests. Assumptions and examples

Common parametric tests include Student's t-test, Pearson's correlation coefficient, linear regression and ANOVAs. Common assumptions for these tests is that observations are independent, and that the samples come from normal distributed populations with common variance. As seen in Listing 5, a correlation test can be used to see whether there is correlation between kilometers traveled and passengers transported. In this case, the null hypothesis is that true correlation is equal to zero. Given how a p-value of $0.003 < 0.05$ was obtained, the null hypothesis can be rejected, concluding there is some correlation between the variables. This can be further supported with the graphic in Figure 1, where kilometers traveled are plotted against passengers transported (x and y axis respectively).

Listing 5: Correlation test.

```
cor.test(km, passng)
#Pearson's product-moment correlation

# data:  km and passng
# t = 3.3106, df = 22, p-value = 0.003181
#alternative hypothesis: true correlation is not equal to 0
#95 percent confidence interval:
#0.2257709 0.7950927
#sample estimates:
#cor
#0.5766491
```

Among the non-parametric tests one can find χ^2 tests, Spearman-Kendall correlation coefficients, and Kruskal-Wallis tests. A χ^2 test is performed in Listing 6 to see if two categorical variables are dependent, by means of a contingency table. The categorical variables are low and high kilometers traveled and energy consumption, defined as being below or above the median. The null hypothesis is that the variables are independent, which cannot be rejected with the obtained p-value (0.219). This is a weak indication of the variables being dependent.

Listing 6: Chi Squared test.

```
df2 <- data.frame(km, kwh)

df2$cat_x <- (df2$km < median(km))

df2$cat_y <- (df2$kwh < median(kwh))
```

```
chisq.test(table(df2$cat_x, df2$cat_y), correct = FALSE)

#Pearson's Chi-squared test

#data:  table(df2$cat_x, df2$cat_y)
#X-squared = 1.5105, df = 1, p-value = 0.2191
```

Other examples of statistical tests are shown next. In Listings 7 and 8, a null hypothesis of whether the number of passengers transported in 2018 has the same mean as the number of passengers transported in 2019 is tested. The difference between these two tests is that the t-test assumes the samples are drawn from a normal distribution, while Wilcoxon rank-sum test does not. In both tests, a p-value larger than 0.05 is obtained, so the null hypothesis cannot be rejected, suggesting both samples may have the same mean.

Listing 7: Two sample t-test.

```
t.test(passng18, y=passng19)

#Welch Two Sample t-test

#data:  passng18 and passng19
#t = -0.087937, df = 21.716, p-value = 0.9307
#alternative hypothesis: true difference in means is not equal to 0
#95 percent confidence interval:
#-5.468207  5.023664
#sample estimates:
#mean of x mean of y
#132.6653  132.8876
```

Listing 8: Wilcoxon rank sum test.

```
wilcox.test(passng18, passng19, alternative = "g")

#Wilcoxon rank sum exact test

# data:  passng18 and passng19
#W = 68, p-value = 0.6006
#alternative hypothesis: true location shift is greater than 0
```

In Listing 9, a null hypothesis of whether the number of passengers transported in 2018 and in 2019 come from the same distribution is tested. A p-value of 0.99 is obtained, which suggests the data indeed comes from the same distribution.

Listing 9: Kolmogorov Smirnov test.

```
ks.test(passng18, passng19)

#Two-sample Kolmogorov-Smirnov test

#data:  passng18 and passng19
# D = 0.16667, p-value = 0.9985
# alternative hypothesis: two-sided
```

2.1.4 Further considerations

The **choosing of significance level** α should depend on how dangerous is to reject H_0 when it is true, with lower α values associated with a lower risk tolerance [9]. Navarro mentions in her book [6] that a **common mistake** is thinking the p-value is the probability of the null hypothesis to be true. **Statistical power** refers to the probability of rejecting the null hypothesis when it is false [3]. It gives a method of discerning between competing tests of the same hypothesis, with the test with the higher power being preferred [3].

3 Conclusion

Statistical hypothesis tests give an important set of tools to use when working with data, as is usually done in many scientific domains. Assumptions and conclusions must be carefully looked at, to avoid reaching to erroneous conclusions.

A Guide

Part of a guide for choosing the correct statistical test, created by the UCLA's Institute for Digital Research and Education [1], is shown in Table 3.

Table 3: Guide for choosing statistical tests.

Number of dependent variables	Nature of Independent Variables (IVs)	Nature of Dependent Variable(s)	Test(s)
1	Zero IVs (1 population)	interval and normal	one-sample t-test
		ordinal or interval	one-sample median
		categorical (two categories)	binomial test
		categorical	Chi-square goodness-of-fit
1	One IV with two levels (independent groups)	interval and normal	two independent sample t-test
		ordinal or interval	Wilcoxon-Mann Whitney test
		categorical	Chi-Square test, Fisher's exact test
1	One IV with two or more levels (independent groups)	interval and normal	one-way ANOVA
		ordinal or interval	Kruskal-Wallis
		categorical	Chi-square test
1	One IV with two levels (dependent / matched groups)	interval and normal	paired t-test
		ordinal or interval	Wilcoxon signed ranks test
		categorical	McNemar
1	Two or more IVs (independent groups)	interval and normal	factorial ANOVA
		ordinal or interval	ordered logistic regression
		categorical (two categories)	factorial logistic regression
1	One interval IV	interval and normal	correlation, simple linear regression
		ordinal or interval	non-parametric correlation
		categorical	simple logistic regression

References

[1] *Choosing the correct statistical test in SAS, STATA, SPSS and R.* `https://stats.idre.ucla.edu/other/mult-pkg/whatstat/`.

[2] G. Casella and R. L. Berger, *Statistical inference*, Thomson Learning, 2nd ed ed., 2002.

[3] B. Everitt, *The Cambridge dictionary of statistics*, Cambridge University Press, 2nd ed ed., 2002.

[4] Instituto Nacional de Estadística y Geografía, *Transporte Urbano de Pasajeros. Principales características del sistema de transporte colectivo metro de la Ciudad de México.* `https://www.inegi.org.mx/app/tabulados/?nc=100100042`.

[5] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al., *Jupyter notebooks—a publishing format for reproducible computational workflows*, in Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing, IOS Press, 2016, p. 87.

[6] D. Navarro, *Learning statistics with R.* `https://learningstatisticswithr.com/`.

[7] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[8] S. M. Ross, *Introductory statistics*, Academic Press/Elsevier, 3rd ed., 2010.

[9] M. R. Spiegel, R. Hernández Heredero, and L. Abellanas Rapun, *Estadística*, McGraw-Hill, 1991.

Curve fitting

G. Palafox

October 20, 2020

Abstract

Various techniques of fitting curves to data are explored. First, computer generated numbers are used to try the models. Afterwards, real data is used.

1 Introduction

It is a common occurrence when handling data to have only observations of phenomena and not an exact relationship between the variables observed. In order to better understand the subject of study at hand, or to make predictions, it is useful to try and fit a curve to the data observed. In this work, performed on a Jupyter notebook [2] with R version 4.0.0 [5], some techniques for fitting curves to data are employed[1]. First, on computer-generated numbers, and then on real data of vehicles in circulation in Mexico, obtained from INEGI's website [1].

2 Curve fitting

The techniques employed here can be found on Navarro's [4] online book, or in the work of Lane et. al. [3]. To begin this work, two hundred numbers between 100 and 500 are generated uniformly, which are taken as the independent x values. Then, different y values dependent on x are generated, to which Gaussian noise N is added. A fragment of these data can be seen in Table 1, while graphics can be seen in Figure 1. An R function `choose_lambda` was created to compute a λ such that the correlation coefficient of x and $\tilde{y}_\lambda$, where

$$\tilde{y}_\lambda := \begin{cases} y^\lambda, & \text{if } \lambda > 0; \\ \log y, & \text{if } \lambda = 0; \\ -(y^\lambda), & \text{if } \lambda < 0, \end{cases} \tag{1}$$

is maximized. This function is shown in Listing 1.

Listing 1: Function for choosing λ in a Tukey transformation.

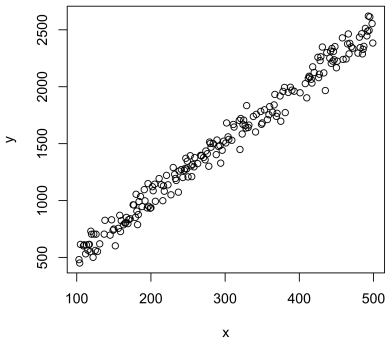
```
choose_lambda <- function(x,y){
if (min(y) < 0){
print("Error. Negative values")
return(NaN)
}

cors <- numeric()
lambdas <- seq(-10, 10, .01)

for (i in lambdas){
if (i == 0)
cors <- c(cors, cor(x, log(y)))
else if (i > 0)
cors <- c(cors, cor(x, y**i))
else
cors <- c(cors, cor(x, -(y**i)))
}

return (lambdas[which.max(cors)])
}
```

For each of the y values generated (see Table 1), the `choose_lambda` function is used to obtain a λ value, and the y values are transformed. Then, using R's `lm` function, a linear regression is performed on the $x, \tilde{y}_\lambda$ values, which gives a linear function $\tilde{y}_\lambda = ax + b$. Finally, an inverse transformation is applied to get a function $y = f(x)$ fitting the original values. The results of this process are plotted in Figure 2.

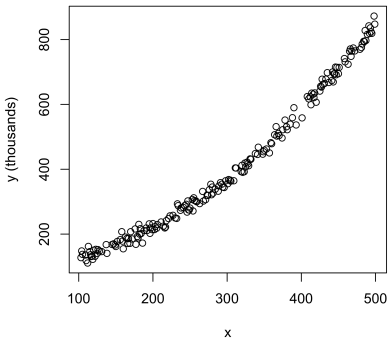
[1]The notebook with the code containing our analysis, as well as this report, can be found in the Github Repository: `https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment7`

Table 1: Fragment of the data generated.

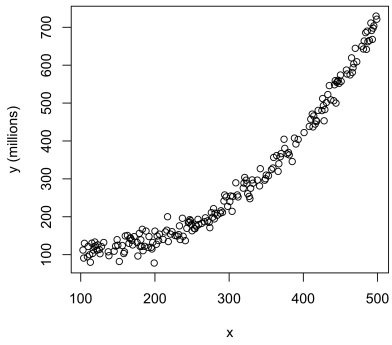
x	$y = 5x + 4 + N$	$y = 3x^2 + 50 + N$	$y = 5x^3 + .4x^2 + 1 + N$	$y = .8\log(x) + 8 + N$	$y = .7\sqrt{x} + 14 + N$
103.00	481.10	127,372.77	112,075,025.75	11.78	21.61
104.00	451.14	148,056.62	90,963,451.34	11.84	23.43
105.00	614.70	137,575.08	129,960,611.26	11.73	23.80
109.00	601.88	135,537.58	94,598,979.05	11.78	19.64
110.00	611.29	118,270.35	121,634,806.71	11.69	21.51
112.00	531.72	110,758.23	99,643,078.53	11.75	21.99



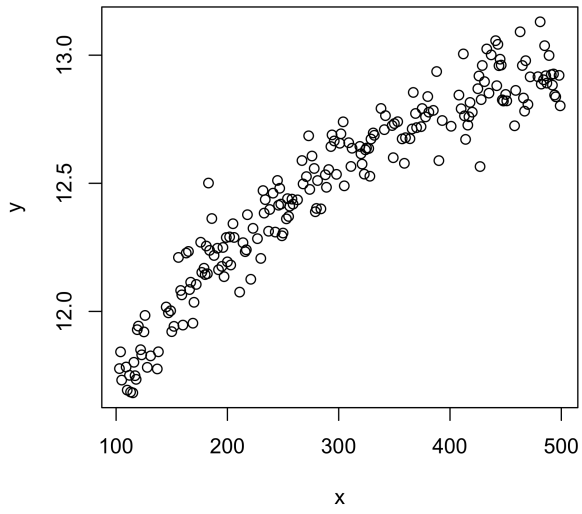
(a) $y = 5x + 4 + N$.



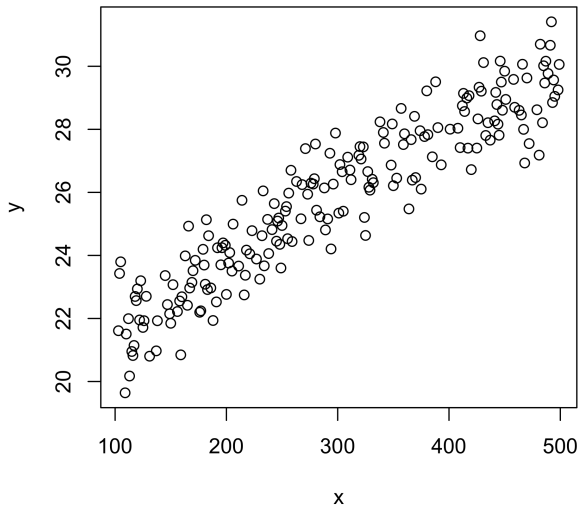
(b) $y = 3x^2 + 50 + N$.



(c) $y = 5x^3 + .4x^2 + 1 + N$.

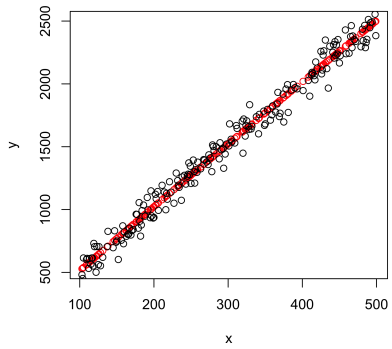


(d) $y = .8\log(x) + 8 + N$.

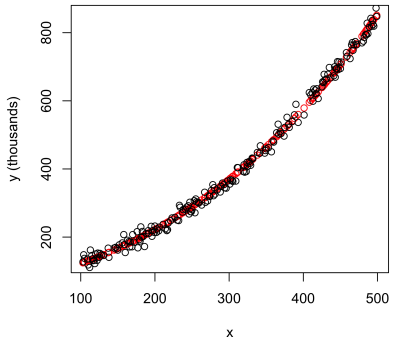


(e) $y = .7\sqrt{x} + 14 + N$.

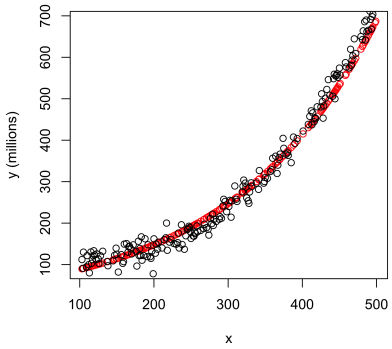
Figure 1: Plots of the data generated.



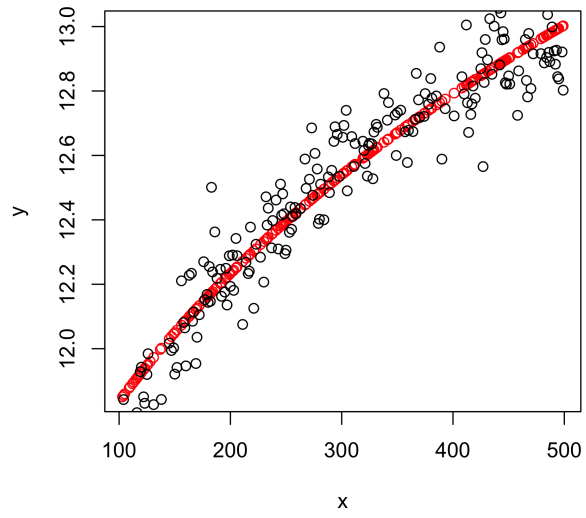
(a) In black, $y = 5x + 4 + N$, and in red, a curve fitted with $\lambda = 1.04$.



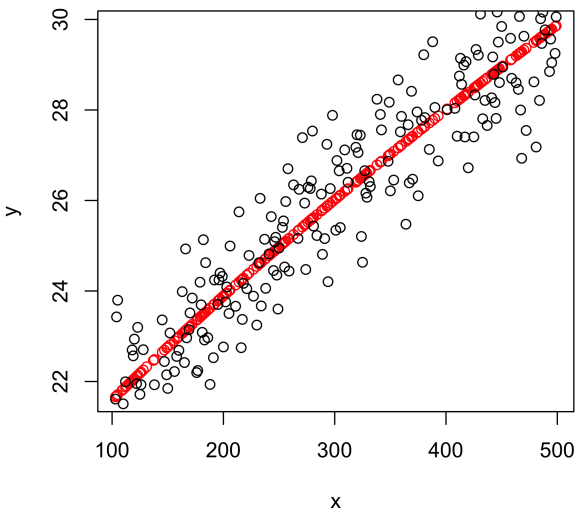
(b) In black, $y = 3x^2 + 50 + N$, and in red, a curve fitted with $\lambda = 0.29$.



(c) In black, $y = 5x^3 + .4x^2 + 1 + N$, and in red, a curve fitted with $\lambda = 0$.



(d) In black, $y = .8 \log(x) + 8 + N$, and in red, a curve fitted with $\lambda = 10$.



(e) In black, $y = .7\sqrt{x} + 14 + N$, and in red, a curve fitted with $\lambda = 1.85$.

Figure 2: Data and curves fitted.

Table 2: Vehicles in circulation in Mexico.

	year	vehicles
1	1981	6,339,836
2	1982	6,695,164
3	1983	6,941,252
4	1984	7,305,066
5	1985	7,725,623
6	1986	7,732,012

The capabilities of R's `lm` function extends to multilinear regression. In order to exemplify this, three different sets of independent variables were created, and a variable dependent on these three was computed. The code where this is done, and its results, are shown in Listing 2.

Listing 2: Multilinear regression with `lm`.

```
x1 <- sample(x = 100:500, size = 200, replace = FALSE)

x2 <- sample(x = 200:600, size = 200, replace = FALSE)

x3 <- sample(x = 100:500, size = 200, replace = FALSE)

my2 <- 3*x1 + .5*log(x2) + 5*x3

lm(my2 ~ x1 + log(x2) + x3)

#Call:
#lm(formula = my2 ~ x1 + log(x2) + x3)

#Coefficients:
#(Intercept)            x1       log(x2)            x3
#2.187e-12     3.000e+00     5.000e-01     5.000e+00
```

2.1 Vehicles in circulation in Mexico

In this subsection, a curve is fitted to real data. The number of vehicles in circulation in Mexico per year, from 1986 to 2019, is downloaded from INEGI's webiste [1]. A fragment of the data can be seen in Table 2. The data is also shown in Figure 3a. Two distinct approaches are taken here. The first approach consists of using the function `choose_lambda`, which gives a value λ to transform the data as was done in Section 2. A value of $\lambda = -0.201$ is obtained. As before, a linear model is fitted for $(x, \tilde{y}_\lambda)$. The function obtained with this method is $y = (0.8393 - 0.0004x)^{-1/0.201}$, and it can be seen in Figure 3b. The second approach consists of assuming the number of vehicles in circulation has exponential growth, and fitting a curve with `lm(log (y) ~ x)`. This gives a model $y = \exp(0.0568 - 97.1138x)$, and can be seen in Figure 3c. This second model is also used to plot a 99% confidence interval for the data, as seen in Figure 3d.

