Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Posgrado en Ingeniería de Sistemas

Doctorado

Portafolio de Evidencias

de

Alberto Martínez Noa

1985271

para el curso de Modelos Probabilistas Aplicados,

con la Dra. Elisa Schaeffer.

Semestre Agosto 2020 - Enero 2021.

https://github.com/Albertomnoa/Tareas_MPA

Tarea 1 de Modelos Probabilistas Aplicados

Sector Construcción

5271

8 de septiembre de 2020

1. Origen de los datos

Los datos utilizados en este trabajo fueros obtenidos en el sitio https://www.inegi.org.mx que pertenece al Instituto Nacional de Estadística y Geografía (INEGI). Se escogió el apartado Construcción donde se muestra información sobre los principales resultados de las Empresas Constructoras, comprende unidades económicas dedicadas principalmente a la edificación; a la construcción de obras de ingeniería civil y a la realización de trabajos especializados de construcción. De este apartado se descargó el tabulado (Valor de producción generado por las empresas constructoras según el tipo de obra) en formato *csv*. En el cuadro 1 de la página 1 se muestra un fragmento de los datos descargados.

Cuadro 1: Fragmento del tabulado utilizado

Periodo	**Año**	**Total**	**Edific**	**Agua_R_S**	**Elect_C**	**Transp**	**Petr_p**	**Otras_C**
Enero	2006	23649637	11562266	1527733	1062658	5237209	2698498	1561273
Febrero	2006	23186956	11513255	1518386	808459	4901882	2931457	1513517
Enero	2007	27801512	13848554	1314745	1011836	6078270	3554327	1993780
Febrero	2007	27019544	13948606	949590	955818	6012856	3197691	1954983
Enero	2020	36031814	17703639	1363041	2418338	7684282	2265532	4596982
Febrero	2020	36626861	18064797	1186512	2189154	8031054	2539702	4615642

2. Análisis y tratamiento de los datos

Para el análisis de los datos del Valor de producción generado por las empresas constructoras según el tipo de obra, se tiene en cuenta que poseen una frecuencia de ocurrencia mensual, en el periodo comprendido entre los años 2006 y 2020. Por lo que se tratan los mismos como una serie de tiempo. El análisis será realizado en el programa R versión 4.0.2 [1] en el entorno de desarrollo Rstudio [2]. Para llevar al formato de serie de tiempo se realizó el código 2 de la página 1.

```r
datos= read.csv("datos_new.csv",header = TRUE)
datos=datos[,-(1:2)]

Valor=ts(datos,start = c(2006,1), frequency = 12)
```

Series.R

2.1. Análisis

En el cuadro 2 de la página 2 se muestra un resumen de los resultados de varias funciones de ajuste de modelos aplicadas a los datos, Aunque la misma es bastante informativa, pero no tan amigable como una representación gráfica. Por lo anterior es conveniente utilizar el diagrama de caja y bigotes para una mejor comprensión de la información. En la figura 1 de la página 2 se muestra la relación entre el valor en miles de pesos corrientes y el tipo de obra.

Cuadro 2: resumen de varias funciones de ajuste de modelos aplicadas

	Edific	**Agua_R_S**	**Elect_C**	**Transp**	**Petr_p**	**Otras_C**
Min.	11513255	868901	769529	4901882	1717099	1277653
1st Qu.	14262184	1400904	1545408	7684372	2594559	2018676
Mediana	15600868	1596256	2267046	8712373	3007558	3458974
Media	16013749	1606719	2280334	8453855	3122855	3351165
3rd Qu.	17527312	1757474	2954560	9544727	3557916	4343621
Max.	22055350	2517616	5320331	1.1E+07	6061249	7027813

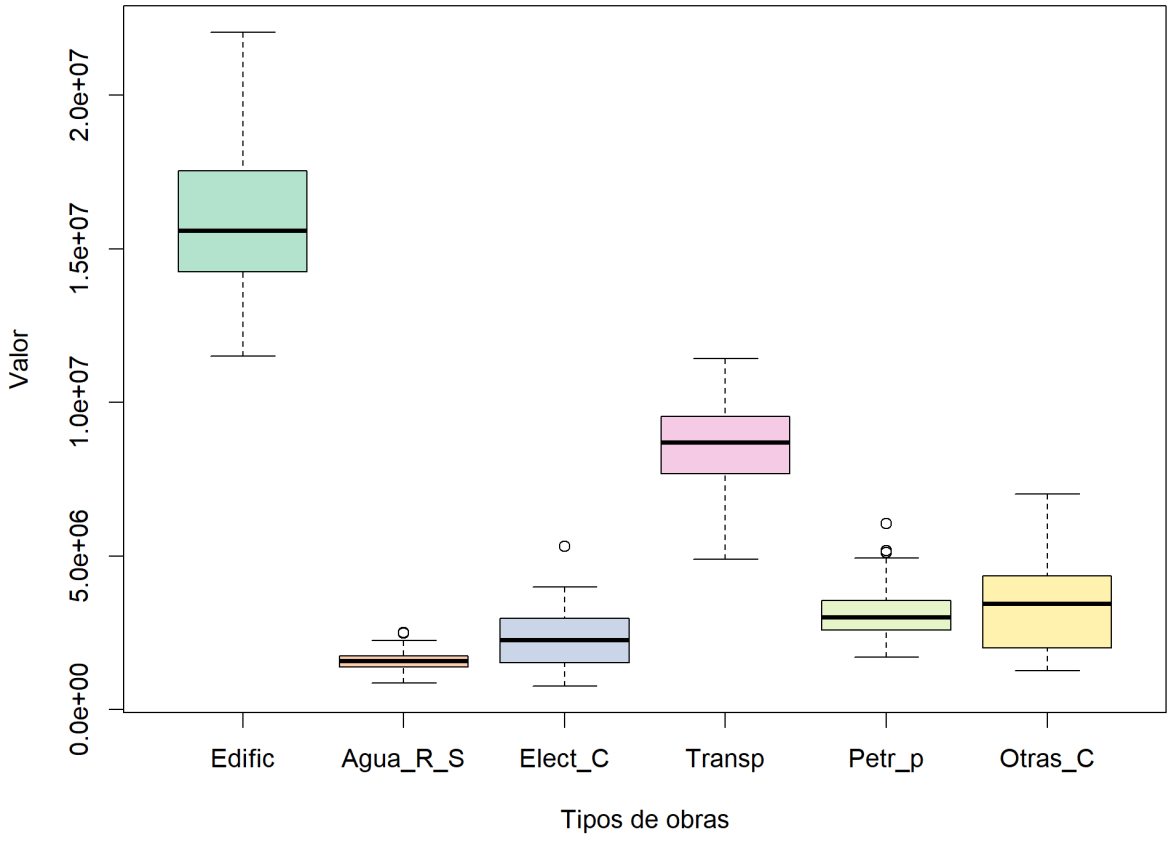


Figura 1: Diagrama de caja y bigotes que relaciona el valor en miles de pesos corrientes con los tipos de obras.

