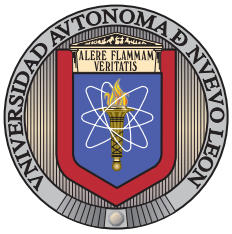
Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Posgrado en Ingeniería de Sistemas

Doctorado



Portafolio de Evidencias

de

Gabriela Sánchez Yepez

1935064

para el curso de Modelos Probabilistas Aplicados,

con la profesora Dra. Satu Elisa Schaeffer.

Semestre Agosto 2020 - Enero 2021.

https://github.com/Saphira3000/MPA

Índice general

Tarea 1
Gabriela Sánchez Y.

En esta actividad se realiza un análisis descriptivo de la prevalencia delictiva en las entidades federativas del país en el periodo 2010-2018.

1. Datos

Los datos utilizados para el análisis se obtienen de la página del INEGI en la sección correspondiente a Seguridad pública y justicia [2]. En este trabajo, únicamente se analiza la información acerca de la tasa de prevalencia delictiva por cada cien mil habitantes de cada una de las 32 entidades federativas del país, en el periodo comprendido entre los años 2010 a 2018.

Los datos se guardan en un archivo CSV y el análisis descriptivo se realiza con la ayuda del lenguaje de programación R versión 4.0.2 [1].

2. Diagramas caja bigote

En la figura 1 se observan los diagramas de caja bigote que resumen la información de la tasa de prevalencia delictiva por cada cien mil habitantes en cada una de las entidades federativas en el periodo 2010-2018. Es de esperarse que el Estado de México presente una mediana mayor al resto ya que es de los estados más poblado del país, le sigue la Ciudad de México y Baja California Norte.

El INEGI también proporciona información de la tasa de prevalencia delictiva por cada cien mil habitantes según el sexo de la víctima, por lo que también es recopilada esa información para los tres estados con mayor tasa delictiva (Estado de México, Ciudad de México y Baja California Norte). Esta última información se recopila directamente en un script de R (T1.R).

Se realiza primero un análisis individual para el Estado de México y, de acuerdo a lo mostrado en la figura 2 se puede concluir que, en promedio, la tasa de prevalencia delictiva en hombres es mayor en el caso de las mujeres.

Finalmente, en el diagrama mostrado en la figura 3, se observa la tasa de prevalencia según el sexo. Podemos notar que en el estado de Baja California Norte, en promedio, la mayoría de las víctimas fueron mujeres.

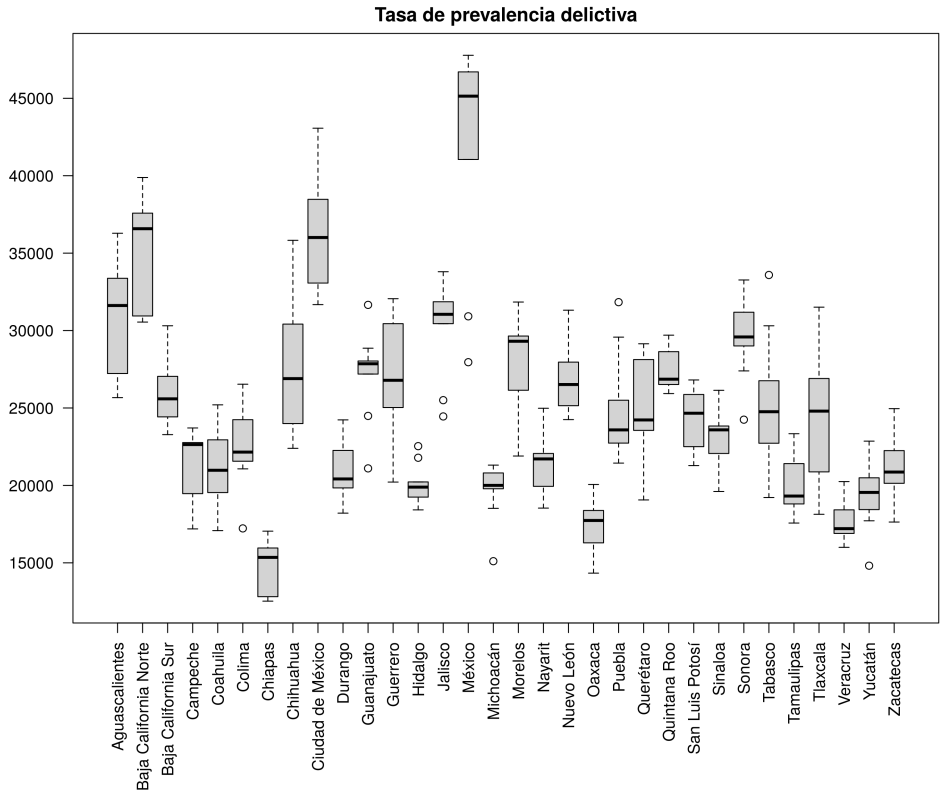


Figura 1: Diagrama caja bigote de la tasa de prevalencia delictiva por cada cien mil habitantes en el periodo 2010-2018.

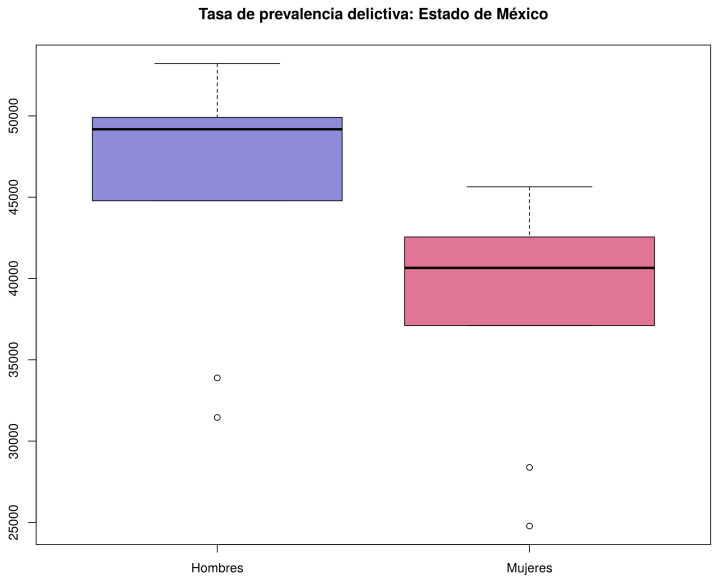


Figura 2: Diagrama caja bigote de la tasa de prevalencia delictiva por cada cien mil habitantes, en el periodo 2010-2018, para el Estado de México según el sexo.

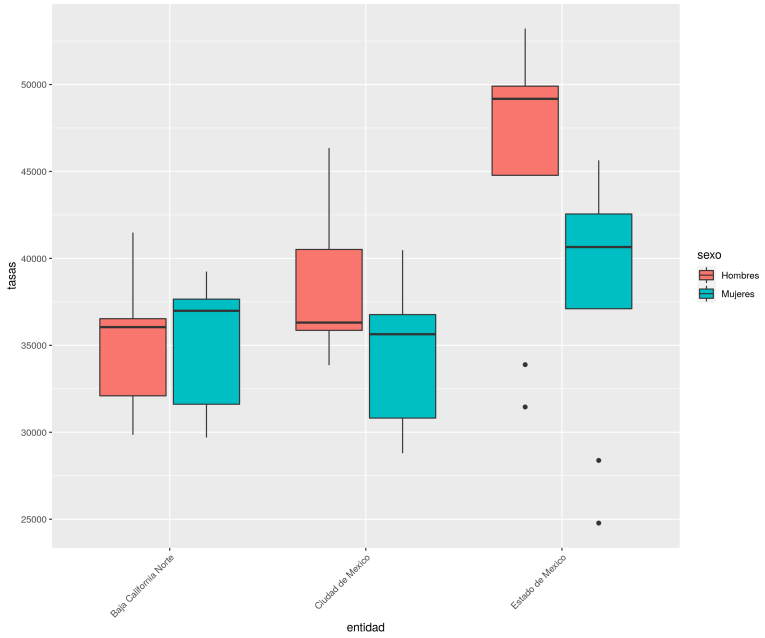


Figura 3: Diagrama caja bigote de la tasa de prevalencia delictiva por cada cien mil habitantes, en el periodo 2010-2018, según el sexo del Estado de México, Ciudad de México y Baja California Norte.

Referencias

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

[2] Instituto Nacional de Estadística y Geografía. Seguridad pública y justicia: victimización. `https://www.inegi.org.mx/temas/victimizacion/`.

Frecuencias

Gabriela Sánchez Y.
5064

En el presente trabajo se realiza un análisis del libro *"Anne of Green Gables"* obtenido del sitio de Project Gutenberg.

1. Introducción

El análisis del libro de texto se realiza con la ayuda del lenguaje de programación R versión 4.0.2 [3], haciendo uso de tres librerías: `gutenbergr` que permite acceder al texto plano del libro y, `tidytext` y `dplyr` que permiten la descomposición del texto.

Dicho análisis se basa en el estudio de las frecuencias de las palabras y letras del texto. El primer paso para poder proceder con el estudio es obtener el texto plano del libro, lo cual es posible mediante la función `gutenberg_download`.

El procedimiento realizado para el análisis puede encontrarse en el script `t2.R` [1].

2. Letras

La descomposición del texto en caracteres se realiza con la función `unnest_tokens`. Ya que es de interés únicamente la frecuencia de las letras, se eliminan todos los caracteres que no lo son. En este caso, el único caracter que no es una letra es "|".

Para mejorar la visualización, las frecuencias se ordenan en forma decreciente. De esta manera es posible observar que las primeras tres letras más usadas en el texto son *e, t* y *a*, mientras que las menos usadas son *x, q* y *z*, tal y como se muestra en la figura 1.

3. Palabras

La descomposición en letras no dice mucho acerca del texto por lo que se procede a realizar una descomposición en palabras. Para esto, nuevamente se usa la función `unnest_tokens`.

Como segundo paso en este análisis, se realiza un filtrado: son eliminadas aquellas palabras que en inglés se conocen como *stop words* (palabras vacías), ya que no serán útiles para el estudio [2]. Son palabras muy comunes en el idioma que pueden eliminarse sin sacrificar el significado de una oración. En inglés algunos ejemplos de palabras vacías son *at, the, is, of, to*.

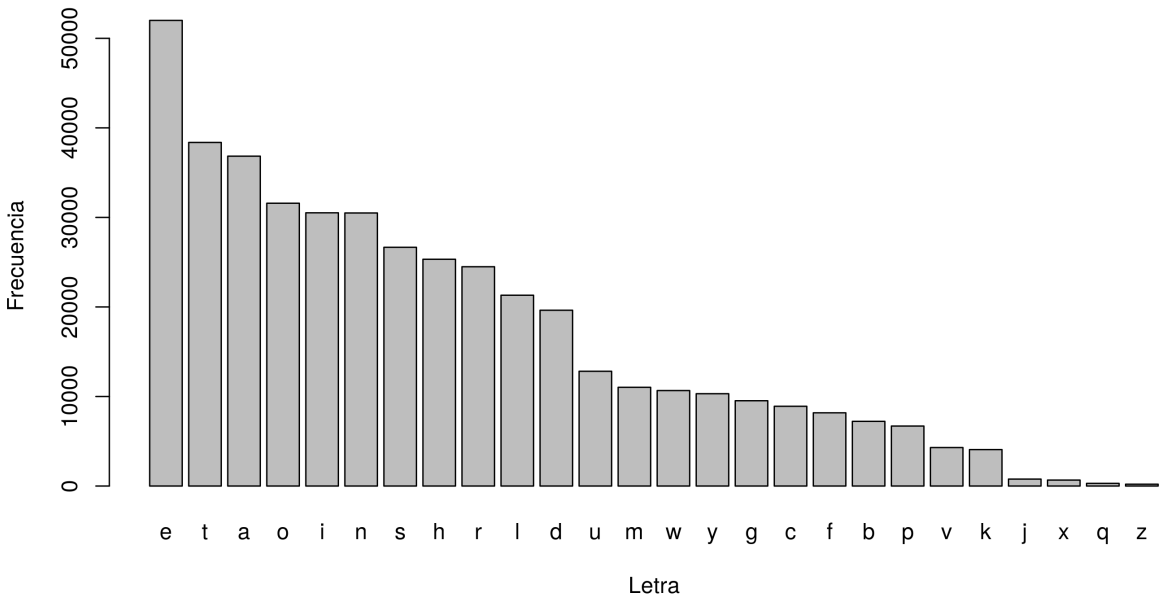


Figura 1: Gráfico de barras de la frecuencia de las letras del abecedario en el texto analizado.

Una vez hecho este filtro, se aplica otro que toma en cuenta únicamente las palabras con una frecuencia mayor a uno. En el cuadro 1 se pueden observar las primeras 10 palabras más frecuentes. Estos resultados permiten inferir que *Anne, Marilla, Diana* y *Matthew* son personajes principales en la novela, siendo *Anne* el principal ya que la frecuencia está muy por encima de los otros.

Continuando con las palabras más frecuentes, en la figura 2 se muestra un gráfico de barras que muestra la frecuencia de las 20 palabras siguientes en frecuencia a las del cuadro 1. Pueden observarse otros nombres como *Gilbert y Jane* y, lo que parecen ser apellidos *Lynde* y *Barry*, por lo que podríamos decir que son personajes secundarios en la novela.

Una persona que ya ha leído el libro sabe que *Barry* es el apellido de *Diana*. Esto advierte

Cuadro 1: Palabras más comunes y su frecuencia.

Palabra	Frecuencia
anne	1107
marilla	797
diana	386
matthew	339
time	178
girl	170
school	152
miss	148
home	144
white	142

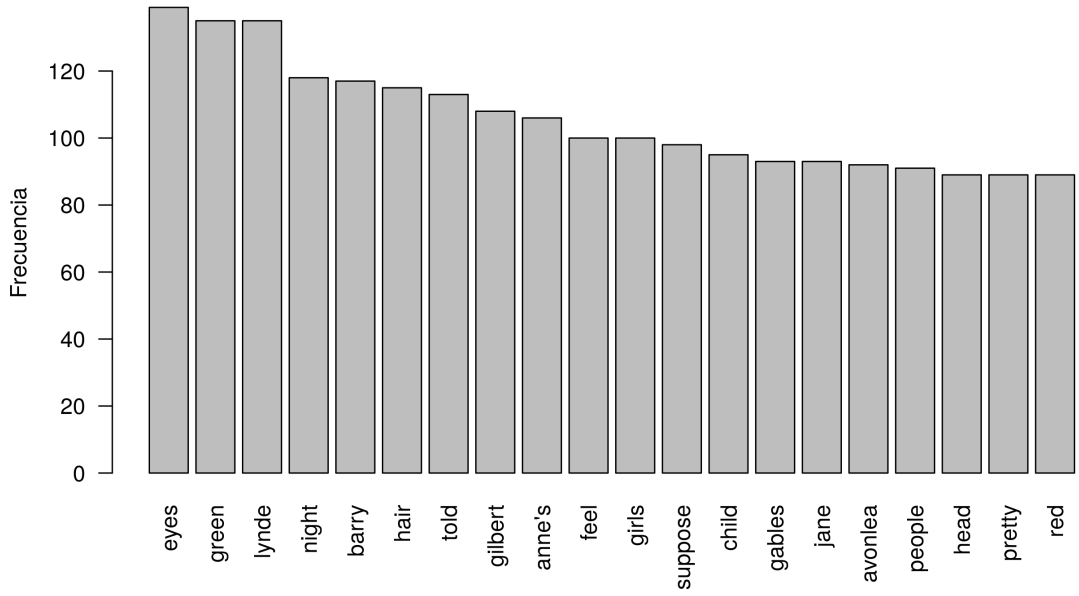


Figura 2: Gráfico de barras de las palabras más frecuentes en el texto, una vez realizados dos filtros.

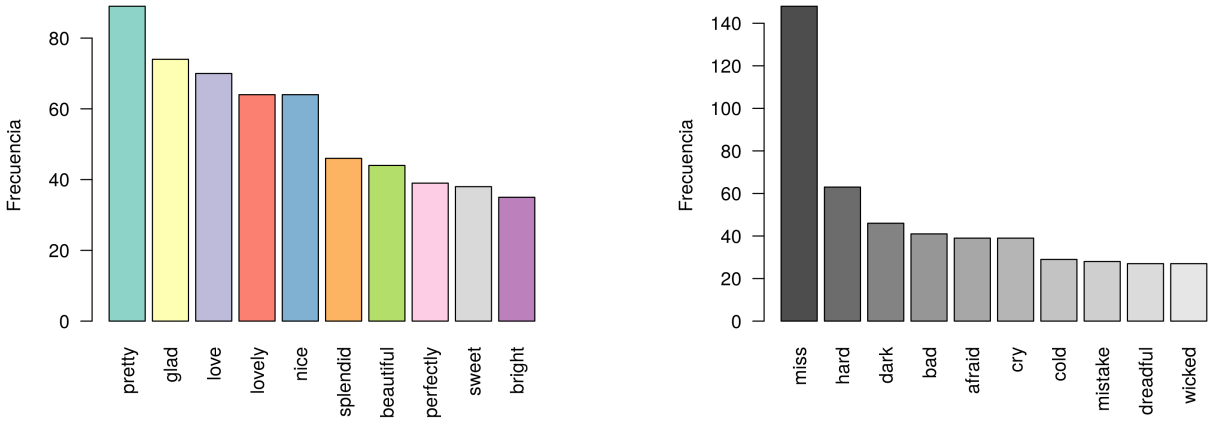
que un análisis de palabras "sueltas" no permite inferir mucho acerca del contenido del texto, una mejor opción es considerar la frecuencia en que aparecen dos o más palabras juntas.

Antes de proceder a analizar conjuntos de palabras, se examinan las palabras positivas y negativas más comunes de acuerdo al léxico `bing`, que es posible obtener a través de la función `get_sentiments`. La figura 3 muestra las 10 palabras más comunes positivas (figura 3a) y negativas (figura 3b).

En el gráfico de barras mostrado en la figura 3b se puede apreciar que la palabra negativa con más frecuencia en el libro es *miss*. Sin embargo, aquí se señala un problema. Se sabe que tanto en español como en inglés existen palabras homónimas, es decir, palabras cuya pronunciación o escritura es igual o similar pero que tienen diferente significado. En este caso, personas que ya han leído el libro, pueden advertir que el personaje principal se refiere a su profesora como *Miss Stacy*, otro personaje relevante en la novela. por lo que no se garantiza que *miss* en su connotación "negativa" realmente tenga una frecuencia más alta que las otras palabras mostradas en el gráfico.

Por último, se analiza la frecuencia en que aparecen las secuencias de dos palabras, también llamado bigrama. El análisis de los brigramas se realiza con los datos que previamente filtraron las palabras vacías, ya que si se omite ese paso, los resultados obtenidos muestran serán los que se muestran en el cuadro 2 y dicha información no aporta sobre el contenido del libro.

En la figura 4 se observan los bigramas más comunes en el texto filtrado. Resalta el nombre de *Green Gables*, lugar donde se desarrolla la mayoría de la trama de la novela y nombres de personajes como *Miss Stacy, Gilbert Blythe, Anne Shirley, Rubby Gillis,* entre otros.



(a) Palabras positivas

(b) Palabras negativas

Figura 3: Gráfico de barras de las palabras positivas y negativas más frecuentes.

Cuadro 2: Bigramas más frecuentes en el texto sin filtro.

Bigrama	Frecuencia
in the	413
to be	325
of the	308
it was	273
to the	240
and i	198
on the	191
i don't	174
going to	165
was a	164

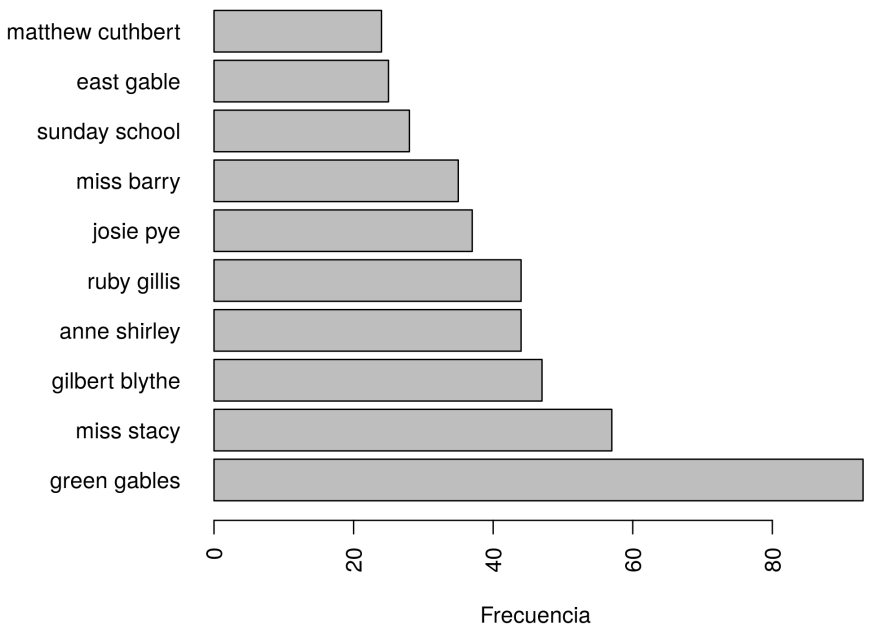


Figura 4: Bigramas con mayor frecuencia en el texto filtrado.

Referencias

[1] Gabriela Sánchez Y. Modelos Probabilistas Aplicados. https://github.com/Saphira3000/MPA.

[2] Julia Silge and David Robinson. Text Mining with R. https://www.tidytextmining.com/tidytext.html.

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

Distribuciones de probabilidad geométrica, binomial y binomial negativa.

Gabriela Sánchez Y.
5064

En el presente trabajo se analiza el tipo de distribuciones que pueden están presentes en el texto del libro *"Anne of Green Gables"* [4] que puede obtenerse de manera gratuita en el sitio de Project Gutenberg.

1. Introducción

El procesamiento del texto se realiza con la ayuda del lenguaje de programación R versión 4.0.2 [2], haciendo uso de distintas librerías:: `gutenbergr` que permite acceder al texto plano del libro y, `tidytext` y `dplyr` que permiten la descomposición del texto.

En el preprocesamiento se eliminan las líneas vacías, el índice y los guiones bajos cuando hay palabras con guiones bajos alrededor para indicar énfasis [6]. Para tener un punto de comparación sobre lo que es propio del autor en la escritura de un texto se elige un cuento cuya fecha de publicación no se aleja mucho de la del libro base. En este caso se utiliza el libro *Peter Pan* [1] que se descarga del mismo sitio [3].

El procedimiento realizado para el análisis puede encontrarse en el código `t3.R` [7].

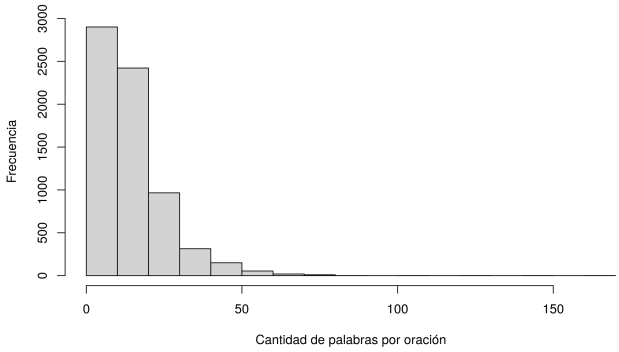
2. Distribuciones en el texto

En esta fase se descompone el texto en oraciones, es decir, se parte después de cada punto. La comparación se realiza a partir de la medición de la cantidad de palabras, y comas que hay en las oraciones de cada texto. Además se analiza la longitud de las palabras usadas.