3 Conclusion

The theory and methods of curve fitting is larger than what was presented here. Many other techniques can be applied and studied further, for example, fitting a model on real data with more than one independent variable. Models including non-uniform independent variables can also be studied.

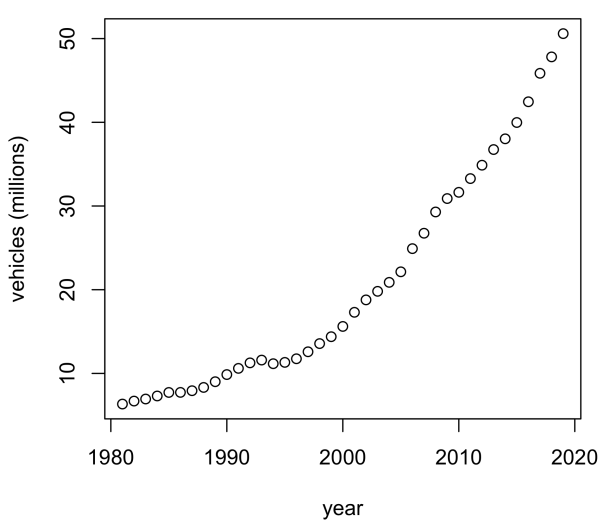
References

[1] Instituto Nacional de Estadística y Geografía, *Total nacional de vehículos. Vehículos de motor registrados en circulación*. `https://www.inegi.org.mx/temas/vehiculos/`.

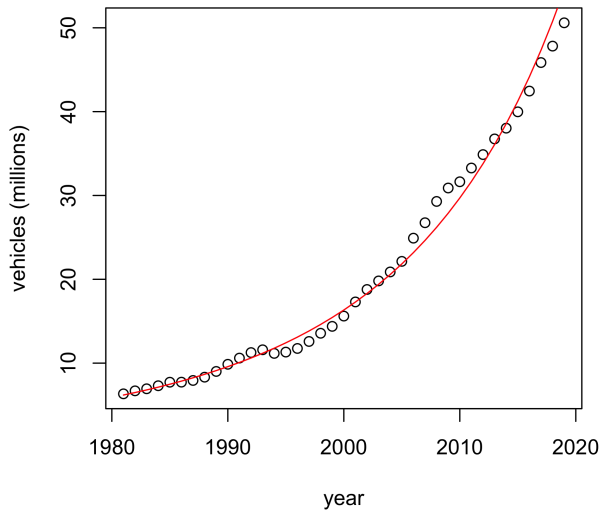
[2] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al., *Jupyter notebooks—a publishing format for reproducible computational workflows*, in Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing, IOS Press, 2016, p. 87.

[3] D. M. Lane, D. Scott, M. Hebl, R. Guerra, D. Osherson, and H. Zimmer, *Introduction to Statistics*, online ed. `http://onlinestatbook.com/Online_Statistics_Education.pdf`.

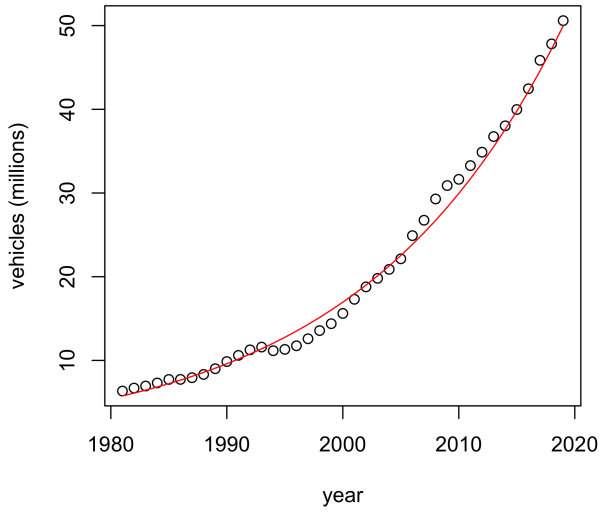
[4] D. Navarro, *Learning statistics with R*. `https://learningstatisticswithr.com/`.



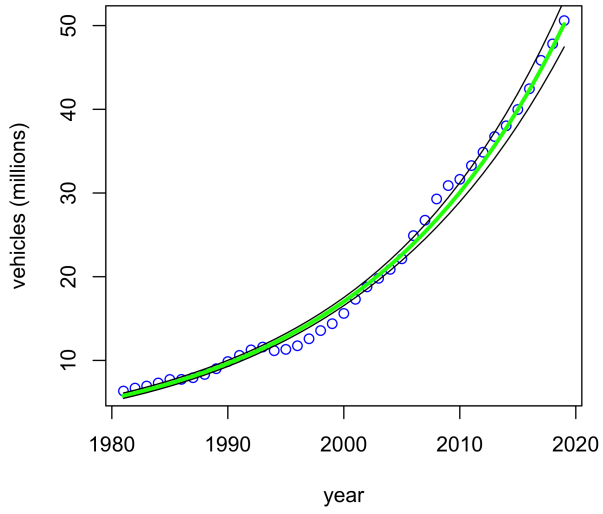
(a) Vehicles in circulation per year.



(b) Vehicle data (black) and $y = (0.8393 - 0.0004x)^{-1/0.201}$ (red).



(c) Vehicle data (black) and $y = \exp(0.0568 - 97.1138x)$ (red).



(d) Vehicle data (blue) and predicted intervals (black and green).

Figure 3: Models for vehicle data.

[5] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020.

Bayes' Theorem

G. Palafox

October 27, 2020

1 Introduction

The use of diagnostic tests is widespread in modern medicine. As Fletcher and Fletcher [2005] mention in their book, *establishing diagnoses is an imperfect process, resulting in a probability rather than a certainty of being right.* Since decisions are made based on the results of these tests, the correct interpretation of their outcome is important, and probability theory can be used to aid with this understanding.

2 SARS-CoV-2

Currently, a strain of coronavirus (SARS-CoV-2, colloquially known as Covid-19) is causing havoc around the world. Diagnostic tests are being used to fight the pandemic, mainly by requiring quarantine for people who test positive, or allowing less movement restrictions for those who test negative (i.e., not quarantining, permitting travel or entrance to places). Given this, it is essential to have a clear understanding of what test results mean. Bayes' theorem, which can be seen in Theorem 1, can help with these interpretations.

Theorem 1 (Bayes' theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\{B_i\}_{i\in\mathbb{N}}$ be a partition of Ω such that $\mathbb{P}(B_i) > 0$ for each i, and let A be any event with positive probability. Then*

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{\infty} \mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}. \tag{1}$$

There are four possible outcomes with any diagnostic test: **true positive**, where a person with the disease tests positive, **false positive**, where a person does not have the disease yet tests positive, **false negative**, where a person with the disease tests negative, and **true negative**, where a person does not have the disease and tests negative. Additionally, with any test, we associate two measurements of how reliable it is: *specificity*, which is the percentage of true negatives out of healthy people, and *sensitivity*, which is the percentage of true positives among people with the disease. Let us denote by $\pm$test the events of having a positive or negative test, and $\pm$cov the events of having or not Covid-19. Using this we can express the sensitivity and specificity of a test as conditional probabilities, namely,

$$\mathbb{P}(+\text{test} \mid +\text{cov}) = \text{sensitivity}, \quad \mathbb{P}(-\text{test} \mid -\text{cov}) = \text{specificity}. \tag{2}$$

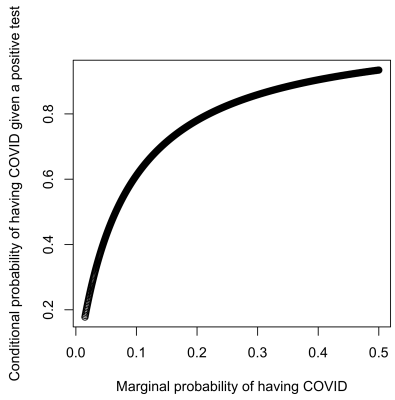
If specificity and sensitivity are know, all is left to know is the marginal probability $\mathbb{P}(+\text{cov})$, and Bayes' theorem can give the probability of having Covid-19 given that a test is positive as

$$\mathbb{P}(+\text{cov} \mid +\text{test}) = \frac{\mathbb{P}(+\text{test} \mid +\text{cov})\ \mathbb{P}(+\text{cov})}{\mathbb{P}(+\text{test} \mid +\text{cov})\ \mathbb{P}(+\text{cov}) + \mathbb{P}(+\text{test} \mid -\text{cov})\ \mathbb{P}(-\text{cov})}. \tag{3}$$

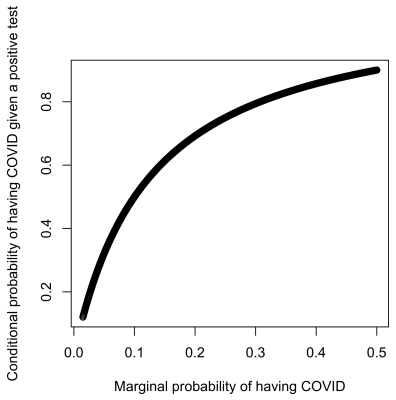
However, two issues arise. The first one is that in the case of Covid-19, sensitivity and specificity are largely unknown for the widely used PCR test [West et al., 2020]. Watson and Whiting [2020] found a sensitivity ranging from 71% to 98%, and use a specificity of 95%. The second issue is the non-trivial calculation of $\mathbb{P}(+\text{cov})$. Ranjan [2020] and Lewis [2020] calculate it as the number of confirmed cases divided by the total population. Schnipper and Sax [2020] consider half the ratio of positive tests in a given population as an estimate of $\mathbb{P}(+\text{cov})$. Others [Ming Chan, 2020, Good et al., 2020] vary $\mathbb{P}(+\text{cov})$, since it can naturally vary depending on who you are testing (random people, people with symptoms, hospital workers, etc.). Personally, the author feels the latter approach gives more insight, since it is adapted easier to different scenarios.

2.1 The case of Nuevo León

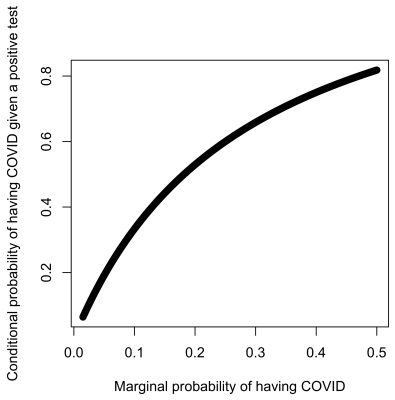
Consider the case of the Mexican state of Nuevo León. As of October 24, 2020, Nuevo Leon had 77807 confirmed Covid cases in a 5.4 million population [Gobierno del Estado de Nuevo León, 2020a, Secretaría de Economía y Trabajo de Nuevo



(a) Sensitivity of 71%, specificity of 95%.



(b) Sensitivity of 90%, specificity of 90%.



(c) Sensitivity of 90%, specificity of 80%.

Figure 1: Marginal vs. conditional probability, varying specificity and sensitivity of tests.

León, 2020]. If $\mathbb{P}(+\text{cov})$ is calculated as confirmed cases over population, that would give $\mathbb{P}(+\text{cov}) = 0.0144$. Assuming a sensitivity of 71% and specificity of 95%, this would give

$$\mathbb{P}(+\text{cov} \mid +\text{test}) = \frac{(.71)(0.0144)}{(.71)(0.0144) + (.05)(.9856)} = 0.17. \tag{4}$$

It may seem counter-intuitive that a positive test gives you only a 17% of having the disease, but this is a consequence of the seemingly low probability of being infected. Consider, on the other hand, that around 40% of tests performed in Nuevo Leon turn out positive [Gobierno del Estado de Nuevo León, 2020b]. If we use half of this value as $\mathbb{P}(+\text{cov})$, Bayes' theorem would give $\mathbb{P}(+\text{cov} \mid +\text{test}) = 0.78$. As it is seen in Figure 1, how widespread the disease is (measured by the marginal probability of being infected) impacts greatly on the interpretation of the test. Specificity and sensitivity, which are not well known for the Covid-19 test, also affect greatly[1].

3 HIV

Unlike the tests for the new coronavirus, tests for HIV have a well-established specificity and sensitivity of over 99% [CDC, 1998]. Of course, the conditional probability when interpreting a positive result still depends on the prevalence of HIV in the community, i.e., in the marginal probability of being HIV positive. However, this allow us to vary only that parameter, fixing the specificity and sensitivity of the test. For reference, considering $\mathbb{P}(+\text{hiv}) = 1.2/327.5 = 0.003$ (cases in the United States over population of the United States, both at the end of 2018 [United States Census Bureau, 2020, CDC, 2020]), would give $\mathbb{P}(+\text{hiv} \mid +\text{test}) = 0.78$, using a sensitivity and specificity of 99%. Varying this marginal probability, with sensitivity and specificity fixed, changes the conditional probability as seen in Figure 2.

References

CDC. Current trends update: Serologic testing for antibody to human immunodeficiency virus. *MMWR*, January 1998. https://www.cdc.gov/mmwr/preview/mmwrhtml/00051681.htm.

CDC. HIV basics. Basic statistics, 2020. https://www.cdc.gov/hiv/basics/statistics.html.

R. H. Fletcher and S. W. Fletcher. *Clinical Epidemiology: The Essentials*. Fourth edition, 2005.

Gobierno del Estado de Nuevo León. Casos de COVID en Nuevo León, 2020a.

Gobierno del Estado de Nuevo León. Monitoreo de indicadores estatales de salud para la reapertura económica semana 42 (11 de octubre – 17 de octubre), 2020b. https://www.nl.gob.mx/presentacion-indicadoresymedidasdemitigacion-covid19-22-10-2020.

C. B. Good, I. Hernandez, and K. Smith. Interpreting COVID-19 test results: a Bayesian approach. *Journal of General Internal Medicine*, 35(8):2490–2491, jun 2020. doi: 10.1007/s11606-020-05918-8. URL https://doi.org/10.1007/s11606-020-05918-8.

[1]The notebook with the code creating these graphics, as well as this report, can be found in the Github Repository: https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment8

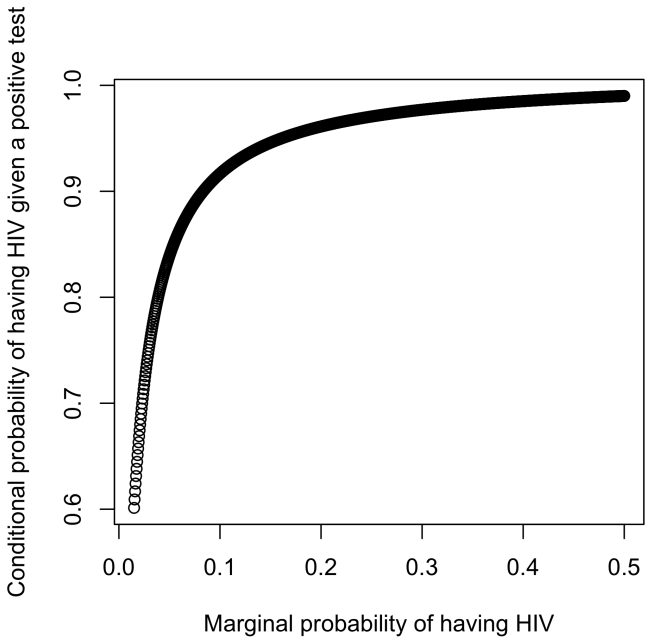


Figure 2: Marginal vs. conditional probability, for a test with specificity and sensitivity of 99%.

M. A. Lewis. Bayes theorem and Covid-19 testing, 2020. https://www.significancemagazine.com/science/660-bayes-theorem-and-covid-19-testing.

G. Ming Chan. Bayes' theorem, COVID19, and screening tests. *The American Journal of Emergency Medicine*, jun 2020. doi: 10.1016/j.ajem.2020.06.054. URL https://doi.org/10.1016/j.ajem.2020.06.054.

A. 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.

J. L. Schnipper and P. E. Sax. Covid-19 test accuracy supplement: the math of Bayes' theorem, 2020. https://www.statnews.com/2020/08/20/covid-19-test-accuracy-supplement-the-math-of-bayes-theorem/.

Secretaría de Economía y Trabajo de Nuevo León. Datos Nuevo León, 2020. http://datos.nl.gob.mx/.

United States Census Bureau. U.S. and world population clock, 2020. https://www.census.gov/popclock/.

J. Watson and P. F. Whiting. Interpreting a covid-19 test result. *The BMJ*, May 2020. doi: 10.1136/bmj.m1808. URL https://www.bmj.com/content/bmj/369/bmj.m1808.full.pdf.

C. P. West, V. M. Montori, and P. Sampathkumar. Covid-19 testing. *Mayo Clinic Proceedings*, 95(6):1127–1129, Jun 2020. ISSN 00256196. doi: 10.1016/j.mayocp.2020.04.004.

Exercises

G. Palafox

November 4, 2020

The following are exercises from the book of Grinstead and Snell [2006].