En el diagrama de caja y bigotes de la figura 1, se puede observar la diferencia entre los valores de los diferentes tipos de obras, además muestra que el tipo de obra Edificaciones es la de mayor valor de producción. Así como las de menor valor son la de Agua, riego y saneamiento y Electricidad y comunicaciones. Lo mismo se puede ver en la figura 2 de la página 3, donde se muestra una gráfica de secuencia que nos describe el comportamiento del valor de los diferentes tipos de obras a lo largo del tiempo. En la misma refleja que las obras en el sector Petróleo y petroquímica a partir de los años 2016 y 2017 tienen un menor valor que Electricidad y comunicaciones.

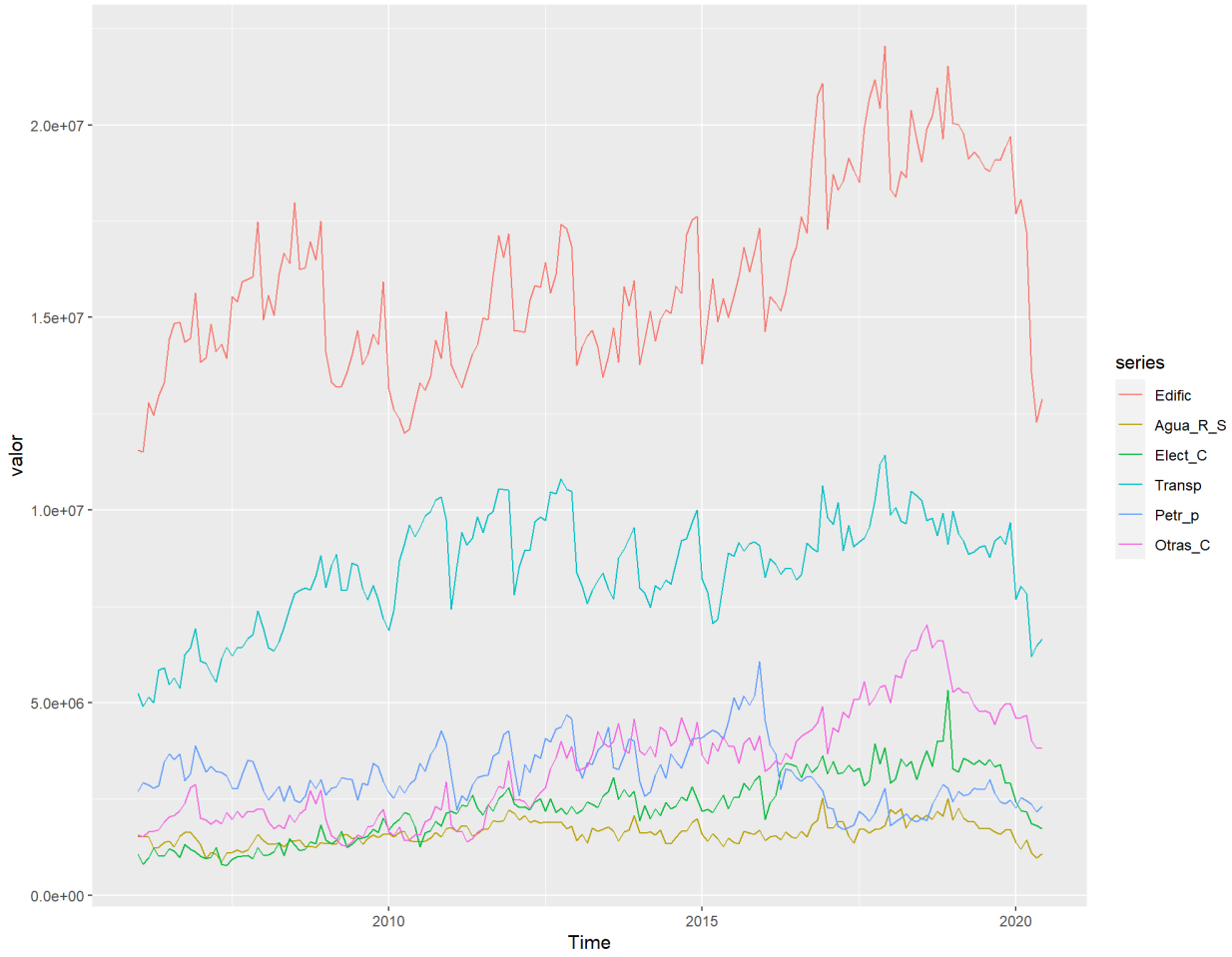


Figura 2: Diagrama de secuencia que muestra el comportamiento del valor de producción por tipo de obras a lo largo del tiempo.

Las figuras 3, 4, 5 6, 7, 8 de las paginas 4, 5, 6, 7, 8, 9 respectivamente, muestran los valores de producción de cada uno de los tipos de obra por meses. De la interpretación de estos diagramas se pueden inferir que los meses donde mayor valor se reporto fueron octubre, noviembre y diciembre de cada año.