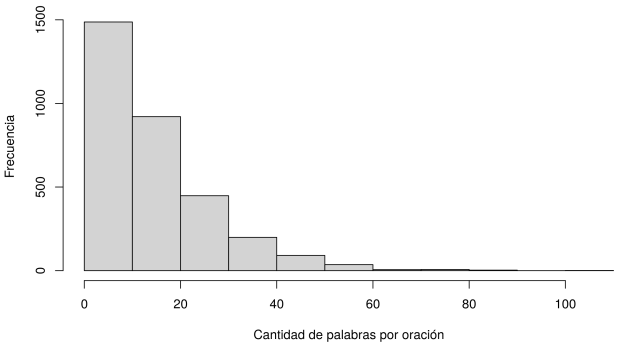
En la figura 1 se puede observar una síntesis de cómo están distribuidos estos elementos en las oraciones de cada texto. En general, tienen un comportamiento similar. Se podría decir que la distribución de la cantidad de palabras por oración en ambos textos (figura 1a y 1b) y la distribución de la cantidad de comas por oración (figura 1c y 1d) es muy similar a una distribución geométrica.

2.1. Experimentos de Bernoulli

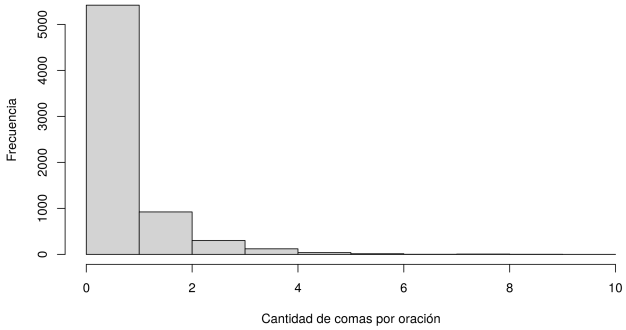
Para esta fase se usa solo el texto de *Anne of Green Gables*. A partir de esa información se realizan distintos experimentos de Bernoulli que buscan imitar el comportamiento de tres distri-



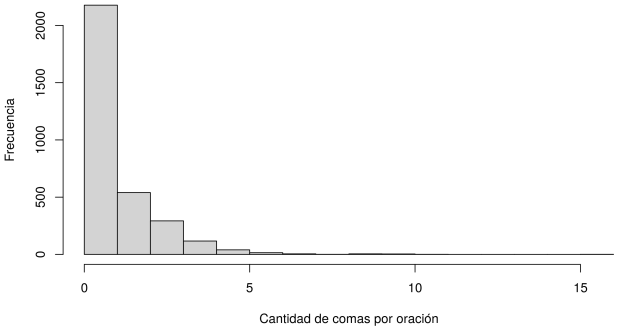
(a) Palabras por oración en *Anne of Green Gables*



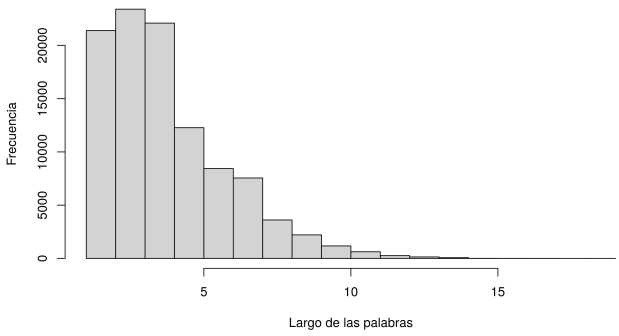
(b) Palabras por oración en *Peter Pan*



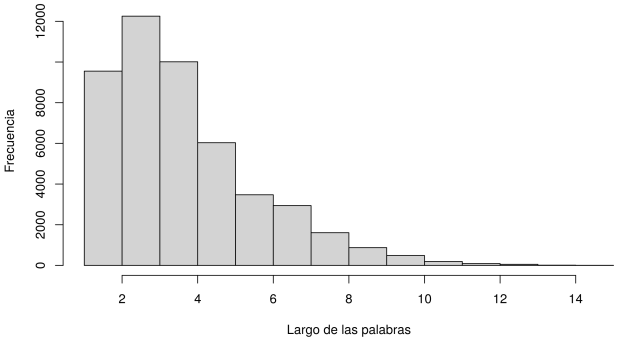
(c) Comas por oración en *Anne of Green Gables*



(d) Comas por oración en *Peter Pan*



(e) Largo de palabras en *Anne of Green Gables*



(f) Largo de palabras en *Peter Pan*

Figura 1: Síntesis de la distribución de las oraciones en los textos.

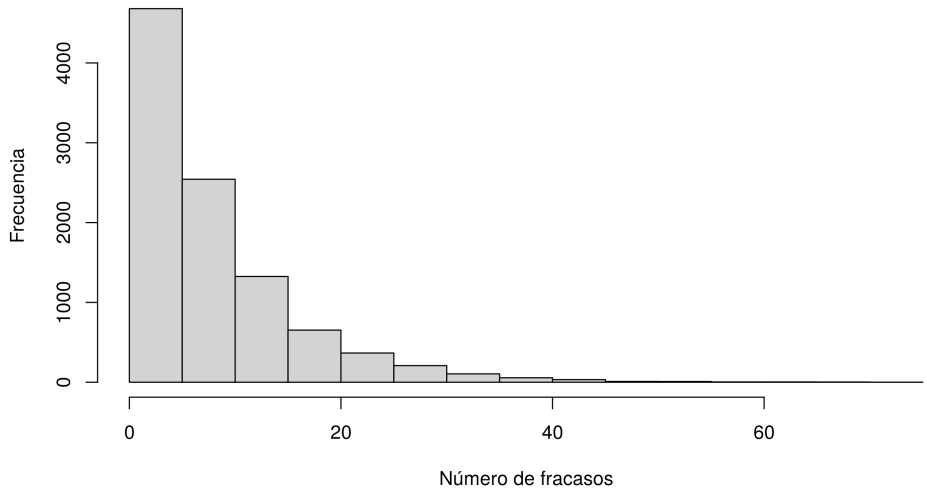


Figura 2: Distribución geométrica.

buciones: geométrica, binomial y binomial negativa.

La distribución geométrica muestra la distribución de la cantidad de repeticiones del experimento que fueron fracasos antes de obtener el primer éxito [5]. Para replicar este comportamiento se define que el éxito se obtiene cuando al elegir al azar una letra del texto esta es una "e", por lo tanto, si se obtiene cualquier otro caracter se considera un fracaso. Al repetir 10,000 veces el experimento se obtienen los resultados de la figura 2. Por la construcción del experimento, es claro que el comportamiento sigue una distribución geométrica.

El siguiente experimento contesta a la pregunta ¿cuántos intentos se necesitan para obtener siete veces el caracter "t", en una elección al azar?. Un éxito, para este experimento, se considera cuando en la elección se obtiene el caracter "t". Esta distribución se conoce como binomial negativa. En la figura 3 se muestran los resultados obtenidos al repetir 10,000 veces el experimento.

La última situación considera un experimento de 30 intentos que se replica 1,000 veces. En cada réplica se cuenta la cantidad de veces que se elige al azar una palabra y ésta tiene una longitud menor o igual a cinco caracteres. La distribución obtenida de este experimento se muestra en la figura 4. La construcción del experimento permite concluir que este comportamiento es el de una distribución binomial.

La distribución en la figura 1e indica que la mayoría de las palabras en el texto tiene una longitud menor o igual a cinco caracteres. De esta forma en el experimento de elegir 30 palabras al azar, se espera que la mayoría de ellas tenga esta restricción en la longitud es por esto que se observa el comportamiento de la figura 4.

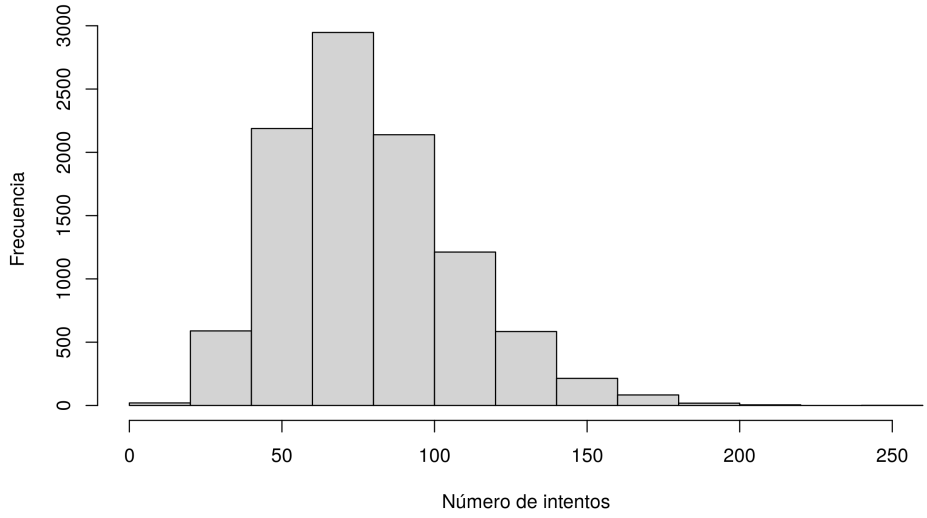


Figura 3: Distribución binomial negativa.

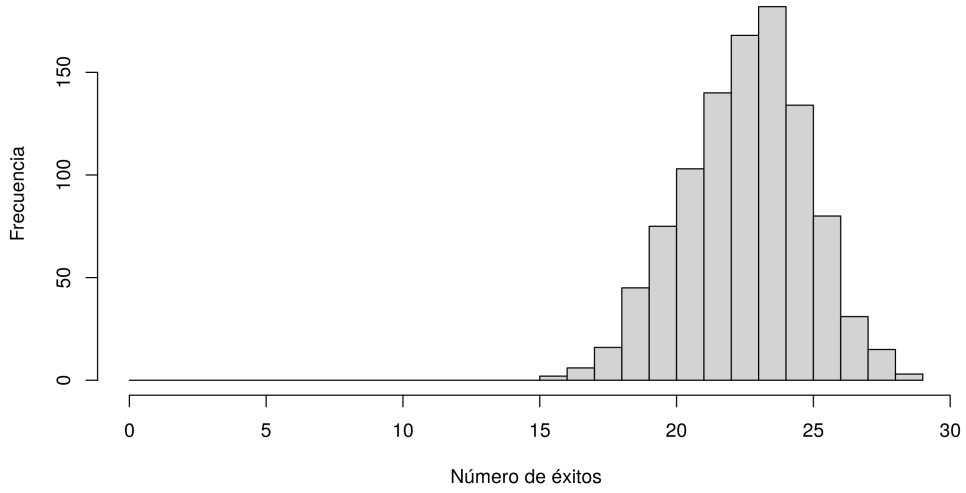


Figura 4: Distribución binomial.

Referencias

[1] James M. Barrie. *Peter Pan (Peter Pan and Wendy)*. Charles Scribner's Sons, 1911.

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

[3] Project Gutenberg. Project Gutenberg: library of free eBooks. http://www.gutenberg.org/.

[4] L. M. Montgomery, M. A. Claus, and W. A. J. Claus. *Anne of Green Gables*. L.C. Page & Co, 1908.

[5] Elisa Schaeffer. Modelos probabilistas aplicados: notas del curso. https://elisa.dyndns-web.com/teaching/prob/pisis/prob.html#ut2.

[6] J. Alberto Benavides Vázquez. Modelos probabilistas aplicados – tarea 3. https://github.com/jbenavidesv87/probabilidad/blob/master/tema3/tarea.ipynb.

[7] Gabriela Sánchez Y. Modelos probabilistas aplicados. https://github.com/Saphira3000/MPA.

Distribución de Poisson

Gabriela Sánchez Y.
5064

En el presente trabajo se realiza un estudio de las formas en que se puede aproximar la distribución de Poisson usando otro tipo de distribuciones. La experimentación para las aproximaciones se realiza con la ayuda del lenguaje de programación R versión 4.0.2 [1]. El código puede encontrarse en el archivo `t4.R` [2].

1. Distribución de Poisson a partir de la distribución binomial

El proceso que se sigue para llegar a una aproximación a la distribución de Poisson a partir de la distribución binomial es el siguiente: se suman valores provenientes de una distribución binomial con parámetros $n = 10000$ y $p = 0.001$, y se cuentan los necesarios para llegar a una *meta*. Los resultados son graficados y se comparan con valores obtenidos con la función `rpois`, con parámetro $\lambda = n \cdot p$.

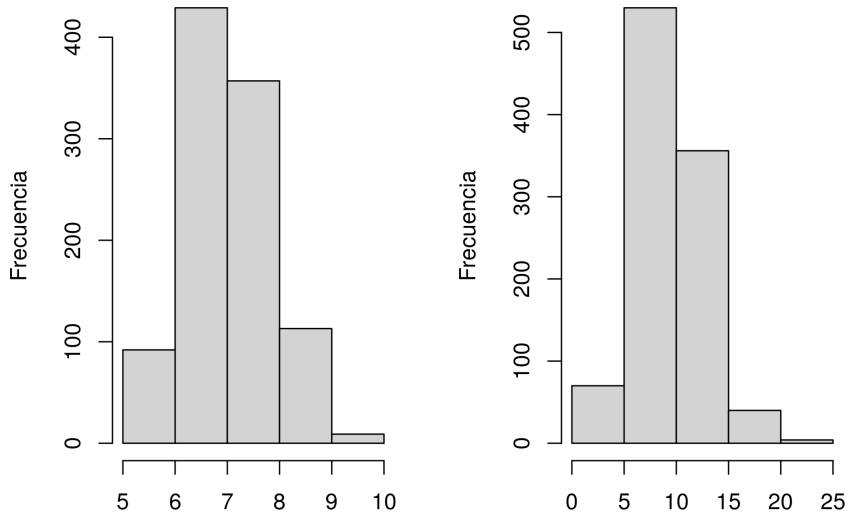
El parámetro de la *meta*, así como el número de veces que se repite el experimento se variaron durante el análisis. La figura 1 muestra dos de los resultados obtenidos con un número fijo de repeticiones $r = 1000$. Las formas se observan parecidas, en especial la mostrada en la figura 1a. Al observar detalladamente los ejes, se advierte que la aproximación parece estar "movida" hacia la izquierda. Es decir, las formas no coinciden en el eje horizontal.

Analíticamente, en clase se estudió que existe una relación entre la distribución de Poisson y la distribución binomial. Si se toman valores de n grandes y valores de p pequeños en la distribución binomial y se define $\lambda = n \cdot p$, se cumple que la distribución binomial aproxima a la distribución de Poisson al hacer $n \to \infty$.

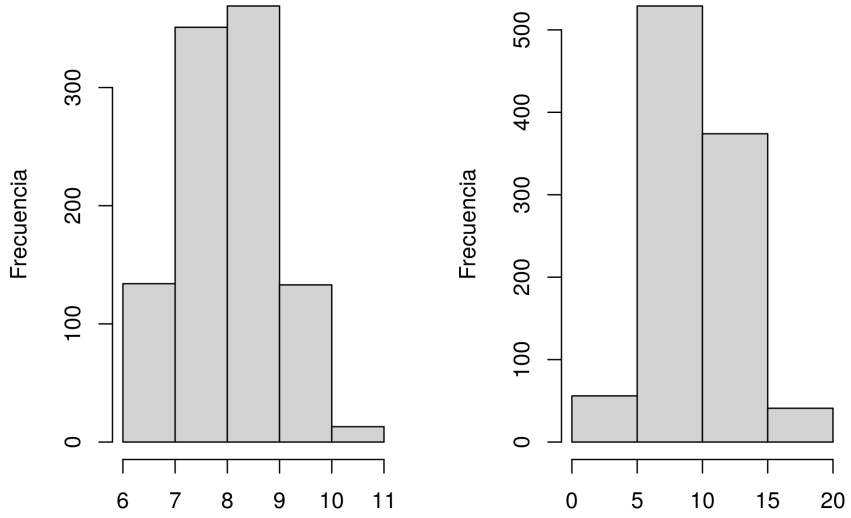
2. Distribución de Poisson a partir de una distribución normal

La idea seguida en la sección anterior se aplica también en este análisis. Es decir, se suman valores provenientes de una distribución normal con parámetros $\mu = \lambda$ y $\sigma^2 = \lambda$, y se cuentan los necesarios para llegar a una *meta*. La elección de estos parámetros fue combinación entre una larga experimentación y lectura sobre ambas distribuciones.

La figura 2 muestra los resultados obtenidos; a la izquierda se muestra la aproximación con la distribución normal y a la derecha la distribución de valores obtenida con la función `rpois(r, l)`. En este caso $\lambda = 100$, $meta = 8000$ y se repite el experimento $r = 1000$ veces.



(a) Aproximación con $meta = 70$.



(b) Aproximación con $meta = 80$.

Figura 1: A la izquierda la aproximación a la distribución de Poisson, a la derecha la distribución de valores con la función `rpois(r, n*p)`.

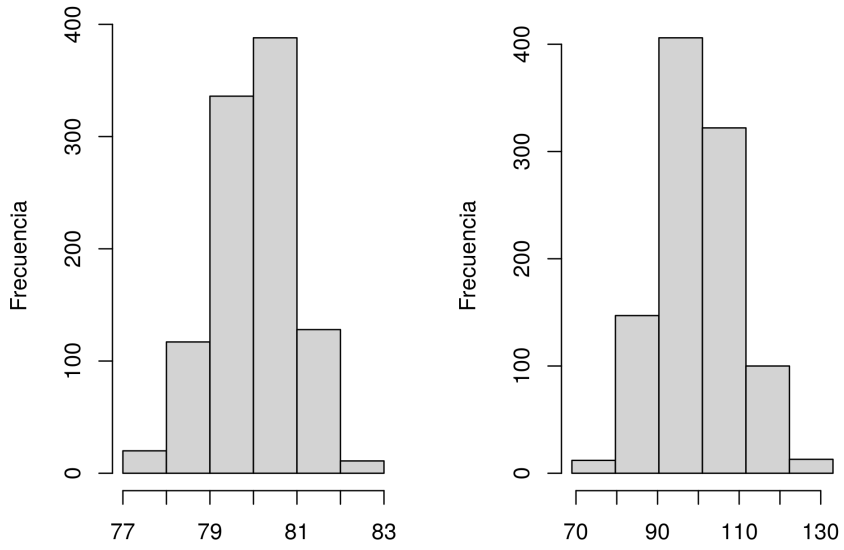


Figura 2: Aproximación a la distribución de Poisson a partir de la distribución normal. A la izquierda se encuentra la aproximación y a la derecha la distribución usando `rpois(r, lambda)`.

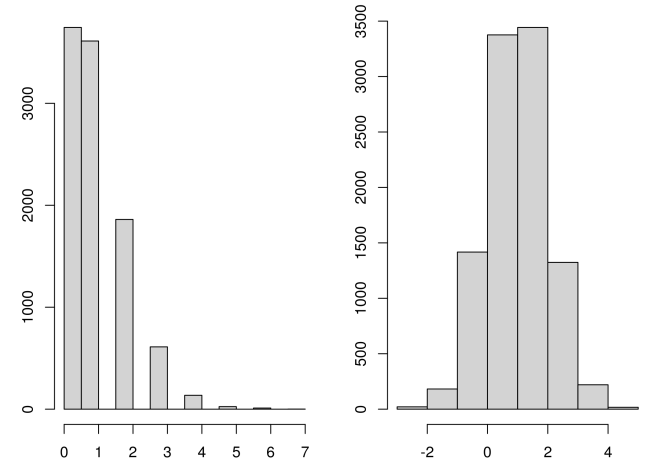
Al analizar las formas de la figura, se puede percibir que ambas distribuciones tienen una forma similar. Sin embargo, en este caso la aproximación también esta "movida" hacia la izquierda. Esto claramente es resultado de la elección de parámetros.

Aunque no se tiene una demostración analítica, en la figura 3 se muestra la distribución de $r = 10000$ valores obtenidos de una distribución de Poisson de parámetro λ y una distribución normal con $\mu = \lambda$ y $\sigma^2 = \lambda$. Nótese que a medida que el parámetro λ aumenta, las distribuciones se asemejan más. Con esto, al menos se puede concluir que la relación entre ambas distribuciones es correcta con la elección de parámetros indicados.

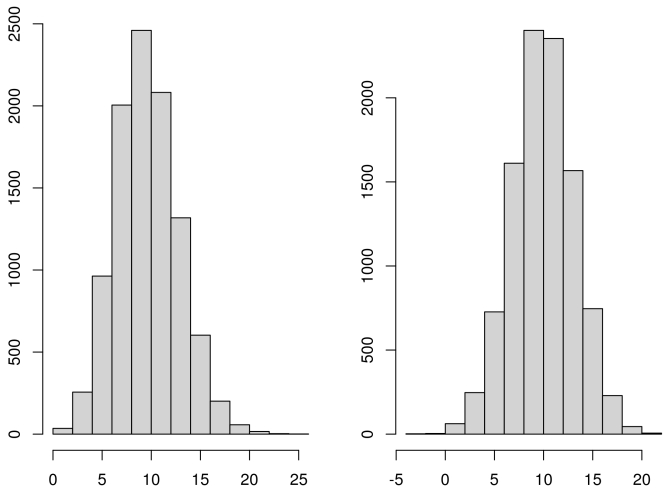
3. Distribución de Poisson a partir de una distribución exponencial

Se continua con la idea aplicada en los casos anteriores. Se suman valores obtenidos de una distribución exponencial y se cuenta cuántos son necesarios para alcanzar una *meta*, que en este caso se fija en la unidad al igual que λ. El número de veces que se repite el experimento es $r = 10000$.

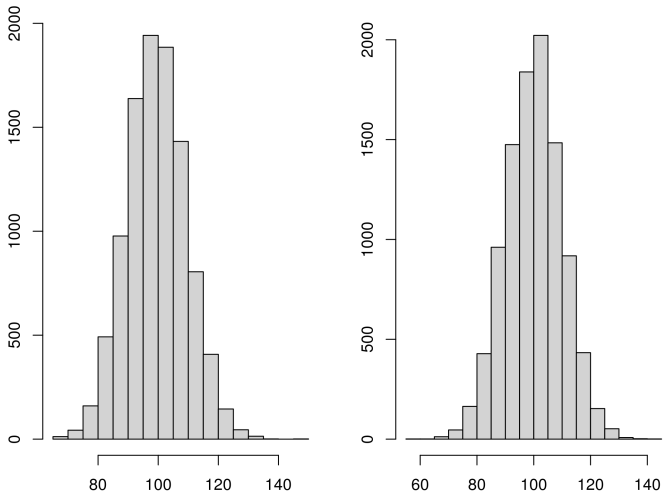
Los resultados se muestran en la figura 4. El procedimiento está sumando valores exponencialmente distribuidos requeridos para completar un valor unitario lo que, por definición, sigue una distribución de Poisson.



(a) $\lambda = 1$.



(b) $\lambda = 10$.



(c) $\lambda = 100$.

Figura 3: A la izquierda distribución de valores obtenidos con la función `rpois(r, lambda)`, a la derecha la distribución de valores con la función `rnorm(r, lambda, sqrt(lambda))`.