Exercise 1 (Ex. 1, p. 247)**.** *A card is drawn at random from a deck consisting of cards numbered 2 through 10. A player wins 1 dollar if the number on the card is odd and loses 1 dollar if the number is even. What is the expected value of his winnings?*

Solution. We have $\Omega = \{2, 3, \ldots, 9, 10\}$ and a random variable

$$X(\omega) = \begin{cases} 1 & \omega \text{ is odd;} \\ -1 & \omega \text{ is even.} \end{cases} \tag{1}$$

We also have $\mathbb{P}(\omega) = 1/9$ for any card in the deck, and have exactly four odd numbers and five even numbers on said deck. Therefore, the expected value of his winnings is

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(w)\mathbb{P}(\omega) = (-1)(1/9) + (1/9) + \cdots + (1/9) = 4(1/9) - 5(1/9) = -1/9. \tag{2}$$

□

Exercise 2 (Ex. 6, p. 247)**.** *A die is rolled twice. Let X denote the sum of the two numbers that turn up, and Y the difference of the numbers (specifically, the number on the first roll minus the number on the second). Show that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. Are X and Y independent?*

Solution. Let $\Omega = \{1, 2, \ldots, 6\} \times \{1, 2, \ldots, 6\}$, and $X(\omega_1, \omega_2) = \omega_1 + \omega_2, Y(\omega_1, \omega_2) = \omega_1 - \omega_2$. Define a new random variable $Z = XY$, that is, $Z(\omega_1, \omega_2) = (\omega_1 + \omega_2)(\omega_1 - \omega_2) = \omega_1^2 - \omega_2^2$. Now, we have

$$\begin{aligned}
\mathbb{E}(XY) = \mathbb{E}(Z) &= \sum_{(\omega_1, \omega_2) \in \Omega} \frac{1}{36}(\omega_1^2 - \omega_2^2) && (3) \\
&= \frac{1}{36}\left((1^2 - 1^2) + (1^2 - 2^2) + \cdots + (2^2 - 1^2) + \cdots + (6^2 - 1^2) + \ldots (6^2 - 6^2)\right) && (4) \\
&= \frac{1}{36}\left[\left(6(1^2) - \sum_{i=1}^{6} i^2\right) + \left(6(2^2) - \sum_{i=1}^{6} i^2\right) + \cdots + \left(6(6^2) - \sum_{i=1}^{6} i^2\right)\right] && (5) \\
&= \frac{1}{36}\left(6(1^2 + \cdots + 6^2) - 6\sum_{i=1}^{6} i^2\right) && (6) \\
&= \frac{1}{36}\left(6\sum_{i=1}^{6} i^2 - 6\sum_{i=1}^{6} i^2\right) && (7) \\
&= 0. && (8)
\end{aligned}$$

k	$\mathbb{P}(X=k)$
0	$\frac{1}{6}$
1	$\frac{5}{6}\cdot\frac{1}{5}=\frac{1}{6}$
2	$\frac{5}{6}\cdot\frac{4}{5}\cdot\frac{1}{4}=\frac{1}{6}$
3	$\frac{5}{6}\cdot\frac{4}{5}\cdot\frac{3}{4}\cdot\frac{1}{3}=\frac{1}{6}$
4	$\frac{5}{6}\cdot\frac{4}{5}\cdot\frac{3}{4}\cdot\frac{2}{3}\cdot\frac{1}{2}=\frac{1}{6}$
5	$\frac{5}{6}\cdot\frac{4}{5}\cdot\frac{3}{4}\cdot\frac{2}{3}\cdot\frac{1}{2}\cdot 1=\frac{1}{6}$

Table 1: Probabilities for Exercise 3

With a similar calculation, we can see that

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{(\omega_1,\omega_2)\in\Omega} \frac{1}{36}(\omega_1-\omega_2) && (9)\\
&= \frac{1}{36}\left((1-1)+(1-2)+\cdots+(2-1)+\cdots+(6-1)+\ldots(6-6)\right) && (10)\\
&= \frac{1}{36}\left[\left(6(1)-\sum_{i=1}^{6} i\right)+\left(6(2)-\sum_{i=1}^{6} i\right)+\ \cdots+\left(6(6)-\sum_{i=1}^{6} i\right)\right] && (11)\\
&= \frac{1}{36}\left(6(1+\cdots+6)-6\sum_{i=1}^{6} i\right) && (12)\\
&= \frac{1}{36}\left(6\sum_{i=1}^{6} i-6\sum_{i=1}^{6} i\right) && (13)\\
&= 0. && (14)
\end{aligned}$$

With this we can see that, regardless of the value of $\mathbb{E}(X)$ we will have

$$\mathbb{E}(X)\mathbb{E}(Y)=\mathbb{E}(X)\cdot 0=0=\mathbb{E}(Z)=\mathbb{E}(XY). \tag{15}$$

The random variables are not independent. To see this, consider $\mathbb{P}(X=2,Y=0)$. There is only one pair of $(\omega_1,\omega_2)\in\Omega$ such that $\omega_1+\omega_2=2$ and $\omega_1-\omega_2=0$, namely $(1,1)$, so $\mathbb{P}(X=2,Y=0)=1/36$. In fact, $\mathbb{P}(X=2)=1/36$ too. However, $\mathbb{P}(Y=0)=6/36$, for every pair (ω_1,ω_2) with $\omega_1=\omega_2$ satisfies, so

$$\mathbb{P}(X=2)\ \mathbb{P}(Y=0)=(1/36)(6/36)\neq\mathbb{P}(X=2,Y=0)=1/36. \tag{16}$$

□

Exercise 3 (Ex. 18, p. 249)**.** *Exactly one of six similar keys opens a certain door. If you try the keys, one after another, what is the expected number of keys that you will have to try before success?*

Solution. Lets denote by 0 choosing a wrong key, and by 1 choosing the correct key. Then

$$\Omega=\{1,01,001,0001,00001,000001\}, \tag{17}$$

and $X(\omega)=$ *number of zeros in* ω. Assuming the keys are chosen uniformly at random, and that once a key is tried we discard it (i.e., *no replacement*), the probabilities of trying k keys before success are as seen in Table 1.

Thus, the expected number of keys to try before success is

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{5}(1/6)k && (18)\\
&= \frac{1}{6}\sum_{k=0}^{5} k && (19)\\
&= \frac{1}{6}\cdot 15 && (20)\\
&= 2.5 && (21)
\end{aligned}$$

□

Exercise 4 (Ex. 19, p. 249)**.** *A multiple choice exam is given. A problem has four possible answers, and exactly one answer is correct. The student is allowed to choose a subset of the four possible answers as his answer. If his chosen subset contains the correct answer, the student receives three points, but he loses one point for each wrong answer in his chosen subset. Show that if he just guesses a subset uniformly and randomly his expected score is zero.*

Solution. Let A be the four-element-set of possible answers for the question. Then our sample space is the power set $\Omega = 2^A$, and $X(\omega) =$ *points received by choosing the subset* ω. Denote by $|\omega|$ the cardinality of a subset $\omega \in \Omega$. Then we have

$$\mathbb{E}(X) = \sum_{\omega\in\Omega} \frac{1}{16}\left(\frac{|w|}{4}\left(3-(|\omega|-1)\right) - \left(1-\frac{|\omega|}{4}\right)|\omega|\right). \tag{22}$$

To explain Equation 22, see that the probability of a subset ω containing the right answer is $|\omega|/4$, in which case, the student would get $3-(|\omega|-1)$ points, three for the right answer and minus one for each of the $|\omega|-1$ wrong answers. On the other hand, the probability of a subset not containing the answer is $1-\frac{|\omega|}{4}$ and the student would lose $|\omega|$ points. Observe that

$$\frac{|w|}{4}\left(3-(|\omega|-1)\right) - \left(1-\frac{|\omega|}{4}\right)|\omega| = \frac{|w|}{4}\left(3-|\omega|+1\right) - \left(\omega - \frac{|\omega|^2}{4}\right) \tag{23}$$

$$= \frac{3|\omega|}{4} - \frac{|\omega|^2}{4} + \frac{|\omega|}{4} - |\omega| + \frac{|\omega|^2}{4} \tag{24}$$

$$= 0. \tag{25}$$

Therefore, the expected score for the student is

$$\mathbb{E}(X) = \sum_{\omega\in\Omega}\frac{1}{16}(0) = 0. \tag{26}$$

□

Exercise 5 (Ex. 31, p. 254)**.** *A large number, N, of people are subjected to a blood test. This can be administered in two ways: (1) Each person can be tested separately, in this case N tests are required, (2) the blood samples of k persons can be pooled and analyzed together. If this test is negative, this one test suffices for the k people. If the test is positive, each of the k persons must be tested separately, and in all, $k+1$ tests are required for the k people. Assume that the probability p that a test is positive is the same for all people and that these events are independent.*

1. *Find the probability that the test for a pooled sample of k people will be positive.*
2. *What is the expected value of the number X of tests necessary under plan (2)? (Assume that N is divisible by k.)*
3. *For small p, show that the value of k which will minimize the expected number of tests under the second plan is approximately $1/\sqrt{p}$.*

Solution. Since each of the k persons have a probability p of being positive, the probability that at least one of them is positive (and hence, the pool) is $p+p+\cdots+p = kp$.

Under plan (2), we have an expected number of tests of

$$\mathbb{E}(X) = \sum_{i=1}^{N/k} kp(k+1) + (1-kp) \tag{27}$$

$$= (N/k)[kp(k+1) + (1-kp)] \tag{28}$$

$$= Np(k+1) + N/k - Np \tag{29}$$

$$= Npk + N/k. \tag{30}$$

To minimize this expected value with a fixed N and p, we consider the expected value as a function of k, $f(k) = Npk + N/k$. The function has derivative $f'(k) = Np - \frac{N}{k^2}$, which is zero in $k = 1/\sqrt{p}$, and second derivative $f''(k) = N/k^3$, which is positive for $k = 1/\sqrt{p}$, so $k = 1/\sqrt{p}$ minimizes the expected value. For this value to make sense, k must be greater than two, so p must be small. □

Exercise 6 (Ex. 1, p. 263)**.** *A number is chosen at random from the set $S = \{-1, 0, 1\}$. Let X be the number chosen. Find the expected value, variance and standard deviation of X.*

Solution. We have

$$\mathbb{E}(X) = (1/3)(-1) + (1/3)(0) + (1/3)(1) = 0, \tag{31}$$

so

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}(X^2) = (1/3)(1) + (1/3)(0) + (1/3)(1) = 2/3, \tag{32}$$

and $D(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{2/3}$. □

Exercise 7 (Ex. 9, p. 264)**.** *A die is loaded so that the probability of a face coming up is proportional to the number on that face. The die is rolled with outcome X. Find $\mathrm{Var}(X)$ and $D(X)$.*

Solution. Since $1+2+\cdots+6=21$,, we make $\mathbb{P}(X=k)=k/21$. With this,

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=1}^{6} k\mathbb{P}(X=k) && (33)\\
&= \sum_{k=1}^{6} k(k/21) && (34)\\
&= (1/21)\sum_{k=1}^{6} k^2 && (35)\\
&= (1/21)\frac{6\cdot 7\cdot 13}{6} && (36)\\
&= (1/21)(91) && (37)\\
&= 13/3 && (38)\\
&\approx 4.333, && (39)
\end{aligned}$$

so

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{k=1}^{6}[(k-13/3)^2(k/21)] && (40)\\
&= \sum_{k=1}^{6}[(k^2-2k(13/3)+(13/3)^2)(k/21)] && (41)\\
&= \sum_{k=1}^{6}[\frac{k^3}{21}-2(\frac{13}{3\cdot 21})k^2+(\frac{13}{3})^2\frac{k}{21}] && (42)\\
&= \frac{1}{21}\sum_{k=1}^{6}k^3-2(\frac{13}{3\cdot 21})\sum_{k=1}^{6}k^2+\frac{1}{21}(\frac{13}{3})^2\sum_{k=1}^{6}k && (43)\\
&= \frac{1}{21}(\frac{36\cdot 49}{4})-\frac{2\cdot 13}{3\cdot 21}(91)+\frac{1}{21}(13/3)^2(21) && (44)\\
&= 9(7/3)-2(13/3)^2+(13/3)^2 && (45)\\
&= 20/9 && (46)\\
&\approx 2.222, && (47)
\end{aligned}$$

and $D(X)=\sqrt{\mathrm{Var}(X)}=\sqrt{20/9}=\sqrt{20}/3\approx 1.49$. □

Exercise 8 (Ex. 12, p. 264)**.** *Let X be a random variable with $\mu=\mathbb{E}(X)$ and $\sigma^2=\mathrm{Var}(X)$. Define $X^*=(X-\mu)/\sigma$. The random variable X^* is called the standardized random variable associated with X. Show that this standardized random variable has expected value 0 and variance 1.*

Solution. Theorem 6.2 from Grinstead and Snell [2006] tells us that, for random variables X, Y with finite expected values, we have

$$\mathbb{E}(X+Y)=\mathbb{E}(X)+\mathbb{E}(Y), \tag{48}$$

and if c is any constant, then

$$\mathbb{E}(cX)=c\mathbb{E}(X). \tag{49}$$

Therefore,

$$\begin{aligned}
\mathbb{E}(X^*) &= \mathbb{E}[(X-\mu)/\sigma] && (50)\\
&= \frac{1}{\sigma}\mathbb{E}(X-\mu) && (51)\\
&= \frac{1}{\sigma}(\mathbb{E}(X)-\mathbb{E}(\mu)) && (52)\\
&= \frac{1}{\sigma}(\mu-\mu) && (53)\\
&= 0, && (54)
\end{aligned}$$

and, writing μ^* for $\mathbb{E}(X^*)$,

$$\begin{aligned}
\mathrm{Var}(X^*) &= \mathbb{E}[(X^* - \mu^*)^2] && (55)\\
&= \mathbb{E}[(X^*)^2] && (56)\\
&= \mathbb{E}[\frac{(X-\mu)^2}{\sigma^2}] && (57)\\
&= \frac{1}{\sigma^2}\mathbb{E}[(X-\mu)^2] && (58)\\
&= \frac{1}{\sigma^2}\mathrm{Var}(X) && (59)\\
&= \frac{1}{\sigma^2}\sigma^2 = 1. && (60)
\end{aligned}$$

□

Exercise 9 (Ex. 3, p. 278)**.** *The lifetime, measured in hours, of the ACME super light bulb is a random variable T with density function $f_T(t) = \lambda^2 t e^{-\lambda t}$, where $\lambda = 0.05$. What is the expected lifetime of this light bulb? What is its variance?*

Solution. The expected lifetime will be given by

$$\mathbb{E}(T) = \int_0^\infty t(\lambda^2 t e^{-\lambda t})dt. \tag{61}$$

First, the indefinite integral is solved. This is rewritten as

$$\int t(\lambda^2 t e^{-\lambda t})\,dt = (-\lambda)\int t^2(-\lambda e^{-\lambda t})\,dt, \tag{62}$$

and integrating by parts gives

$$\int t^2(-\lambda e^{-\lambda t})\,dt = t^2 e^{-\lambda t} - \int (2t)e^{-\lambda t}dt. \tag{63}$$

Substituting Equation 63 in 62 we obtain:

$$\int t(\lambda^2 t e^{-\lambda t})\,dt = -\lambda t^2 e^{-\lambda t} - 2\int t(-\lambda e^{-\lambda t})\,dt. \tag{64}$$

The integral in Equation 64 can be done via integration parts too, as

$$\int t(-\lambda e^{-\lambda t})\,dt = te^{\lambda t} - \int e^{-\lambda t}\,dt = te^{-\lambda t} + \frac{1}{\lambda}e^{-\lambda t}. \tag{65}$$

Putting the result of Equation 65 in 64 gives

$$\int t(\lambda^2 t e^{-\lambda t})\,dt = -\lambda t^2 e^{-\lambda t} - 2te^{-\lambda t} - \frac{2}{\lambda}e^{-\lambda t}. \tag{66}$$

Denoting the resulting function of Equation 66 as $F(t)$, we have

$$\mathbb{E}(T) = \lim_{t\longrightarrow\infty} F(t) - F(0) = \frac{2}{\lambda}. \tag{67}$$

Therefore, the **expected lifetime of this light bulb** is $2/0.05 = 40$ hours. To find the variance, first $\mathbb{E}(T^2)$ will be obtained. This is equal to

$$\int_0^\infty t^2(\lambda t e^{-\lambda t})\,dt = \int_0^\infty \lambda^2 t^3 e^{-\lambda t}\;\,dt. \tag{68}$$

As before, the indefinite integral is calculated first. First, it is observed that

$$\int \lambda^2 t^3 e^{-\lambda t}\;\,dt = -\lambda\int t^3(-\lambda e^{-\lambda t})\,dt, \tag{69}$$

and integrating by parts gives

$$\int t^3(-\lambda e^{-\lambda t})\,dt = t^3 e^{-\lambda t} - \int 3t^2 e^{-\lambda t}\,dt. \tag{70}$$

Therefore, substituting the result of Equation 70 in Equation 69, we obtain

$$\int \lambda^2 t^3 e^{-\lambda t}dt = -\lambda t^3 e^{-\lambda t} + 3\int \lambda t^2 e^{-\lambda t}\,dt. \tag{71}$$

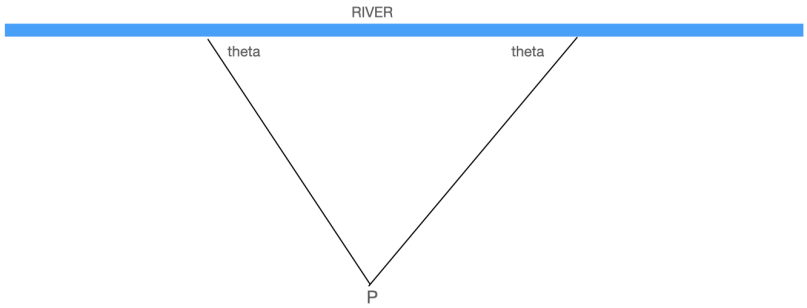


Figure 1: Figure for Exercise 10

From Equation 66, we know

$$\int \lambda t^2 e^{-\lambda t}\, dt = \frac{1}{\lambda}\int \lambda^2 t^2 e^{-\lambda t}\, dt = \frac{1}{\lambda}F(t), \tag{72}$$

thus

$$\int \lambda^2 t^3 e^{-\lambda t}\, dt = -\lambda t^3 e^{-\lambda t} + \frac{3}{\lambda}F(t), \tag{73}$$

and

$$\mathbb{E}(T^2) = \lim_{t\longrightarrow\infty}(-\lambda t^3 e^{-\lambda t} + \frac{3}{\lambda}F(t) + \frac{3}{\lambda}\frac{2}{\lambda}) = \frac{6}{\lambda^2}. \tag{74}$$

Now, by Theorem 6.15 [Grinstead and Snell, 2006], $\mathrm{Var}(T) = \mathbb{E}(T^2) - \mathbb{E}(T)^2$, so we get $\mathrm{Var}(T) = \frac{6}{\lambda^2} - (\frac{2}{\lambda})^2$. In this particular example, $\lambda = 0.05$, which gives a variance $\mathrm{Var}(T) = 2400 - 1600 = 800$. □

Exercise 10 (Ex. 26, p. 284)**.** *Suppose you are standing on the bank of a straight river.*

1. *Choose, at random, a direction which will keep you on dry land, and walk 1 km in that direction. Let P denote your position. What is the expected distance from P to the river?*
2. *Now suppose you proceed as in part (1), but when you get to P, you pick a random direction (from among all directions) and walk 1 km. What is the probability that you will reach the river before the second walk is completed?*

Solution. In part 1, the possible directions to stay on dry land are given by $\theta, 0 \leq \theta \leq \pi$. By basic trigonometry, the distance between the river and the point P resulting from walking 1 km in a θ-radian direction is $\sin\theta$. Since this θ is chosen uniformly at random, the expected distance is

$$\int_0^\pi \frac{1}{\pi}\sin(\theta)\, d\theta = \frac{2}{\pi}. \tag{75}$$

In part 2, we need to know in which directions is the river less than 1 km away. These are the directions inside the triangle in Figure 1, which forms a $\pi - 2\theta$ radians angle. The probability of choosing any of these directions is $\frac{\pi-2\theta}{2\pi} = \frac{1}{2} - \frac{\theta}{\pi}$. □

Exercise 11 (Ex. 27, p. 284)**.** *A game is played as follows: A random number X is chosen uniformly from $[0,1]$. Then a sequence $Y_1, Y_2, \ldots$ of random numbers is chosen independently and uniformly from $[0,1]$. The game ends the first time that $Y_i > X$. You are then paid $(i-1)$ dollars. What is a fair entrance fee for this game?*

Solution. First, let's establish that if Z is the payout from playing the game, we will consider $E(Z)$ a fair entrance fee, since anything larger than $E(Z)$ would be disadvantageous to the player, and anything less than $E(Z)$ would be unfair to whomever is running the game. Now, to calculate the expected gain, see that for any Y_i, $\mathbb{P}(Y_i < X) = X$ and

$\mathbb{P}(Y_i \geq X) = 1 - X$. Hence, the probability of ending the game in k steps would be $X^{k-1}(1-X)$, therefore

$$E(Z) = \sum_{k=1}^{\infty}(k-1)X^{k-1}(1-X) \tag{76}$$

$$= \sum_{k=0}^{\infty} kX^k(1-X) \tag{77}$$

$$= \sum_{k=0}^{\infty} kX^k - \sum_{k=0}^{\infty} kX^{k+1} \tag{78}$$

$$= \frac{X}{(X-1)^2} - \frac{X^2}{(X-1)^2} \tag{79}$$

$$= \frac{X - X^2}{(X-1)^2} \tag{80}$$

$$= \frac{-X}{X-1} = \frac{X}{1-X}. \tag{81}$$

Since $\mathbb{E}(X) = 1/2$, then $E(Z) = \frac{1/2}{1-1/2} = 1$ would be a fair entry fee. □

References

C. M. Grinstead and J. L. Snell. *Introduction to Probability*. 2006.

Computer simulated exercises

G. Palafox

November 9, 2020

In the following, exercises from the book of Grinstead and Snell [2006] are simulated computationally on a Jupyter [Kluyver et al., 2016] notebook[1] with R [R Core Team, 2020].