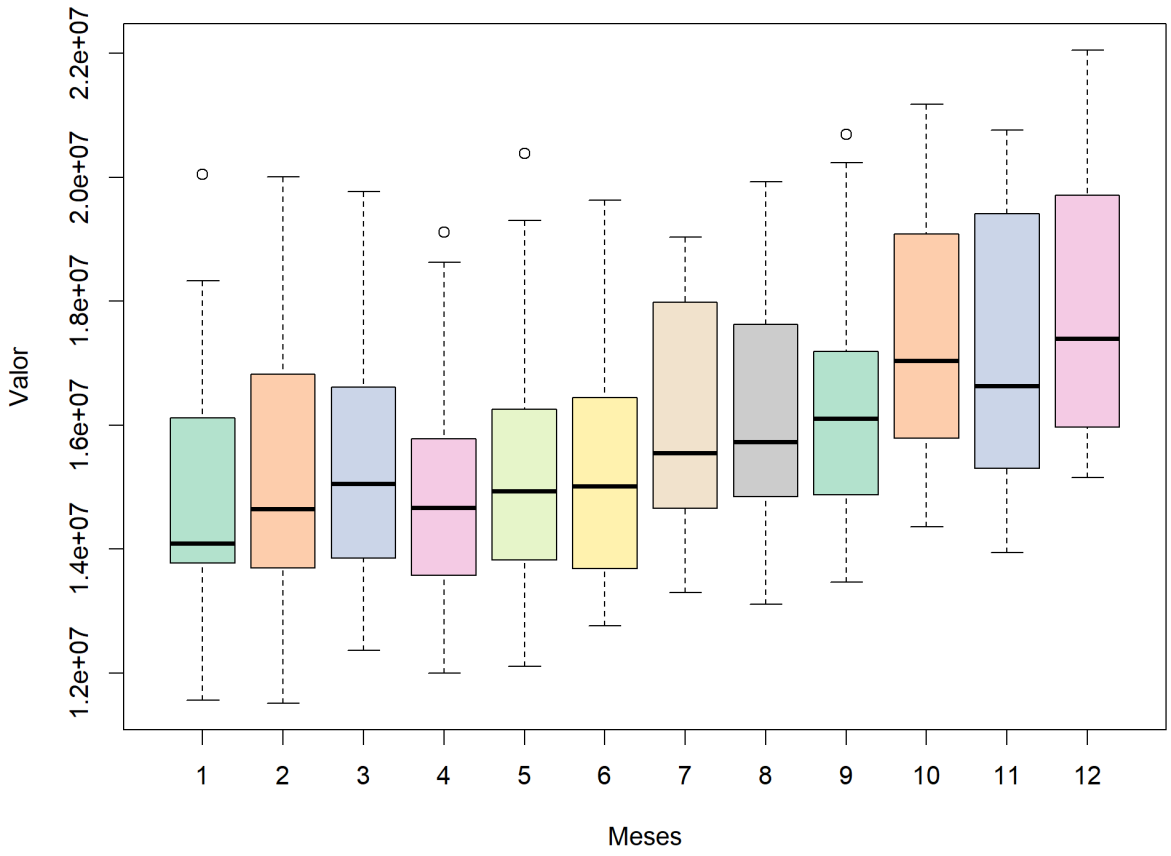


Figura 3: Diagrama de caja y bigotes del valor de producción del tipo de obra Edificación por meses.

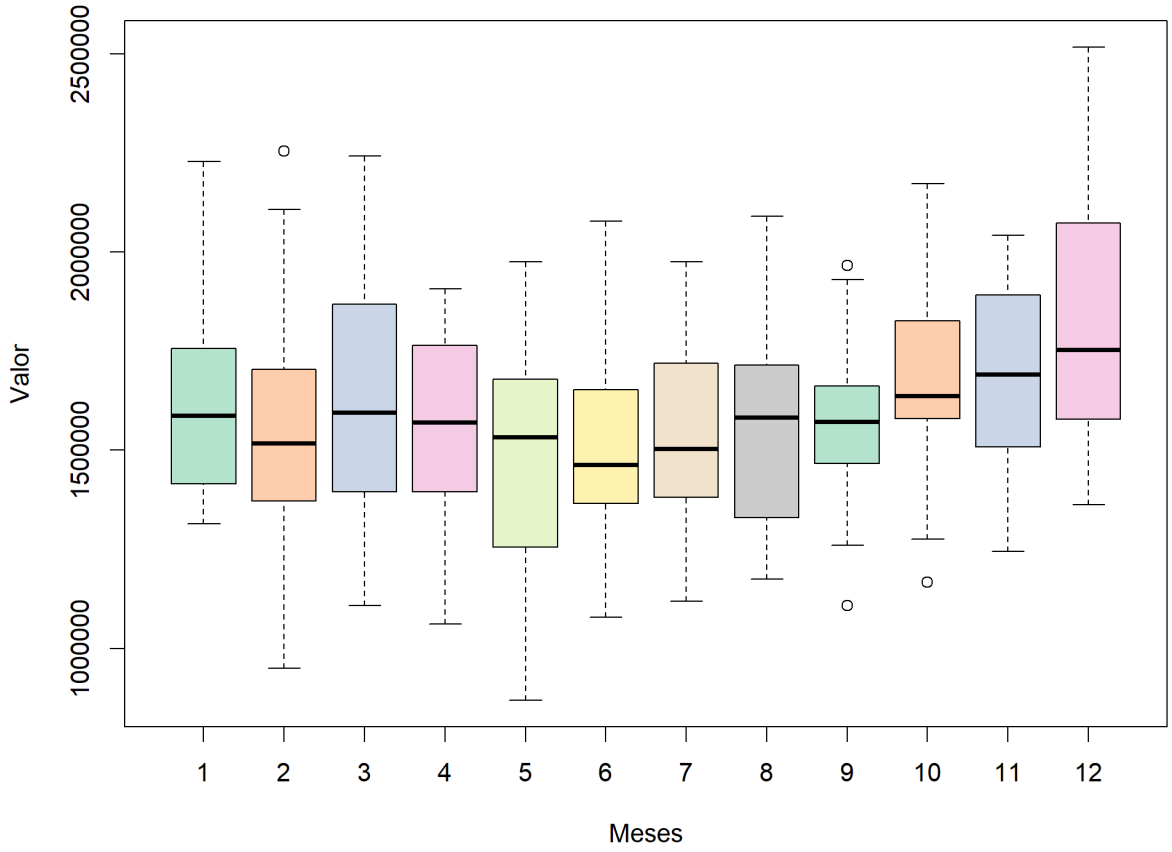


Figura 4: Diagrama de caja y bigotes del valor de producción del tipo de obra Agua, riego y saneamiento por meses.

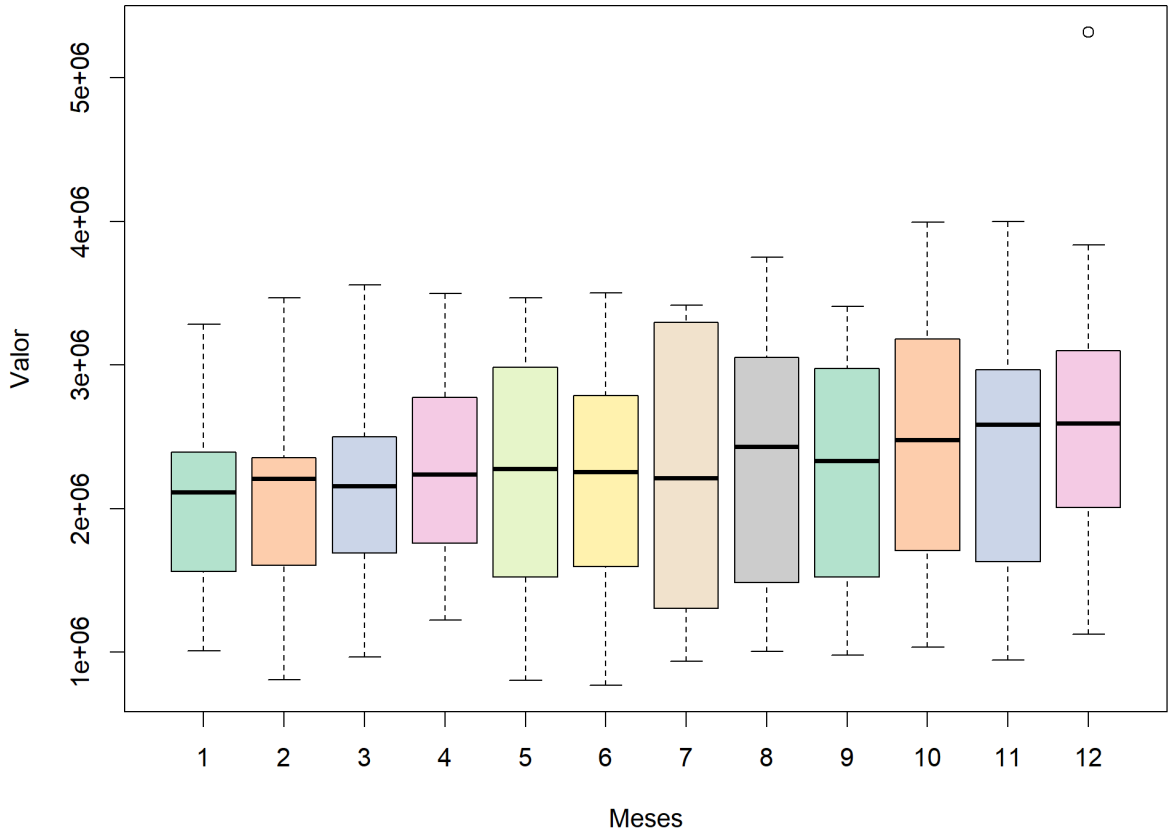


Figura 5: Diagrama de caja y bigotes del valor de producción del tipo de obra Electricidad y comunicaciones por meses.