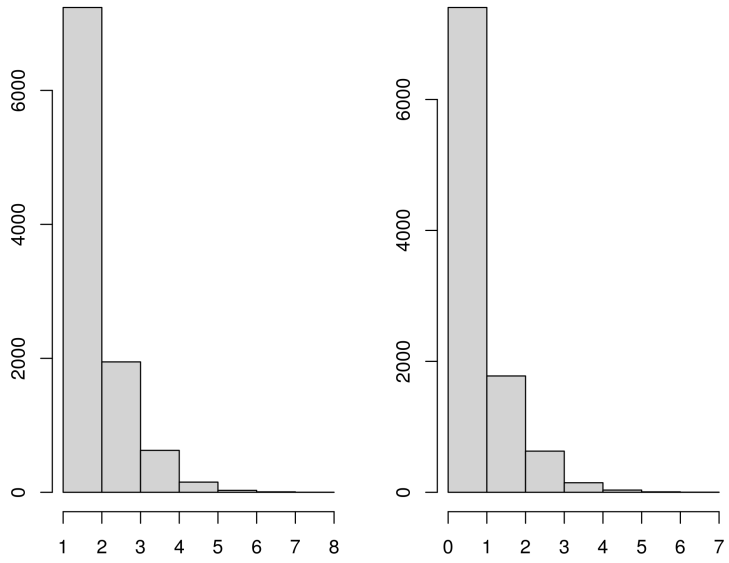


Figura 4: Aproximación a la distribución de Poisson a partir de la distribución exponencial. A la izquierda se encuentra la aproximación y a la derecha la distribución usando `rpois(r, 1)`

Referencias

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

[2] Gabriela Sánchez Y. Modelos probabilistas aplicados. https://github.com/Saphira3000/MPA.

Distribución normal y generación pseudoaleatoria

Gabriela Sánchez Y.
5064

En el presente trabajo se realiza un estudio del generador lineal congruencial [4], algoritmo que permite obtener una secuencia de números pseudoaleatorios y el método de Box-Muller [3], un método de generación de números con distribución normal. Además se hace analiza el efecto que tiene el usa del generador lineal congruencial en la generación de los números del método de Box-Muller. La experimentación se realiza con la ayuda del lenguaje de programación R versión 4.0.2 [1]. El código puede encontrarse en el archivo `t5.R` [5].

1. Generador lineal congruencial

El generador lineal congruencial es un algoritmo que permite obtener una secuencia de números pseudoaleatorios mediante la relación de recurrencia (1), donde m es el módulo, a el multiplicador, c el incremento y X_0 la semilla o valor inicial. Por lo tanto para poder obtener la secuencia son necesarios estos cuatro parámetros.

$$X_{n+1} = (a \cdot X_n + c) \bmod m. \tag{1}$$

Es claro que esta relación permitirá generar a lo más $m - 1$ valores diferentes, pero si no se eligen bien los coeficientes a y c el número se reduce.

Un primer experimento analiza el efecto de usar solo números primos de uno, dos, tres y cuatro dígitos en los parámetros. Con excepción de los primos de un dígito, todos los demás se eligieron de forma "aleatoria". Considerando una semilla $X_0 = 103$ se generan secuencias de $n = 1000$ números y se verifica con la prueba de *Shapiro* si cumplen con la uniformidad o no. En el cuadro 1 se muestran los resultados que se obtienen del p-valor al realizar la prueba y el periodo que se logra con la configuración.

La primera configuración de parámetros genera una secuencia con un periodo muy pequeño, es decir, la secuencia solo tiene seis valores diferentes por lo que no es de extrañar que no pase la prueba de normalidad. Un caso interesante es la segunda y tercera configuración, aunque el módulo y el incremento son primos entre sí en ambos casos, la variación en el valor del multiplicador tiene

Cuadro 1: Resultados obtenidos en el periodo y p-valor para las diferentes configuraciones de parámetros.

Parámetros a, c, m	p-valor	Periodo
3, 5, 7	0	6
11, 43, 97	7.495585e-72	48
59, 43, 97	0.12690289	96
613, 919, 857	0.3345382	428
1291, 3821, 5449	0.71023907	1816

un efecto considerable en la secuencia que se genera. Nótese que con $a = 11$ la secuencia logra un periodo de 48 mientras que con $a = 59$ alcanza el periodo máximo, lo que le permite pasar la prueba de *Shapiro*.

Con estos resultados se puede concluir que el método es bastante sensible a los parámetros para el módulo, multiplicador e incremento. ¿Qué pasa si se varía el valor de la semilla? ¿qué efecto tiene?.

Utilizando las segunda, tercera y cuarta configuración mostradas en el cuadro 1, ahora se varía el valor de la semilla en 5, 59, 103 y 1117, también números primos elegidos al "azar", para analizar el efecto que tiene en la generación de la secuencia.

La figura 1 muestra la distribución de los valores generados. A simple vista no se observa mucha diferencia al variar el valor de la semilla, con excepción de la configuración $a = 11$, $c = 43$ y $m = 97$. Aunque es importante recordar que la configuración no crea una buena secuencia (en términos de uniformidad). La figura permite corroborar el resultado previamente obtenido con la prueba de uniformidad, los valores no parecen distribuirse de manera uniforme.

El diagrama caja-bigote de la figura 2 muestra la variación en las secuencias generadas con la configuración $a = 613$, $c = 919$ y $m = 857$, cambiando el valor de la semilla. Claramente se observa que la variación entre los resultados es minúscula. Sin embargo, para comprobar estadísticamente si el valor de la semilla tiene un efecto considerable en la secuencia o no, se usa la prueba de *Kruskal-Wallis* [2] ya que se sabe que los datos no cumplen con la condición de normalidad que un análisis ANOVA requiere.

La prueba de *Kruskal-Wallis* considera como hipótesis nula que *todas las muestras provienen de una misma distribución*. El resultado de aplicar la prueba a los resultados de la configuración $a = 613$, $c = 919$ y $m = 857$ se muestra a continuación.

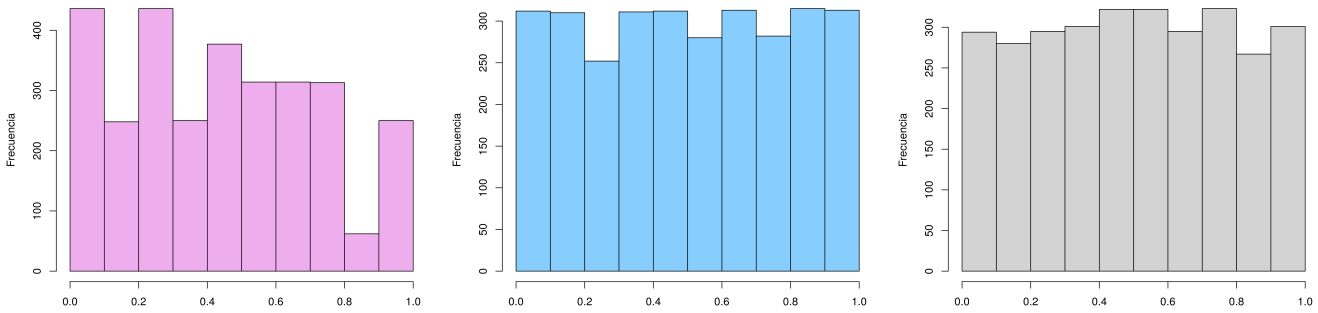
```
Kruskal-Wallis rank sum test

data:  valores by semilla
Kruskal-Wallis chi-squared = 0.61394, df = 3, p-value = 0.8932
```

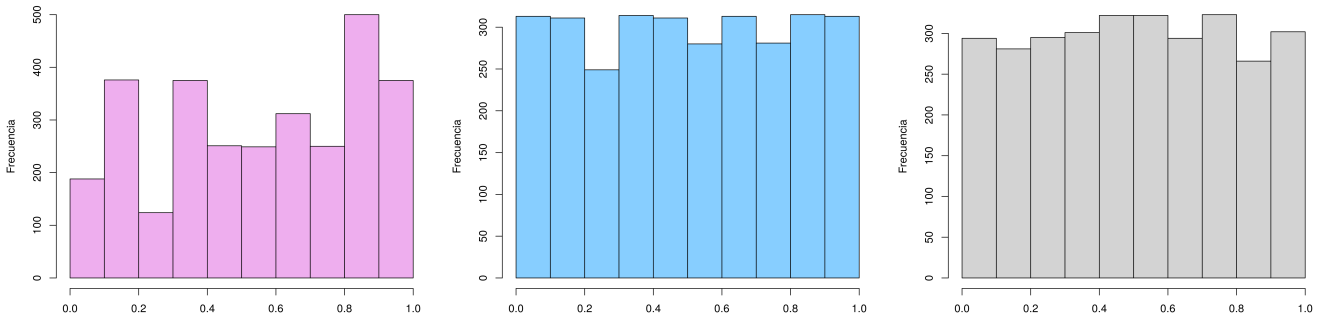
El p-valor obtenido nos permite concluir que la variación en la semilla no tiene un efecto significativo en la secuencia generada. Al aplicar la prueba para la configuración $a = 59$, $c = 43$ y $m = 97$, se llega al mismo resultado.

2. Método de Box-Muller

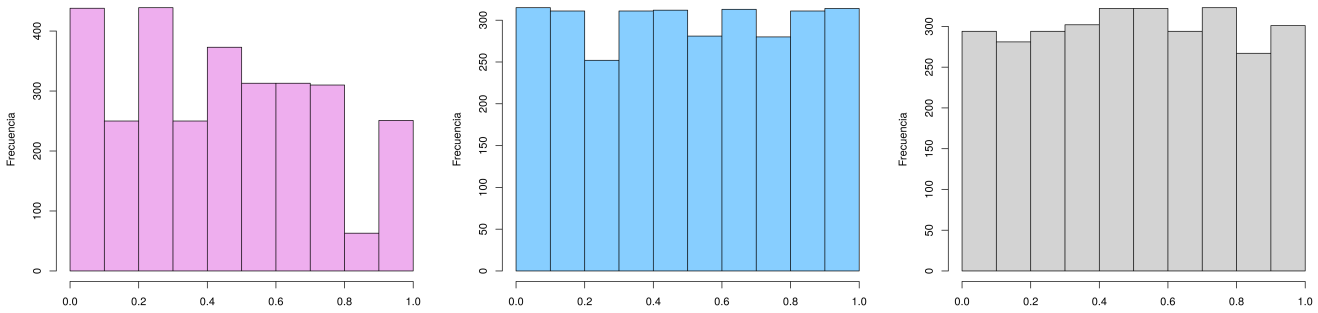
El método Box-Muller genera pares de números aleatorios con distribución normal de media cero y desviación estándar igual a uno. Para la generación utiliza como fuente números aleatorios uniformemente distribuidos.



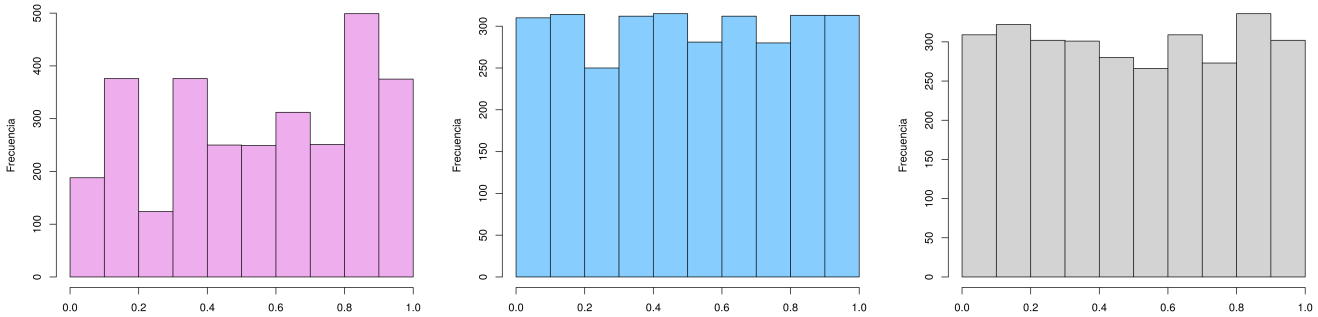
(a) $X_0 = 5$.



(b) $X_0 = 59$.



(c) $X_0 = 103$.



(d) $X_0 = 1117$.

Figura 1: Distribución de la secuencia generada con las distintas configuraciones de parámetros. En rosa la generación con $a = 11, c = 43$ y $m = 97$; en azul con $a = 59$, $c = 43$ y $m = 97$ y, en gris con $a = 613$, $c = 919$ y $m = 857$.

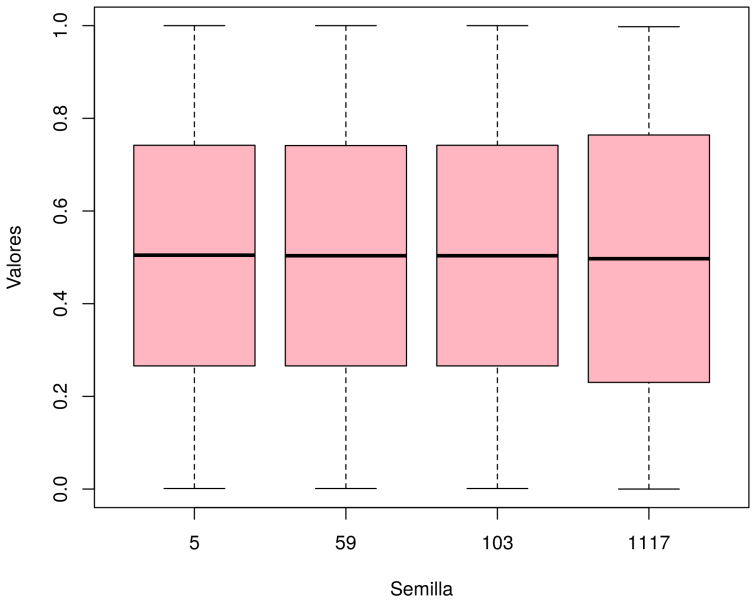


Figura 2: Variación de las secuencias generadas usando $a = 613$, $c = 919$ y $m = 857$, cambiando el valor de la semilla.

Siendo u_1 y u_2 dos números independientes obtenidos de una distribución uniforme, la generación mediante el método se realiza con las ecuaciones (2) y (3).

$$z_0 = \left(\sqrt{-2\ln u_1}\cdot\cos(2\pi u_2)\right)\sigma + \mu, \tag{2}$$

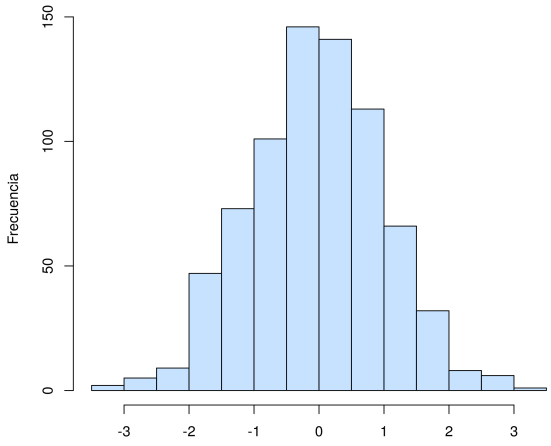
$$z_1 = \left(\sqrt{-2\ln u_1}\cdot\sin(2\pi u_2)\right)\sigma + \mu. \tag{3}$$

Un primer análisis estudia si usar sólo uno de estos valores genera diferencias en los datos. La figura 3 muestra las diferentes distribuciones que se generan con el método al usar sólo uno de los valores y ambos. Las tres distribuciones pasan la prueba de normalidad de *Shapiro* con un p-valor de 0.5854184, 0.1078563, 0.2750796, usando sólo z_0, z_1 y ambos, respectivamente.

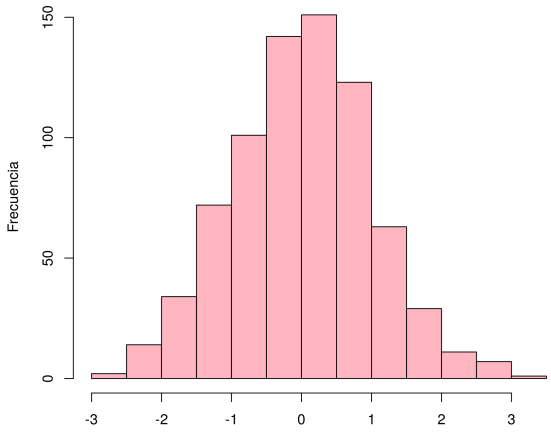
En la sección anterior se estudió un método para generar números uniformes, así que se hace una combinación con el método de Box-Muller. En lugar de usar la función `runif` de R para obtener los valores u_1 y u_2, se usa la función `uniforme.R` que es un generador lineal congruencial.

El análisis previo del generador lineal congruencial, permitió identificar dos configuraciones que dan "buenas" y "malas" secuencias de números uniformemente distribuidos. La pregunta ahora es, si se usan buenos y malos valores de u_1 y u_2, ¿qué pasa con la calidad de la secuencia que se genera? ¿pierde la normalidad?.

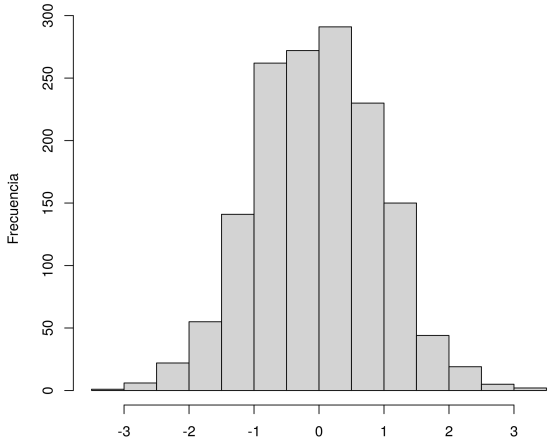
El experimento que se plantea para realizar este estudio es el siguiente: se crea la secuencia de $n = 1000$ números mediante generador lineal congruencial con una de las configuraciones previas y se crea la secuencia con el método de Box-Muller tomando u_1 y u_2 aleatorios de la secuencia generada en el paso anterior. Esta ultima también de tamaño $n = 1000$. Se realizan diez réplicas del experimento, para cada una de ellas se verifica la normalidad con la prueba de *Shapiro*.



(a) Sólo z_0.



(b) Sólo z_1.



(c) Ambos.

Figura 3: Distribuciones generadas por el método usando sólo z_0, z_1 o ambos valores.

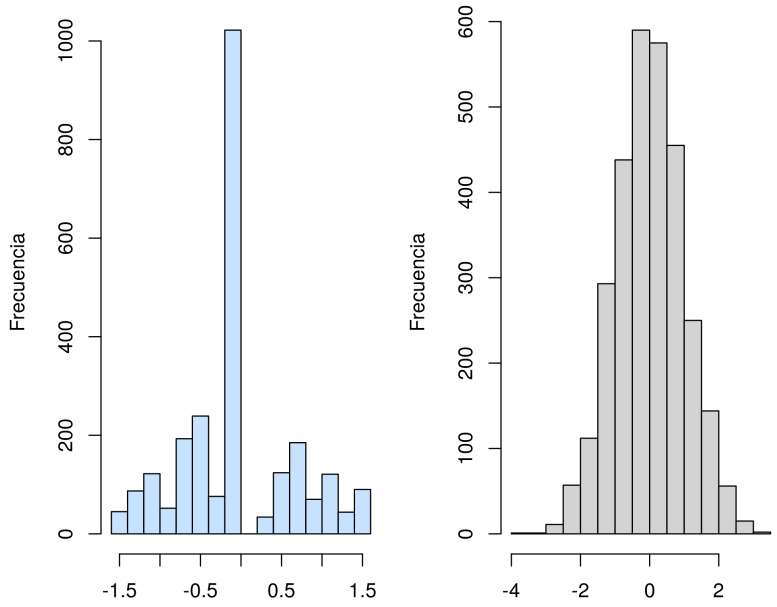


Figura 4: Distribución normal. En azul, la distribución generada por el método de Box-Muller usando el generador lineal congruencial con parámetros "malos" y en gris la distribución generada por `rnorm(n)`.

Se sabe que la configuración $x_0 = 1117$, $a = 3$, $c = 5$ y $m = 7$ no da una secuencia mala de números ya que sólo tiene seis valores diferentes, entre los cuales se encuentra el valor cero. Esto representa un problema para el método de Box-Muller porque para calcular z_0 y z_1 se usa la función logaritmo que no está definida para ese valor.

La figura 4 muestra la distribución obtenida con el método de Box-Muller usando el generador lineal congruencial con la configuración de parámetros $x_0 = 1117$, $a = 3$, $c = 5$ y $m = 7$ en contraste con la distribución generada a partir de la función `rnorm`. Se puede observar que una mala calidad de valores u_1 y u_2 generan una mala calidad de valores z.

Finalmente, la figura 5 muestra la distribución obtenida de una sola réplica con el método de Box-Muller usando el generador lineal congruencial con la configuración de parámetros $x_0 = 1117$, $a = 1291$, $c = 3821$ y $m = 5449$ en contraste con la distribución generada a partir de la función `rnorm`. En este caso se puede observar que el generador lineal congruencial proporciona un buen punto de partida para el método de Box-Muller.

Los resultados del p-valor al aplicar la prueba de normalidad a las diez réplicas se muestran en el cuadro 2. De las diez réplicas, unicamente dos no pasaron la prueba de normalidad.

Referencias

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

[2] Joaquín Amat Rodrigo. Kruskal-wallis test. https://rpubs.com/Joaquin_AR/219504.

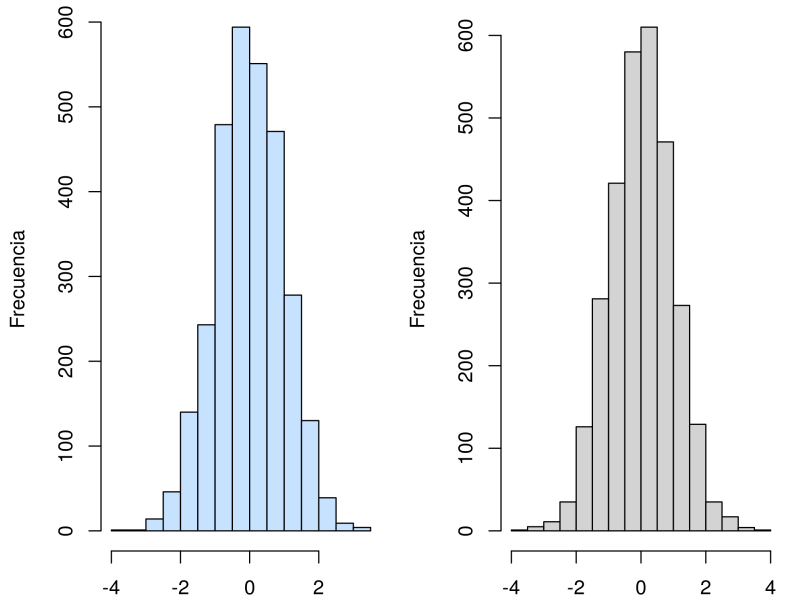


Figura 5: Distribución normal. En azul, la distribución generada por el método de Box-Muller usando el generador lineal congruencial con parámetros "buenos" y en gris la distribución generada por `rnorm(n)`.

Cuadro 2: Resultados del p-valor al aplicar la prueba de *Shapiro* a los datos generados con `uniforme.R`.

Réplica	p-valor
1	0.167831102
2	0.722710035
3	0.388615457
4	0.193132625
5	0.001657073
6	0.407205524
7	0.174958154
8	0.683738471
9	0.016105365
10	0.189060094

[3] Wikipedia. Box–muller transform. https://en.wikipedia.org/wiki/Box%E2%80%93Muller_transform.

[4] Wikipedia. Linear congruential generator. https://en.wikipedia.org/wiki/Linear_congruential_generator.

[5] Gabriela Sánchez Y. Modelos probabilistas aplicados. https://github.com/Saphira3000/MPA.