Exercise 1 (Ex. 19, p. 249)**.** *A multiple choice exam is given. A problem has four possible answers, and exactly one answer is correct. The student is allowed to choose a subset of the four possible answers as his answer. If his chosen subset contains the correct answer, the student receives three points, but he loses one point for each wrong answer in his chosen subset. Show that if he just guesses a subset uniformly and randomly his expected score is zero.*

Computer simulation. The exercise is simulated as follows. A vector of possible answers, `a, b, c, d` is created. One of these is randomly selected as the correct answer. Then a random subset is taken from the possible answers, and a score is computed as per the rules laid out in the exercise. This is repeated 10,000 times to calculate the average score. The computation was done using the code in Listing 1. On average, the score obtained is -0.0129, which is close to zero. A boxplot of the scores can be seen in Figure 1.

Listing 1: Code for Exercise 1.

```
answers = c('a', 'b', 'c', 'd')

scores <- numeric()

for(i in 1:10000){

correct_answer <- sample(answers, 1)

answerset <- sample(answers, sample(c(1,2,3,4), 1))

if (correct_answer %in% answerset){
  score <- 3 - (length(answerset)-1)
}else{
  score <- -length(answerset)
}

scores <- c(scores, score)
}
```

[1]The notebook with the code for the experiments, as well as this report, can be found in the Github Repository: https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment10

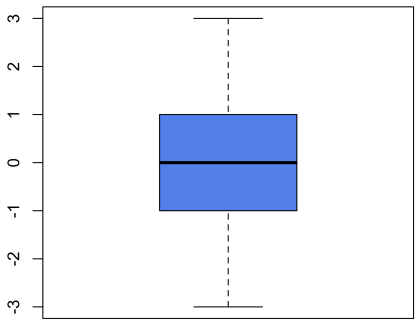


Figure 1: Scores obtained in experiment for Exercise 1.

Exercise 2 (Ex. 12, p. 264). *Let X be a random variable with $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$. Define $X^* = (X - \mu)/\sigma$. The random variable X^* is called the standardized random variable associated with X. Show that this standardized random variable has expected value 0 and variance 1.*

Computer simulation. A thousand numbers following some specific distribution are generated. Their mean and variance is calculated. Then, the numbers are *standardized* as the exercise suggests, and the mean and variance are re-calculated. The mean and variance of the original and standardized values are stored, and this is repeated a thousand times. The code in Listing 2 shows a function for standardizing a vector, and the experiment performed for uniformly distributed numbers. The experiment was also performed for normal (mean 1, standard deviation 0.5) distributed numbers, and for exponential distributed numbers with rate 10. It can be seen in Figure 2 that the mean always shifts to zero, and variance always shifts to one, as expected.

Listing 2: Code for Exercise 2.

```
standardize <- function(vector){
  return ((vector - mean(vector))/sqrt(var(vector)))
}

means_unif <- numeric()
means_unif_std <- numeric()

var_unif <- numeric()
var_unif_std <- numeric()

for(i in 1:1000){
  x <- runif(1000)

  means_unif <- c(means_unif, mean(x))
  means_unif_std <- c(means_unif_std, mean(standardize(x)))

  var_unif <- c(var_unif, var(x))
  var_unif_std <- c(var_unif_std, var(standardize(x)))

}
```

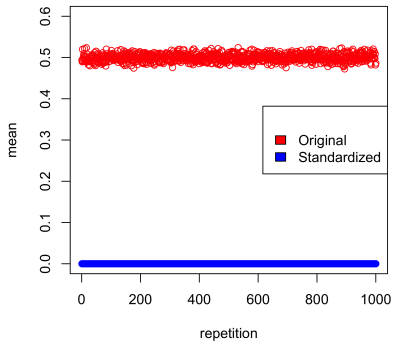
Exercise 3 (Ex. 3, p. 278). *The lifetime, measured in hours, of the ACME super light bulb is a random variable T with density function $f_T(t) = \lambda^2 t e^{-\lambda t}$, where $\lambda = 0.05$. What is the expected lifetime of this light bulb? What is its variance?*

Computer simulation. Analytically, it can be shown the light bulb has an expected lifetime of 40 hours, with a variance of 800. In order to simulate the lifetime of the light bulb, a thousand random numbers with density $f_T(t) = \lambda^2 t e^{-\lambda t}$ were generated, and its mean and variance computed. This was repeated a thousand times, storing the mean and variance in each repetition. On average, the mean was 40.138 and the variance 798.26. The computations were performed with the code in Listings 3. The mean and variance obtained in each repetition are shown in the boxplot in Figure 3.

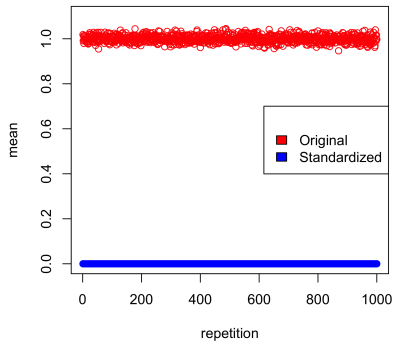
Listing 3: Code for Exercise 3.

```
library(distr)
p    <- function(x){ (0.05**2) * (x*exp(-0.05*x)) } # probability density function
dist <-AbscontDistribution(d=p)  # signature for a dist with pdf ~ p
rdist <- r(dist)                 # function to create random variates from p

mean_x <- numeric()
var_x <- numeric()

for (i in 1:1000){
  X <- rdist(1000)
  mean_x <- c(mean_x, mean(X))
  var_x <- c(var_x, var(X))
}
```

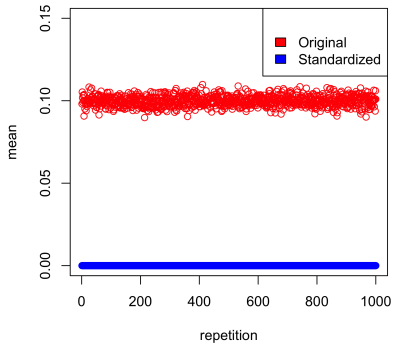
Exercise 4 (Ex. 1, p. 247). *A card is drawn at random from a deck consisting of cards numbered 2 through 10. A player wins 1 dollar if the number on the card is odd and loses 1 dollar if the number is even. What is the expected value of his winnings?*



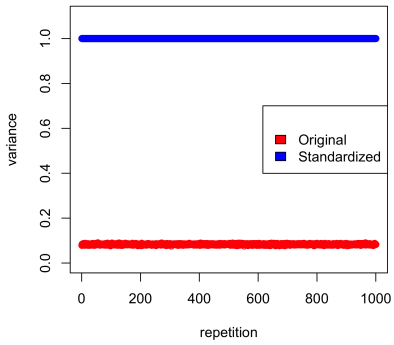
(a) Mean, with original numbers uniformly distributed.



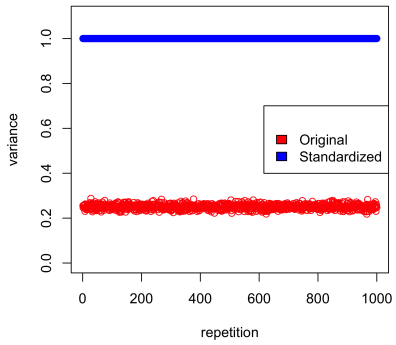
(b) Mean, with original numbers normally distributed.



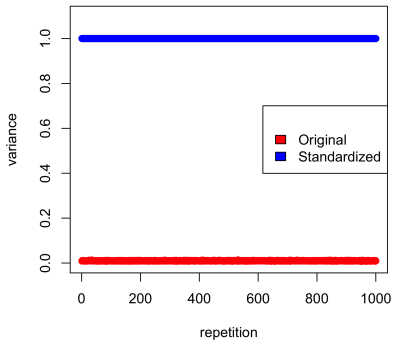
(c) Mean, with original numbers exponentially distributed.



(d) Variance, with original numbers uniformly distributed.

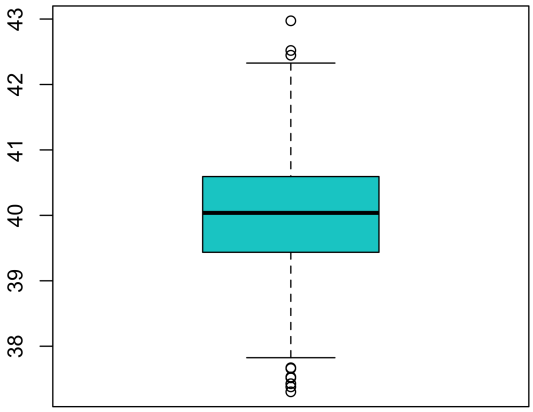


(e) Variance, with original numbers normally distributed.

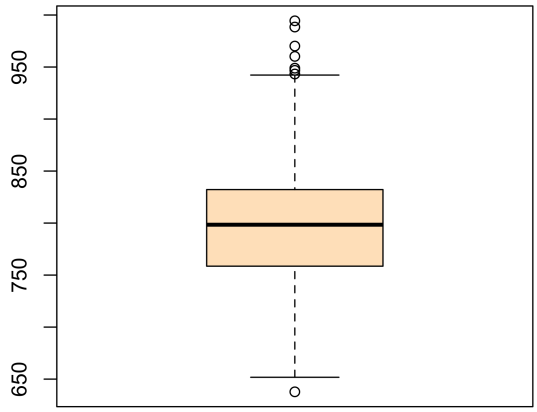


(f) Variance, with original numbers exponentially distributed.

Figure 2: Mean and variance before and after standardizing, Exercise 2.



(a) Mean in repetitions.



(b) Variance in repetitions.

Figure 3: Mean and variance in the experiment repetitions for Exercise 3.

Computer simulation. We simulate 50,000 repetitions of the experiment, drawing a random card and storing the corresponding score. A mean of -0.113 is obtained from this experiment, while analytically the expected value is $-1/9 \approx -0.111$. The code used for the simulation is in Listing 4.

Listing 4: Code for Exercise 4.

```
cards = c(2, 3, 4, 5, 6, 7, 8, 9, 10)

winnings <- numeric()

for (i in 1:50000){
  drawn_card <- sample(cards,1)
  winnings <- c(winnings, 2*(drawn_card %% 2) - 1)
}
```

References

C. M. Grinstead and J. L. Snell. *Introduction to Probability*. 2006.

T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al. Jupyter notebooks—a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing*, page 87. IOS Press, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

Convolutions, variance and covariance

G. Palafox

November 15, 2020

1 Convolutions

A convolution is, in the more general sense, an operator between two functions which produces a third function. In this section, an application of convolutions of probability functions on finite groups is explored. In particular, convolutions are used to obtain the distribution of a random walk on a finite group. After the basic theory is presented, a practical example is given. Most of what is discussed in this section, and further topics regarding random walks on finite groups, can be found in the books of Steinberg [2012] or Diaconis [1988].

1.1 Random walks on finite groups

Let G be a finite group [Herstein, 1975]. A probability on G is a function $P : G \longrightarrow [0, 1]$ such that

$$\sum_{g \in G} P(g) = 1. \tag{1}$$

For a subset $A \subseteq G$, one defines $P(A) := \sum_{g \in A} P(g)$. Now, suppose P, Q are probabilities on G, and $X \sim P$, $Y \sim Q$ are chosen independently at random. What is the probability of $XY = g$ for some g? If $Y = h$, it must be that $X = gh^{-1}$ for XY to occur. Because of independence, the probability of this happening is $P(gh^{-1})Q(h)$. Summing the probabilities over all possible h gives us the probability of $XY = g$ equal to

$$\sum_{h \in G} P(gh^{-1})Q(h) = P * Q(g), \tag{2}$$

where $*$ denotes the convolution operator. Thus, if X, Y are independent and $X \sim P, Y \sim Q$, it follows that $XY \sim P * Q$. This in turn can be used to model a random walk on G as follows. Starting at the identity e of G, a walker chooses $X_1 \in G$ at random according to a probability P, and moves to X_1. Then, the walker chooses X_2 according to P and moves to X_2X_1. Following this process, the walker is choosing a sequence of independent, identically distributed random variables $X_1, X_2, \ldots$ with common distribution P, landing at $X_kX_{k-1} \cdots X_1$ at step k. Let

$$\delta_g(h) = \begin{cases} 1 & \text{if } h = g; \\ 0 & \text{if } h \neq g. \end{cases} \tag{3}$$

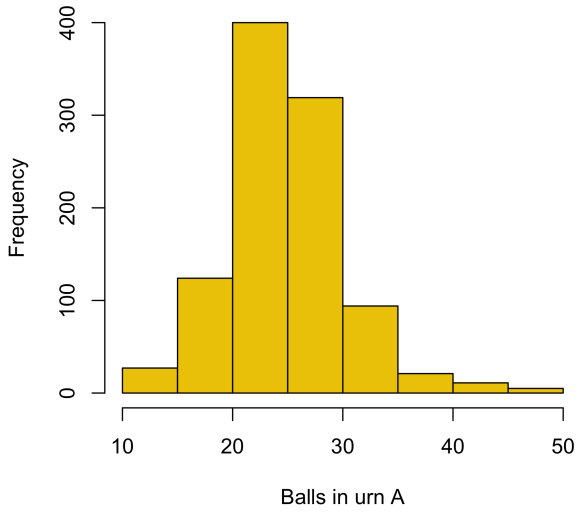
With this notation, one can let $Y_0 \sim \delta_e$ (so $Y_0 = e$ with probability 1), and $Y_k = X_kY_{k-1}$ for $k \geq 1$. The random variable Y_k gives the position of the walker in the k-th step, and by the preceding discussion, $Y_k \sim P^{*k}$, where P^{*k} denotes the convolution of P with itself k times.

1.1.1 Ehrenfest's urn

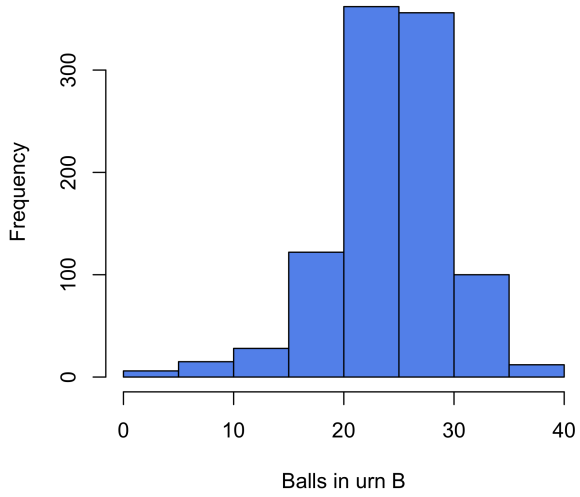
For a concrete example, the following is given. Suppose there are two urns, A and B, and n balls. At time 0, all balls are in urn A. At each step, one ball is chosen uniformly at random, and moved to the other urn. Denote by $\mathbb{Z}/2\mathbb{Z}$ the group of integers modulo 2. The state at step t can be encoded in a vector $v = (c_1, \ldots, c_n) \in (\mathbb{Z}/2\mathbb{Z})^n$, where $c_i = 0$ if and only if ball i is in urn A. Denote by e_i the element of $(\mathbb{Z}/2\mathbb{Z})^n$ having a 1 in its i-th coordinate, and 0 in the rest. If at an arbitrary time the state of the process is encoded by a vector v as previously described, changing ball i to a different urn consists of adding e_i to the state vector v. That is, the state changes to $v + e_i$. Thus, starting at identity $e = (0, 0, \ldots, 0)$ (all balls in urn A), the process of interchanging balls between the urns corresponds to a random walk on $(\mathbb{Z}/2\mathbb{Z})^n$ driven by probability

$$P(g) = \begin{cases} 1/n & \text{if } g \in \{e_1, e_2, \ldots, e_n\}; \\ 0 & \text{else.} \end{cases} \tag{4}$$

A thousand steps of this process, with fifty balls total, was simulated with R [R Core Team, 2020] on a Jupyter notebook [Kluyver et al., 2016]. The average amount of balls in urns A and B where 25.25 and 24.75 respectively, which suggests



(a) Distribution of balls in urn A.



(b) Distribution of balls in urn B.

Figure 1: Histogram for ball distribution in a thousand steps of an Erhenfest process with fifty balls.

the urns are balanced on average. Histograms showing the distribution of the balls in the urns are shown in Figure 1. Additionally, using ImageMagick [The ImageMagick Development Team], a GIF animation[1] was created for twenty steps of this process with six balls total.