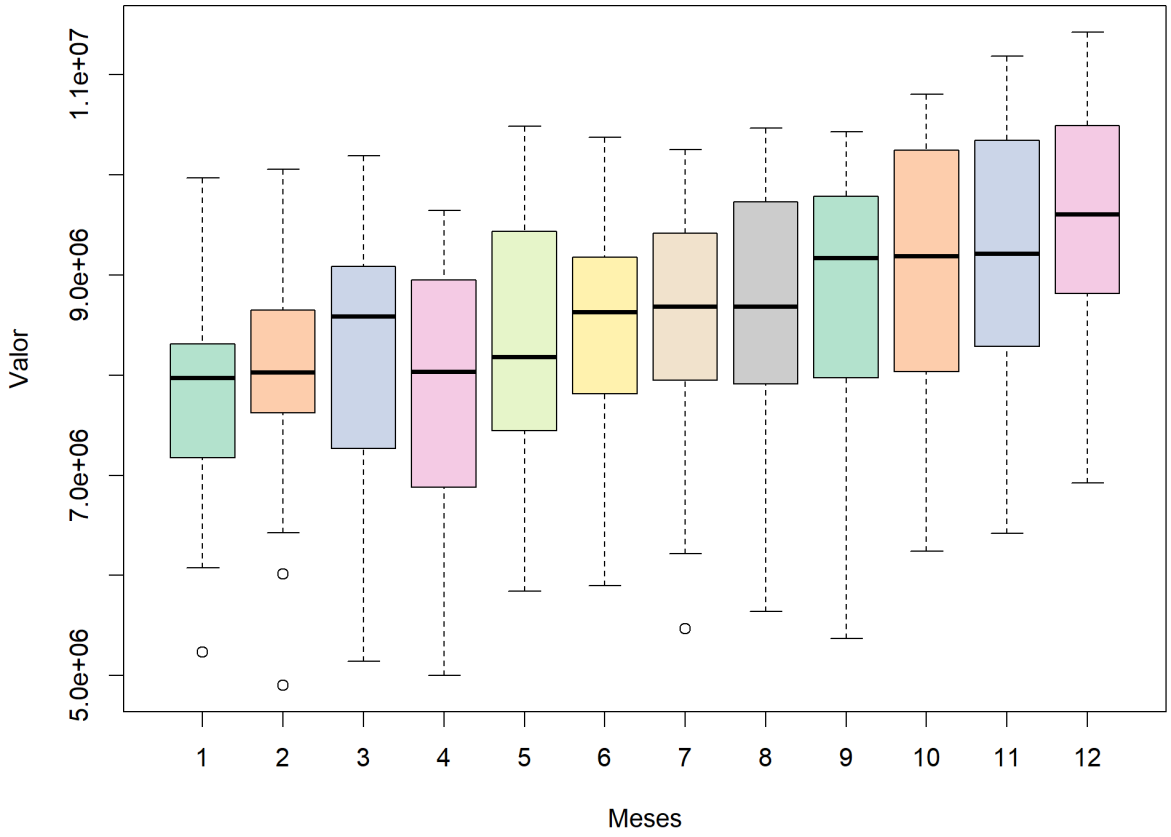


Figura 6: Diagrama de caja y bigotes del valor de producción del tipo de obra Transporte por meses.

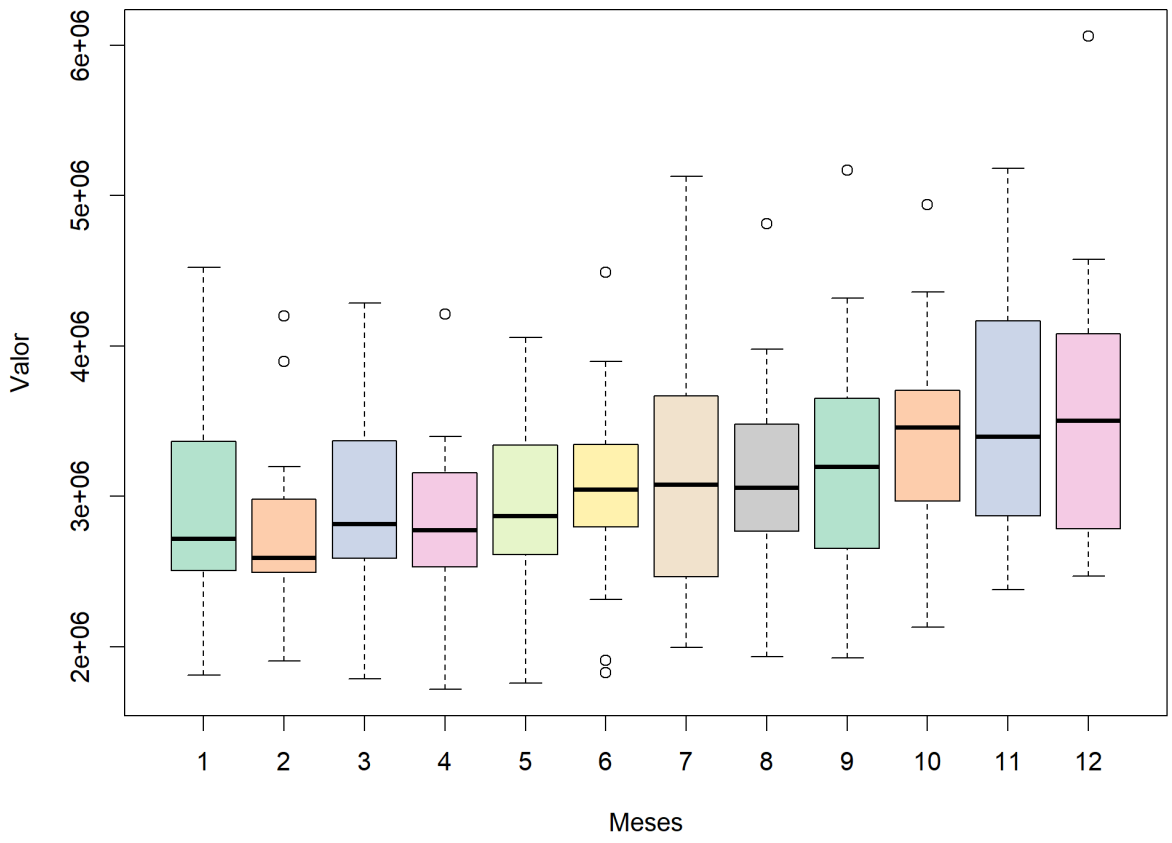


Figura 7: Diagrama de caja y bigotes del valor de producción del tipo de obra Petróleo y petroquímica por meses.