Pruebas estadísticas

Gabriela Sánchez Y.
5064

Esta práctica se divide en dos partes, una teórica y una práctica. La parte teórica consiste en contestar una serie de preguntas relacionadas con pruebas estadísticas, mientras que la práctica muestra la aplicación de las mismas al un conjunto de datos de *seguridad pública y justicia* obtenidos del INEGI [1].

1. Parte teórica

En esta sección se responde una serie de preguntas realizando una lectura previa sobre pruebas estadísticas en tres sitios: Centro de Ayuda XLSTAT, EcuRed y Máxima formación.

Relación entre contraste de hipótesis y pruebas estadísticas

Ambas formulan hipótesis y evalúan la evidencia estadística que proporcionan los datos para aceptar o rechazar dichas hipótesis.

¿Qué indicaría rechazar la hipótesis nula?

Para poder rechazar una hipótesis nula, es necesario que el p-valor obtenido de la prueba estadística aplicada sea menor que el nivel de significación α de la misma. Al rechazar la hipótesis nula, se acepta la hipótesis alternativa.

¿Cómo se interpreta la salida de una prueba estadística?

Antes de realizar la prueba se debe especificar un nivel de significación α, este valor indica la probabilidad de rechazar la hipótesis nula cuando es verdadera (error de tipo I). Para interpretar la salida de la prueba, se compara el p-valor con el valor de α: si p-valor $< \alpha$, se rechaza la hipótesis nula H_0; en cambio, si p-valor $> \alpha$ no se rechaza la hipótesis nula. Este último resultado no necesariamente implica que se debe aceptar la hipótesis nula, más bien se tiene que no existe suficiente evidencia estadística para rechazar dicha hipótesis.

¿Cómo seleccionar el valor de α?

Anteriormente se mencionó que $\alpha \in [0, 1]$ indica la probabilidad de rechazar la hipótesis nula cuando ésta es verdadera, por lo que la elección de su valor depende de los datos que se estudian. El qué tan peligroso es cometer un error de este tipo, determinará el valor de α a usar. Por ejemplo, si se desea analizar el efecto de algún tratamiento médico, el valor elegido debe ser muy pequeño.

¿Cuáles son los errores frecuentes de interpretación del p-valor?

Tomar un valor de significación no adecuado, conduciría a cometer el error de rechazar la

hipótesis nula cuando ésta es verdadera.

¿Qué es la potencia estadística y para qué sirve?

La potencia de una prueba estadística es la capacidad de la prueba para rechazar la hipótesis nula cuando la hipótesis alternativa es verdadera. En otras palabras, la potencia es la probabilidad de no cometer un error de tipo II. Es por esto que su valor es $1 - \beta$, donde β es la probabilidad de aceptar la hipótesis nula cuando es falsa (error de tipo II).

Tener un buen valor para la potencia estadística, indica que la prueba detectará un efecto que realmente existe en los datos.

¿Cuáles son los supuestos para aplicar técnicas paramétricas?

Se usa una prueba paramétrica si los datos son cuantitativos. Un requisito indispensable es que los datos tengan una distribución de probabilidad normal, además deben tener varianzas iguales o similares (homocedasticidad).

Ejemplos de pruebas estadísticas paramétricas y no paramétricas

Dentro de las pruebas paramétricas más comunes se encuentran la prueba t de Student que se utiliza para comparar las medias de dos poblaciones, el análisis de varianza (ANOVA), el análisis de covarianza (ANCOVA), entre otros.

Algunos ejemplos de pruebas no paramétricas son la prueba de Wilcoxon que se puede usar para comparar las medias de dos grupos que no siguen una distribución normal, Kruskal-Wallis (el equivalente a un ANOVA de una vía) y para comparar dos grupos de variables cualitativas nominales y ordinales se puede usar la prueba Chi-cuadrada.

Guía para encontrar la prueba estadística que buscas

Si se tiene una sola muestra y se quiere determinar si la media de la misma tiene un valor específico, se puede seguir el diagrama de la figura 1, en este caso lo único que se debe verificar para elegir el tipo de prueba es si los datos o no siguen una distribución normal.

En la figura 2 se presenta una guía para determinar el tipo de prueba a usar cuando se desea comparar la media de dos muestras. En este diagrama se usa ND para preguntar si los datos son normalmente distribuidos.

2. Parte práctica

En esta sección se aplican distintas pruebas estadísticas usando el lenguaje de programación R [2] a los datos sobre *seguridad pública y justicia* obtenidos de la página del Instituto Nacional de Estadística y Geografía [1]. Para todas las pruebas se fija el nivel de significación $\alpha = 0.05$.

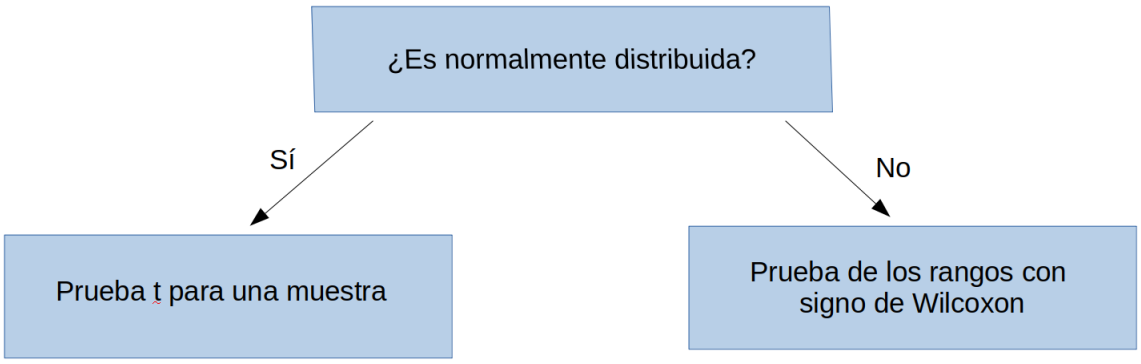


Figura 1: Guía para determinar el tipo de prueba para una sola muestra.

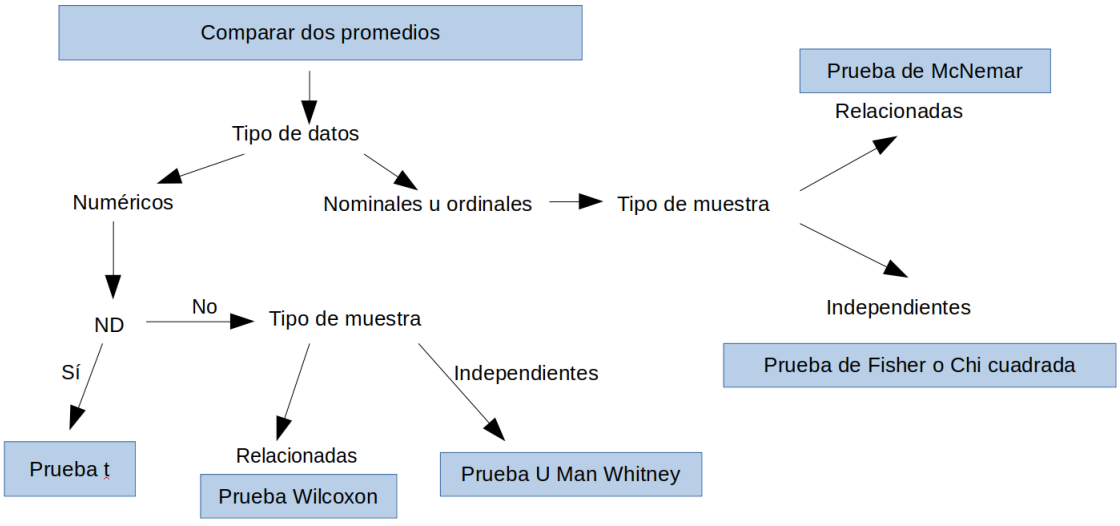


Figura 2: Guía para determinar el tipo de prueba para comparar la media de dos muestras.

2.1. Prueba de Shapiro

La prueba de *Shapiro* es usada para determinar si una muestra sigue una distribución normal. La hipótesis nula H_0 de esta prueba es que los datos de la muestra están normalmente distribuidos. Si el p-valor es mayor que α, no se rechaza la hipótesis nula y en este caso se confirma que la muestra cumple con la normalidad. En las siguientes subsecciones se muestran algunos ejemplos prácticos del uso de esta prueba.

2.2. Prueba t de una muestra

Es una prueba paramétrica usada para determinar si la media de una muestra (con distribución normal), podría tener un valor específico.

Se toman los datos de la tasa de prevalencia delictiva por cada cien mil habitantes en las distintas entidades federativas del país en los años 2010 a 2013 y se verifica con la prueba si se cumple la hipótesis H_0 que la población tiene una media igual a 23,600. Antes de aplicar la prueba se verifica que los datos sigan una distribución normal con la prueba de *Shapiro*. A continuación se muestran los resultados obtenidos de ambas pruebas.

```
Shapiro-Wilk normality test

data:  delitos
W = 0.97897, p-value = 0.1259
--------------------------------------
One Sample t-test

data:  delitos
t = 0.022397, df = 95, p-value = 0.9822
alternative hypothesis: true mean is not equal to 23600
95 percent confidence interval:
22456.15 24769.96
sample estimates:
mean of x
23613.05
```

El p-valor obtenido de la prueba de *Shapiro* indica normalidad. Por otra parte, el p-valor obtenido de la prueba t indica que no se puede rechazar la hipótesis nula, sin embargo esto no quiere decir que se acepte. Es decir, esto no afirma que la media de los datos sea 23,600.

2.3. Prueba de los rangos con signo de Wilcoxon

Es la alternativa a la prueba anterior cuando no se cumple con la distribución normal.

Se analiza si la media de los datos de la tasa de prevalencia delictiva en el periodo comprendido desde 2010 hasta 2019 en todas las entidades federativas del país es igual a 23,600. Antes de aplicar la prueba, se verifica si hay normalidad en los datos. La salida que se obtiene es la siguiente:

```
Shapiro-Wilk normality test

data:  todos_delitos
W = 0.94654, p-value = 9.857e-09
----------------------------------------

Wilcoxon signed rank test with continuity correction

data:  todos_delitos
V = 23976, p-value = 0.02514
alternative hypothesis: true location is not equal to 23600
```

El p-valor obtenido es menor a α, por lo que se rechaza la hipótesis nula y se acepta la hipótesis alternativa que la media de los datos no es igual a 23,600.

2.4. Prueba t para dos muestras y prueba de suma de rangos de Wilcoxon

Las dos pruebas son usadas para comparar las medias de dos muestras. La diferencia radica en que la prueba t requiere que se cumpla la normalidad, mientras que la prueba de *Wilcoxon* no.

Ahora se analiza si la media de la tasa de prevalencia delictiva en el periodo comprendido desde 2010 hasta 2019 en la zona norte del país es igual a la de la zona centro. Para la zona norte se consideran únicamente los estados que están en la frontera del país; es decir, Baja California Norte, Coahuila, Chihuahua, Nuevo León, Sonora y Tamaulipas. Para la zona centro se consideran también seis entidades federativas: Ciudad de México, Guanajuato, Estado de México, Michoacán, Querétaro y Tlaxcala.

Como primer paso se revisa si ambas muestras cumplen con la normalidad y de acuerdo a los p-valores obtenidos, se concluye que una de ellas no lo cumple. Es por esto que se utiliza la prueba de suma de rangos de *Wilcoxon*.

```
Shapiro-Wilk normality test

data:  norte
W = 0.96785, p-value = 0.1548
------------------------------
Shapiro-Wilk normality test
```

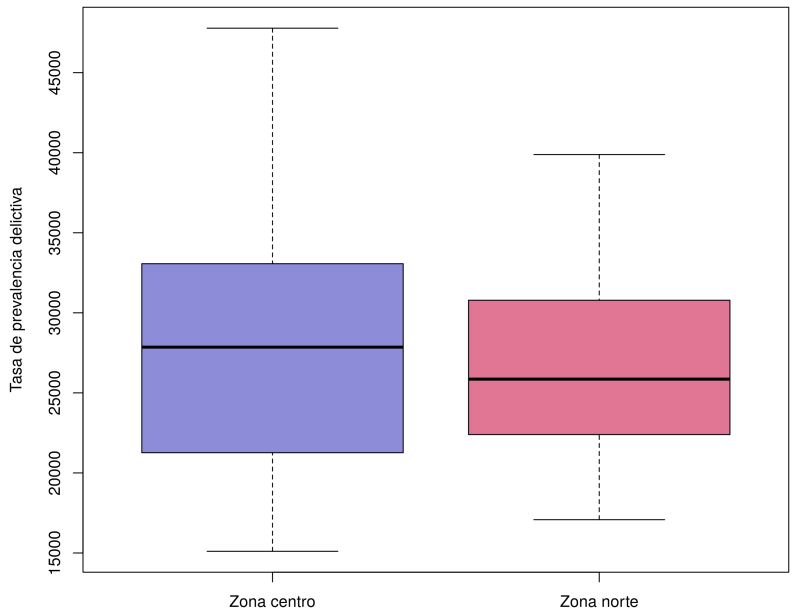


Figura 3: Diagrama caja-bigote de la variación de la tasa de prevalencia delictiva en distintas zonas del país.

```
data:  centro
W = 0.92417, p-value = 0.002158
```

El p-valor obtenido de la prueba de suma de rangos de *Wilcoxon* es mayor que el valor de α por lo que se concluye que no existe suficiente evidencia estadística para rechazar dicha hipótesis nula.

```
Wilcoxon rank sum test with continuity correction

data:  norte and centro
W = 1298, p-value = 0.3271
alternative hypothesis: true location shift is not equal to 0
```

Sin embargo, si se grafica un diagrama de caja-bigote de ambas muestras tal y como se observa en la figura 3, es claro que las muestras no tienen la misma media.

2.5. Prueba de Kolmogorov-Smirnov

Esta prueba es utilizada para determinar si dos muestras siguen la misma distribución. La hipótesis nula H_0 de esta prueba es precisamente que ambas muestras provienen de la misma distribución.

Cuadro 1: Percepción de la efectividad del trabajo de diferentes tipos de autoridades.

	Muy efectivo	Algo efectivo
Policía federal	16.3	49.7
Policía estatal	8.7	45.1

2.6. Prueba de Fisher

Es usado para determinar si dos muestras tiene la misma varianza. Para un ejemplo de su uso, se retoman las muestras utilizadas en la subsección 2.4. El resultado obtenido de aplicar la prueba es el siguiente:

```
F test to compare two variances

data:  centro and norte
F = 2.1669, num df = 53, denom df = 53, p-value = 0.00564
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
1.257380 3.734393
sample estimates:
ratio of variances
2.166922
```

Vea que se cumple que p-valor $< \alpha$, por lo que se rechaza la hipótesis nula y se acepta la hipótesis alterna que las varianzas son distintas.

2.7. Prueba Chi cuadrada

Esta prueba se usa para determinar si dos características son independientes o tienen alguna asociación.

Para ejemplificar esta prueba se toman los resultados de percepción de la población mayor de 18 años, por tipo de autoridad. Se tienen dos tipos de percepción *algo efectivo* y *muy efectivo* y se analizan sólo tres tipos de autoridad: policía federal, estatal y municipal. Los datos se muestran en el cuadro 1.

Con estos datos, se desea saber si el tipo de autoridad afecta al nivel de percepción. La hipótesis nula es que no hay asociación entre estas dos variables, es decir, que el tipo de autoridad no se asocia con la percepción de efectividad. Los resultados obtenidos se muestran a continuación.

```
Pearson's Chi-squared test

data:  prueba
```

```
X-squared = 1.3047, df = 1, p-value = 0.2534
```

El p-valor es mayor que el nivel de significación, lo que significa que no se puede rechazar la hipótesis nula. Para analizar el valor de *chi cuadrada*, primero se aclara que se tiene una tabla de valores de 2×2, es decir, se tienen dos grados de libertad. En este caso el valor crítico con el que se compara el valor de *chi cuadrada* es 3.841. Observe que *chi cuadrada* ¡3.841, de acuerdo a esto, la hipótesis nula no se rechaza.

Referencias

[1] Instituto Nacional de Estadística y Geografía. Seguridad pública y justicia. https://www.inegi.org.mx/datos/?t=0230.

[2] Selva Prabhakaran. R-statistics. http://r-statistics.co/Statistical-Tests-in-R.html.

Teorema de Bayes

Gabriela Sánchez Y.
5064

1. Teorema de Bayes

Sean $H_1, H_2, \ldots, H_m$ un conjunto de eventos disjuntos por pares llamados *hipótesis*, tales que el espacio muestral satisface que $\Omega = H_1 \cup H_2 \cup \cdots \cup H_m$. Además sea E un evento llamado *evidencia* que proporciona información sobre cuál hipótesis es correcta.

Antes de recibir la evidencia se tiene el conjunto de probabilidades previas $P(H_1), \ldots, P(H_m)$ para las hipótesis. Si se conoce la hipótesis correcta entonces se conoce $P(E \mid H_i)$ para todo i. Es de interés calcular las probabilidades para las hipótesis dada la evidencia. Es decir, se desea encontrar las probabilidades posteriores. Estas probabilidades se encuentran mediante la fórmula de Bayes [6] expresada en la ecuación (1)

$$P(H_i \mid E) = \frac{P(H_i)P(E \mid H_i)}{\sum_{k=1}^{m} P(H_k)P(E \mid H_k)}. \tag{1}$$

En este trabajo se desea analizar esta fórmula para el caso específico de pruebas para la detección de alguna enfermedad. En el contexto del problema se tienen dos hipótesis: estar enfermo que se denotará como C^+ y no estar enfermo denotado por C^-. La evidencia, será el resultado de la prueba, *positivo* o *negativo*.

Si se desea saber cuál es la probabilidad de que una persona realmente esté enferma dado que el resultado de la prueba es positivo (también llamado *valor predictivo positivo*), se reescribe la fórmula (1) y se obtiene la ecuación (2)

$$P(C^+ \mid +) = \frac{P(C^+)P(+ \mid C^+)}{P(C^+)P(+ \mid C^+) + P(C^-)P(+ \mid C^-)}, \tag{2}$$

donde $P(+ \mid C^+)$ indica la probabilidad de que la prueba dé un resultado positivo cuando la persona tiene la enfermedad, es decir, la probabilidad de obtener un *verdadero positivo* y, $P(+ \mid C^-)$ indica la probabilidad de que una persona sin la enfermedad obtenga un resultado positivo en la prueba, es decir, la probabilidad de obtener un *falso positivo*.

De manera análoga, la probabilidad de que una persona no esté enferma dado que el resultado de la prueba es negativo, (*valor predictivo negativo*) se puede obtener mediante la ecuación (3)

$$P(C^- \mid -) = \frac{P(C^-)P(- \mid C^-)}{P(C^-)P(+ \mid C^-) + P(C^+)P(- \mid C^+)}, \tag{3}$$

donde $P(- \mid C^-)$ es la probabilidad de que una persona sin la enfermedad obtenga un resultado negativo en la prueba (*verdadero negativo*) y $P(- \mid C^+)$ es la probabilidad de que una persona enferma obtenga un resultado negativo en la prueba (*falso negativo*).

Las probabilidades condicionadas $P(+ \mid C^+)$ y $P(- \mid C^-)$ expresan la *sensibilidad* (capacidad de la prueba para detectar la enfermedad) y *especificidad* de la prueba (capacidad de la prueba de detectar a los individuos sanos) [7]. De manera que aún se pueden reescribir [4] las ecuaciones (2) y (3) como (4) y (5)

$$P(C^+ \mid +) = \frac{(\text{prevalencia}) \cdot \text{sensibilidad}}{(\text{prevalencia}) \cdot \text{sensibilidad} + (1 - \text{prevalencia}) \cdot (1 - \text{especificidad})}, \tag{4}$$

$$P(C^- \mid -) = \frac{(1 - \text{prevalencia}) \cdot \text{especificidad}}{(1 - \text{prevalencia}) \cdot \text{especificidad} + \text{prevalencia} \cdot (1 - \text{sensibilidad})}. \tag{5}$$

Además se puede determinar la *precisión* y *exactitud* [7] de la prueba mediante las ecuaciones (6) y (7)

$$\text{Precisión} = \frac{\text{verdaderos positivos}}{\text{verdaderos positivos} + \text{falsos positivos}}, \tag{6}$$

$$\text{Exactitud} = \frac{\text{verdaderos positivos} + \text{verdaderos negativos}}{\text{total de resultados}}. \tag{7}$$

2. Pruebas Covid-19

Existen dos diferentes tipos de pruebas para la detección del virus SARS-CoV-2, causante de la enfermedad Covid-19: las pruebas de diagnóstico, como la prueba RT-PCR que diagnostica una infección activa de coronavirus y las pruebas de anticuerpos, que muestran si una persona ha sido infectada por el coronavirus en el pasado [3].

Utilizando sólo la información de la prueba RT-PCR, se desea analizar las implicaciones del teorema de Bayes para el caso específico de la detección de Covid-19. Usando la notación descrita previamente, C^+ representará la población enferma y C^- la población sana. De esta manera, conociendo la especificidad y sensibilidad de esta prueba, se pueden determinar el valor predictivo positivo y negativo.

No se encontró un valor específico para la especificidad de la prueba sin embargo se estima que es muy cercana al 100 %. Para fines prácticos se fijará en un 99 %. Tampoco se encontró un valor para la sensibilidad y el rango de valores es variado además de que depende del lugar de la muestra y de la carga viral. De acuerdo a la información de diversas fuentes [1, 5], este valor oscila entre 31 % – 70 %.

Con esta información y los casos de Covid-19 en el estado de Michoacán [2] se determinarán los valores predictivos positivo y negativo.

La información proporcionada al 26 de octubre indica que de un total de 57050 pruebas realizadas, 24,499 han resultado positivas y 32,551 han resultado negativas. Si la población del estado es alrededor de 5 millones, la prevalencia será entonces de un 0.049. Usando estos datos y considerando una especificidad del 70 %, se calcula el valor predictivo positivo y negativo a partir de las

ecuaciones (4) y (5):

$$P(C^+ \mid +) = \frac{(0.049)\cdot(0.7)}{(0.049)\cdot(0.7)+(0.951)\cdot(0.01)} = 0.78,$$

$$P(C^- \mid -) = \frac{(0.951)\cdot(0.99)}{(0.951)\cdot(0.99)+(0.049)\cdot(0.3)} = 0.98.$$

Lo que está indicando el primer resultado es la probabilidad de que una persona que haya dado positivo en la prueba realmente esté enferma, mientras que el segundo resultado indica la probabilidad de que una persona que haya dado negativo en la prueba realmente esté sana.

Note que según los resultados obtenidos, la probabilidad de obtener un falso negativo $[1 - P(C^- \mid -)] = 0.02$ es baja, mientras que la probabilidad de obtener un falso positivo es más alta $[1 - P(C^- \mid -)] = 0.22$.

Referencias

[1] Gar Ming Chan. Bayes' theorem, COVID19, and screening tests. *The American Journal of Emergency Medicine*, 2020.

[2] Gobierno del estado de Michoacán. Michoacán coronavirus Covid-19. https://michoacancoronavirus.com/.

[3] FDA. Conceptos básicos de las pruebas para el coronavirus. https://www.fda.gov/consumers/articulos-en-espanol/conceptos-basicos-de-las-pruebas-para-el-coronavirus, 2020.

[4] Raúl Fernández Regalado. El teorema de Bayes y su utilización en la interpretación de las pruebas diagnósticas en el laboratorio clínico. *Revista Cubana de Investigaciones Biomédicas*, 28, 2009.