2 Goodness of fit

In this section, we perform a goodness of fit test to determine whether the degree distribution of Pennsylvania's road network [Leskovec and Krevl, 2014] follows a Poisson distribution. The network has $|V| = 1,088,092$ nodes, and an average degree of $\lambda_d = 2.834$. A χ^2 goodness of fit test is performed to see if the road "is random". Given the large size of the road, if nodes were connected to each other at random, a Poisson degree distribution would be expected [Newman, 2018]. To perform the test, first the nodes of degree $k = 1, 2, \ldots, 14$ are counted. Then, the expected Poisson distribution is obtained computationally, generating $|V|$ numbers following a Poiss(λ_d) distribution. The expected and observed degrees are displayed in Table 1. Writing O_k for the observed nodes with degree k, and E_k for the expected nodes with degree k, a statistic

$$\sum_{k=1}^{14} = \frac{(O_i - E_i)^2}{E_i} = 672,672.24. \tag{5}$$

is obtained. The corresponding p-value is 0, so it is concluded that the degree distribution is not Poisson.

3 Theoretical results

To conclude this work, two theorems concerning variance and covariance are presented. Theorem 1 was tested computationally for random integers a, b, c, d between one and two hundred, and a pair of one thousand number pseudo-random vectors X, Y, where $X \sim \text{Unif}(0,1), Y \sim \text{Exp}(0.5)$; $X \sim N(0,1), Y \sim \text{Exp}(0.5)$; and $X \sim \text{Geom}(0.33), Y \sim \text{Poiss}(1)$. In a thousand repetitions, the equality always held. Theorem 2 was also verified computationally with the same pair of vectors X, Y, and it too held true in each of the one thousand repetitions.

Theorem 1. *For constants a, b, c, d and random variables X, Y,*

$$\text{Cov}\left(aX + b, cY + d\right) = ac\,\text{Cov}\left(X, Y\right). \tag{6}$$

[1]The notebook with the code for all the simulations in this report, as well the GIF animation, can be found in the Github Repository: https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment11.

Table 1: Observed and expected degree of nodes.

Degree	Expected	Observed
1	245,208	188,317
2	256,577	90,740
3	243,159	532,686
4	171,512	267,256
5	97,772	7,759
6	45,727	1,237
7	18,825	80
8	6,526	13
9	2,039	4
10	550	0
11	147	0
12	41	0
13	7	0
14	2	0

Proof. By definition, it is seen that

$$\begin{aligned}
\operatorname{Cov}(aX+b, cY+d) &= \mathbb{E}[(aX+b)(cY+d)] - \mathbb{E}[aX+b]\,\mathbb{E}[cY+d] && (7)\\
&= \mathbb{E}[acXY + adX + bcY + bd] - (a\,\mathbb{E}[x]+b)(c\,\mathbb{E}[y]+d) && (8)\\
&= ac\,\mathbb{E}[XY] + ad\,\mathbb{E}[X] + bc\,\mathbb{E}[Y] + bd - ac\,\mathbb{E}[X]\,\mathbb{E}[Y] - ad\,\mathbb{E}[X] - bc\,\mathbb{E}[Y] - bd && (9)\\
&= ac(\,\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]) && (10)\\
&= ac\operatorname{Cov}(X,Y)\,. && (11)
\end{aligned}$$

□

Theorem 2. *For random variables X, Y,*

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y)\,. \tag{12}$$

Proof. Since $\operatorname{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, it follows that

$$\begin{aligned}
\operatorname{Var}(X+Y) &= \mathbb{E}[(X+Y)^2] - \mathbb{E}[X+Y]^2 && (13)\\
&= \mathbb{E}[X^2 + 2XY + Y^2] - (\,\mathbb{E}[X] + \mathbb{E}[Y])^2 && (14)\\
&= \mathbb{E}[X^2] + 2\,\mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\,\mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[Y]^2 && (15)\\
&= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2(\,\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]) && (16)\\
&= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y)\,. && (17)
\end{aligned}$$

□

References

P. Diaconis. *Group representations in probability and statistics*. Lecture notes-monograph series. Institute of Mathematical Statistics, 1988. ISBN 978-0-940600-14-0.

I. N. Herstein. *Topics in algebra*. Xerox College Pub, 2nd edition, 1975. ISBN 978-0-536-01090-2.

T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al. Jupyter notebooks—a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing*, page 87. IOS Press, 2016.

J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

M. Newman. *Networks*, volume 1. Oxford University Press, Oct 2018. ISBN 978-0-19-880509-0. doi: 10.1093/oso/9780198805090.001.0001. URL https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198805090.001.0001/oso-9780198805090.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

B. Steinberg. *Representation theory of finite groups: an introductory approach*. Universitext. Springer, 2012. ISBN 978-1-4614-0775-1.

The ImageMagick Development Team. Imagemagick. URL https://imagemagick.org.

Exercises

G. Palafox

November 24, 2020

The following are exercises from the book of Grinstead and Snell [2006].

Exercise 1 (Ex. 1, p. 392)**.** *Let $Z_1, Z_2, \ldots, Z_N$ describe a branching process in which each parent has j offspring with probability p_j. Find the probability d that the process eventually dies out if*

1. $p_0 = 1/2, p_1 = 1/4, p_2 = 1/4$.
2. $p_0 = 1/3, p_1 = 1/3, p_2 = 1/3$.
3. $p_0 = 1/3, p_1 = 0, p_2 = 2/3$.
4. $p_j = 1/2^{j+1}$, *for* $j = 0, 1, 2, \ldots$.
5. $p_j = (1/3)(2/3)^j$, *for* $j = 0, 1, 2, \ldots$.
6. $p_j = e^{-2}2^j/j!$, *for* $j = 0, 1, 2, \ldots$ *(estimate d numerically).*

Solution to 1.1. We have $m = \sum kp_k = 0(1/2) + 1(1/4) + 2(1/4) = 3/4 \leq 1$ so $d = 1$. □

Solution to 1.2. Here, $m = \sum kp_k = 0(1/3) + 1(1/3) + 2(1/3) = 1 \leq 1$ so $d = 1$. □

Solution to 1.3. Since $m = \sum kp_k = 0(1/3) + 0(0) + 2(2/3) = 4/3 > 1$, to find d we need to compute the roots of $h(x) = x$, where $h(x) = \sum p_k x^k$. In this case, $h(x) = (1/3) + (2/3)x^2$, and

$$\begin{aligned}(1/3) + (2/3)x^2 = x &\Leftrightarrow 1 + 2x^2 = 3x &(1)\\ &\Leftrightarrow 2x^2 - 3x + 1 = 0 &(2)\\ &\Leftrightarrow (2x-1)(x-1) = 0 &(3)\\ &\Leftrightarrow x = 1/2, x = 1, &(4)\end{aligned}$$

so $d = 1/2$. □

Solution to 1.4. Knowing $\sum_{k=1}^{\infty} kx^k = \frac{x}{(1-x)^2}$ and $\sum_{k=1}^{\infty} x^k = \frac{x}{1-x}$ when $|x| < 1$, we see that

$$\begin{aligned}m = \sum_{j=0}^{\infty} j\left(\frac{1}{2^{j+1}}\right) = \sum_{j=1}^{\infty}(j-1)\left(\frac{1}{2^j}\right) &(5)\\ = \sum_{j=1}^{\infty}\frac{j}{2^j} - \sum_{j=1}^{\infty}\frac{1}{2^j} &(6)\\ = 2 - 1 = 1. &(7)\end{aligned}$$

Given how it is equal to one, we conclude $d = 1$. □

Solution to 1.5. Proceeding as before, we see that

$$\begin{aligned}m = \sum_{j=0}^{\infty} j(1/3)(2/3)^j &(8)\\ = (1/3)\frac{2/3}{(1-(2/3))^2} &(9)\\ = (1/3)\frac{(2/3)}{(1/3)^2} = \frac{(2/3)}{(1/3)} = 2. &(10)\end{aligned}$$

Again, having $m = 2 > 1$, we must find x such that

$$\sum_{j=0}^{\infty}(1/3)(2/3)^j x^j = x. \tag{11}$$

Rearranging the terms, we must find x such that

$$(1/3)\sum_{j=0}^{\infty}(\frac{2}{3}x)^j = x. \tag{12}$$

Given that $|x| < 1$, then $|\frac{2}{3}x| < 1$ and

$$\sum_{j=0}^{\infty}(\frac{2}{3}x)^j = \frac{1}{1-(\frac{2}{3}x)}, \tag{13}$$

so we need x such that

$$(1/3)\frac{1}{1-(\frac{2}{3}x)} = (1/3)\left(\frac{3}{3-2x}\right) = \frac{1}{3-2x} = x. \tag{14}$$

This x is found by solving the equation $2x^2 - 3x + 1 = 0$, which we did in the solution to 1.3. Therefore $d = 1/2$. □

Solution to 1.6. Starting by finding m one gets

$$m = \sum_{j=0}^{\infty}\frac{e^{-2}2^j}{j!}j \tag{15}$$

$$= e^{-2}\sum_{j=0}^{\infty}\frac{2^j}{j!}j \tag{16}$$

$$= e^{-2}\sum_{j=1}^{\infty}2\frac{2^{j-1}}{(j-1)!} \tag{17}$$

$$= 2e^{-2}\sum_{j=1}^{\infty}\frac{2^{j-1}}{(j-1)!} \tag{18}$$

$$= 2e^{-2}\sum_{j=0}^{\infty}\frac{2^j}{j!} \tag{19}$$

$$= 2e^{-2}e^2 = 2. \tag{20}$$

Since $m > 1$, to obtain d we must find x such that

$$\sum_{j=0}^{\infty}\frac{e^{-2}2^j}{j!}x^j = x. \tag{21}$$

Given that

$$\sum_{j=0}^{\infty}\frac{e^{-2}2^j}{j!}x^j = e^{-2}\sum_{j=0}^{\infty}\frac{2^j}{j!}x^j = e^{-2}e^{2x} = e^{2x-2}, \tag{22}$$

we numerically estimate with Wolfram Alpha [Wolfram Research Inc.] that $e^{2x-2} = x$ for $x = 1$ and $x \approx 0.203$, so $d \approx 0.203$. □

Exercise 2 (Ex. 3, p. 392)**.** *In the chain letter problem (see Example 10.14) find your expected profit if*

1. $p_0 = 1/2, p_1 = 0$, *and* $p_2 = 1/2$.
2. $p_0 = 1/6, p_1 = 1/2$, *and* $p_2 = 1/3$.

Solution to 2.1. We see in Grinstead and Snell [2006] that the expected profit is $50m + 50m^{12} - 100$, where $m = p_1 + 2p_2$. Here, $m = 0 + 2(1/2) = 1$, so the expected profit is $50(1) + 50(1)^{12} - 100 = 100 - 100 = 0$. □

Solution to 2.2. Here, $m = (1/2) + 2(1/3) = 7/6$, so the expected profit is $50(7/6) + 50(7/6)^{12} - 100 \approx 276.26$. If $p_0 > 1/2$, then $p_1 + p_2 < 1/2$, so $2p_1 + 2p_2 < 1$ and $p_1 + 2p_2 < 1 - p_1 \leq 1$. Therefore, $50(p_1 + 2p_2) < 50$ and $50(p_1 + 2p_2)^{12} < 50$, so $50(p_1 + 2p_2) + 50(p_1 + 2p_2)^{12} < 100$, so the expected profit is negative. □

Exercise 3 (Ex. 3, p. 401)**.** *Let X be a continuous random variable with values in $[0,2]$ and density f_X. Find the moment generating function $g(t)$ for X if*

1. $f_X(x) = 1/2$.
2. $f_X(x) = (1/2)x$.
3. $f_X(x) = 1 - (1/2)x$.
4. $f_X(x) = |1 - x|$.
5. $f_X(x) = (3/8)x^2$.

Solution. We calculate these with the equation $g(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx$. In this particular case, since the random variable has values in $[0,2]$, the moment generating function will be $g(t) = \int_0^2 e^{tx} f_X(x)\,dx$. For each of the densities f_X from number 1. to 5., the corresponding generating function will be denoted by $g_k(t), k = 1, 2, \ldots, 5$. First,

$$
\begin{aligned}
g_1(t) &= \int_0^2 e^{tx}(1/2)\,dx && (23)\\
&= (1/2)\int_0^2 e^{tx}\,dx && (24)\\
&= (1/2)[(1/t)e^{tx}\Big|_0^2] && (25)\\
&= (1/2)[(1/t)e^{2t} - (1/t)e^0] && (26)\\
&= (1/2)[\frac{e^{2t}-1}{t}] && (27)\\
&= \frac{1}{2t}(e^{2t} - 1) && (28)
\end{aligned}
$$

is obtained. Then we compute

$$
\begin{aligned}
g_2(t) &= \int_0^2 e^{tx}(1/2)x\,dx && (29)\\
&= (1/2)[\frac{1}{t}xe^{tx}\Big|_0^2 - \int_0^2 \frac{1}{t}e^{tx}\,dx\,] && (30)\\
&= (1/2)[\frac{1}{t}xe^{tx}\Big|_0^2 - \frac{1}{t^2}e^{tx}\Big|_0^2] && (31)\\
&= (1/2)[(\frac{1}{t}2e^{2t} - \frac{1}{t^2}e^{2t}) - (0 - \frac{1}{t^2})] && (32)\\
&= (1/2)[\frac{2}{t}e^{2t} - \frac{1}{t^2}e^{2t} + \frac{1}{t^2}] && (33)\\
&= \frac{1}{t}e^{2t} - \frac{1}{2t^2}e^{2t} + \frac{1}{2t^2}\ . && (34)\\
& && (35)
\end{aligned}
$$

For the third, one gets

$$
\begin{aligned}
g_3(t) &= \int_0^2 e^{tx}(1 - (1/2)x)\,dx && (36)\\
&= \int_0^2 e^{tx}\,dx - \int_0^2 (1/2)xe^{tx}\,dx && (37)\\
&= \frac{1}{t}e^{tx}\Big|_0^2 - (\frac{1}{t}e^{2t} - \frac{1}{2t^2}e^{2t} + \frac{1}{2t^2}) && (38)\\
&= \frac{1}{t}e^{2t} - \frac{1}{t} - \frac{1}{t}e^{2t} + \frac{1}{2t^2}e^{2t} - \frac{1}{2t^2} && (39)\\
&= \frac{1}{2t^2}e^{2t} - \frac{1}{2t^2} - \frac{1}{t}. && (40)\\
& && (41)
\end{aligned}
$$

Afterwards, we proceed to compute

$$
g_4(t) = \int_0^2 |1-x|e^{tx}\,dx = \int_0^1 (1-x)e^{tx}\,dx + \int_1^2 (x-1)e^{tx}\,dx. \tag{42}
$$

Computing the integrals one gets

$$\int_0^1 (1-x)e^{tx}\,dx = \int_0^1 e^{tx}\,dx - \int_0^1 xe^{tx}\,dx \tag{43}$$

$$= \frac{1}{t}e^{tx}\Big|_0^1 - [\frac{1}{t}xe^{tx} - \frac{1}{t^2}e^{tx}]\Big|_0^1 \tag{44}$$

$$= (\frac{1}{t}e^t - \frac{1}{t}) - (\frac{1}{t}e^t - \frac{1}{t^2}e^t + \frac{1}{t^2}) \tag{45}$$

$$= \frac{1}{t^2}e^t - \frac{1}{t^2} - \frac{1}{t}, \tag{46}$$

and

$$\int_1^2 (x-1)e^{tx}\,dx = \int_1^2 xe^{tx}\,dx - \int_1^2 2e^{tx}\,dx \tag{47}$$

$$= (\frac{1}{t}xe^{tx} - \frac{1}{t^2}e^{tx})\Big|_1^2 - \frac{1}{t}e^{tx}\Big|_1^2 \tag{48}$$

$$= \frac{2}{t}e^{2t} - \frac{1}{t^2}e^{2t} - \frac{1}{t}e^t + \frac{1}{t^2}e^t - \frac{1}{t}e^{2t} + \frac{1}{t}e^t \tag{49}$$

$$= \frac{1}{t}e^{2t} - \frac{1}{t^2}e^{2t} + \frac{1}{t^2}e^t. \tag{50}$$

$$\tag{51}$$

Therefore

$$g(t) = (\frac{1}{t^2}e^t - \frac{1}{t^2} - \frac{1}{t}) + (\frac{1}{t}\ e^{2t} - \frac{1}{t^2}e^{2t} + \frac{1}{t^2}e^t) \tag{52}$$

$$= (\frac{1}{t} - \frac{1}{t^2})e^{2t} + \frac{2}{t^2}e^t - \frac{1}{t^2} - \frac{1}{t}. \tag{53}$$

Finally, we compute $g_5(t)$ as

$$g_5(t) = \int_0^2 (3/8)x^2e^{tx}\,dx \tag{54}$$

$$= (3/8)\int_0^2 x^2e^{tx}\,dx \tag{55}$$

$$= (3/8)\left((\frac{x^2}{t}\ e^{tx})\Big|_0^2 - \int_0^2 2x(\frac{1}{t}e^{tx})\,dx\ \right) \tag{56}$$

$$= (3/8)\left(\frac{4}{t}e^{2t} - \frac{2}{t}(\frac{2}{t}e^{2t} - \frac{1}{t^2}e^{2t} + \frac{1}{t^2})\right) \tag{57}$$

$$= (3/8)\left(\frac{4}{t}e^{2t} - \frac{4}{t^2}e^{2t} + \frac{2}{t^3}e^{2t} - \frac{2}{t^3}\right) \tag{58}$$

$$= \frac{3}{2t}e^{2t} - \frac{3}{2t^2}e^{2t} + \frac{3}{4t^3}e^{2t} - \frac{3}{4t^3}. \tag{59}$$

□

Exercise 4 (Ex. 6, p. 402)**.** *Let X be a continuous random variable whose characteristic function $k_X(\tau)$ is*

$$k_X(\tau) = e^{-|\tau|}, \quad -\infty < \tau < \tau.$$

Show directly that the density f_X of X is

$$f_X(x) = \frac{1}{\pi(1+x^2).}$$

Proof. We know that

$$f_X(x) = \frac{1}{2\,\pi}\int_{-\infty}^{\infty} e^{-i\tau x}k_X(\tau)\,d\tau, \tag{60}$$

where we have

$$k_X(\tau) = e^{-|\tau|} = \begin{cases} e^{\tau} & \tau \le 0; \\ e^{-\tau} & \tau > 0. \end{cases} \tag{61}$$

Additionally, we know that $e^{i\theta} = \cos\theta + i\sin\theta$. Thus, substituting this in Equation 60 we get