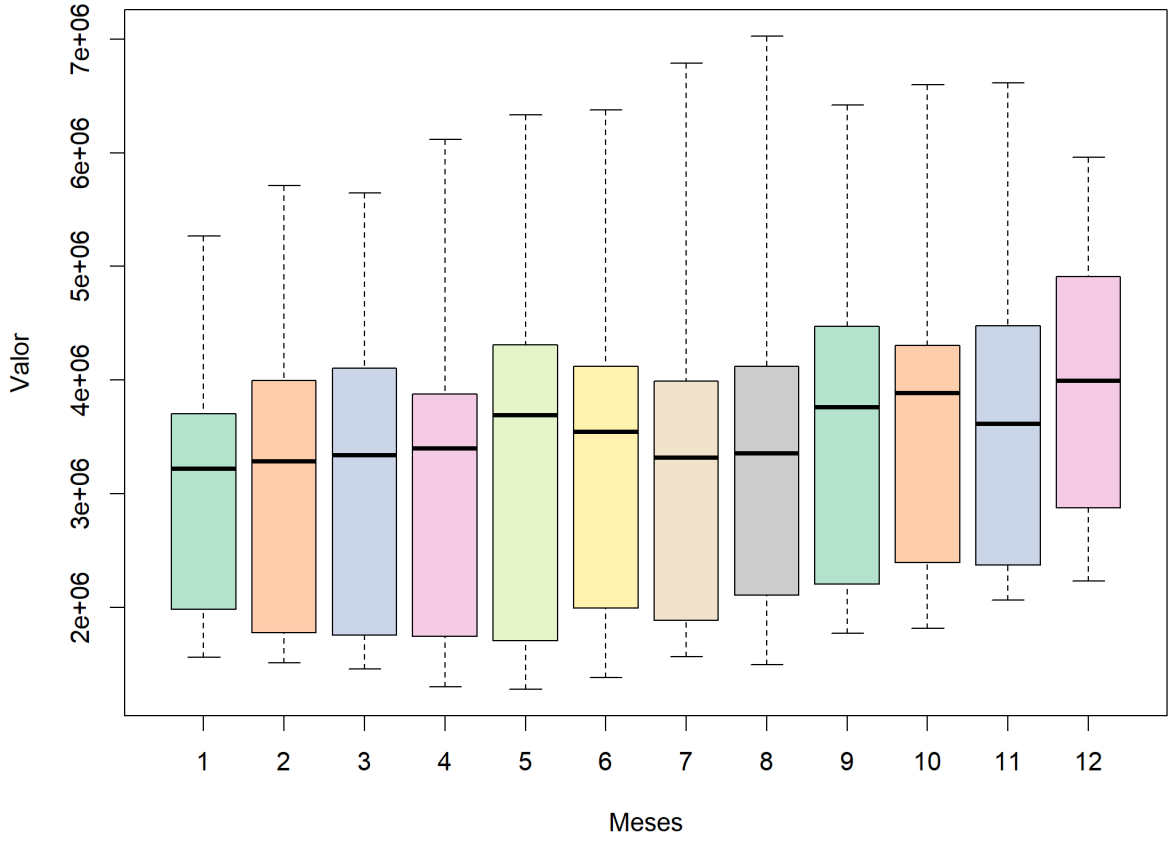


Figura 8: Diagrama de caja y bigotes del valor de producción del tipo de obra Otras obras por meses.

Para la realización del resumen y las gráficas se elaboró un el código de R 2.1 de la página 10, el código general se encuentra disponible en el repositorio https://github.com/Albertomnoa/Tareas_MPA/Tarea1

```
Sumario = summary(Valor)
capture.output(Sumario, file="Sumario.csv")

png(filename = "series.png",width = 2000, height = 1600, res =200)
autoplot(Valor, ts.colour = "blue", ts.linetype = "dashed")
dev.off()

png(filename = "boxplot.png",width = 2000, height = 1600, res =250)
boxplot(valor, col=palette("Pastel 2"),xlab="Tipos de obras",ylab = "Valor")
dev.off()

for (x in c(1:6)) {
  png(filename = paste("boxplot",x,".png", sep=""),width = 2000, height = 1600, res
    =250)
  boxplot(Valor[,x]~cycle(Valor[,x]),col=palette("Pastel 2"),ylab="Valor",xlab = "
    Meses")
  dev.off()
}
```

Graficar.R

3. Conclusiones

- El tipo de obra con mayor valor de producción es la de Edificaciones y las de menor valor son Agua, riego y saneamiento y Electricidad y comunicaciones.
- A partir de los años 2016 y 20017 el tipo de obra Petróleo y petroquímicas bajo su valor de producción con respecto al sector Electricidad y comunicaciones.
- El valor de todos los tipos de obras presentan una caída a partir del año 2019 con respecto a la tendencia creciente de las mismas hasta el 20018.
- Los meses donde mayores valores de producción se reportaron fue en los del ultimo trimestre de cada año.

Referencias

[1] R Core Team. *R: R: Un lenguaje y un entorno para la informática estadística.* R Foundation for Statistical Computing, Vienna, Austria, 2020.

[2] RStudio Team. *RStudio: Entorno de desarrollo integrado para R.* RStudio, PBC., Boston, MA, 2020.

Tarea 2 de Modelos Probabilistas Aplicados

Frecuencias y histogramas

5271

10 de enero de 2021

1. Origen de los datos

En este trabajo se utilizo como fuente de los datos el libro "The Adventures of Sherlock Holmes", del escritor y médico británico Arthur Conan Doyle. Este libro se encuentra disponible en la biblioteca virtual gratuita Project Gutenberg, al que se puede acceder desde el siguiente enlace: https://www.gutenberg.org.

2. Sobre el libro

Este libro es una colección de doce cuentos de Arthur Conan Doyle , los cuales fueron publicados por primera vez el 14 de octubre de 1892. El mismo agrupa los primeros cuentos con el detective consultor Sherlock Holmes , que se habían publicado en doce números mensuales de The Strand Magazine de Julio de 1891 a junio de 1892.

3. Análisis y tratamiento de los datos

Al libro seleccionado se le realiza un estudio de frecuencia de ocurrencia tanto de las letras como de las palabras que compone el texto. El análisis será realizado en el programa R versión 4.0.2 [1] en el entorno de desarrollo Rstudio [2].

3.1. Letras

De le libro en cuestión se extrajeron todas las letras que componen el texto, teniendo el mismo 432064 caracteres alfanuméricos que fueron almacenados en un *Data frame*, como se muestra en el cuadro 1 de la página 2. De los datos obtenidos solo necesitamos las letras por lo cual procedemos a un filtrado del *Data frame*. A partir de los datos filtrados se realiza un análisis de frecuencia de ocurrencia de las letras en el texto, como se muestra en el cuadro 2 de la página 2. Para una mejor comprensión de los datos obtenidos se realiza una representación gráfica de los mismos mediante un histograma, como se puede observar en la figura 1 de la página 3.