[5] Chester B Good, Inmaculada Hernandez, and Kenneth Smith. Interpreting COVID-19 Test Results: a Bayesian Approach. *Journal of General Internal Medicine*, 2020.

[6] Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability*. AMS, 2003.

[7] Towards Data Science. COVID-19, Bayes' theorem and taking probabilistic decisions. https://towardsdatascience.com/covid-19-bayes-theorem-and-taking-data-driven-decisions-part-1-b61e2c2b3bea, 2020.

Expected value and variance

Gabriela Sánchez Y.
5064

In this activity a serie of exercises of the book Introducción to Probability [1] are solved.

Exercise 1, page 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 if even. What is the expected value of his winnings?*

Let X be the card selected and $\phi(X)$ define as in equation (1).

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

The sample space of X is the set $\{2, 3, \ldots, 10\}$ and $P(X = 2) = \cdots = P(X = 10) = \frac{1}{9}$. Therefore

$$\begin{aligned} E[\phi(X)] &= \sum_{x \in \Omega} \phi(x) \cdot P(X = x) \\ &= -1 \cdot \left(\frac{1}{9}\right) + 1 \cdot \left(\frac{1}{9}\right) - 1 \cdot \left(\frac{1}{9}\right) + \ldots - 1 \cdot \left(\frac{1}{9}\right) \\ &= -1 \cdot \left(\frac{5}{9}\right) + 1 \cdot \left(\frac{4}{9}\right) \\ &= -\frac{1}{9}. \end{aligned}$$

Excercise 6, page 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 $E[XY] = E[X]E[Y]$. Are X and Y independent?

Let D_1 and D_2 the outcomes on the first and second roll, respectively. Then $X = D_1 + D_2$ and $Y = D_1 - D_2$.

D_1 and D_2 have the same outcomes with the same probabilities so it is clear that $E[D_1] = E[D_2]$. Therefore,

$$\begin{aligned} E[XY] &= E[(D_1 + D_2) \cdot (D_1 - D_2)] \\ &= E[(D_1)^2 - (D_2)^2] \\ &= E[(D_1)^2] - E[(D_2)^2] = 0, \end{aligned}$$

and $E[Y] = E[D_1] - E[D_2] = 0$. Then $E[X]E[Y] = 0$, and $E[XY] = E[X]E[Y]$.

Even though $E[XY] = E[X]E[Y]$, X and Y are not independent. Two ramdom variables are independent if equation (2) is satisfied for any $x \in X$ and $y \in Y$.

$$P(X = x, Y = y) = P(X = x)P(Y = y). \tag{2}$$

The possible outcomes of the rolles are

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

each one with probability of 1/36.

Note that $P(X = 6, Y = 0) = 1/36$ because this can only happend when $D_1 = 3$ and $D_2 = 3$, and

$$P(X = 6)P(Y = 0) = \frac{5}{36} \cdot \frac{6}{36} = \frac{5}{36} \cdot \frac{1}{6} = \frac{5}{216}.$$

Therefore $P(X = x, Y = y) \neq P(X = x)P(Y = y)$ for $x = 6$ and $y = 0$, so it has been proven that X and Y are not independent.

Exercise 15, page 249

A box contains two gold balls and three silver balls. You are allowed to choose successively balls from the box at random. You win 1 dollar each time you draw a gold ball and lose 1 dollar each time you draw a silver ball. After a draw, the ball is not replaced. Show that, if you draw until you are ahead by 1 dollar or until there are no more gold balls, this is a favorable game.

The possible outcomes of the draw when the player is ahead by 1 dollar or until there are no more gold balls is shown in Table 1, where S and G represents an outcome of a silver and gold ball, respectively.

Then

$$\begin{aligned} E[X] &= 1 \cdot \frac{2}{5} + 1 \cdot \frac{1}{10} + 1 \cdot \frac{1}{10} + 2 \cdot \frac{1}{10} + 0 \cdot \frac{1}{10} + 0 \cdot \frac{1}{10} - 1 \cdot \frac{1}{10} - 1 \cdot \frac{1}{10} - 1 \cdot \frac{1}{10} \\ &= \frac{2}{5} + \frac{2}{10} + \frac{2}{10} - \frac{3}{10} \\ &= \frac{2}{5} + \frac{4}{10} - \frac{3}{10} \\ &= \frac{2}{5} + \frac{1}{10} = \frac{1}{2}. \end{aligned}$$

Table 1: Possible outcomes.

Outcome	**Probability**	**Winnings**
G	$\frac{2}{5}$	1
SGG	$\frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} = \frac{1}{10}$	1
GSG	$\frac{2}{5} \cdot \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{10}$	1
GG	$\frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}$	2
SGSG	$\frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{10}$	0
SSGG	$\frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{10}$	0
SSGSG	$\frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{10}$	-1
SGSSG	$\frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{10}$	-1
SSSGG	$\frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} \cdot 1 \cdot 1 = \frac{1}{10}$	-1

Given that $E[X] > 0$, playing until the ahead by 1 dollar is kept or until there are no more gold balls, the game will be favorable.

Exercise 18, page 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?

It is assumed that once a key is tried it is discarted. Let X be the number of failures before success. Having two failures before success means that the sequence of tried keys was FFS, where F represents a failure and S a success. The probability of this sequence (i.e., having two failures before success) can be calculated using the tree in Figure 1, so $P(X = 2) = \frac{5}{6} \cdot \frac{4}{5} \cdot \frac{1}{4} = \frac{1}{6}$.

The possible values of X are $\{0, 1, \ldots, 5\}$ and $P(X = 0) = P(X = 1) = \ldots = P(X = 5) = \frac{1}{6}$ (see Figure 1). Thus

$$\begin{aligned} E[X] &= 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} \\ &= \frac{15}{6} = \frac{5}{2}. \end{aligned}$$

Exercise 19, page 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.

It does not make any sense that the student do not choose a subset, so he can choose subsets of one, two, three or four possible answers.

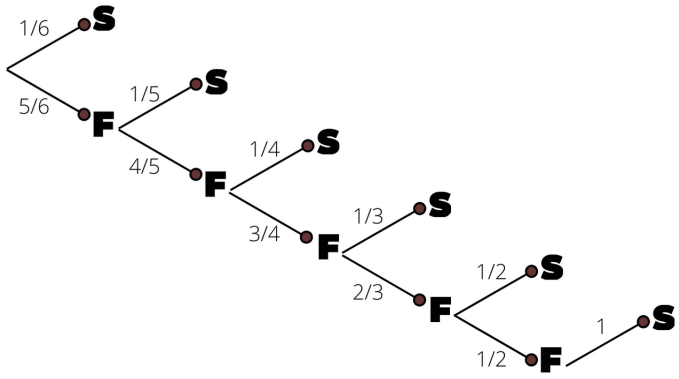


Figure 1: Tree for excercise 18, page 249.

Let R_1, R_2, R_3, R_4 the four possible answers and without loss of generality let R_1 be the right one.

- Subsets of one answer

In this scenario are four options $\{R_1\}, \{R_2\}, \{R_3\}$ and $\{R_4\}$ each one of them with a probability of been choose equal to $\frac{1}{4}$. Then the espected value is

$$3\left(\frac{1}{4}\right) - 1\left(\frac{1}{4}\right) - 1\left(\frac{1}{4}\right) - 1\left(\frac{1}{4}\right) = 0.$$

- Subsets of two answers

The number of possible subsets with two answers are $\binom{4}{2} = 6$: $\{R_1, R_2\}, \{R_1, R_3\}$, $\{R_1, R_4\}$, $\{R_2, R_3\}$, $\{R_2, R_4\}$ and $\{R_3, R_4\}$. The right answer R_1 is in three of these subsets. The espected value in this scenario is

$$3\left(\frac{3}{6}\right) - 1\left(\frac{3}{6}\right) - 2\left(\frac{1}{6}\right) - 2\left(\frac{1}{6}\right) - 2\left(\frac{1}{6}\right) = 0.$$

- Subsets of three answers

In this case the possible subsets are $\binom{4}{3} = 4$: $\{R_1, R_2, R_3\}, \{R_1, R_3, R_4\}, \{R_2, R_3, R_4\}$ and $\{R_1, R_2, R_4\}$. Again, the right answer is in three of these subsets. The espected value is

$$3\left(\frac{3}{4}\right) - 2\left(\frac{3}{4}\right) - 3\left(\frac{1}{4}\right) = 0.$$

- Subsets of four answers

There is only one possible subset $\{R_1, R_2, R_3, R_4\}$. The espected value is

$$3 \cdot (1) - 3 \cdot (1) = 0.$$

Exercise 1, page 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.

$$E[X] = -1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = 0.$$

Recalling that $V[X] = E[X^2] - \mu^2$, where $\mu = E[X]$, then

$$V[X] = E[X^2] - (0)^2 = (-1)^2 \cdot \frac{1}{3} + (0)^2 \cdot \frac{1}{3} + (1)^2 \cdot \frac{1}{3} = \frac{2}{3},$$

and $D[X] = \sqrt{V[X]} = \sqrt{\frac{2}{3}}$.

Exercise 9, page 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 $V(X)$ and $D(X)$.

Let X be the outcome of the die, then $P(X = x) = x \cdot p$. Recall that $\sum_{i=1}^{6} i \cdot p = 1$, therefore $p = \frac{1}{21}$.

To find $V[X]$ it is necessary to find $E[X]$ and $E[X^2]$ first, so

$$\begin{aligned} E[X] &= 1 \cdot p + 2 \cdot 2p + 3 \cdot 3p + 4 \cdot 4p + 5 \cdot 5p + 6 \cdot 6p \\ &= p(1 + 4 + 9 + 16 + 25 + 36) \\ &= 91p \\ &= \frac{91}{21}, \end{aligned}$$

and

$$\begin{aligned} E[X^2] &= \sum_{i=1}^{6} i^2 \cdot (i \cdot p) \\ &= p \sum_{i=1}^{6} i^3 \\ &= p \left(\frac{6^2 \cdot 7^2}{4} \right) \\ &= p \cdot \frac{1764}{4} \\ &= \frac{441}{21} = 21. \end{aligned}$$

Then

$$\begin{aligned} V[X] &= E[X^2] - (E[X])^2 \\ &= 21 - \left(\frac{91}{21}\right)^2 \\ &= \frac{20}{9}, \end{aligned}$$

and $D[X] = \sqrt{V[X]} = \sqrt{\frac{20}{9}} = \frac{2}{3}\sqrt{5}$.

Exercise 12, page 264

Let X be a random variable with $\mu = E(X)$ and $\sigma^2 = V(X)$. 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.

Bearing in mind the properties of linearity of the expected value, the following is obtained

$$\begin{aligned} E[X^*] &= E\left[\frac{X-\mu}{\sigma}\right] \\ &= E\left[\frac{1}{\sigma}(X-\mu)\right] \\ &= \frac{1}{\sigma} \cdot E[X-\mu] \\ &= \frac{1}{\sigma} \cdot (E[X] - E[\mu]) \\ &= \frac{1}{\sigma} \cdot (\mu - \mu) = 0, \end{aligned}$$

and

$$\begin{aligned} V[X^*] &= E[(X^* - 0)^2] \\ &= E\left[\left(\frac{X-\mu}{\sigma}\right)^2\right] \\ &= E\left[\frac{1}{\sigma^2} \cdot (X-\mu)^2\right] \\ &= \frac{1}{\sigma^2} \cdot E[(X-\mu)^2] \\ &= \frac{1}{\sigma^2} \cdot V[X] \\ &= \frac{1}{\sigma^2} \cdot \sigma^2 = 1. \end{aligned}$$

Exercise 3, page 278

The lifetime, measure 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?

The expected lifetime of the light bulb is given by the integral in equation (3)

$$E[T] = \int_0^\infty t \cdot \left(\lambda^2 t e^{-\lambda t}\right)\, dt. \tag{3}$$

One can rewrite integral in equation (3) as follows

$$E[T] = \lambda^2 \int_0^\infty t^2 e^{-\lambda t}\, dt. \tag{4}$$

To compute the expected lifetime of the light bulb, first the integral

$$\int t^2 e^{-\lambda t}\, dt$$

is solved integrating by parts. Thus

$$\int t^2\, e^{-\lambda t} dt = -\frac{t^2}{\lambda} e^{-\lambda t} + \frac{2}{\lambda} \int t e^{-\lambda t}\, dt. \tag{5}$$

The integral $\int te^{-\lambda t} dt$ is solved using integration by parts again

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

The last integral can be solve directly by formulas, then

$$\int te^{-\lambda t} dt = -\frac{t}{\lambda} e^{-\lambda t} - \frac{1}{\lambda^2} e^{-\lambda t}. \tag{6}$$

The result found in (6) is replaced into equation (5), so

$$\begin{aligned} \int t^2\, e^{-\lambda t} dt &= -\frac{t^2}{\lambda} e^{-\lambda t} + \frac{2}{\lambda}\left(-\frac{t}{\lambda} e^{-\lambda t} - \frac{1}{\lambda^2} e^{-\lambda t}\right) \\ &= -\frac{t^2}{\lambda} e^{-\lambda t} - \frac{2t}{\lambda^2} e^{-\lambda t} - \frac{2}{\lambda^3} e^{-\lambda t} \\ &= -e^{-\lambda t}\left(\frac{\lambda^2 t^2 + 2\lambda t + 2}{\lambda^3}\right). \end{aligned} \tag{7}$$

Finally the result in (7) is replaced in equation (4)

$$\begin{aligned} E[T] &= \lambda^2 \left[-e^{-\lambda t}\left(\frac{\lambda^2 t^2 + 2\lambda t + 2}{\lambda^3}\right)\right]\Bigg|_0^\infty \\ &= \lambda^2 \cdot \frac{2}{\lambda^3} \\ &= \frac{2}{\lambda}. \end{aligned}$$

To compute the variance it is neccesary to solve the integral in equation (8)

$$\int t^2 \left(\lambda^2 t e^{-\lambda t}\right)\, dt = \lambda^2 \int t^3 e^{-\lambda t}\, dt. \tag{8}$$

The integral on the right side of equation (8) can be solved integrating by parts

$$\int t^3 e^{-\lambda t}\, dt = -\frac{t^3}{\lambda} e^{-\lambda t} + \frac{3}{\lambda} \int t^2\, e^{-\lambda t}\, dt. \tag{9}$$

Note that the integral in equation (9) was already calculated so

$$\begin{aligned} \int t^3 e^{-\lambda t}\, dt &= -\frac{t^3}{\lambda} e^{-\lambda t} + \frac{3}{\lambda} \cdot \left[-e^{-\lambda t} \left(\frac{\lambda^2 t^2 + 2\lambda t + 2}{\lambda^3} \right) \right] \\ &= -\frac{t^3}{\lambda} e^{-\lambda t} - e^{-\lambda t} \left(\frac{3\lambda^2 t^2 + 6\lambda t + 6}{\lambda^4} \right) \\ &= -e^{-\lambda t} \left(\frac{\lambda^3 t^3 + 3\lambda^2 t^2 + 6\lambda t + 6}{\lambda^4} \right). \end{aligned} \tag{10}$$

Using the result in equation (10) the variance is

$$\begin{aligned} V[T] &= \int_0^\infty t^2 \left(\lambda^2 t e^{-\lambda t} \right)\, dt - \left(\frac{2}{\lambda} \right)^2 \\ &= \lambda^2 \int_0^\infty t^3 e^{-\lambda t}\, dt - \frac{4}{\lambda^2} \\ &= \lambda^2 \left[-e^{-\lambda t} \left(\frac{\lambda^3 t^3 + 3\lambda^2 t^2 + 6\lambda t + 6}{\lambda^4} \right) \right] \Bigg|_0^\infty - \frac{4}{\lambda^2} \\ &= \lambda^2 \cdot \frac{6}{\lambda^4} - \frac{4}{\lambda^2} \\ &= \frac{2}{\lambda^2}. \end{aligned}$$

Bearing in mind that $\lambda = 0.05$ the expected lifetime of the light bulb is $E[T] = 40$ hours and its variance $V[T] = 800$ hours.

Exercise 12, page 280

Find $E[X^Y]$, where X and Y are independent random variables which are uniform on [0, 1]. Then verify your answer by simulation.

$$\begin{aligned} E[X^Y] &= \int_0^1 \int_0^1 x^y dx dy \\ &= \int_0^1 \left[\frac{x^{y+1}}{y+1} \right] \Bigg|_0^1 dy \\ &= \int_0^1 \left(\frac{1^{y+1}}{y+1} - \frac{0^{y+1}}{y+1} \right) dy \\ &= \int_0^1 \left(\frac{1}{y+1} \right) dy \\ &= \ln(y+1)|_0^1 \\ &= \ln(2) \approx 0.6931 \end{aligned}$$

References

[1] Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability.* AMS, 2003.

Valor esperado y varianza

Gabriela Sánchez Y.
5064

En la presente actividad se simulan problemas previamente que previamente se resolvieron de forma analítica [3], con el objetivo de encontrar evidencia numérica que apoye el resultado analítico. Las simulaciones se realizan con el apoyo del lenguaje de programación R [1]. El código puede consultarse en el archivo `t10.R` [2].

Ejercicio 1, página 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 if even. What is the expected value of his winnings?*

El resultado analítico de este problema es $-\frac{1}{9}$. Como primera aproximación para apoyar dicho resultado se realizan dos simulaciones del juego, en cada una se guarda la frecuencia y la frecuencia relativa con que cada resultado ocurre, además se calcula el promedio de las ganancias. En el cuadro 1 se pueden observar estos resultados.

Cuadro 1: Frecuencias para el juego de las cartas.

Ganancia	$n = 100$		$n = 10,000$	
	Frecuencia	**Frecuencia relativa**	**Frecuencia**	**Frecuencia relativa**
-1	54	0.54	5578	0.5578
1	46	0.46	4422	0.4422
Ganancia promedio		**-0.08**		**-0.1156**

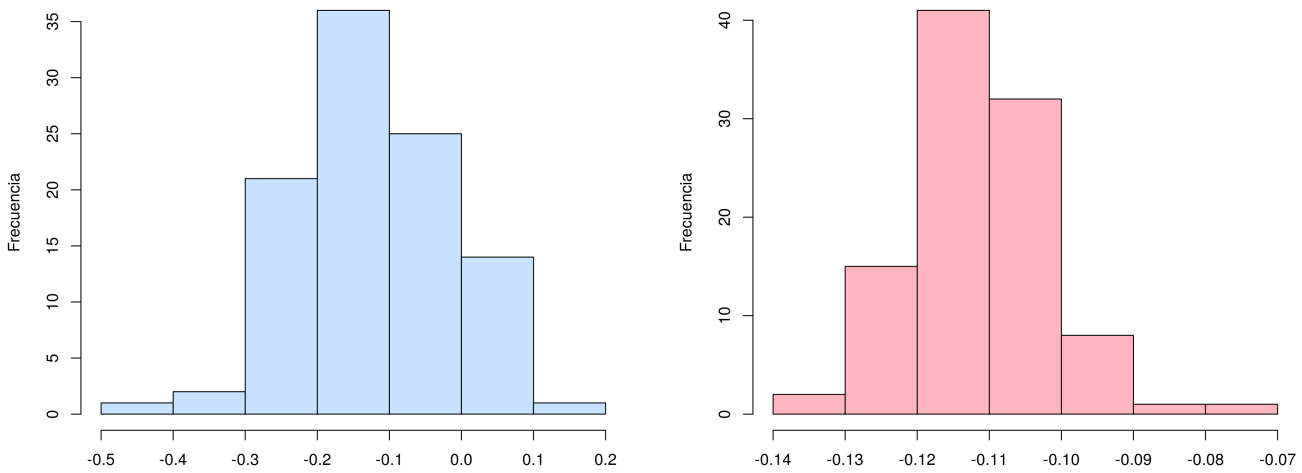
La ganancia promedio al jugar cien veces es de -0.08 y al jugar 10,000 veces de -0.1156. Se puede observar que este último resultado concuerda un poco mejor con el valor esperado de la ganancia.

Es claro que estos resultados pueden variar, por lo que se realizan réplicas del juego para observar cómo varía la ganancia promedio. Los resultados obtenidos para cien réplicas se muestran en la figura 1. La ganancia promedio se encuentra en su mayoría entre -0.1 y -0.2 cuando se juega cien veces y, entre -0.12 y -0.10 cuando se juega 10,000 veces. Es decir, los resultados se acercan al valor real.

Ejercicio 15, página 249

A box contains two gold balls and three silver balls. You are allowed to choose successively balls from the box at random. You win 1 dollar each time you draw a gold ball and lose 1 dollar each time you draw a silver ball. After a draw, the ball is not replaced. Show that, if you draw until you are ahead by 1 dollar or until there are no more gold balls, this is a favorable game.

Previamente se determinó que el valor esperado jugando con las condiciones indicadas es 0.2. El



(a) $n = 100$ (b) $n = 10,000$

Figura 1: Ganancia promedio del juego de cartas.

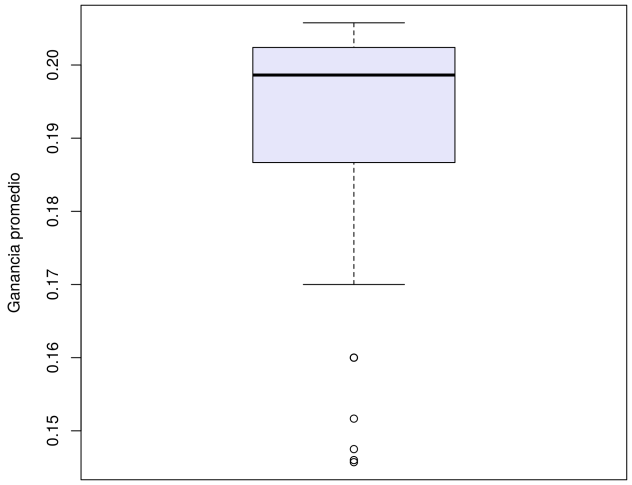


Figura 2: Diagrama de caja bigote de las ganancias obtenidas en cien réplicas del experimento.

experimento diseñado para apoyar ese resultado es el siguiente: se plantea un experimento donde se juega cien veces. Ya que la elección es al azar, un sólo resultado del experimento no aporta mucha información por lo que se realizan cien réplicas. Los resultados obtenidos de la ganancia promedio se muestran en el diagrama de caja bigote de la figura 2. Note que la media de las ganancias obtenidas en las distintas réplicas es muy cercana a 0.2, de esta forma se tiene evidencia numérica que respalda el resultado obtenido previamente.

Es interesante analizar qué tanto cambia el valor esperado al modificar las condiciones de juego. Suponga que ahora el jugador se retira cuando lleva una ventaja de 1 dólar o luego de sacar dos pelotas, ¿cuál es el valor esperado de las ganancias en este caso?.

Jugar con esas condiciones reduce el espacio muestral a tres casos: G, SG y SS, cada uno con probabilidad 2/5, 3/10 y 3/10, respectivamente; donde G representa una pelota dorada y S una

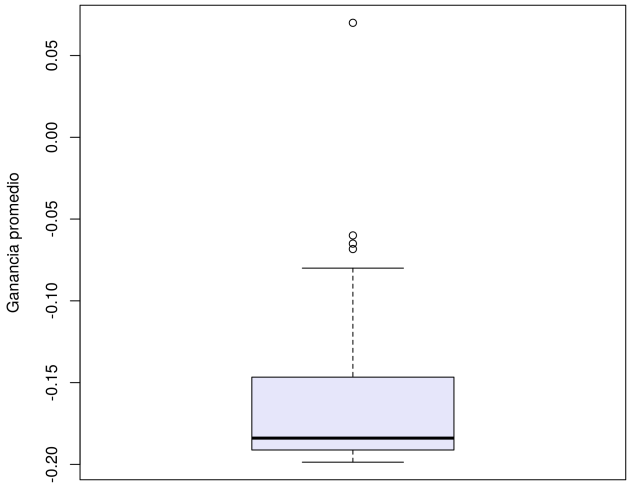


Figura 3: Ganancia promedio cuando el jugador se retira con una ventaja de 1 dólar o luego de sacar dos pelotas.

plata. El valor esperado es $E[X] = 1(2/5) + 0(3/20) - 2(3/10) = -1/5$.