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}(\cos(-\tau x) + i\sin(-\tau x))k_X(\tau)\,d\tau \tag{62}$$

$$= \frac{1}{2\pi}\left[\int_{-\infty}^{\infty}\cos(-\tau x)k_X(\tau)\,d\tau + i\int_{-\infty}^{\infty}\sin(-\tau x)k_X(\tau)\,d\tau\right] \tag{63}$$

$$= \frac{1}{2\pi}\left[\left\{\int_{-\infty}^{0}\cos(-\tau x)e^{\tau}\,d\tau + \int_{0}^{\infty}\cos(-\tau x)e^{-\tau}\,d\tau\right\} + i\left\{\int_{-\infty}^{0}\sin(-\tau x)e^{\tau}\,d\tau + \int_{0}^{\infty}\sin(-\tau x)e^{-\tau}d\tau\right\}\right] \tag{64}$$

Using the fact that the cosine function is even, changing variables and rearranging the integration limits, we see that

$$\int_{-\infty}^{0}\cos(-\tau x)e^{\tau}\,d\tau = \int_{0}^{\infty}\cos(\tau x)e^{-\tau}\,d\tau \tag{65}$$

and

$$\int_{0}^{\infty}\cos(-\tau x)e^{-\tau}\,d\tau = \int_{0}^{\infty}\cos(\tau x)e^{-\tau}\,d\tau. \tag{66}$$

Similarly, seeing how sine is odd, we get

$$\int_{-\infty}^{0}\sin(-\tau x)e^{\tau}\,d\tau = \int_{0}^{\infty}\sin(\tau x)e^{-\tau}\,d\tau \tag{67}$$

and

$$\int_{0}^{\infty}\sin(-\tau x)e^{-\tau}\,d\tau = -\int_{0}^{\infty}\sin(\tau x)e^{-\tau}\,d\tau. \tag{68}$$

Therefore,

$$f_X(x) = \frac{1}{2\pi}\left[\left\{\int_{0}^{\infty}\cos(\tau x)e^{-\tau}\,d\tau + \int_{0}^{\infty}\cos(\tau x)e^{-\tau}\,d\tau\right\} + i\left\{\int_{0}^{\infty}\sin(\tau x)e^{-\tau}\,d\tau - \int_{0}^{\infty}\sin(\tau x)e^{-\tau}\,d\tau\right\}\right] \tag{69}$$

$$= \frac{1}{2\pi}2\int_{0}^{\infty}\cos(\tau x)e^{-\tau}\,d\tau, \tag{70}$$

where integrating by parts twice we get

$$\int_{0}^{\infty}\cos(\tau x)e^{-\tau}\,d\tau = -e^{-\tau}\cos(\tau x)\Big|_{0}^{\infty} - \int_{0}^{\infty}x\sin(\tau x)e^{-\tau}\;\,d\tau \tag{71}$$

$$= -e^{-\tau}\cos(\tau x)\Big|_{0}^{\infty} + x\sin(\tau x)e^{-\tau}\Big|_{0}^{\infty} - \int_{0}^{\infty}x^2\cos(\tau x)e^{-\tau}\,dt \tag{72}$$

$$= \frac{x\sin(-\tau x) - \cos(\tau x)}{1+x^2}e^{-\tau}\Big|_{0}^{\infty} \tag{73}$$

$$= 0 - \frac{x\sin 0 - \cos 0}{1+x^2} \tag{74}$$

$$= \frac{1}{1+x^2}, \tag{75}$$

therefore

$$f_X(x) = \frac{1}{\pi(x^2+1)}. \tag{76}$$

□

Exercise 5 (Ex. 10, p. 403)**.** *Let $X_1, \ldots, X_n$ be an independent trials process with density*

$$f(x) = \frac{1}{2}e^{-|x|}, \quad -\infty < x < \infty.$$

1. *Find the mean and variance of $f(x)$.*
2. *Find the moment generating function for X_1, S_n, A_n, S_n^*.*
3. *What can you say about the moment generating function of S_n^* as $n \to \infty$.*
4. *What can you say about the moment generating function of A_n as $n \to \infty$.*

Solution to 5.1. We know from Grinstead and Snell [2006] that

$$k_X(\tau) = g_X(i\tau) = \int_{-\infty}^{\infty} e^{i\tau x} f_X(x)dx = \frac{1}{2}\int_{-\infty}^{\infty} e^{i\tau x} e^{-|x|}\, dx. \tag{77}$$

Making a $u = -x$ substitution in the integral from Exercise 4 one sees that

$$\int_{-\infty}^{\infty} e^{i\tau x} e^{-|x|}\, dx = \int_{-\infty}^{\infty} e^{-i\tau x} e^{-|x|}\, dx = \frac{2}{1+\tau^2}, \tag{78}$$

so $k_X(\tau) = \frac{1}{1+\tau^2}$, and from the relation $k_X(\tau) = g_X(i\tau)$, one sees $g_X(t) = \frac{1}{1-t^2}$. Knowing this, the mean is obtained as

$$\frac{dg_X(t)}{dt}\Big|_{t=0} = \frac{2t}{(1-t^2)^2}\Big|_{t=0} = 0, \tag{79}$$

and the variance as

$$\frac{d^2 g_X(t)}{dt^2}\Big|_{t=0} = \left[\frac{2}{(1-t^2)^2} + \frac{8t^2}{(1-t^2)^3}\right]\Big|_{t=0} = 2. \tag{80}$$

□

Solution to 5.2. The moment generating function for X_1, and the rest of the X_i, is as obtained in the solution to 5.1, and is $g_X(t) = \frac{1}{1-t^2}$. Then,

$$\begin{aligned} g_{S_n}(t) = \mathbb{E}\left[e^{S_n t}\right] &= \mathbb{E}\left[e^{(X_1+\cdots+X_n)t}\right] && (81)\\ &= \mathbb{E}\left[e^{X_1 t} e^{X_2 t}\cdots e^{X_n t}\right] && (82)\\ &= \mathbb{E}\left[e^{X_1 t}\right]\cdots\mathbb{E}\left[e^{X_n t}\right] \quad \text{(because of independence)} && (83)\\ &= g_{X_1}(t)\cdots g_{X_n}(t) && (84)\\ &= \left(\frac{1}{1-t^2}\right)^n. && (85) \end{aligned}$$

For $A_n = S_n/n$, see that

$$g_{A_n}(t) = g_{\frac{S_n}{n}} = \mathbb{E}\left[e^{\frac{S_n}{n}t}\right] = \mathbb{E}\left[e^{S_n \frac{t}{n}}\right] = g_{S_n}(t/n), \tag{86}$$

so

$$g_{A_n}(t) = \left(\frac{1}{1-(\frac{t}{n})^2}\right)^n. \tag{87}$$

For $S_n^* = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{S_n}{\sqrt{2n}}$ (substituting the mean and variance found), proceeding as was done with A_n, we obtain

$$g_{S_n^*}(t) = g_{\frac{S_n}{\sqrt{2n}}}(t) = g_{S_n}(t/\sqrt{2n}) = \left(\frac{1}{1-(\frac{t}{\sqrt{2n}})^2}\right)^n. \tag{88}$$

□

Solution to 5.3. Here, and in the solution of Exercise 5.4, the following result found in Casella and Berger [2002], will be used: *If a sequence* (a_n) *converges to* a*, then* $(1+\frac{a_n}{n})^n$ *converges to* e^a. To find the limit of $g_{S_n^*}$ as $n \longrightarrow \infty$, see that

$$\lim_{n\to\infty}\left(\frac{1}{1-(\frac{t}{\sqrt{2n}})^2}\right)^n = \frac{\lim_{n\to\infty} 1^n}{\lim_{n\to\infty}(1-(\frac{t}{\sqrt{2n}})^2)^n} \tag{89}$$

if both limits exist. Given how

$$1 - \left(\frac{t}{\sqrt{2n}}\right)^2 = 1 + \left(\frac{-t^2}{2n}\right) = 1 + \left(\frac{\frac{-t^2}{2}}{n}\right), \tag{90}$$

and the sequence $a_n = -t^2/2$ converges to $-t^2/2$ as n goes to infinity, the aforementioned theorem tells us

$$\lim_{n\to\infty}\left(1-(\frac{t}{\sqrt{2n}})^2\right)^n = \lim_{n\to\infty}\left(1+\left(\frac{\frac{-t^2}{2}}{n}\right)\right)^n = e^{\frac{-t^2}{2}}, \tag{91}$$

so

$$\lim_{n\to\infty} g_{S_n^*} = \lim_{n\to\infty}\left(\frac{1}{1-(\frac{t}{\sqrt{2n}})^2}\right)^n = \frac{\lim_{n\to\infty} 1^n}{\lim_{n\to\infty}(1-(\frac{t}{\sqrt{2n}})^2)^n} = \frac{1}{e^{\frac{-t^2}{2}}} = e^{\frac{t^2}{2}}. \tag{92}$$

□

Solution to 5.4. As before, we see that

$$\lim_{n\to\infty}\left(\frac{1}{1-(\frac{t}{n})^2}\right)^n = \frac{\lim_{n\to\infty} 1^n}{\lim_{n\to\infty}\left(1-(\frac{t}{n})^2\right)^n} \tag{93}$$

if both limits exist, and

$$1-(\frac{t}{n})^2 = 1+\left(\frac{-t^2}{n^2}\right) = 1+\frac{\frac{-t^2}{n}}{n}. \tag{94}$$

The sequence $a_n = -t^2/n$ converges to zero, so

$$\lim_{n\to\infty}\left(1-\left(\frac{t}{n}\right)^2\right)^n = \lim_{n\to\infty}\left(1+\frac{\frac{-t^2}{n}}{n}\right)^n = e^0 = 1, \tag{95}$$

which gives

$$\lim_{n\to\infty} g_{A_n} = \lim_{n\to\infty}\left(\frac{1}{1-(\frac{t}{n})^2}\right)^n = \frac{\lim_{n\to\infty} 1^n}{\lim_{n\to\infty}\left(1-(\frac{t}{n})^2\right)^n} = 1. \tag{96}$$

□

Consider the Poisson process with waiting times $T_1, T_2, \ldots$. These are independent, exponential random variables with parameter λ and N_t is the Poisson process.

References

G. Casella and R. L. Berger. *Statistical inference*. Thomson Learning, 2nd edition, 2002. ISBN 978-0-534-24312-8.

C. M. Grinstead and J. L. Snell. *Introduction to Probability*. 2006.

Wolfram Research Inc. Wolfram Alpha. URL https://www.wolframalpha.com/. Champaign, IL, 2019.

Law of large numbers

Gerardo Palafox

December 1, 2020

Abstract

In this work, the law of large numbers is presented, along some applications of the same.

1 Introduction

Two theorems will be given, known as the weak and strong law of large numbers. Both of the statements, as well as the definitions preceding it, can be found in the work of Casella and Berger [2002].

Definition 1. *A sequence of random variables $X_1, X_2, \ldots$, converges in probability to a random variable X if, for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) = 0 \quad \textit{or, equivalently} \quad \lim_{n\to\infty} \mathbb{P}\left(|X_n - X| < \varepsilon\right) = 1 \tag{1}$$

Definition 2. *A sequence of random variables $X_1, X_2, \ldots$, converges almost surely to a random variable X if, for every $\varepsilon > 0$,*

$$\mathbb{P}\left(\lim_{n\to\infty} |X_n - X| < \varepsilon\right) = 1 \tag{2}$$

Theorem 1 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then, $\bar{X_n} = (1/n)\sum_{i=1}^n X_i$ converges in probability to μ.*

Theorem 2 (Strong law of large numbers)**.** *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then, $\bar{X_n} = (1/n)\sum_{i=1}^n X_i$ converges almost surely to μ.*

2 Renewal processes

One of the many applications of the law of large numbers is in a kind of stochastic process called renewal processes. The contents of this section can be found in the book of Lawler [2006].

Definition 3. *Let $T_1, T_2, \ldots$ be independent, identically distributed, non-negative random variables with distribution $F(x)$. The renewal process associated with $\{T_i\}$ is the process that counts the number of events that have occurred by time t. More precisely, the renewal process N_t is defined by*

$$N_t = \begin{cases} 0 & t < T_1; \\ \max\{n : T_1 + \cdots + T_n \leq t\} & \textit{else.} \end{cases} \tag{3}$$

The random variables T_i are thought of as being the lifetimes of a component, or as the times between occurrences of some event. Time 0 is assumed as the beginning of a lifetime. The random variables T_i are assumed to have finite, positive mean $\mu = \mathbb{E}[T_i]$.

Example 1. *Consider the Poisson process with waiting times $T_1, T_2, \ldots$. These are independent, exponential random variables with parameter λ and N_t is the Poisson process.*

Example 2. *Consider a queue with a single server where customers arrive according to a Poisson process with rate λ. Assume the service times for customers are independent, identically distributed random variables with mean μ. Let Y_t denote the number of people in the queue at time t. Let*

$$R_1 = \inf\{t > 0 : Y_t = 1\}, \tag{4}$$

$$S_1 = \inf\{t > 0 : Y_{R_1+t} = 0\}, \tag{5}$$

$$T_1 = R_1 + S_1, \tag{6}$$

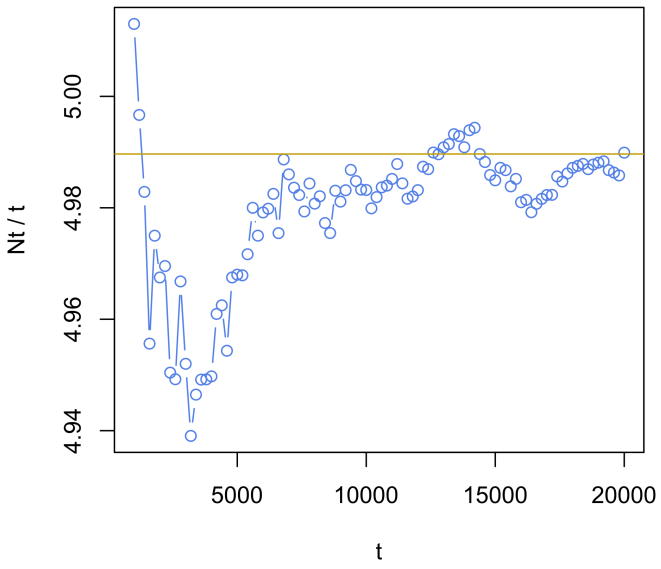


Figure 1: In blue, the value of N_t/t. In gold, the value $1/\mu$.

and for $i > 1$,

$$R_i = \inf\{t > 0 : Y_{T_1+\cdots+T_{i-1}+t} = 1\}, \tag{7}$$
$$S_i = \inf\{t > 0 : Y_{T_1+\cdots+T_{i-1}+R_i+t} = 0\}, \tag{8}$$
$$T_i = R_i + S_i. \tag{9}$$

The time represented by R_i *can be thought of as idle times, and the time represented by* S_i *as the busy times. The random variables* $T_1, T_2, \ldots$ *give rise to a renewal process* N_t.

Using the Strong Law of Large Numbers (Theorem 2), the following analogue of the law of large numbers for renewal processes can be derived:

Proposition 1. *For a renewal process* N_t*, with probability one,*

$$\lim_{t\to\infty} \frac{N_t}{t} = \frac{1}{\mu}. \tag{10}$$

A Poisson process (see Example 1) was simulated[1] with R version 4.0.0 in a Jupyter notebook [R Core Team, 2020, Kluyver et al., 2016]. In Figure 1 it is observed how $N_t/t \to 1/\mu$ as t grows.

3 Epidemics on networks

In the work of Janson et al. [2014], a law of large numbers is established for an epidemic process on a random graph. In particular, a random graph with a given degree sequence is considered. Given the graph, the epidemic evolves as a continuous-time Markov chain [Liggett, 2010] where each infectious vertex recovers at rate $\rho \geq 0$ and also infects each susceptible neighbor at a rate $\beta > 0$. The graph has n vertices, of which initially n_S, n_I, n_R are susceptible, infectious and recovered respectively. It is assumed the fractions of initially susceptible, infectious and recovered vertices converge to some $\alpha_S, \alpha_R, \alpha_I \in [0, 1]$, with $\alpha_S > 0$. Additionally, it is assumed the degree of a randomly chosen susceptible vertex converges to a probability distribution $(p_k)_{k=0}^{\infty}$, that is,

$$n_{S,k}/n \to p_k, \tag{11}$$

where $n_{S,k}$ is the number of susceptible vertices of degree k. Furthermore, the average degree over all vertices converges to $\mu > 0$. Let S_t, I_t, R_t be the number of susceptible, infectious and recovered vertices at time t, and T_0 the time in which

[1] The code, as well as this report, can be found in https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment13.

the fraction of susceptible individuals has fallen from about α_S to some fixed smaller s_0. Janson et al. [2014] prove that when the quantity

$$R_0 := \left(\frac{\beta}{\beta+\rho}\right)\left(\frac{\alpha_S}{\mu}\right)\sum_{k=0}^{\infty}(k-1)kp_k, \tag{12}$$

interpreted as the basic reproduction number of the epidemic, is greater than one, then for every $\varepsilon > 0$,

$$\mathbb{P}\left(T_0 = \infty \text{ and } S_0 - S_\infty > \varepsilon n\right) \to 0. \tag{13}$$

This represents a small outbreak. However, it is also proved that if $T_0 < \infty$, the outbreak is large.

3.1 Vaccination

In their work, Janson et al. [2014] also study the effects of different vaccination strategies. A perfect vaccine is assumed, so a vaccinated vertex never becomes infected and in practice behaves as recovered. It is assumed each initially susceptible vertex of degree k is vaccinated with probability $\pi_k \in [0, 1)$, independently of the others. As an example, consider the *uniform vaccination* strategy. Here, every susceptible vertex is vaccinated with the same probability $v \in [0, 1)$, independently of all the others. Using the law of large numbers, it is proven that V/n_S converges in probability to v, where V is the total number V of vaccinations.

References

G. Casella and R. L. Berger. *Statistical inference*. Thomson Learning, 2nd edition, 2002. ISBN 978-0-534-24312-8.

S. Janson, M. Luczak, and P. Windridge. Law of large numbers for the SIR epidemic on a random graph with given degrees. 2014. URL https://arxiv.org/abs/1308.5493.

T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al. Jupyter notebooks—a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing*, page 87. IOS Press, 2016.

G. F. Lawler. *Introduction to Stochastic Processes*. Taylor and Francis/CRC Press, 2nd edition, 2006.

T. M. Liggett. *Continuous time Markov processes: an introduction*. Graduate studies in mathematics. American Mathematical Society, 2010. ISBN 978-0-8218-4949-1.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

Central Limit Theorem

Gerardo Palafox

December 8, 2020

Abstract

In this work, the central limit theorem is presented, along some applications of the same.