Cuadro 1: fragmento del *Data frame* que contiene todos los caracteres que aparecen en el texto.

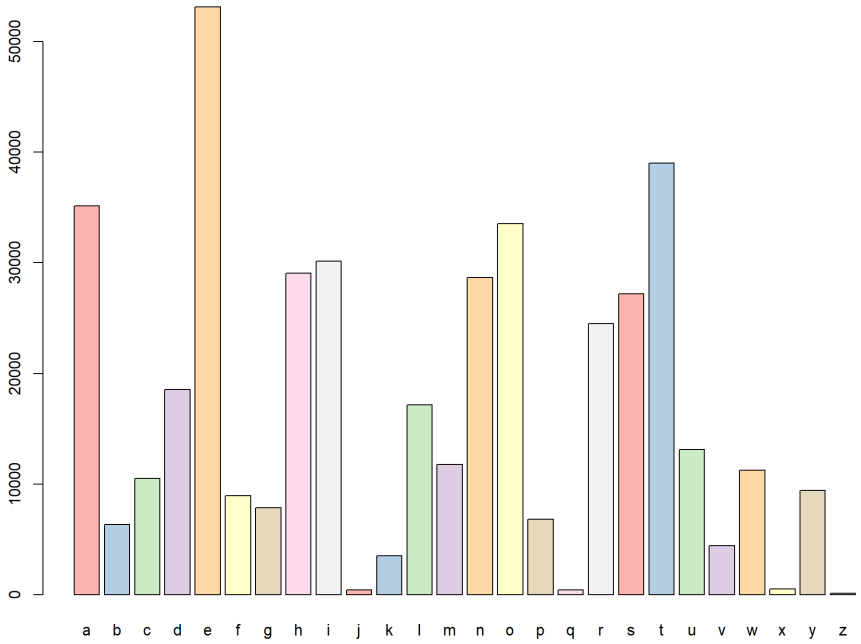
	gutenberg_id	letra
1	1661	t
2	1661	h
3	1661	e
4	1661	a
5	1661	d
6	1661	v
7	1661	e
8	1661	n
9	1661	t
10	1661	u

Cuadro 2: fragmento de *Data frame* que contiene la frecuencia de ocurrencia de cada letra en el texto

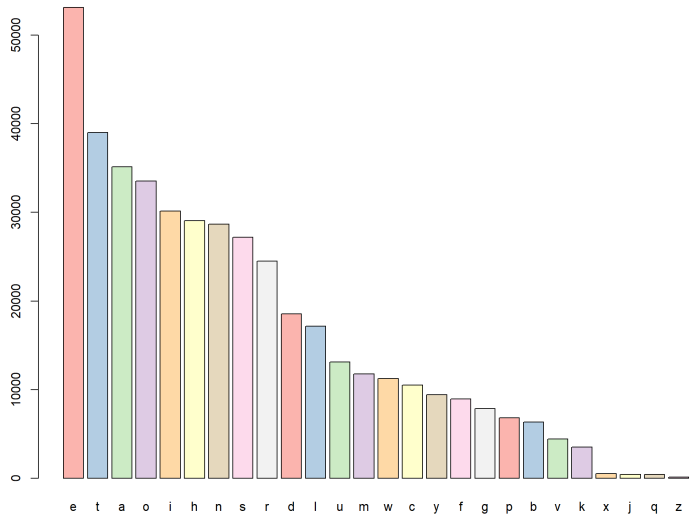
	Var1	Freq
11	a	35137
12	b	6362
13	c	10499
14	d	18563
15	e	53111
16	f	8975
17	g	7887
18	h	29047
19	i	30140
20	j	452
21	k	3543
22	l	17145
23	m	11787

```r
Conan = gutenberg_download(c(1661))

letras = Conan %>% unnest_tokens(letra, text, "characters")
write.csv(letras, "letras.csv")

palabras = Conan %>% unnest_tokens(palabra, text, "words")
write.csv(palabras,"palabras.csv")

oraciones = Conan %>% unnest_tokens(oraciones, text, "sentences")
write.csv(oraciones,"oraciones.csv")

ngrams= Conan %>% unnest_tokens(ngram, text,"ngrams", n = 2)



barplot(sort(table(palabras$palabra), decreasing=TRUE),log="y")

frcl = as.data.frame(table(letras$letra))
```

Tarea2.R



(a) Frecuencia de ocurrencia de las letras en el texto



(b) Frecuencia de ocurrencia de las letras en el texto ordenada de manera decreciente

Figura 1: Histogramas de frecuencia de ocurrencia de las letras en el texto

Cuadro 3: fragmento del *Data frame* que contiene todas las palabras que aparecen en el texto.

	Gutenberg_id	Palabra
1	1661	the
2	1661	adventures
3	1661	of
4	1661	sherlock
5	1661	holmes
6	1661	by
7	1661	sir
8	1661	arthur
9	1661	conan
10	1661	doyle

Cuadro 4: fragmento de *Data frame* que contiene la frecuencia de ocurrencia de cada palabra en el texto

	Var1	Freq
79	8s	1
80	9	1
81	90	1
82	9th	2
83	a	2641
84	abandoned	3
85	abandons	1
86	abbots	1
87	aberdeen	2

3.2. Palabras

Para el trabajo con las palabras, se extraen todas las presentes en el texto, teniendo el mismo 432064 palabras que fueron almacenadas en un *Data frame*, como se muestra en el cuadro 3 de la página 4. A partir de los datos obtenidos se realiza un análisis de frecuencia de ocurrencia de las palabras en el texto, como se muestra en el cuadro 4 de la página 4. Para una mejor comprensión de los datos obtenidos se realiza una representación gráfica de los mismos mediante un histograma, como se puede observar en la figura 2 de la página 5.

Como se puede observar en la figura 2, hay palabras que aparecen en pocas ocasiones en el texto y otras que se repiten gran cantidad de veces, estos extremos nos impiden hacer un análisis mas a fondo del contenido del texto, por lo que se realiza un filtrado de las palabras por frecuencia de ocurrencia tomando solamente aquellas que aparecen más de 200 y menos de 450 veces, el resultado de este filtro se muestra en el cuadro 5 de la página 5 y gráficamente en la figura 3 de la página 6 y en la nube de palabras que se muestra en la figura 4 de la página 7.