Para responder a esta pregunta de forma numérica, se plantea el mismo experimento realizado previamente con la única diferencia de las condiciones de paro en el juego. Los resultados de la ganancia promedio se muestran en el diagrama de caja bigote de la figura 3. Se puede concluir que el juego ha dejado de ser favorable, ahora la media de las ganancias es muy cercana a -0.2. Nótese que el resultado se aproxima al analítico que se ha calculado anteriormente.

Ejercicio 18, página 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?

La simulación de este problema sigue un criterio similar al planteado para los ejercicios anteriores. El experimento consiste en realizar la prueba de las llaves hasta llegar a la correcta un total de cien veces y se replica 500. El promedio de los intentos fallidos antes de llegar al correcto obtenido en los experimentos se muestra en el diagrama de caja bigote de la figura 4. La media de la simulación casi coincide con el valor analítico de 0.5.

Ejercicio 3, página 278

The lifetime, measure 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?

La simulación para este ejercicio consiste en obtener cien lotes de 500 bombillas cada uno. La vida de cada una de estas bombillas debe ser tal que siga la distribución de densidad $f_T(t) = \lambda^2 t e^{-\lambda t}$. Para obtener un aproximado al valor esperado de la vida útil y la varianza, se calcula el promedio de vida de las bombillas y la varianza en cada lote, respectivamente.

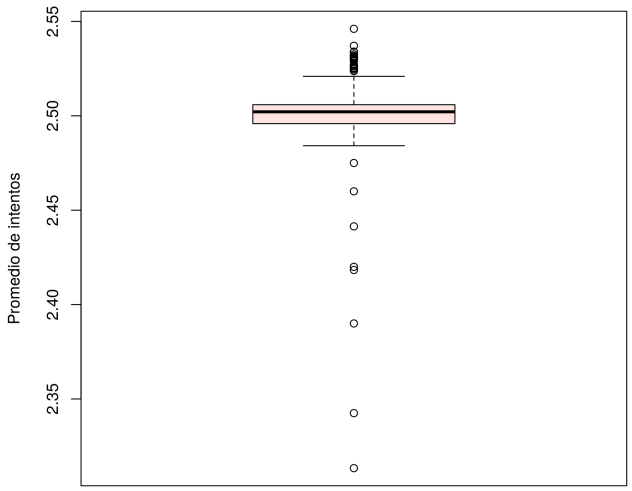
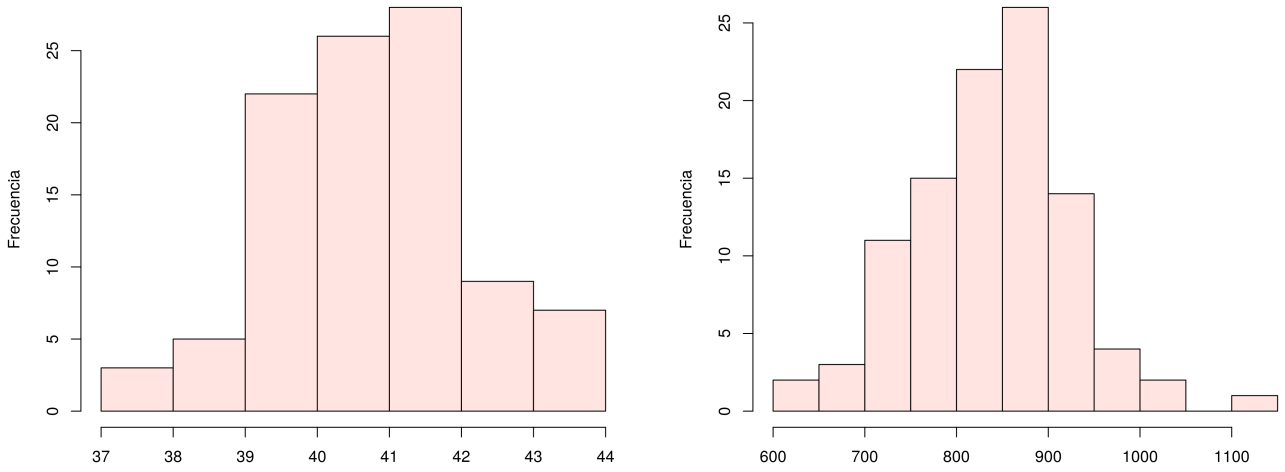


Figura 4: Promedio de intentos fallidos antes de encontrar la llave correcta.



(a) Media

(b) Varianza

Figura 5: Media y varianza de la vida útil de la bombilla.

La figura 5 muestra los resultados obtenidos. En la figura 5a se muestra un histograma con las frecuencias en el promedio de vida útil en los distintos lotes, noté que en su mayoría se encuentran entre 39 y 42, bastante cercanos al valor exacto de 40. Por su parte, la figura 5b muestra que la varianza se encuentra en su mayoría entre 800 y 900. Recuerde que el valor analítico de la varianza es de 800, por lo que la simulación puede ayudar a respaldar dicho resultado.

Referencias

[1] Selva Prabhakaran. R-statistics. http://r-statistics.co/Statistical-Tests-in-R.html.

[2] Gabriela Sánchez Y. Modelos probabilistas aplicados. https://github.com/Saphira3000/MPA.

[3] Gabriela Sánchez Y. Modelos probabilistas aplicados: tarea 9. https://github.com/Saphira3000/MPA/blob/master/t9/t9.pdf.

Prueba de Chi cuadrada y covarianza

Gabriela Sánchez Y.
5064

El presente trabajo se divide en secciones, cada una de ellas se enfoca en un objetivo en particular. La sección 1 presenta una aplicación de la prueba de Chi cuadrada y finalmente la sección 2 valida de forma numérica y analítica dos propiedades relacionadas a la covarianza.

Los experimentos realizados para encontrar los resultados numéricos se realizan con el apoyo del lenguaje de programación R [4] cuyo código puede consultarse en el archivo `t11.R` [5].

1. Prueba de Chi cuadrada

Existen varios tipos de pruebas de Chi cuadrada: la prueba de bondad de ajuste y las pruebas de asociación e independencia. La prueba de bondad de ajuste se utiliza para probar qué tan bien se ajusta una muestra de datos categóricos a una distribución teórica. Las pruebas de asociación e independencia, tal y como su nombre lo dice, se usan para determinar si una variable está asociada con otra y para determinar si el valor observado de una variable depende del valor observado de otra, respectivamente [1].

Por el momento no fue posible determinar una aplicación de esta prueba a los datos actuales de la tesis por lo que se usan los datos proporcionados por el INEGI [2] sobre Seguridad pública y justicia. Específicamente datos sobre la percepción de la población mayor de 18 años, por tipo de autoridad. Se tienen dos tipos de percepción *algo efectivo* y *muy efectivo* y se analizan sólo dos tipos de autoridad: policía federal y estatal. Los datos se muestran en el cuadro 1.

Cuadro 1: Percepción de la efectividad del trabajo de diferentes tipos de autoridades.

	Muy efectivo	Algo efectivo
Policía federal	16.3	49.7
Policía estatal	8.7	45.1

Con estos datos, se aplica la prueba de asociación de Chi cuadrada para determinar si el tipo de autoridad afecta al nivel de percepción. La hipótesis nula es que no hay asociación entre estas dos variables, es decir, que el tipo de autoridad no se asocia con la percepción de efectividad. Para poder analizar el valor de *chi cuadrada*, primero se aclara que se tiene una tabla de valores de 2×2, es decir, se tienen dos grados de libertad. En este caso el valor crítico con el que se compara el valor de *chi cuadrada* es 3.841.

Los resultados obtenidos se muestran a continuación:

```
Pearson's Chi-squared test
```

```
data:  prueba
X-squared = 1.3047, df = 1, p-value = 0.2534
```

El p-valor es mayor que el nivel de significación, lo que indica que no se puede rechazar la hipótesis nula. Observe también que *chi cuadrada* < 3.841, de acuerdo a esto, la hipótesis nula no se rechaza.

2. Covarianza

En esta sección se desea validar de forma numérica y analítica dos propiedades relacionadas a la covarianza. Para lograr el objetivo, primero es necesario definir la covarianza, recordar el concepto de varianza y algunas propiedades del valor esperado. En adelante μ_x y μ_y representan el valor esperado de las variables aleatorias X y Y, respectivamente

Sean X y Y dos variables aleatorias, la covarianza de éstas se define mediante la ecuación 1 o, de manera alternativa, mediante la ecuación 2

$$\begin{aligned} \text{Cov}[X,Y] &= E[(X-\mu_x)(Y-\mu_y)] && (1)\\ &= E[XY]-\mu_x\mu_y. && (2)\end{aligned}$$

Por su parte, la varianza de una variable aleatoria X se define por la ecuación 1

$$V[X] = E[(X-\mu_x)^2] = E[X^2]-\mu_x^2. \quad (3)$$

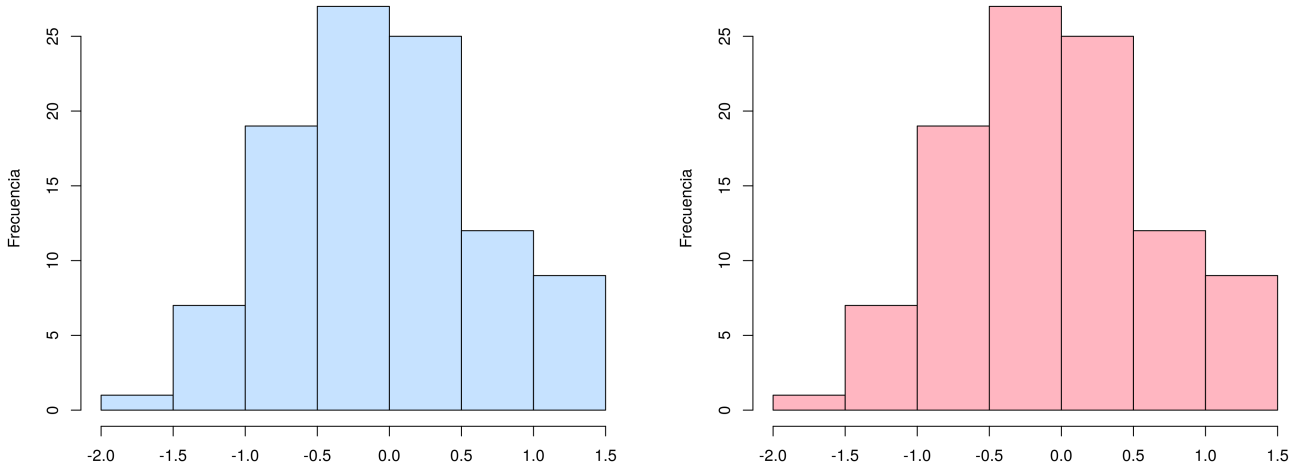
El valor esperado cumple con diferentes propiedades [3]: el valor esperado de la suma de dos variables aleatorias es la suma de los valores esperados y el valor esperado de una variable aleatoria multiplicada por una constante es esa constante multiplicada por el valor valor esperado de la variable aleatoria. Es decir,

$$\begin{aligned} E[X+Y] &= E[X]+E[Y], && (4)\\ E[cX] &= cE[X]. && (5)\end{aligned}$$

2.1. Propiedad 1

Primero se demuestra que $\text{Cov}[aX+b, cY+d] = ac\cdot\text{Cov}[X,Y]$ de forma analítica. De acuerdo a la definición de covarianza, es necesario calcular el valor esperado de las variables $aX+b$ y $cY+d$. De esta manera, utilizando las propiedades del valor esperado, se tiene que

$$E[aX+b] = E[aX]+E[b] = aE[X]+b = a\mu_x+b,$$



(a) $\text{Cov}[aX + b, cY + d]$ (b) $ac \cdot Cov[X, Y]$

Figura 1: Demostración numérica de la propiedad $\text{Cov}[aX + b, cY + d] = ac \cdot \text{Cov}[X, Y]$.

procediendo de manera análoga se puede determinar que $E[cY + d] = c\mu_y + d$. Entonces

$$\begin{aligned}
\text{Cov}[aX + b, cY + d] &= E[\{aX + b - (a\mu_x + b)\}\{cY + d - (c\mu_y + d)\}] \\
&= E[(aX - a\mu_x)(cY - c\mu_y)] \\
&= E[acXY - ac\mu_x Y - ac\mu_y X + ac\mu_x\mu_y] \\
&= ac \cdot E[XY - \mu_x Y - \mu_y X + \mu_x\mu_y] \\
&= ac \cdot E[(X - \mu_x)(Y - \mu_y)] \\
&= ac \cdot Cov[X, Y].
\end{aligned}$$

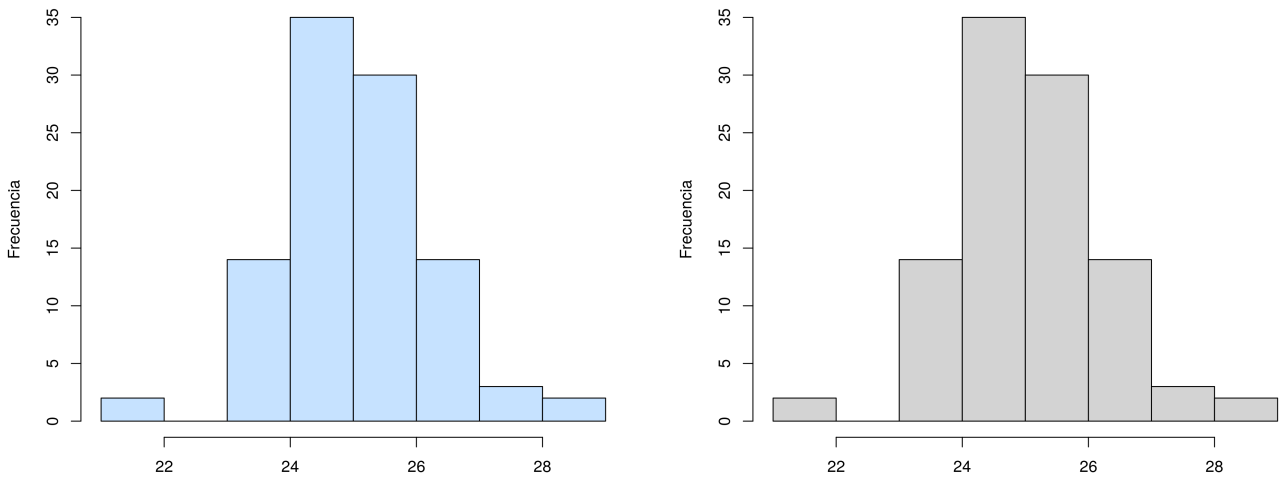
La demostración numérica consiste en tomar, para la variable X mil elementos que siguen una distribución de probabilidad normal con media cero y desviación estándar igual a cinco. Mientras que los elementos de la variable Y siguen una distribución uniforme. Después se crean las variables aleatorias $aX + b$ y $cY + d$.

Sin perder la generalidad se definen $a = 3$, $b = 1$, $c = 5$ y $d = 2$. Se calculan por separado $\text{Cov}[aX + b, cY + d]$ y $ac \cdot Cov[X, Y]$ mediante el uso de la función `cov` de R. El procedimiento se repite cien veces y los resultados se muestran en los histogramas de la figura 2. Prácticamente se tiene el mismo histograma por lo que se puede concluir que la propiedad sí se cumple.

2.2. Propiedad 2

Se desea demostrar que $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$. Dada la definición de varianza, es necesario encontrar el valor esperado de la suma de las variables aleatorias X y Y:

$$E[X + Y] = E[X] + E[Y] = \mu_x + \mu_y.$$



(a) $\text{Var}[X+Y]$ (b) $\text{Var}[X]+\text{Var}[Y]+2\text{Cov}[X,Y]$

Figura 2: Demostración numérica de la propiedad $\text{Var}[X+Y] = \text{Var}[X]+\text{Var}[Y]+2\text{Cov}[X,Y]$.

Con este resultado, se procede a la demostración de la propiedad

$$\begin{aligned}
\text{Var}[X+Y] &= E[(X+Y-\mu_x-\mu_y)^2] \\
&= E[X^2+XY-\mu_x X-\mu_y X+YX+Y^2-\mu_x Y-\mu_y Y-\mu_x X-\mu_x Y+\mu_x^2+ \\
&\quad \mu_x\mu_y-\mu_y X-\mu_y Y+\mu_x\mu_y+\mu_y^2] \\
&= E[X^2+2XY-2\mu_x X-2\mu_y X-2\mu_x Y-2\mu_y Y+2\mu_x\mu_y+\mu_x^2+\mu_y^2+Y^2] \\
&= E[X^2-2\mu_x X+\mu_x^2]+E[Y^2-2\mu_y Y+\mu_y^2]+E[2XY-2\mu_y X-2\mu_x Y+2\mu_x\mu_y] \\
&= E[(X-\mu_x)^2]+E[(Y-\mu_y)^2]+2E[(X-\mu_x)(Y-\mu_y)] \\
&= \text{Var}[X]+\text{Var}[Y]+2\text{Cov}[X,Y].
\end{aligned}$$

La demostración numérica sigue una idea similar a la planteada para demostrar la propiedad 1. Nuevamente se definen $X \sim Norm(0,5)$ y $Y \sim Norm(0,1)$. Mediante las funciones `var` y `cov` de R se realiza el cálculo de $\text{Var}[X+Y]$ y $\text{Var}[X]+\text{Var}[Y]+2\text{Cov}[X,Y]$. El proceso se repite un total de cien veces cuyos resultados se presentan en los histogramas de la figura 1. En este experimento, nuevamente se observa que se obtienen los mismos resultados en cada cálculo de manera que se apoya la demostración de la propiedad 2.

Referencias

[1] Minitab 18. ¿qué es una prueba de chi-cuadrada? https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/tables/supporting-topics/chi-square/what-is-a-chi-square-test/.

[2] Instituto Nacional de Estadística y Geografía. Seguridad pública y justicia. https://www.inegi.org.mx/datos/?t=0230.

[3] Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability*. AMS, 2003.

[4] Selva Prabhakaran. R-statistics. http://r-statistics.co/Statistical-Tests-in-R.html.

[5] Gabriela Sánchez Y. Modelos probabilistas aplicados. https://github.com/Saphira3000/MPA.

Generating functions

Gabriela Sánchez Y.
5064

In this activity some exercises of the book Introduction to Probability [1] are solved.

Exercise 1, page 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

a) $p_0 = 1/2$, $p_1 = 1/4$, $p_2 = 1/4$.

b) $p_0 = 1/3$, $p_1 = 1/3$, $p_2 = 1/3$.

c) $p_0 = 1/3$, $p_1 = 0$, $p_2 = 2/3$.

d) $p_j = 1/2^{j+1}$, *for* $j = 0, 1, 2, \ldots$

e) $p_j = (1/3)(2/3)^j$, *for* $j = 0, 1, 2, \ldots$

f) $p_j = (e^{-2}2^j)/j!$, *for* $j = 0, 1, 2, \ldots$ *(estimate d nummerically).*

Let d the probability that the process will ultimately die out. Theorem 10.2 from page 380 [1] says that if the mean number m of offspring produced by a single parent is ≤ 1, then $d = 1$ and the process dies out with probability 1. But if $m > 1$ then $d < 1$ and the process dies out with probability d.

In the particular case of a), b) and c), the mean number m of offspring produced by a single parent is $m = p_1 + 2p_2 = 1 - p_0 + p_2$. If $m > 1$, d can be easily calculated by $d = p_0/p_2$.

a) $p_0 = 1/2$, $p_1 = 1/4$, $p_2 = 1/4$

The mean number m of offspring produced by a single parent is

$$m = \frac{1}{4} + 2\left(\frac{1}{4}\right) = \frac{3}{4} < 1.$$

Then, by theorem 10.2, follows that the process dies out with probability 1.

b) $p_0 = 1/3$, $p_1 = 1/3$, $p_2 = 1/3$

For this exercise $m = \frac{1}{3} + 2\left(\frac{1}{3}\right) = 1$. Therefore the process dies out with probability 1.

c) $p_0 = 1/3$, $p_1 = 0$, $p_2 = 2/3$

The mean number m of offspring produced by a single parent in this case is

$$m = 0 + 2\left(\frac{2}{3}\right) = \frac{4}{3} > 1.$$

The process dies out with probability $d = p_0/p_2 = \frac{1}{3}/\frac{2}{3} = 0.5$.

To solve d) and e) it is neccesary to remember that $h(z)$, the ordinary generating function for the p_i, is

$$h(z) = p_0 + p_1 z + p_2 z^2 + \cdots$$

and $m = h'(1)$. If $m \leq 1$, the process will surely die out and $d = 1$. To find the probability d when $m > 1$ one must find a root $d < 1$ of the equation

$$z = h(z).$$

d) $p_j = 1/2^{j+1}$, for $j = 0, 1, 2, \ldots$

The ordinary generating function of the problem is

$$\begin{aligned} h(z) &= \frac{1}{2} + \frac{1}{2^2}z + \frac{1}{2^3}z^2 + \ldots \\ &= \frac{1}{2}\left(1 + \frac{1}{2}z + \frac{1}{2^2}z^2 + \ldots\right) \\ &= \frac{1}{2}\left[\left(\frac{1}{2}z\right)^0 + \left(\frac{1}{2}z\right)^1 + \left(\frac{1}{2}z\right)^2 + \ldots\right] \\ &= \frac{1}{2}\left(\frac{1}{1-\frac{1}{2}z}\right) \\ &= \frac{1}{2-z}. \end{aligned}$$

To get this result it has been used that $1 + r + r^2 + \ldots = \frac{1}{1-r}$. Then

$$\begin{aligned} h'(z) &= \frac{d}{dz}\left(\frac{1}{2-z}\right) \\ &= \frac{d}{dz}(2-z)^{-1} \quad \text{chain rule} \\ &= \frac{1}{(2-z)^2}, \end{aligned}$$

and $m = h'(1) = \frac{1}{(2-1)^2} = 1 \leq 1$, therefore $d = 1$.

e) $p_j = (1/3)(2/3)^j$, for $j = 0, 1, 2, \ldots$

The ordinary generating function is

$$h(z) = \frac{1}{3}\left(\frac{2}{3}\right)^0 + \frac{1}{3}\left(\frac{2}{3}\right)^1 z + \frac{1}{3}\left(\frac{2}{3}\right)^2 z^2 + \frac{1}{3}\left(\frac{2}{3}\right)^3 z^3 + \ldots$$

$$
\begin{aligned}
&= \frac{1}{3}\left[1+\left(\frac{2}{3}z\right)^1+\left(\frac{2}{3}z\right)^2+\ldots\right] \\
&= \frac{1}{3}\left(\frac{1}{1-\frac{2}{3}z}\right) \\
&= \frac{1}{3-2z}.
\end{aligned}
$$

Then, one can calculate $h'(z)$:

$$
\begin{aligned}
h'(z) &= \frac{d}{dz}\left(\frac{1}{3-2z}\right) \\
&= \frac{d}{dz}(3-2z)^{-1} \quad \text{chain rule} \\
&= \frac{2}{(3-2z)^2}.
\end{aligned}
$$

from which $m = h'(1) = \frac{2}{(3-2)^2} = 2$ and $d < 1$. To find the probability d we need to solve the equation $z = h(z)$. Using the previous result found for $h(z)$ we have

$$2z^2 - 3z + 1 = 0.$$

The roots of this equation are $z_1 = 1$ and $z_2 = 1/2$. Therefore, the probability d that the process eventually dies out is 0.5.