1 Introduction

First, a definition concerning convergence of random variables is given. With this, the central limit theorem (CLT) can be stated. The proof is omitted but can be seen in the work of Casella and Berger [2002]. In Section 2, an application of the CLT to Markov chains is explained. Then, an application of the central limit theorem to defining and detecting hierarchical community structures in networks is given in Section 3. Finally, in Section 4, a CLT for an SIR epidemics in a configuration model is discussed.

1.1 Basic theory

Definition 1. *A sequence of random variables $X_1, X_2, \ldots$, converges in distribution to a random variable X if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \tag{1}$$

at all points x where $F_X(x)$ is continuous.

Theorem 1 (Central limit theorem [Casella and Berger, 2002])**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cumulative distribution function of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then, for any x,*

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy; \tag{2}$$

that is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converges in distribution to a standard normal random variable.

2 Markov Chains

A stochastic process is a sequence of random variables $\{X_i\}_{i\in I}$. A Markov chain is a stochastic process that takes on a finite or countable number of possible values, called states, and such that

$$\mathbb{P}\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\} = \mathbb{P}\{X_{n+1} = j | X_n = i\}. \tag{3}$$

Markov chains have several applications, for example, modeling of epidemic processes, such as the one described in Section 4. More about the basic theory of Markov chains can be found in the texts of Ross [2000], Feller [1964], Lawler [2006]. For a state i, let f_i denote the probability that, starting at state i, the process will ever reenter state i. If $f_i = 1$, state i is said to be *recurrent*, and if $f_i < 1$, the state is said to be *transient* . If state i is recurrent, starting in state i, the process will enter state i infinitely often. Let k be a recurrent state in a finite Markov chain. Let N_n denote the number of passages up to time n of the system through state k. Then, N_n is normally distributed as $n \to \infty$ [Feller, 1964]. For example, consider a Markov chain with states $\{0, 1\}$ and probabilities $\mathbb{P}\{X_{n+1} = j | X_n = i\} = P_{ij}$ given by the entries of the matrix

$$P = \begin{bmatrix} .3 & .7 \\ .6 & .4 \end{bmatrix}. \tag{4}$$

Then the number N_n of passages up to time n of the system through the state 0 is normally distributed as $n \to \infty$. This Markov chain was simulated[1] in R [R Core Team, 2020] on a Jupyter notebook [Kluyver et al., 2016]. The number N_n for $n = 10,000$ was computed a thousand times. The histogram of these results is shown in Figure 1.

[1]The code of this simulation, as well as this report, can be found in the Github repository https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment14

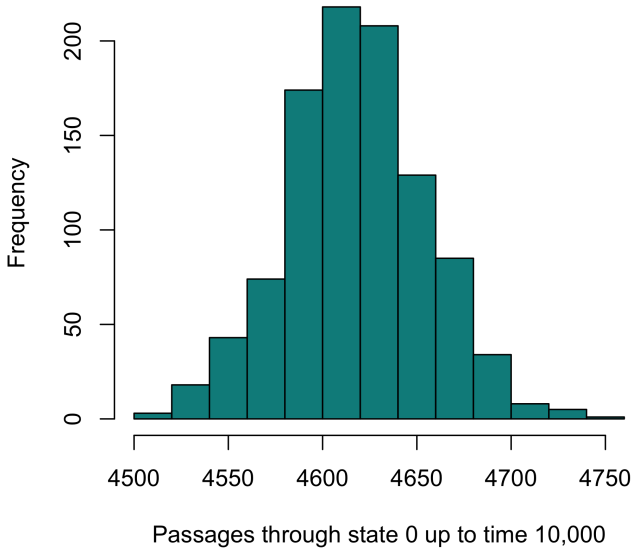


Figure 1: Histogram of N_n for large n.

3 Hierarchical community structure in networks

Schaub and Peel [2020] studied hierarchical community structures in networks. In particular, their work addresses how to define hierarchies of communities, how to determine if such hierarchical structure exists in a network, and how to detect such structures efficiently. To define similarities between nodes, some terminology is introduced. Groups of nodes r and s are called *stochastically equivalent* if any node in group r has the same probability Ω_{rs} of linking to any node in group s. Then, the stochastic block model (SBM) is used to represent the community structure of a network. The SBM defines the probability of a link existing between two nodes depending on their community assignment, with a group indicator binary matrix H, where $H_{ir} = 1$ if node i is assigned to group r and $H_{ir} = 0$ otherwise. Denoting by $H_{i\cdot}$ the ith row of H, A the adjacency matrix of the network, and Ω the affinity matrix, the probability of nodes i and j being linked is given by

$$\mathbb{P}\{A_{ij} = 1\} = H_{i\cdot}\Omega H_{j\cdot}^{\top}. \tag{5}$$

The SBM provides a parametric probability distribution over adjacency matrices. The expected adjacency matrix of this distribution can be calculated from the affinity matrix Ω and group indicator matrix H as

$$\mathbb{E}[A] = H\Omega H^{\top}. \tag{6}$$

Nodes in the network $\mathbb{E}[A]$ which are in the same group are caled *structurally equivalent*. A partition of an adjacency matrix A such that every node in a group r has the same number of links to nodes in a group s is called an *equitable partition*. A *stochastic equitable partition* is an equitable partition in expectation. Partitions that are equitable only between different groups are called *externally equitable partitions* (EEP). A *stochasitc externally equitable partition* (sEEP) is a partition that is externally equitable in expectation. A hierarchical partition is a *valid hierarchy* if at each level the partition is a sEEP and is not degenerate. In detecting hierarchies via spectral methods, authors approximate the true affinity matrix via an estimated affinity matrix $\hat{\Omega}$. To measure how well a partition of $\hat{\Omega}$ approximates an EEP of Ω, the authors use the central limit theorem to conclude the spectral properties of $\hat{\Omega}$ will closely approximate the true Ω, since for large n the entry $\hat{\Omega}_{ij}$ will be approximated by a normal $\mathrm{N}(\mu_{ij}, \sigma_{ij})$ random variable, and $\mu_{ij} = \Omega_{ij}$, $\sigma_{ij}^2 = \Omega_{ij}(1 - \Omega_{ij})/n_i n_j$, where n_i, n_j are the number of nodes in group i and j respectively.

4 Central limit theorems for SIR epidemics on random graphs

Ball [2018] develops central limit theorems for a stochastic susceptible - infectious - recovered epidemic defined on a configuration model [Newman, 2018] random graph. Graphs where degrees of individuals are deterministic (Molloy-Reed) and where degrees are i.i.d. (Newman-Strogatz-Watts) are considered. A population of n individuals, labeled $1, 2, \ldots, n$, is considered. Let $T_i^{(n)}$ be the total number of degree-i susceptibles infected by the epidemic, and $T^{(n)} = \sum_{i=0}^{d_{\max}} T_i^{(n)}$ the final

size of the epidemic, where $d_{\max}$ is the maximum degree among all nodes. Upon a suitable standardization, it is proven that the final size of the epidemic converges in distribution to a normal random variable. Let $a_i^{(n)}$ be the number of degree i initial infectious, and $a^{(n)} = \sum_i a_i^{(n)}$ the total number of initial infectious. Let $\epsilon^{(n)} = n^{-1}a^{(n)}, \epsilon_i^{(n)} = n^{-1}a_i^{(n)}, \epsilon := \lim_{n\to\infty} \epsilon^{(n)}$, and let ϵ_i be such that $\lim_{n\to\infty} \sqrt{n}(\epsilon_i^{(n)} - \epsilon_i) = 0$. If the degree distribution is a random variable D, let $\mathbb{P}\{D = i\} = p_i$ and $\mu_D = \mathbb{E}[D]$. Let p_I be the probability that the neighbor of an infectious individual is contacted, and $q_I = 1 - p_I$ the probability that an infectious fails to contact a given neighbor. Let $f_{D_\epsilon}(s) = \sum_{i=0}^{d_{\max}} (p_i - \epsilon_i)s^i$. Define $z \in [0, 1)$ as the unique solution of

$$z - q_I = \mu_D^{-1} p_I f_{D_\epsilon}(z), \tag{7}$$

and

$$\rho = 1 - \epsilon - f_{D_\epsilon}(z). \tag{8}$$

It is proven [Ball, 2018] that $\sqrt{n}(n^{-1}T^{(n)} - \rho)$ converges in distribution to a normal random variable.

References

F. Ball. Central limit theorems for SIR epidemics and percolation on configuration model random graphs. 2018. URL https://arxiv.org/abs/1812.03105v1.

G. Casella and R. L. Berger. *Statistical inference*. Thomson Learning, 2nd edition, 2002. ISBN 978-0-534-24312-8.

W. Feller. *An introduction to probability theory and its applications. Vol I*. John Wiley and Sons, Inc., 2nd edition, 1964.

T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al. Jupyter notebooks—a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing*, page 87. IOS Press, 2016.

G. F. Lawler. *Introduction to Stochastic Processes*. Taylor and Francis/CRC Press, 2nd edition, 2006.

M. Newman. *Networks*. Oct 2018. ISBN 978-0-19-880509-0. doi: 10.1093/oso/9780198805090.001.0001. URL https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198805090.001.0001/oso-9780198805090.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

S. M. Ross. *Introduction to probability models*. Harcourt/Academic Press, 7th edition, 2000. ISBN 978-0-12-598475-1.

M. T. Schaub and L. Peel. Hierarchical community structure in networks. 2020. URL https://arxiv.org/abs/2009.07196.

Project proposals

Gerardo Palafox

December 15, 2020

This document presents different work avenues the author wishes to pursue as a final project for an Applied Probability Models course. Each section represents one such avenue, in which the justification for choosing such project is given.

1 Epidemic modeling

The modeling of infectious diseases is of upmost importance, as can be appreciated greatly these days. In this project, we pretend to study the stochastic spread of a contagion process. Questions that can be addressed by these studies include: What is the probability of a major outbreak? How long is the disease likely to persist? [Britton, 2010]. Both theoretical and computational results are to be presented in an attempt to bring light to these questions.

2 Random graphs

Networks arise in many scientific and technological fields [Newman, 2018]. The internet, social networks, electrical networks, are among many available examples. To study network processes, sometimes it is convenient to have a model which preserves the essential characteristics of the network. A random graph is a model network in which the values of certain properties are fixed, but the network is in other respects random [Newman, 2018]. For example, a number n nodes and m edges could be fixed, but edges between any two nodes placed at random. The aim of this project is to do a theoretical and computational study of random graphs, and analyze how closely some of these resemble real world networks [Leskovec and Krevl, 2014].

3 Community detection

The problem of finding clusters or communities in graphs is omnipresent in data mining [Alamgir and von Luxburg, 2010]. Several approaches to finding these communities involve the use of probabilistic models, such as random walks [Alamgir and von Luxburg, 2010, Pons and Latapy, 2005, Lambiotte et al., 2014, Zhang et al., 2018] or stochastic block models [Schaub and Peel, 2020]. In these project, the goal is to implement some of these methods in real world networks [Leskovec and Krevl, 2014].

4 Stock pricing models

The random-walk hypothesis states that a random walk model provides a good explanation of the variation of stock market prices [Godfrey et al., 1964]. In this project, we will explore some of these models, such as the geometric Brownian motion model [Dunbar], and compare it to the performance of real world stocks.

References

M. Alamgir and U. von Luxburg. Multi-agent random walks for local clustering on graphs. page 18–27, Dec 2010. doi: 10.1109/ICDM.2010.87. URL http://ieeexplore.ieee.org/document/5693955/.

T. Britton. Stochastic epidemic models: A survey. *Mathematical Biosciences*, 225(1):24–35, May 2010. ISSN 00255564. doi: 10.1016/j.mbs.2010.01.006.

S. R. Dunbar. Stochastic pocesses and advanced mathematical finance. Models of stock market prices. https://www.math.unl.edu/~sdunbar1/MathematicalFinance/Lessons/StochasticCalculus/StockMarketModel/stockmarketmodel.pdf.

M. D. Godfrey, C. W. J. Granger, and O. Morgenstern. The random-walk hypothesis of stock market behavior. *Kyklos*, 17(1):1–30, Feb 1964. ISSN 0023-5962, 1467-6435. doi: 10.1111/j.1467-6435.1964.tb02458.x.

R. Lambiotte, J.-C. Delvenne, and M. Barahona. Random walks, markov processes and the multiscale modular organization of complex networks. *IEEE Transactions on Network Science and Engineering*, 1(2):76–90, Jul 2014. ISSN 2327-4697. doi: 10.1109/TNSE.2015.2391998. arXiv: 1502.04381.

J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

M. Newman. *Networks*, volume 1. Oct 2018. ISBN 978-0-19-880509-0. doi: 10.1093/oso/9780198805090.001.0001. URL https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198805090.001.0001/oso-9780198805090.

P. Pons and M. Latapy. Computing communities in large networks using random walks (long version). *arXiv:physics/0512106*, Dec 2005. URL http://arxiv.org/abs/physics/0512106. arXiv: physics/0512106.

M. T. Schaub and L. Peel. Hierarchical community structure in networks. 2020. URL https://arxiv.org/abs/2009.07196.

W. Zhang, F. Kong, L. Yang, Y. Chen, and M. Zhang. Hierarchical community detection based on partial matrix convergence using random walks. *Tsinghua Science and Technology*, 23(1):35–46, Feb 2018. ISSN 1007-0214. doi: 10.26599/TST.2018.9010053.

Reseñas a proyectos

Gerardo Palafox

15 de diciembre del 2020

Las siguientes son observaciones que se le hicieron a compañeros del posgrado respecto a sus proyectos.

A Fabiola, sobre modelado multi agente de epidemias. El proyecto se ve bastante víable e interesante. Sería de provecho quizá comparar los resultados de este modelo con otro tipo de planteamientos, como las epidemias modeladas con cadenas de Markov u otros procesos estocásticos. Así mismo, explorar las capacidades del enfoque multiagente para epidemias con poblaciones no homogeneas.

A Fabiola, sobre redes bayesianas. No me queda claro en qué aspecto específico de la economía piensas aplicar el concepto de red bayesiana, ni de qué forma. Si bien son una herramienta fuerte, si no está claro en qué precisamente las usarás, te puedes perder a la hora de llevar el proyecto.

A Gabriela, sobre análisis de índices delictivos. Me parece algo general el hacer estadística descriptiva a los datos de inseguridad. ¿Tienes algunas ideas sobre qué específicamente buscas encontrar? Por ejemplo, las distintas tendencias delictivas que haya en las diversas regiones del país. Una idea que a mi me parecería interesante es buscar alguna correlación de estas tendencias con parámetros de índole social, como los que suele proporcionar el INEGI.

Submitted to the Annals of Applied Probability

STOCHASTIC EPIDEMIC MODELS

BY GERARDO PALAFOX[1],

[1]*Facultad de Ingeniería Mecánica y Eléctrica, Universidad Autónoma de Nuevo León, gerardo.palafoxcstl@uanl.edu.mx*

In this work, a stochastic model for the spread of infectious diseases is presented. Theoretical results and simulations are shown. Additionally, data from the ongoing pandemic caused by the SARS-CoV-2 virus is incorporated to discuss the implications of the model.

1. Introduction. The goal of this work is to present one of the most common stochastic epidemic models, alongside computational simulations of this. The terminology and theoretical preliminaries for later sections are detailed in Section 2. In Section 2.1, related work is presented. Models and variations are later presented in Sections 3, 4 and 5. Finally, data from the current COVID-19 pandemic, caused by the SARS-CoV-2 virus, is discussed in Section 6.

2. Preliminaries. The probability of an event A will be denoted by $\mathbb{P}\{A\}$. If X is a random variable, the expected value of X and its variance are denoted by $\mathbb{E}[X]$ and $\operatorname{Var}(X)$ respectively. A sequence of random variables $X_1, X_2, \ldots$, converges in probability to a random variable X if, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\{|X_n - X| \geq \varepsilon\} = 0 \text{ or, equivalently, } \lim_{n\to\infty} \mathbb{P}\{|X_n - X| < \varepsilon\} = 1. \tag{1}$$

A sequence of random variables $X_1, X_2, \ldots$, converges in distribution to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \tag{2}$$

at all points x where $F_X(x)$ is continuous, where F_X is the distribution of the random variable X. A stochastic process is a collection $\{X(t) : t \in J\}$ of random variables. Index t is often interpreted as time, and $X(t)$ is referred to as the state of the process at time t. The set J is called the index set of the process, and determines whether the process is discrete-time or continuous-time, depending on whether J is countable or an interval of the real line. A Markov chain is a stochastic process where the random variables take a finite or countable number of possible values, and with the property that

$$\mathbb{P}\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} = \mathbb{P}\{X_{n+1} = j \mid X_n = i\}, \tag{3}$$

for any such states i_k and any $n > 0$.

2.1. *Related Work.* The mathematical study of epidemics has been long studied, with early models focusing on specific diseases [5, 23]. More general studies have been performed since then, being Kermack and McKendrick [16] some of the first in doing so, and Bailey [2], Frauenthal [13], Brauer and Castillo-Chávez [6] more recently. This work will focus on models with stochastic components [1, 7], but deterministic models have also been studied [14]. Models where the disease spreads over a network have been studied thoroughly, both in their deterministic [20] and stochastic [8, 26] form. Agent-based simulations have also been used to model epidemics [15].

MSC2020 subject classifications: Primary 60J05, 60J27; secondary 92D30.
Keywords and phrases: Stochastic processes, Epidemics.

3. Reed-Frost model. One of the simplest stochastic models of disease spreading is the Reed-Frost model, which is included in this work for reference. This model is a discrete-time Markov chain, were an individual in a finite population is classified as susceptible, infected or removed; these are called SIR models. The model in Section 4 is also an SIR model. At time t, there are S_t individuals susceptible to infection, and I_t currently infectious. The incubation period of the disease, as well as the recovery time, is assumed to occur between time jumps. That is, susceptible people exposed at time t will be infectious at time $t+1$, and people infectious at time t will be removed from the process at time $t+1$. Any two individuals have a probability $p = 1 - q$ of coming in contact at time t, with encounters being independent of each other. As such, the distribution of S_{t+1} is binomial $\mathrm{Bin}(s, q^i)$. Observe that $I_{t+1} = S_t - S_{t+1}$, since in the step from t to $t+1$, all the infectious from time t are removed, and the decrease in susceptible individuals is precisely the amount of new infectious people. Thus, $I_{t+1} \sim \mathrm{Bin}(s, 1-q^i)$. More of this process and its asymptotic properties can be read in a work by Von Bahr and Martin-Löf [27].