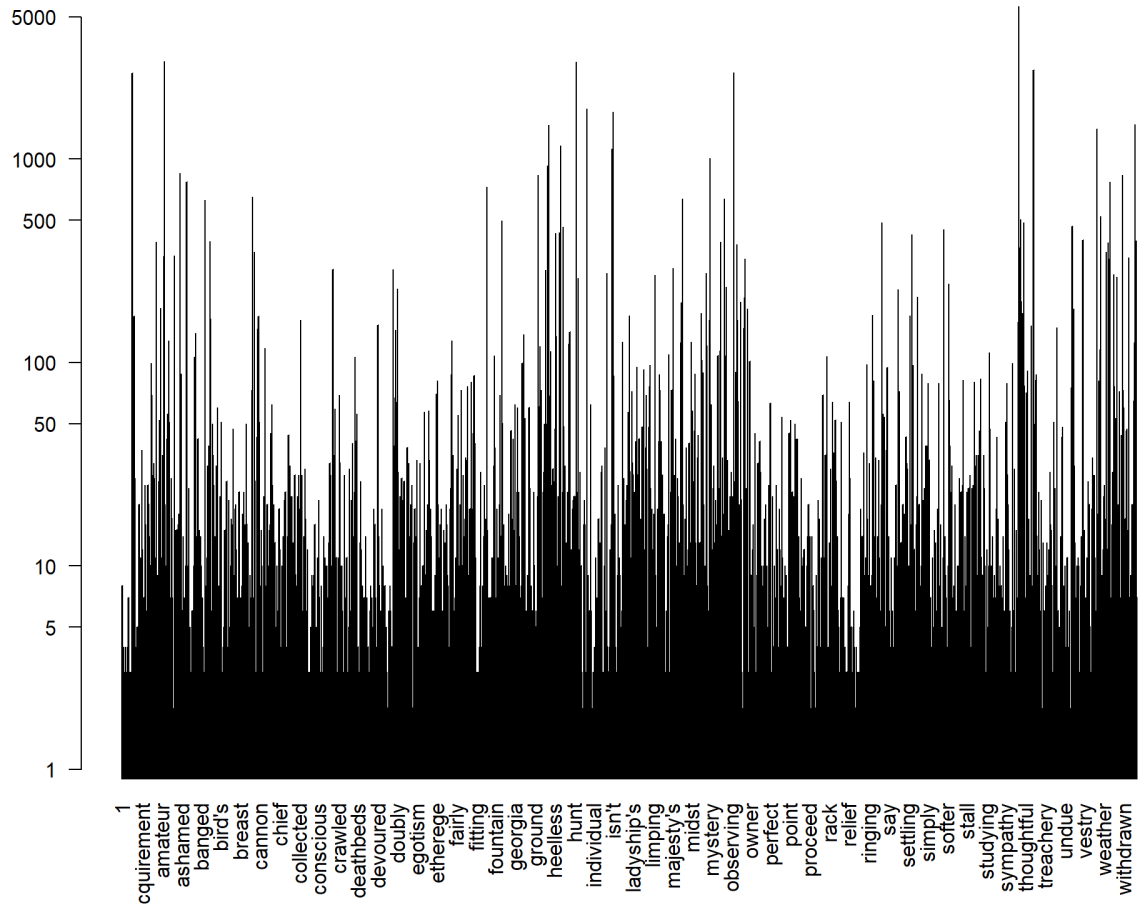
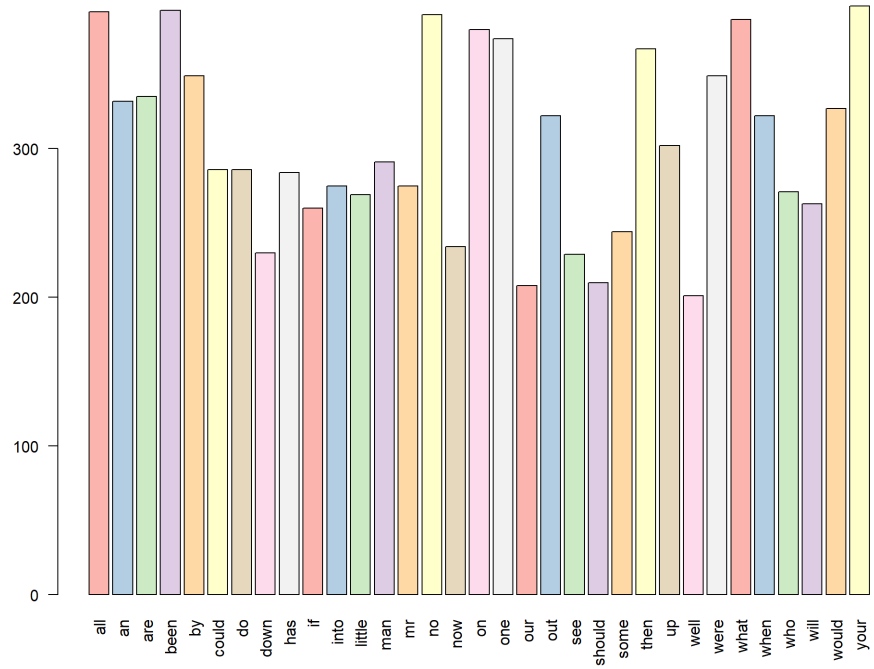


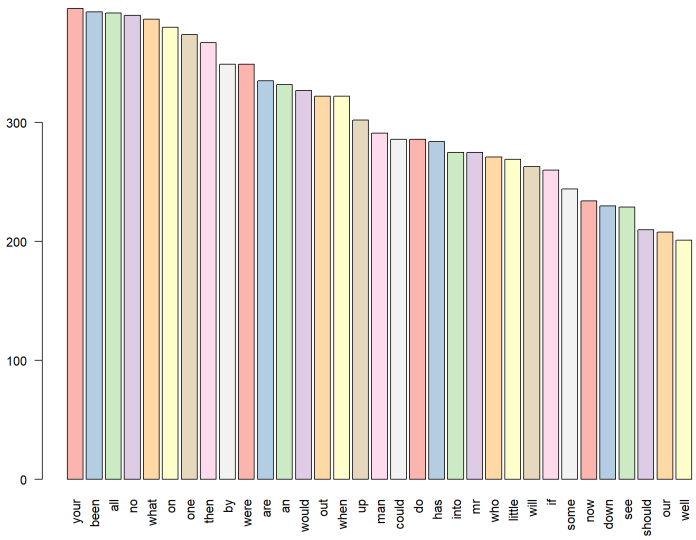
Figura 2: Histogramas de frecuencia de ocurrencia de las palabras en el texto en el texto, usando la escala logarítmica para mejor comprensión del gráfico

Cuadro 5: fragmento de *Data frame* frecuencia de ocurrencia de las palabras en el texto en un rango de 200 – 450 veces

	Var1	**Freq**
274	all	392
333	an	332
417	are	335
697	been	393
1046	by	349
1667	could	286
2143	do	286
2177	down	230
3350	has	284



(a) Frecuencia de ocurrencia de las palabras en el texto en un rango de 200–450 veces



(b) Frecuencia de ocurrencia de las palabras en el texto en un rango 200 – 450 veces, ordenada de manera decreciente

Figura 3: Frecuencia de ocurrencia de las palabras en el texto en un rango 200 – 450 veces