Excercise 3, page 392

In the chain letter problem (see Example 10.14) find your expected profit if

a) $p_0 = 1/2$, $p_1 = 0$, $p_2 = 1/2$.

b) $p_0 = 1/6$, $p_1 = 1/2$, $p_2 = 1/3$.

Show that if $p_0 > 1/2$, *you cannot expect to make a profit.*

The expected profit of the chain letter problem can be found by the expression $50m + 50m^{12}$, where $m = p_1 + 2p_2$.

a) $p_0 = 1/2$, $p_1 = 0$, $p_2 = 1/2$.

In this particular case $m = 0 + 2\left(\frac{1}{2}\right) = 1$. Then, the expected profit is: $50(1 + 1^{12}) - 100 = 0$.

b) $p_0 = 1/6$, $p_1 = 1/2$, $p_2 = 1/3$.

For this problem $m = \frac{1}{2} + 2\left(\frac{1}{3}\right) = \frac{7}{6}$ and the expected profit is

$$50\left[\frac{7}{6} + \left(\frac{7}{6}\right)^{12}\right] - 100 \approx 376.26 - 100 = 276.26.$$

Now, if $p_0 > 1/2$ then $p_0 > p_2$ and $d = p_0/p_2 > 1$. But, if $d > 1$ then $m \leq 1$. The condition to the problem to be favorable is $m + m^{12} > 2$, considering that $m \leq 1$, the condition it is not satisficed. Therefore if $p_0 > 1/2$, you cannot expect to make a profit.

Exercise 1, page 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

a) $f_X(x) = \frac{1}{2}$.

b) $f_X(x) = \frac{1}{2}x$.

c) $f_X(x) = 1 - \frac{1}{2}x$.

d) $f_X(x) = |1 - x|$.

e) $f_X(x) = \frac{3}{8}x^2$.

The *moment generating function* $g(t)$ for X is define by Equation (1)

$$g(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx. \tag{1}$$

The values of the variable X are in the interval $[0, 2]$, therefore the moment generating funcion will be define by the integral in Equation (2)

$$g(t) = \int_{0}^{2} e^{tx} f_X(x)\, dx. \tag{2}$$

a) $f_X(x) = \frac{1}{2}$.

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

b) $f_X(x) = \frac{1}{2}x$.

$$g(t) = \int_0^2 e^{tx} \left(\frac{1}{2}x\right) dx$$

$$
\begin{aligned}
&= \frac{1}{2}\int_0^2 xe^{tx}\,dx && \text{i.b.p}\\
&= \frac{1}{2}\left[\frac{x}{t}\cdot e^{tx} - \frac{1}{t^2}\cdot e^{tx}\right]\Big|_0^2\\
&= \frac{1}{2}\cdot\frac{2te^{2t} - e^{2t} + 1}{t^2}.
\end{aligned}
$$

c) $f_X(x) = 1 - \frac{1}{2}x$.

$$
\begin{aligned}
g(t) &= \int_0^2 e^{tx}\left(1 - \frac{1}{2}x\right)dx\\
&= \int_0^2 e^{tx}\,dx - \int_0^2 e^{tx}\left(\frac{1}{2}x\right)dx\\
&= \int_0^2 e^{tx}\,dx - \frac{1}{2}\int_0^2 xe^{tx}\,dx
\end{aligned}
$$

Note that these integrals were already calculated in a) and b), then

$$
\begin{aligned}
g(t) &= \int_0^2 e^{tx}\,dx - \frac{1}{2}\int_0^2 xe^{tx}\,dx\\
&= \frac{e^{2t}-1}{t} - \frac{1}{2}\cdot\frac{2te^{2t} - e^{2t} + 1}{t^2}\\
&= \frac{3te^{2t} - 3t - 2te^{2t} + e^{2t} - 1}{2t^2}\\
&= \frac{e^{2t} - 2t + 1}{2t^2}.
\end{aligned}
$$

d) $f_X(x) = |1 - x|$.

Following the definition of absolute value, the density function f_X can be define by Equation (3)

$$
f_X(x) = \begin{cases} 1 - x, & \text{if} \quad x \leq 1\\ -1 + x, & \text{if} \quad x > 1. \end{cases} \tag{3}
$$

Therefore the moment generating function will be define by

$$
\begin{aligned}
g(t) &= \int_0^2 e^{tx}|1 - x|\,dx\\
&= \int_0^1 e^{tx}(1 - x)\,dx + \int_1^2 e^{tx}(-1 + x)\,dx\\
&= \int_0^1 e^{tx}\,dx - \int_0^1 xe^{tx}\,dx - \int_1^2 e^{tx}\,dx + \int_1^2 xe^{tx}\,dx\\
&= \left[\frac{1}{t}e^{tx}\right]\Big|_0^1 - \left[\frac{x}{t}\cdot e^{tx} - \frac{1}{t^2}\cdot e^{tx}\right]\Big|_0^1 - \left[\frac{1}{t}e^{tx}\right]\Big|_1^2 + \left[\frac{x}{t}\cdot e^{tx} - \frac{1}{t^2}\cdot e^{tx}\right]\Big|_1^2
\end{aligned}
$$

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

e) $f_X(x) = \frac{3}{8}x^2$.

$$\begin{aligned} g(t) &= \int_0^2 e^{tx}\left(\frac{3}{8}x^2\right)\,dx \\ &= \frac{3}{8}\int_0^2 x^2 e^{tx}\,dx. \end{aligned}$$

Integrating by parts twice, the following result is obtained

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

Exercise 6, page 402

Let X be a continuous random variable whose characteristic function $k_X(\tau)$ is $k_X(\tau) = e^{-|\tau|}$, $-\infty < \tau < \infty$. Show directly that the density f_X of X is

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

Having the characteristic function k_X, it is possible to determine the density function f_X by Equation (4)

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

The characteristic function of the problem is define by an absolute value, therefore

$$k_X(\tau) = \begin{cases} -\tau, & \text{if} \quad \tau \geq 0 \\ \tau, & \text{if} \quad \tau < 0. \end{cases} \tag{5}$$

Using this result, the density function f_X will be

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ix\tau}e^{-|\tau|}\,d\tau \\ &= \frac{1}{2\pi}\left(\int_{-\infty}^{0} e^{-ix\tau}e^{\tau}\,d\tau\right) + \frac{1}{2\pi}\left(\int_0^{\infty} e^{-ix\tau}e^{-\tau}\,d\tau\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2\pi}\left(\int_{-\infty}^{0} e^{\tau(1-ix)}\,d\tau\right) + \frac{1}{2\pi}\left(\int_{0}^{\infty} e^{-\tau(1+ix)}\,d\tau\right) \\
&= \frac{1}{2\pi}\lim_{R\to\infty}\left(\int_{-R}^{0} e^{\tau(1-ix)}\,d\tau\right) + \frac{1}{2\pi}\lim_{R\to\infty}\left(\int_{0}^{R} e^{-\tau(1+ix)}\,d\tau\right) \\
&= \frac{1}{2\pi}\lim_{R\to\infty}\left[\left(\frac{1}{1-ix}\right)e^{\tau(1-ix)}\right]\Bigg|_{-R}^{0} + \frac{1}{2\pi}\lim_{R\to\infty}\left[\left(-\frac{1}{1+ix}\right)e^{-\tau(1+ix)}\right]\Bigg|_{0}^{R} \\
&= \frac{1}{2\pi}\left(\frac{1}{1-ix} + \frac{1}{ix+1}\right) \\
&= \frac{1}{2\pi}\cdot\frac{1+ix+1-ix}{(1-ix)(1+ix)} \\
&= \frac{1}{2\pi}\cdot\frac{2}{1-i^2x^2} \\
&= \frac{1}{\pi(1+x^2)}.
\end{aligned}
$$

Exercise 10, page 403

Let $X_1, X_2, \ldots, X_n$ be an independent trials process with density

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

a) Find the mean and variance of $f(x)$.

b) Find the moment generating function for $X_1, S_n, A_n,$ and S_n^*.

c) What can you say about the moment generating function of S_n^* as $n \to \infty$?

d) What can you say about the moment generating function of A_n as $n \to \infty$?

References

[1] Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability.* AMS, 2003.

Ley de los grandes números

Gabriela Sánchez Y.
5064

1. Ley de los grandes números

Existen dos versiones de la ley de los grandes números, la ley débil y la ley fuerte. Ambas se establecen en los siguientes teoremas.

Teorema 1 (Ley débil de los grandes números) *Sea $X_1, X_2, \ldots, X_n$ una secuencia de variables aleatorias independientes e idénticamente distribuidas, cada una con una media $E[X_i] = \mu$ y desviación estándar σ. Se define $\bar{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$. Para todo $\epsilon > 0$ se tiene que*

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0.$$

Note que $\bar{X}_n$ es un promedio de los resultados individuales, es por esto que la ley de los grandes números es a veces llamada *ley de los promedios*.

En otras palabras, la ley de los grandes números dice que mientras más se repita un experimento aleatorio el promedio de los resultados se acercará al valor esperado exacto.

Teorema 2 (Ley fuerte de los grandes números) *Sea $X_1, X_2, \ldots, X_n$ una secuencia de variables aleatorias independientes e idénticamente distribuidas, cada una con una media $E[X_i] = \mu$ y desviación estándar σ, entonces*

$$P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1.$$

2. Ejemplos

En esta sección se explica la ley débil de los grandes números con dos ejemplos básicos: el lanzamiento de una moneda y el lanzamiento de un dado.

2.1. Lanzamiento de una moneda

Considere el lanzamiento de una moneda. La variable aleatoria puede tomar dos valores: 0 si el resultado es águila y si el resultado es sol. El valor esperado de esta variable aleatoria es $E[X] = 0 \cdot (\frac{1}{2}) + 1 \cdot (\frac{1}{2}) = 0.5$. Es decir, se espera que el resultado de los lanzamientos sea mitad sol y mitad águila.

Cuadro 1: Promedio del lanzamiento de una moneda.

Lanzamientos	**Promedio**
10	0.8000
100	0.5200
1,000	0.4989
10,000	0.4980
100,000	0.4950

Se formuló un experimento en R [1] que simulara el lanzamiento de una moneda un número variable de veces. El cuadro 1, muestra los resultados del experimento. Se observa que, a medida que se incrementa el número de lanzamientos, el promedio tiende a acercarse al valor esperado 0.5.

2.2. Lanzamiento de un dado

Sea X la variable aleatoria correspondiente al lanzamiento de un dado. En una actividad previa se calculó que el valor esperado del lanzamiento de un dado es $E[X] = 3.5$. De acuerdo a la ley de los grandes números, a medida que se aumente la cantidad de lanzamientos, el promedio de los mismos se acercará al valor 3.5. Esto puede comprobarse fácilmente con un experimento sencillo. Nuevamente se utiliza el lenguaje de programación R [1] que ayuda a simular el lanzamiento del dado un número variable de veces. Los resultados obtenidos se muestran en el cuadro 2 y confirman lo estipulado.

Cuadro 2: Promedio de los lanzamientos de un dado.

Lanzamientos	**Promedio**
10	5.000
100	3.230
1,000	3.525
10,000	3.492
100,000	3.503

3. Desigualdad de Chebyshev

Teorema 3 (Desigualdad de Chebyshev) *Sea X una variable aleatoria discreta con valor esperado $E[X] = \mu$ y sea $\epsilon > 0$ un número entero positivo. Entonces*

$$P(|X - \mu| \geq \epsilon) \leq \frac{V[X]}{\epsilon^2}$$

Si se combina la ley de los grandes números y la desigualdad de Chebyshev se obtiene un resultado interesante:

$$P(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Referencias

[1] Selva Prabhakaran. R-statistics. http://r-statistics.co/Statistical-Tests-in-R.html.

Teorema del límite central

Gabriela Sánchez Y.
5064

1. Teorema del límite central

El teorema del límite central es otro de los teoremas fundamentales de la probabilidad. El enunciado formal de acuerdo a Grinstead y Snell [1] es el siguiente.

Teorema 1 (Teorema del límite central) *Sea $X_1, X_2, \ldots, X_n$ una secuencia de variables aleatorias independientes e idénticamente distribuidas y sea $S_n = X_1 + X_2 + \ldots + X_n$. Para cada n la media y varianza de X_n se denotan por μ_n y σ_n^2, respectivamente. Defina la media y varianza de S_n como m_n y s_n^2, respectivamente y asuma que $s_n \to \infty$. Si existe una constante A tal que $|X_n| \leq A$ para todo n, entonces para $a < b$,*

$$\lim_{n\to\infty} P\left(a < \frac{S_n - m_n}{s_n} < b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{x^2/2}\, dx.$$

En esencia lo que el teorema dice es que el promedio de las medias de muestras será la media de la población. Es decir, si se suman las medias de todas las muestras y se determina el promedio, ese promedio será la media de la población real.

2. Aplicaciones

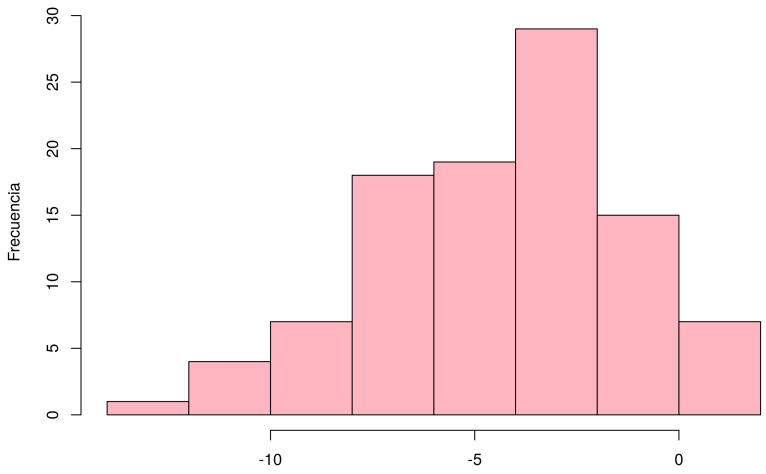
2.1. Experimentación con datos de la tesis

Se tienen soluciones de un modelo con dos métodos distintos. Uno de ellos, un optimizador y el otro una metaheurística. Para cada resultado se hace una comparación para saber qué tan alejados están uno del otro, tomando como base el valor obtenido por el optimizador. En otras palabras, se calcula un gap de la siguiente manera

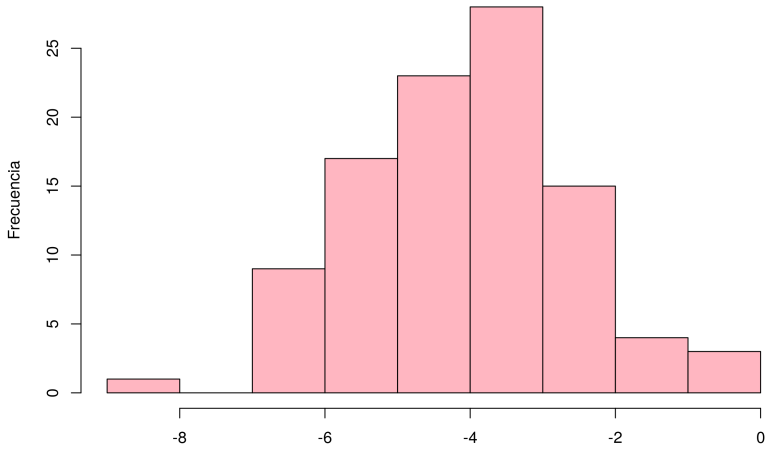
$$gap = \frac{z_{opt} - z_{met}}{z_{opt}} \times 100\,\%,$$

en total se tienen 3510 resultados y se desea saber si habría un comportamiento tal y como lo dice el teorema.

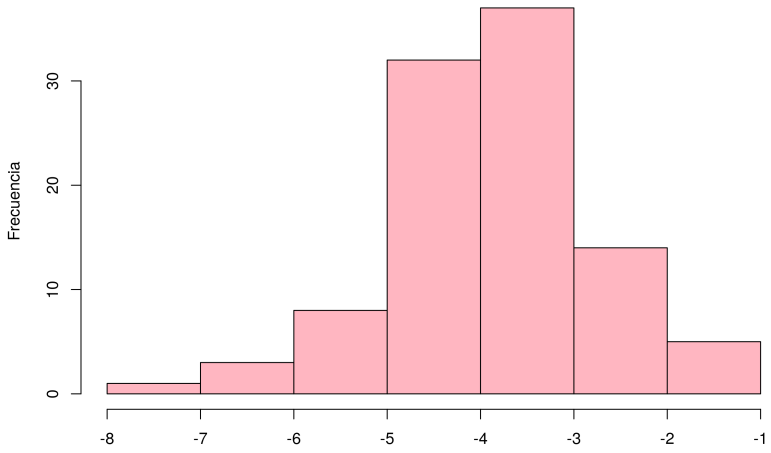
Se toman cien muestras de tamaño n variable, con $n = 10$, 50 y 80. La distribución que tienen las medias de las muestras se puede observar en los histogramas de la figura 1. En las figuras 1b y 1c se observa un poco mejor el comportamiento de *campana*, de estos resultados se puede concluir que la media de la población (los gap de todas las instancias) es cercano a -4. Este resultado se comprobó comparando con la media real de la población que es de -4.052034.



(a) $n = 10$



(b) $n = 50$



(c) $n = 80$

Figura 1: Distribución de las medias de cien muestras de tamaño variable.

Referencias

[1] Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability*. AMS, 2003.

Propuestas de proyecto integrador

Gabriela Sánchez Y.
5064

1. **Calibración de parámetros:** El método de solución propuesto en el trabajo de tesis requiere diferentes parámetros entre los cuales se encuentran el número de iteraciones del algoritmo, el periodo de reevaluación de probabilidades y el criterio de paro de la búsqueda local. Distintas combinaciones de los mismos dan lugar a mejores o peores soluciones, el objetivo es determinar, mediante un diseño de experimentos, la combinación de parámetros que logra los mejores resultados para las distintas clases de instancias.

2. **Comparación de soluciones:** Como parte del trabajo de tesis se tienen datos sobre las soluciones obtenidas con dos formulaciones diferentes, se desea analizar dichas soluciones para verificar si hay diferencias significativas entre ambas. Como primera instancia se pretende usar estadística descriptiva y después verificar con pruebas de hipótesis que sean aplicables a los datos.

3. **Incidencia delictiva:** Analizar los reportes de incidencia delictiva en las entidades federativas del país en el periodo comprendido del año 2015 a octubre 2020 que proporciona el Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública [1], usando estadística descriptiva y pruebas de hipótesis. Además ya que se cuenta con una base de datos amplia, puede aplicarse la ley de los grandes números y el teorema del límite central.

4. **Efectos del covid en eduación:** debido a la contingencia sanitaria que atravesamos a causa del coronavirus diversas actividades han tenido que ser suspendidas o modificar la forma en que se llevan a cabo. Tal es el caso de la educación en línea. Me gustaría analizar el efecto que ha tenido la contingencia sanitaria en la educación principalmente en el estado de Michoacán.

Referencias

[1] Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública. Incidencia delictiva. https://www.gob.mx/sesnsp/acciones-y-programas/incidencia-delictiva-87005?idiom=es.

Retroalimentación propuestas de proyecto integrador

Gabriela Sánchez Y.
5064

1. Propuesta Erick

En mi investigación de tesis, se esta analizando una variante del problema de enrutamiento de vehículos, en el cuál, los vehículos salen del depósito inicial y deben moverse a algún nodo en donde se requiera satisfacer la demanda de algún cliente, para nuestro caso el primer movimiento es hacia algún hotel, por lo que se desea analizar la frecuencia/probabilidad con la que los nodos asignados a los hoteles están conectados en la ruta de algún vehículos o no lo están en las diferentes soluciones encontradas, ya que esto permitiría ver la pertinencia de analizar o no un modelo en dónde los hoteles sean tratados como depósitos iniciales

1.1. Retroalimentación

Me parece interesante lo que deseas hacer pero ¿cuál es la razón por la que te interesa ese caso particular? ¿qué otro tipo de nodos tienes?

2. Propuesta Johana

En el tema de tesis, se cuenta con un modelo matemático y una metaheurística para encontrar la solución del problema planteado. Por medio de la prueba de hipótesis de medias de diferencia se pretende comprobar que existe un ahorro entre la metaheurística y la solución ofrecida por el modelo matemático en un 95 % y determinar que tanto mejora la solución considerando intervalos de confianza del 90 % y 95 %. Estas pruebas se aplicarán tanto por tamaño de instancias (pequeñas, medianas tipo 1, medianas tipo 2 y grandes) como para el total de instancias. Además, se aplicaría la prueba de hipótesis para proporciones para determinar la proporción de mejores soluciones encontradas mediante el uso de la metaheurística.

2.1. Retroalimentación

Podrías revisar primero si tus dos poblaciones de resultados son independientes o no y así sabrías si puedes aplicar más pruebas y de qué tipo deben ser.

3. Propuesta Mayra

Law of Large Numbers. – For this subject, we wanted to explore some of the qualities of different optimizers used in training convolutional neural networks. Each one of them has different features that try to correct the failings of its predecessors. And it is because all of this different versions that there is not one optimizer that is perfect for a certain problem. With the LLN we want to see if the optimizers are affected in a good or a bad way. If it helps them converge to the closest to optimal value, or if in some cases it becomes flawed and lands in over training.

3.1. Retroalimentación

¿Realizarás el análisis para un problema en particular? Podrías realizar el análisis en conjunto con el teorema del límite central.

Optimización de la planificación de servicios: análisis de soluciones

Gabriela Sánchez Yepez

[a]*Posgrado en Ingeniería de Sistemas, Universidad Autónoma de Nuevo León, gabriela.sanchezypz@uanl.edu.mx,*

Abstract

Se han propuesto dos modelos para resolver un problema de toma de decisiones. En objetivo del presente trabajo es aplicar herramientas estadísticas que ayuden a determinar si hay diferencias significativas entre las soluciones que proporcionan ambos modelos.

Keywords:
Modelo matemático, GRASP reactivo, gap

1. Introducción

En el presente trabajo se analizan los resultados obtenidos al resolver un problema de planificación de servicios de telecomunicaciones que consiste en asignar órdenes de servicio a un conjunto de cuadrillas de trabajadores, así como en determinar la secuencia en que deben realizarse dichos servicios, con el fin de balancear el salario de las cuadrillas sujeto a diversas restricciones. Cada orden tiene asignado un puntaje que depende de la dificultad de la misma, por este motivo el salario de los trabajadores depende directamente de la cantidad y el tipo de servicio que realicen.

Ya que no existe una única forma de plantear el balance del salario se proponen dos formulaciones matemáticas. Una de ellas aborda el balance maximizando el mínimo puntaje colectado por las cuadrillas (s_{min}), por lo tanto se referirá a esta formulación como modelo *Max-Min*. La segunda formulación toma como base el caso ideal, que consiste en considerar que todas las cuadrillas obtienen un puntaje igual a la media de los puntos de las órdenes (μ). Ésta busca que los puntajes colectados por las cuadrillas se alejen lo menos posible de esta media. Es decir, el segundo modelo, al cual se referirá

como modelo *Min-Max* tiene como función objetivo minimizar la máxima de las desviaciones respecto a μ. Más detalles acerca del problema así como las formulaciones completas se encuentran en el trabajo de Sánchez-Yepez (2019).