4. Standard stochastic epidemic model. The construction and details of this model can be found in the book of Andersson and Britton [1]. A description of the properties presented in Subsection 4.1, among more, can be found in a survey by Britton [7]. A fixed population of size n will be considered. It is assumed there are no births, deaths or migration; this is what is known as a closed population. Additionally, population is assumed well-mixed and homogeneous: this means any two elements of the population can be in contact with each other, and everyone is affected by the disease in the same way. Each individual will be either susceptible (S), infectious (I), or recovered (R, also called removed). Let $S(t), I(t), R(t)$ be the number of susceptible, infectious and recovered individuals at time t. At $t = 0$, one has $S(0) = n - m, I(0) = m, R(0) = 0$. The contacts an infectious individual has with others follows a Poisson process with intensity λ [1, 18]. Each such contact is with an individual selected uniformly at random from the population, with contacts of different infectious individuals being mutually independent. When the contact occurs between an infectious individual and a susceptible individual, the susceptible individual is assumed to be infected, and it is immediately infectious. Every infectious individual remains so for a random time $I \sim F_I$, called the *infectious period*, where F_I has mean $1/\gamma$. These infectious periods are independent and identically distributed, and are independent from the contact process. Epidemic starts at time 0, and goes until a time T where $I(T) = 0$. Thus the final state of the epidemic is given by $S(T) = n - R(T), I(T) = 0$ and $R(T) = m + Z$; i.e., Z people where infected in the course of the epidemic. While, regardless of the distribution F_I, there is no closed form expression for the time dynamics of the process, it is possible to derive a recursive formula for the final size of the epidemic, see the survey of Britton [7]. However, this becomes difficult for even moderately large n [7].

When $F_I \sim \mathrm{Exp}(\gamma)$, the model becomes a continuous-time Markov chain, known as the *stochastic general epidemic model*, which was first considered by Bartlett [4]. There is no epidemiological reason to assume the infectious period follows an exponential distribution, but doing so is of mathematical convenience, as it allows diffusion approximations valid for large values of n [1]. If the infectious period is not random but constant, the resulting model is a continuous-time version of the Reed-Frost model [1] (see Section 3 for a definition of this model).

4.1. *Properties of the model.* The aim of this section is to explore key questions revolving around an epidemic process, such as the probability of having a major outbreak, or how many people will be infected in such an outbreak. An approximation of the early stages of the epidemic process is also considered. Everything presented will be independent of the distribution F_I, i.e. it will not rely on the Markovian case.

Britton [7] details how, when n is sufficiently large, the initial phase of the epidemic can be approximated by a homogeneous branching process with birth-rate λ and life-distribution $I \sim F_I$ [18] having Laplace transform $\phi(s)$. The epidemic and branching process will agree at least until there has been k contacts, with k small in relation to $\sqrt{n}$. The offspring distribution of the branching process has mean $\lambda \mathbb{E}[I] = \lambda/\gamma$ [7], a quantity denoted by R_0 from here on. Also mentioned by Britton [7] is the fact that when $R_0 \leq 1$ the final size of the epidemic is bounded in probability, and it is not when $R_0 > 1$. Additionally, if $R_0 > 1$, the epidemic will be minor with probability q^m and mayor with probability $\rho = 1 - q^m$, where q is the smallest solution to

$$q = \phi(\lambda(1-q)). \tag{4}$$

A more detailed treatment of the properties just stated can be found in the work of Ball [3]. The number R_0 is known in the literature as the *basic reproduction number*, and, as it is seen, plays an important role in whether the epidemic dies out "rapidly"or there is a major outbreak.

A rigorous study of the final size of the epidemic can be found in the work of Scalia-Tomba [24, 25]. Britton [7] gives a balance equation to find the fraction of people not getting infected. Particularly, one can find in his work [7] that if z is the fraction of people infected, by the law of large numbers the limiting fraction infected in case of a major outbreak is a solution to the equation

$$1 - z = e^{-R_0 z}. \tag{5}$$

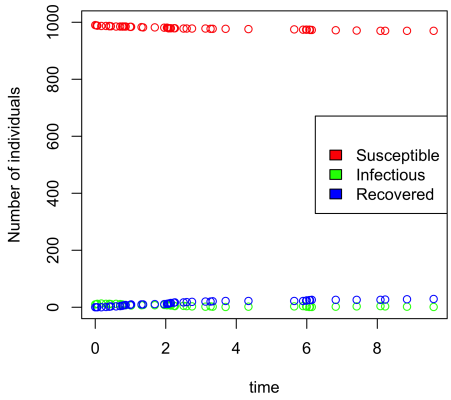
The value $z = 0$, which corresponds to a minor outbreak, is always a solution to this equation. Also, when $R_0 > 1$, there exists an additional unique solution z^* in the unit interval. Furthermore, denoting by Z_n the final number of infected individuals for a population of size n, one has that if $R_0 \leq 1$, then $\bar{Z}_n := Z_n/n \to 0$ in probability as $n \to \infty$ [7]. If, however, $R_0 > 1$, then $\bar{Z}_n \to \zeta$, where ζ is a two point distribution defined by $\mathbb{P}\{\zeta = 0\} = q^m$, $\mathbb{P}\{\zeta = z^*\} = 1 - q^m$ [7], with q as seen in Equation 4 and z^* being the positive solution to Equation 5 previously mentioned. Defining $r^2 = \mathrm{Var}(I)/(\mathbb{E}[I])^2$, when $\bar{Z}_n \to z^*$ one has

$$\sqrt{n}(\bar{Z}_n - z^*) \to \mathcal{N}\left(0, \frac{z^*(1-z^*)(1 + r^2(1-z^*)R_0^2)}{(1-(1-z^*)R_0)^2}\right), \tag{6}$$

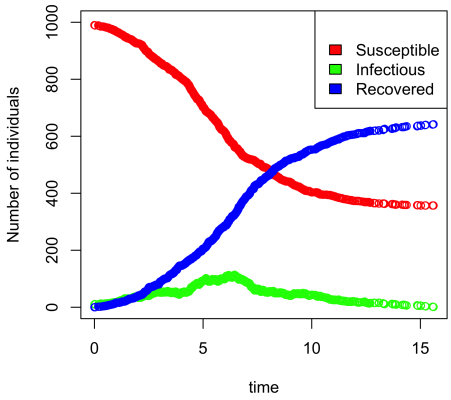
where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution of mean μ and variance σ^2.

4.2. *Simulations.* The code for the simulations in this and further sections is publicly available in a GitHub repository[1]. The simulations were performed using Gillespie's direct method algorithm [11]. In Figure 1, two trajectories of the stochastic general epidemic model are shown. Both consist of a population of 1000 with 10 initial infected. The trajectory in Figure 1a has an $R_0 = .5$ and the one in Figure 1b has an $R_0 = 1.5$. In Figure 2, final size distributions of these models are displayed; Figure 2a shows the distribution of the final size of the model with $R_0 = .5$, while Figure 2b shows the distribution of the final size of the model with $R_0 = 1.5$.

[1] https://github.com/palafox794/AppliedProbabilityModels/tree/master/FinalProject

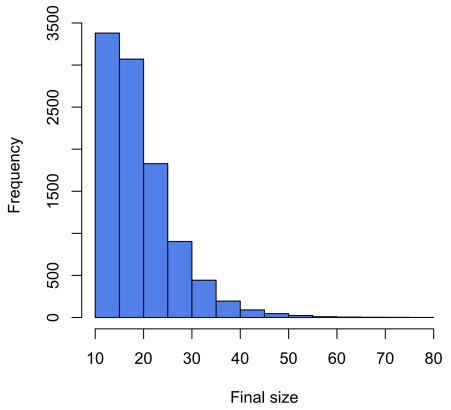


(a) Model with $R_0 = 0.5$.

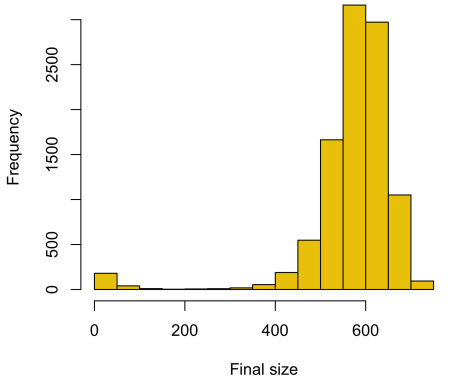


(b) Model with $R_0 = 1.5$.

Fig 1: Trajectories of a stochastic general epidemic model.



(a) Final size distribution, $R_0 = 0.5$.



(b) Final size distribution, $R_0 = 1.5$.

Fig 2: Final size distribution of 10,000 simulations in a general stochastic epidemic model.

5. Network models. One of the assumptions of the standard epidemic model, is that any two individuals have an equal probability of making contact. While this makes it simpler, it is of course less realistic. In reality, some people are more prone to being exposed (hospital or school workers, students), and more likely to infect those who are in close contact with them (coworkers, household members). A way to model this situations is to consider the members of the population as nodes in a network (see the book of Newman [21] for basic network definitions). Such epidemic models abound in the scientific literature [10, 12, 17, 22]. As an example, we define two such models, as they appear in the article of Britton [8]. The interested reader is invited to read the survey by Pastor-Satorras et al. [22] or the work of Britton [8].

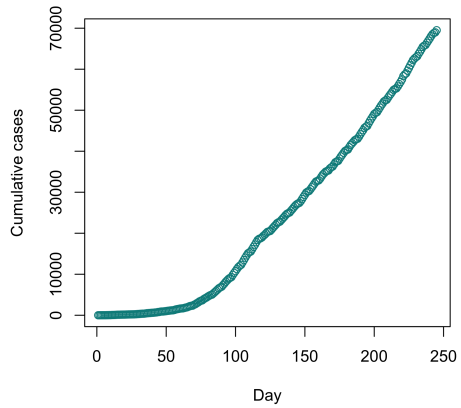
DEFINITION 5.1 (Discrete time Reed-Frost model on a network). This model is a network generalization of the one described in Section 3. Having an arbitrary network, a randomly chosen node is set as infectious at time $t = 0$, while the rest are susceptible. At time t, infectious individuals will infect susceptible neighbors (i.e., adjacent nodes) independently with probability p. The susceptible individuals that where infected at time t will be infectious at time $t + 1$, and infectious individuals at time t will be removed at time $t + 1$. This goes on until a time T when there are no new infections.

DEFINITION 5.2 (Continuous-time Markov chain model on a network). This model is a network generalization of the stochastic general epidemic model described in Section 4. Having an arbitrary network, a randomly chosen node is set as infectious at time $t = 0$, and the rest are susceptible. Infectious individuals have contacts with each susceptible neighbor (i.e. adjacent nodes) randomly in time according to an independent Poisson processes with rate λ [18]. Each infected individual remains infectious for a period $I \sim \mathrm{Exp}(\gamma)$, after which it is removed from the epidemic (considered recovered and immune). Both infectious periods and contact processes are defined independently. This goes on until a time T when there are no new infectious nodes.

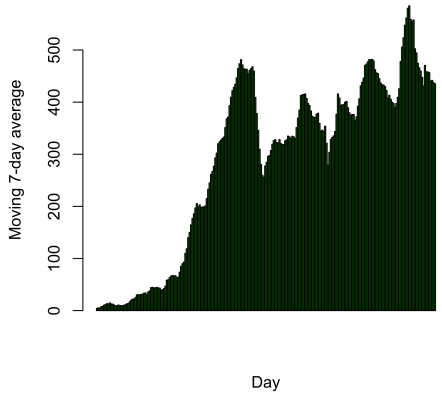
As with the standard epidemic model in Section 4, the distribution of the infectious periods can be chosen arbitrarily, at the expense of the Markovian property. Similarly, setting a constant infectious period gives rise to a continuous-time Reed-Frost model [8].

6. COVID-19 pandemic. At the moment this work is being written, a pandemic caused by the SARS-CoV-2, colloquially known as COVID-19, is widespread in most countries of the world. Various authors have tried to estimate this disease's R_0; a review by Liu et al. [19] has found that the mean R_0 estimated is of 3.28, with a median of 2.79. In the Mexican northern state of Nuevo León, the first cases were reported in March 2020. In Figure 3, two graphics are shown regarding the spread of the epidemic process in Nuevo León from April 1st, 2020 to November 30th, 2020, with information provided by the Mexican government [9]. In particular, a graph of the cumulative cases in the state is shown in Figure 3a. A 7-day rolling average of cases per day is seen in Figure 3b. In a stochastic general epidemic model, as the presented in Section 4, an R_0 of 3 would give a limiting fraction of .94 infected in the population, by solving Equation 5. This can be seen in the trajectory in Figure 4, a simulated trajectory of an epidemic in a population of 1000 which has $R_0 = 3$. This of course should not be interpreted as a suggestion that the current pandemic will infect 94% of the population, since the model is quite weak in its assumptions. It is, however, a simplified example of how quickly these processes can grow out of control when there is nothing stopping its spread.

7. Conclusions. In this work, two common epidemic models were presented, as well as generalizations to network models. In particular, the standard stochastic epidemic model was studied in more detail. Simulations were performed and displayed. Finally, some discussion of how these relate to a current, ongoing pandemic was presented. The standard stochastic epidemic model can be generalized beyond its SIR presentation. Models for further study include those were there is an incubation period after infection, before becoming infectious (SEIR) or those with waning immunity (SIRS). Models with dynamic population, taking into account births and deaths or migration, were also lacking in this work. While the network models in Section 5 remove the well-mixed assumption of the population, individuals are still homogeneous. Multi-type models [1] address this limitations, and can be studied in further work.



(a) Cumulative COVID-19 cases in Nuevo León.



(b) Rolling 7-day average of COVID-19 cases in Nuevo León.

Fig 3: COVID-19 cases in the state of Nuevo León from 2020-04-01 to 2020-11-30.

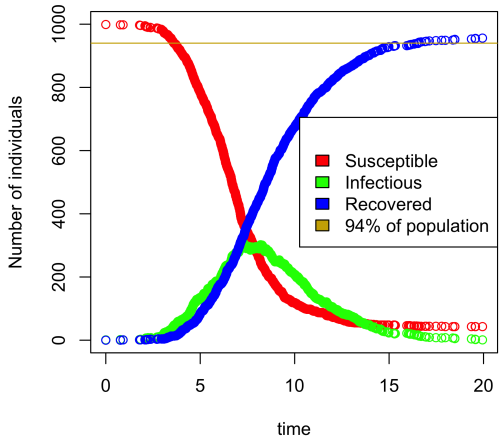


Fig 4: General stochastic model with $R_0 = 3$.

Acknowledgements. The author thanks the financial support of CONACyT, and the suggestion of Fabiola Vázquez to incorporate in this work data from a real epidemic.

REFERENCES

[1] ANDERSSON, H. and BRITTON, T. (2000). *Stochastic epidemic models and their statistical analysis*. *Lecture notes in statistics*. Springer.

[2] BAILEY, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications*, 2nd ed. Griffin.

[3] BALL, F. (1983). The threshold behaviour of epidemic models. *Journal of Applied Probability* **20** 227–241.

[4] BARTLETT, M. S. (1949). Some Evolutionary Stochastic Processes. *Journal of the Royal Statistical Society: Series B (Methodological)* **11** 211–229.

[5] BERNOULLI, D. (1760). Essai d'une nouvelle analyse de la mortalite causee par la petite verole, et des avantages de l'inoculation pour la prevenir. *Histoire de l'Acad., Roy. Sci. (Paris) avec Mem* 1–45.

[6] BRAUER, F. and CASTILLO-CHÁVEZ, C. (2012). *Mathematical models in population biology and epidemiology*, 2nd ed. *Texts in applied mathematics*. Springer.

[7] BRITTON, T. (2010). Stochastic epidemic models: A survey. *Mathematical Biosciences* **225** 24–35.

[8] BRITTON, T. (2019). Epidemic models on social networks – with inference. arXiv:1908.05517.

[9] GOBIERNO DE MÉXICO Covid-19 México. https://datos.covid-19.conacyt.mx/#DownZCSV.

[10] DEIJFEN, M. (2011). Epidemics and vaccination on weighted graphs. arXiv:1101.4154.

[11] DRAKE, J. M. and ROHANI, P. Stochastic Models. https://daphnia.ecology.uga.edu/drakelab/wp-content/uploads/2016/07/sismid-stochastic-lecture.pdf.

[12] FRANSSON, C. and TRAPMAN, P. (2019). SIR epidemics and vaccination on random graphs with clustering. *Journal of Mathematical Biology* **78** 2369–2398.

[13] FRAUENTHAL, J. C. (1980). *Mathematical modeling in epidemiology*. *Universitext*. Springer.

[14] HETHCOTE, H. W. (2000). The Mathematics of Infectious Diseases. *SIAM Review* **42** 599–653.

[15] HOERTEL, N., BLACHIER, M., BLANCO, C., OLFSON, M., MASSETTI, M., RICO, M. S., LIMOSIN, F. and LELEU, H. (2020). A stochastic agent-based model of the SARS-CoV-2 epidemic in France. *Nature Medicine* **26** 1417–1421.

[16] KERMACK, W. O. and MCKENDRICK, A. G. (1927). A contribution to the mathematical theory of epidemics. **115** 700–721.

[17] KISS, I. Z., MILLER, J. C. and SIMON, P. L. (2017). *Mathematics of Epidemics on Networks: From Exact to Approximate Models*, 1st ed. *Interdisciplinary Applied Mathematics*. Springer International.

[18] LAWLER, G. F. (2006). *Introduction to Stochastic Processes*, 2nd ed. Taylor and Francis/CRC Press.

[19] LIU, Y., GAYLE, A. A., WILDER-SMITH, A. and ROCKLÖV, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine* **27**.

[20] MEI, W., MOHAGHEGHI, S., ZAMPIERI, S. and BULLO, F. (2017). On the dynamics of deterministic epidemic propagation over networks. *Annual Reviews in Control* **44** 116–128.

[21] NEWMAN, M. (2018). *Networks* **1**. Oxford University Press.

[22] PASTOR-SATORRAS, R., CASTELLANO, C., VAN MIEGHEM, P. and VESPIGNANI, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics* **87** 925–979.

[23] ROSS, R. (1910). *The Prevention of Malaria*. J. Murray.

[24] SCALIA-TOMBA, G. (1985). Asymptotic final-size distribution for some chain-binomial processes. *Advances in Applied Probability* **17** 477–495.

[25] SCALIA-TOMBA, G. (1990). On the asymptotic final size distribution of epidemics in heterogeneous populations. In *Stochastic Processes in Epidemic Theory* (J.-P. GABRIEL, C. LEFÈVRE and P. PICARD, eds.) 189–196. Springer Berlin Heidelberg.

[26] TRAPMAN, P. and BOOTSMA, M. C. J. (2009). A useful relationship between epidemiology and queueing theory: The distribution of the number of infectives at the moment of the first detection. *Mathematical Biosciences* **219** 15–22.

[27] VON BAHR, B. and MARTIN-LÖF, A. (1980). Threshold limit theorems for some epidemic processes. *Advances in Applied Probability* **12** 319–349.