El objetivo del proyecto es determinar si al resolver ambos modelos utilizando como método de solución una metaheurística basada en un GRASP reactivo (Sánchez-Yepez, 2019), hay diferencias significativas entre las soluciones. Para cumplir con dicho objetivo se emplean distintos test estadísticos para los cuales se ha seleccionado un nivel de significancia de 0.05.

El resto del documento se distribuye de la siguiente manera: en la sección 2 se describe cómo se obtienen los datos a analizar, la sección 3 presenta y analiza los datos utilizando distintas herramientas de estadística descriptiva, mientras que en la sección 4 se describe el uso de pruebas estadísticas para justificar si hay diferencias entre los datos. Finalmente, la sección 5 presenta las conclusiones del trabajo.

2. Datos

Se plantean dos formulaciones distintas para resolver un problema de toma de desiciones. Para validar las mismas y contar con un punto de referencia al resolver con la metaheurística GRASP reactiva, los modelos se resuelven con el optimizador CPLEX versión 12.8 usando un solo hilo y considerando dos criterios de paro: un tiempo de cómputo máximo de 7200 segundos o un gap de 0.0 % (soluciones óptimas).

Las instancias utilizadas en la experimentación son adaptadas de las propuestas en la literatura (Chao et al., 1996). El conjunto se divide en siete clases teniendo en total 353 instancias disponibles. Las instancias pertenecientes a una misma clase contienen el mismo grafo y varían en la cantidad de cuadrillas disponibles y tiempo límite para la duración de las rutas. Las características de cada clase de instancias se especifica en el cuadro 1.

Debido a la naturaleza aleatoria del GRASP reactivo, se proporcionan diez soluciones a cada instancia para obtener poblaciones de tamaño 3530. Teniendo en mente el punto de referencia, los conjuntos de datos que se analizan son diferencias porcentuales (**gap**) entre el valor objetivo de las soluciones obtenidas con la metaheurística y el valor objetivo de la mejor solución encontrada por CPLEX, para un modelo en particular.

Por ejemplo, el gap para el modelo Max-Min queda determinado por la

Cuadro 1: Instancias.

Clase	**Órdenes**	**Instancias**			
		Cuadrillas			**Total**
		2	**3**	**4**	
I	21	11	11	11	33
II	32	17	16	15	48
III	33	20	20	20	60
IV	64	11	8	5	24
V	66	25	25	24	74
VI	100	20	19	17	56
VII	102	20	19	19	58

ecuación (1)

$$gap = 100 \times \frac{Z_{\text{cplex}} - Z_{\text{grasp}}}{Z_{\text{cplex}}}, \tag{1}$$

donde Z_{cplex} y Z_{grasp} representan el valor de la función objetivo obtenida por CPLEX y el GRASP reactivo, respectivamente.

Es claro que el gap puede ser tanto positivo como negativo. Si es positivo indica que el valor objetivo encontrado con CPLEX es mejor que el encontrado por la metaheurística, en cambio si es negativa indica que el valor objetivo encontrado con la metaheurística es mejor que el determinado por el optimizador. Para tener la misma interpretación al analizar los resultados del modelo Min-Max, el gap determinado por ecuación 1 se multiplica por (-1).

Es de interés determinar el comportamiento que siguen las soluciones obtenidas con un modelo específico al evaluar con una función objetivo distinta. En este caso las soluciones del modelo Max-Min son evaluadas en la función objetivo del modelo Min-Max y las soluciones del modelo Min-Max son evaluadas en la función objetivo del modelo Max-Min.

Considerando toda la información descrita en la sección, en total se analizan cuatro conjuntos de soluciones: las del modelo Max-Min, las soluciones del modelo Min-Max evaluadas en la función objetivo del modelo Max-Min, las del modelo Min-Max y las soluciones del modelo Max-Min evaluadas en la función objetivo del modelo Min-Max, identificadas por s_1, s_{21}, s_2 y s_{12}, respectivamente. El cálculo del gap en cada uno de estos conjuntos da lugar a los cuatro distintos conjuntos de datos que se analizan en el trabajo.

3. Estadística descriptiva

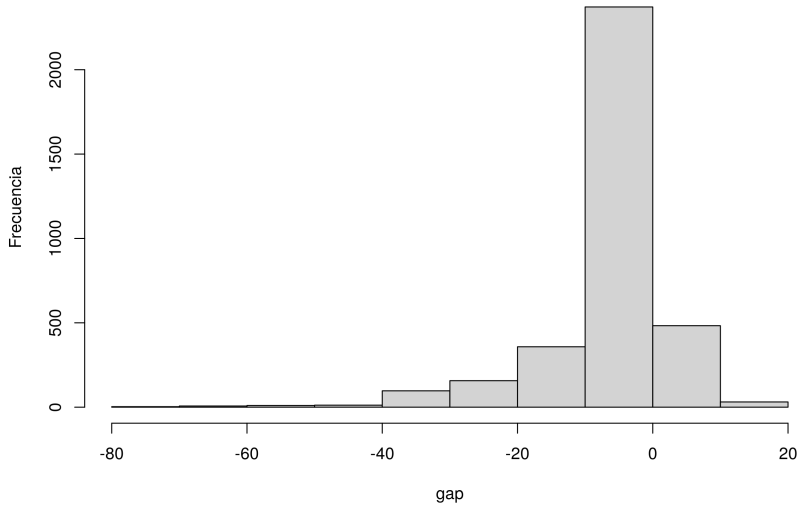
La distribución de los resultados del gap en cada conjunto de datos se presentan en las figuras 1 y 2. El apoyo visual de estas figuras permite identificar que existe un gran número de replicas para las cuales no hay diferencia entre los dos métodos de solución empleados (metaheurística y optimizador). Esto es de esperarse ya que el optimizador logra encontrar soluciones óptimas para aproximadamente un 50 % del total de las instancias en ambos modelos y los resultados de la metaheurística se mantienen muy similares. De forma muy particular en casi el 100 % de las instancias de las clases I y II, los resultados son los mismos, esto puede observarse más a detalle en los diagramas de caja bigote de las figuras 3 y 4.

Es importante destacar que en adelante el análisis se realiza por pares ya que el interés es determinar si es que alguno de los modelos da mejores resultados. Recuerde que la diferencia entre las formulaciones radica en la función objetivo.

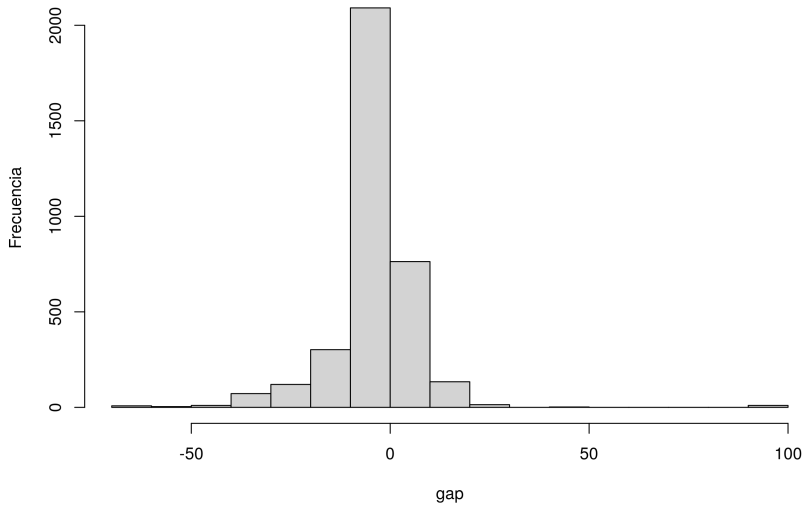
Los pares de conjuntos que se analizan son el el gap al evaluar las soluciones del modelo Max-Min en la función objetivo Min-Max (s_{12}) junto con el gap de las soluciones s_1 y, el gap al evaluar las soluciones del modelo Min-Max en la función objetivo Max-Min s_{21} y el gap de las soluciones s_2. Si se analizan los diagramas de caja bigote de acuerdo a los pares antes mencionados, se puede observar que los resultados son similares con mayor frecuencia en las clases más pequeñas. La figura 5 permite observar las densidades de los resultados separados de acuerdo a la función objetivo. En esta figura son más notorias las similitudes que las diferencias.

Si se comparan directamente las medias de los conjuntos de datos se observa que la media del gap obtenido de las soluciones s_1 es -4.0173 mientras que la media del gap de las soluciones s_{21} es -1.7953. En ambos casos es negativo lo que indica que, en promedio, las soluciones de la metaheurística son mejores que las del optimizador. Ya que la media de las soluciones s_1 es más pequeña, se podría decir que en promedio el modelo Max-Min mostró mejor desempeño. Al realizar el mismo análisis con las soluciones s_2 y s_12 los resultados obtenidos para las medias son 0.0313 y -1.9195, lo que permite concluir que en promedio nuevamente las soluciones del modelo Max-Min muestran mejores resultados.

El análisis descriptivo muestra que sí hay ligeras diferencias entre los conjuntos. Sin embargo, no es suficiente para justificar si en general la diferencia será significativa o no.

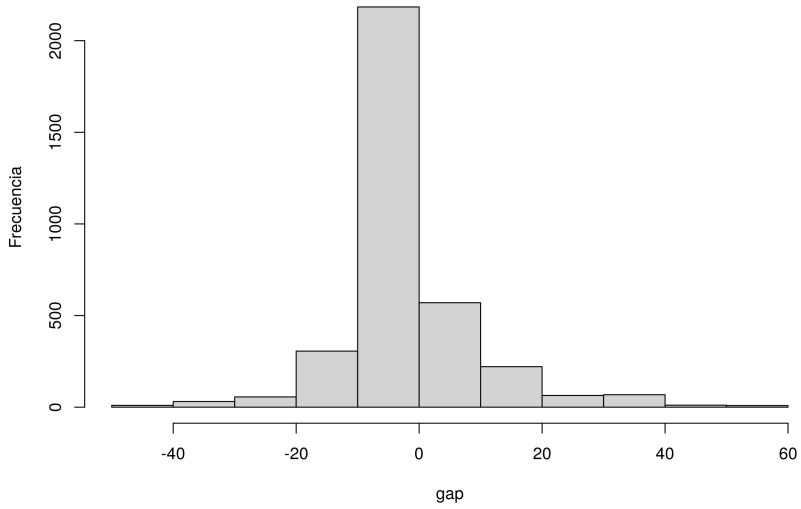


(a) Soluciones s_1

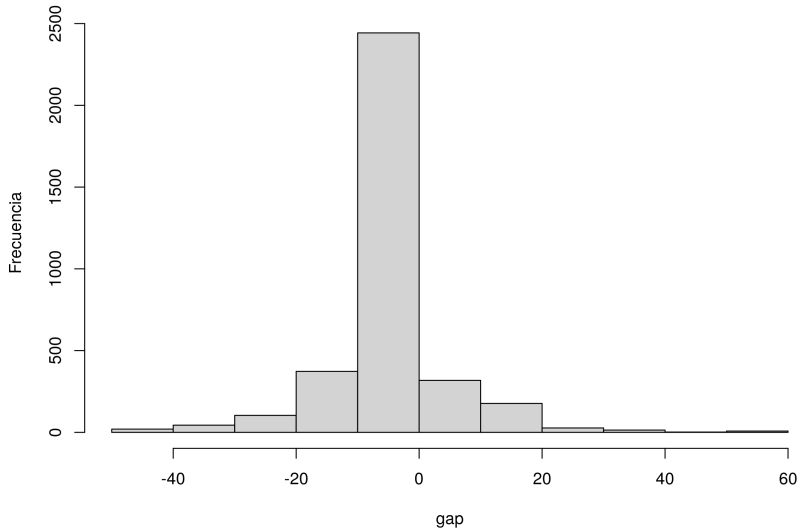


(b) Soluciones s_{21}

Figura 1: Gap para la función objetivo Max-Min.

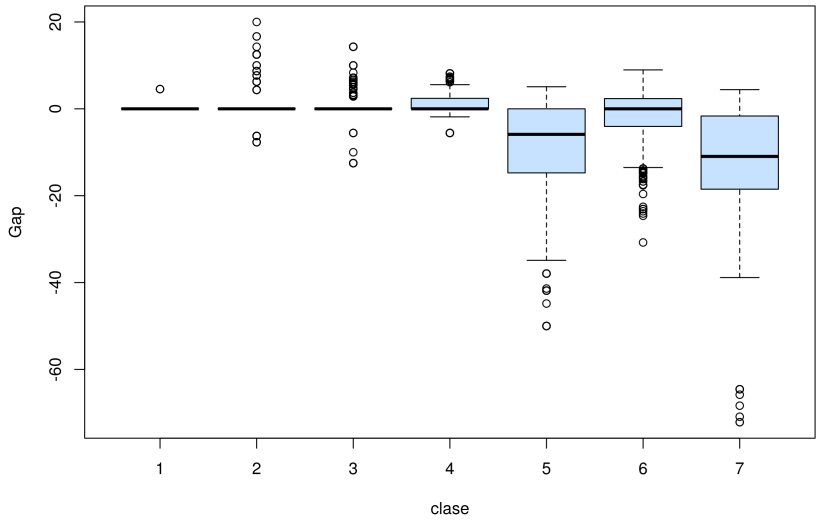


(a) Soluciones s_2

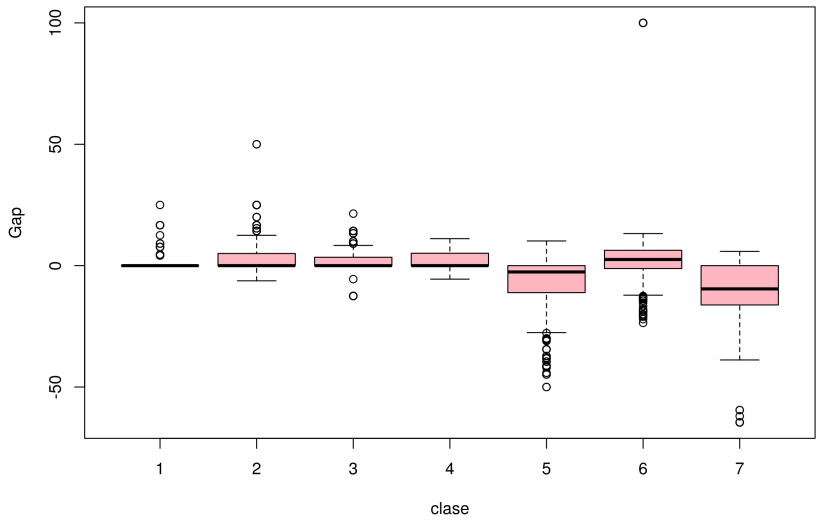


(b) Soluciones s_{12}

Figura 2: Gap para la función objetivo Min-Max.

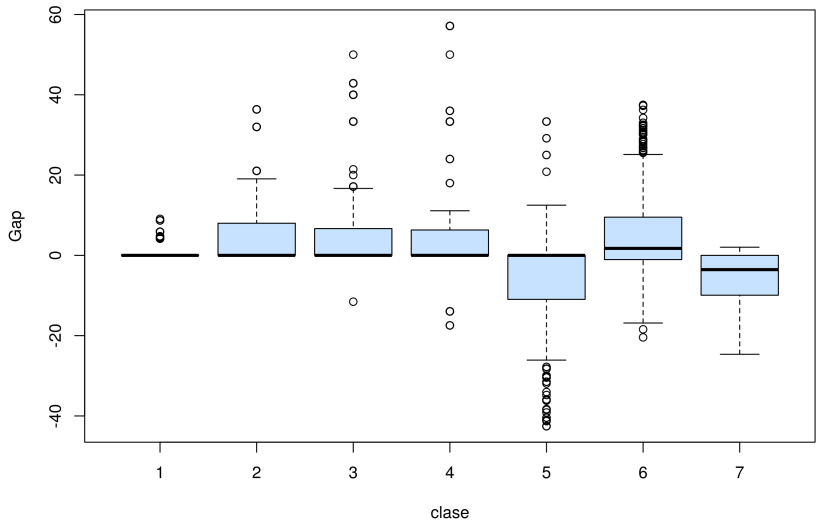


(a) Soluciones s_1

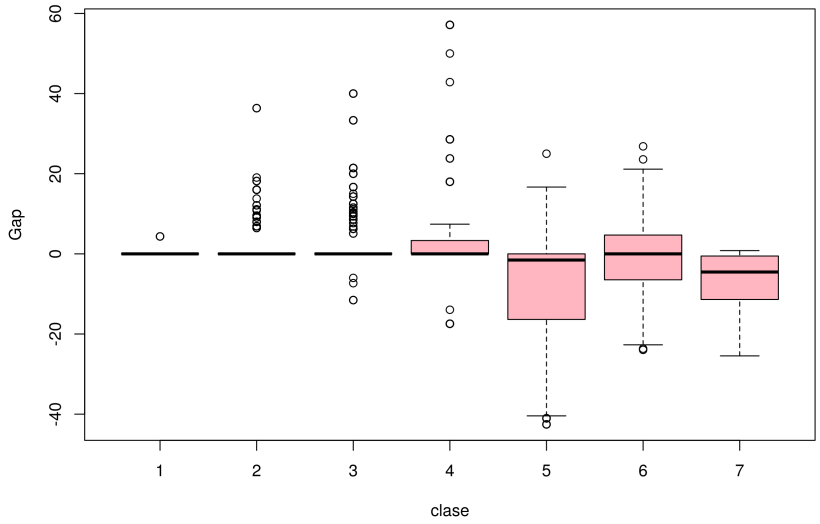


(b) Soluciones s_{21}

Figura 3: Gap para la función objetivo Max-Min separado por clase.

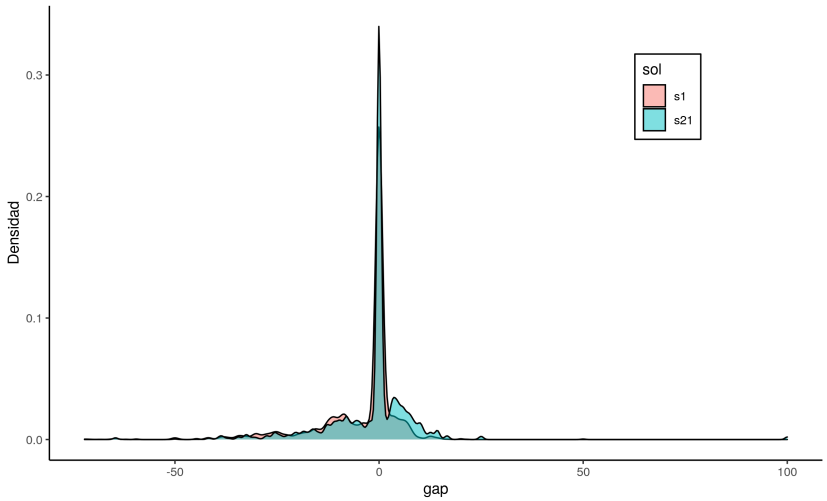


(a) Soluciones s_2

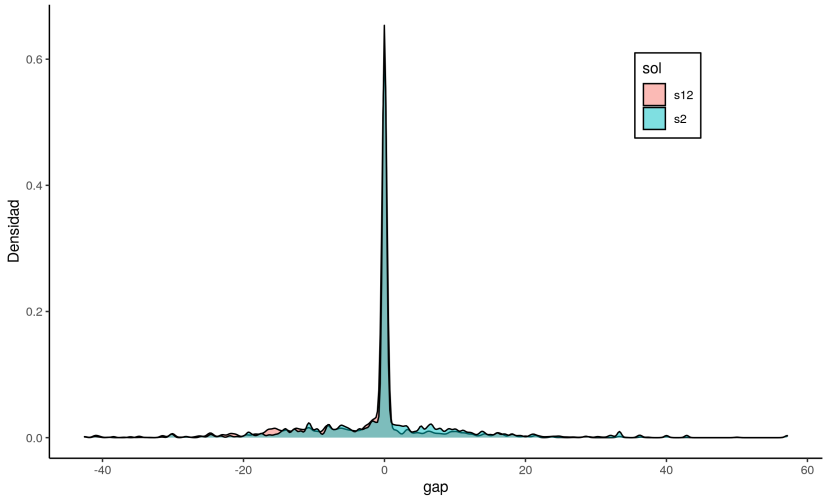


(b) Soluciones s_{12}

Figura 4: Gap para la función objetivo Min-Max separado por clase.



(a) Función objetivo Max-Min



(b) Función objetivo Min-Max

Figura 5: Densidad del gap separado por función objetivo.

4. Pruebas estadísticas

En esta sección se emplean pruebas estadísticas con el objetivo de justificar diferencias significativas en los resultados. Todas las pruebas usadas se realizan utilizando como apoyo el lenguaje de programación R (The R Foundation). Como primer paso, es importante verificar si las muestras cumplen con la normalidad o no, ya que ayudará a determinar si deben usarse pruebas paramétricas o no paramétricas.

4.1. Normalidad

Para determinar si los conjuntos de datos siguen una distribución normal se emplea la prueba de *Shapiro-Wilk*. Este test plantea la hipótesis nula de que la muestra proviene de una distribución normal, por lo tanto si el p-valor es menor que 0.05 se rechaza la hipótesis nula y se concluye que los datos no provienen de una distribución normal. Los resultados de los cuatro distintos conjuntos de datos indican que ninguno cumple con la normalidad por lo tanto sólo son utilizadas pruebas no paramétricas para analizar las diferencias entre los datos.

4.2. Prueba de los rangos con signo de Wilcoxon

Se resalta que el conjunto de instancias usadas para resolver ambos modelos es el mismo, por lo tanto se tienen datos *pareados*. La prueba de los rangos con signo de Wilcoxon permite analizar si hay diferencia entre las muestras pareadas. La hipótesis nula de la prueba es que la mediana de las diferencias de cada par es cero.

Al realizar la prueba para los dos pares de conjuntos de datos de estudio, el p-valor permite rechazar la hipótesis nula ya que es menor que el nivel de significancia. Se debe hacer énfasis en que existen varios pares en los cuales ambos datos son iguales, eso ocasiona lo que se conoce como *ties* o *ligaduras*. En este caso `wilcox.test()` devuelve un p-valor aproximado por lo que no se puede hacer una conclusión al respecto.

Se optó por hacer una tercera prueba pareada en la que se consideran como datos las evaluaciones cruzadas, es decir el gap obtenido de las soluciones s_{12} y s_{21}. El resultado del p-valor en este panorama no permite rechazar la hipótesis nula. Sin embargo, sucede lo mismo que en los casos anteriores, hay pares de datos iguales que generan ligaduras por lo que tampoco se puede concluir con seguridad un resultado.

5. Conclusiones

El empleo de la prueba para muestras pareadas no permite hacer conclusiones respecto a los datos ya que los datos generan ligaduras. Como trabajo a futuro se puede implementar la estrategia descrita por DeGroot and Schervish (2012) para tratar con estos casos. Otro punto a resaltar en esta prueba es que de cada instancia hay diez replicas por lo que podría haber diferencias en el resultado del p-valor si los pares de conjuntos de datos se consideran de forma diferente. Una estrategia a implementar es reducir el tamaño de las muestras al usar únicamente los mejores resultados obtenidos para cada una de las instancias.

Referencias

Chao, I.M., Golden, B.L., Wasil, E.A., 1996. The team orienteering problem. European Journal of Operational Research 88, 464 – 474.

DeGroot, M., Schervish, M., 2012. Probability and Statistics. Addison-Wesley.

Sánchez-Yepez, G., 2019. Optimización de la planificación de servicios. Master's thesis. Posgrado en Ingeniería de Sistemas, Universidad Autónoma de Nuevo León.

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