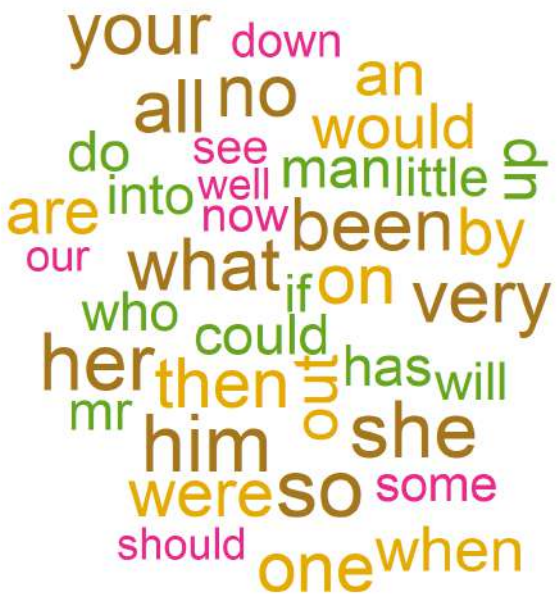


Figura 4: Nube de palabras utilizadas en el texto en un rango 200 – 450

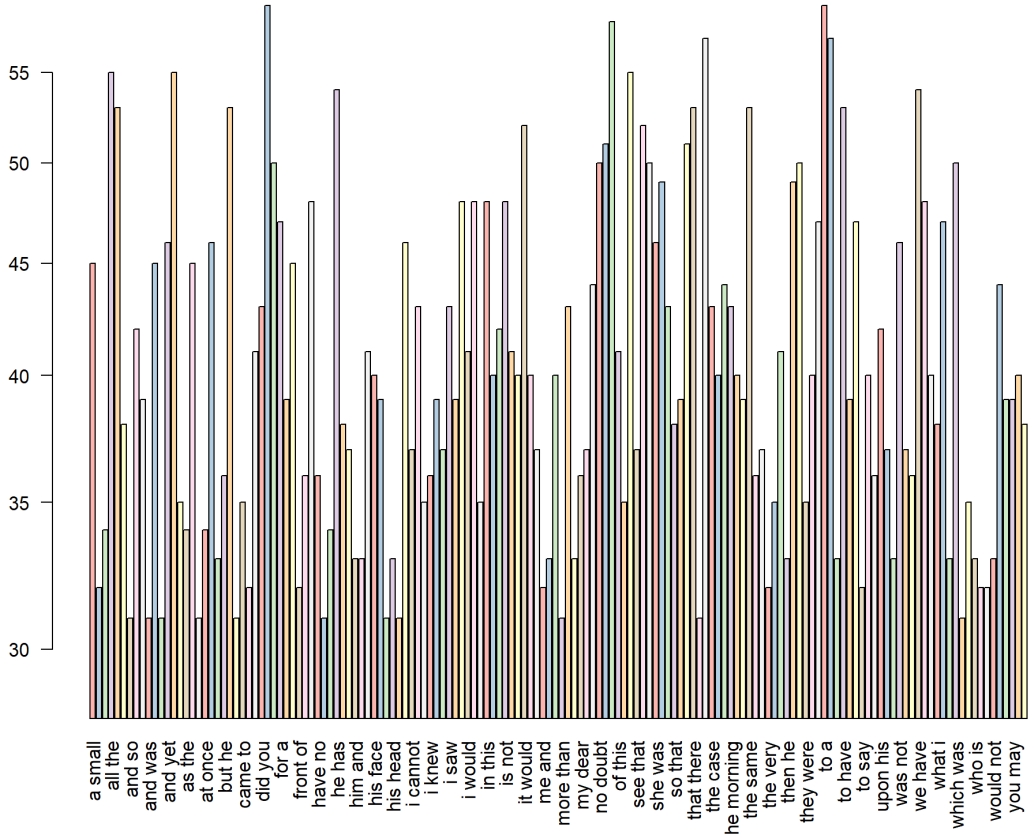


Figura 5: histograma de pares de palabras usadas en el texto en un rango de 30 – 60 veces

Además se analizan los pares de palabras que aparecen juntas en el texto, como se muestra en la figura 3 de la página 6

El código general se encuentra disponible en el repositorio:

https://github.com/Albertomnoa/Tareas_MPA/Tarea2

Referencias

[1] R Core Team. R: R: Un lenguaje y un entorno para la informática estadística, 2020.

[2] RStudio Team. Rstudio: Entorno de desarrollo integrado para r, 2020.

Tarea 3 de Modelos Probabilistas Aplicados

Distribuciones de Probabilidad

5271

22 de septiembre de 2020

1. Origen de los datos

En este trabajo se utiliza como fuente de los datos el libro "The Adventures of Sherlock Holmes"[1], del escritor y médico británico Arthur Conan Doyle. Este libro se encuentra disponible en la biblioteca virtual gratuita Project Gutenberg, con el siguiente enlace: https://www.gutenberg.org.

2. Sobre el libro

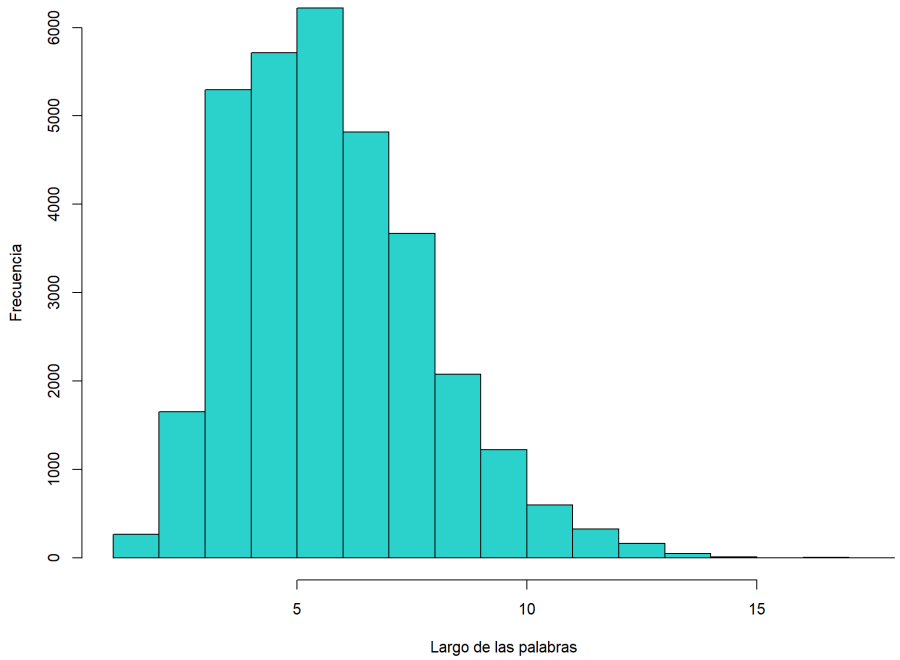
Este libro es una colección de doce cuentos de Arthur Conan Doyle, los cuales fueron publicados por primera vez el 14 de octubre de 1892. El mismo agrupa los primeros cuentos con el detective consultor Sherlock Holmes, que se habían publicado en doce números mensuales de The Strand Magazine de Julio de 1891 a junio de 1892.

3. Análisis y tratamiento de los datos

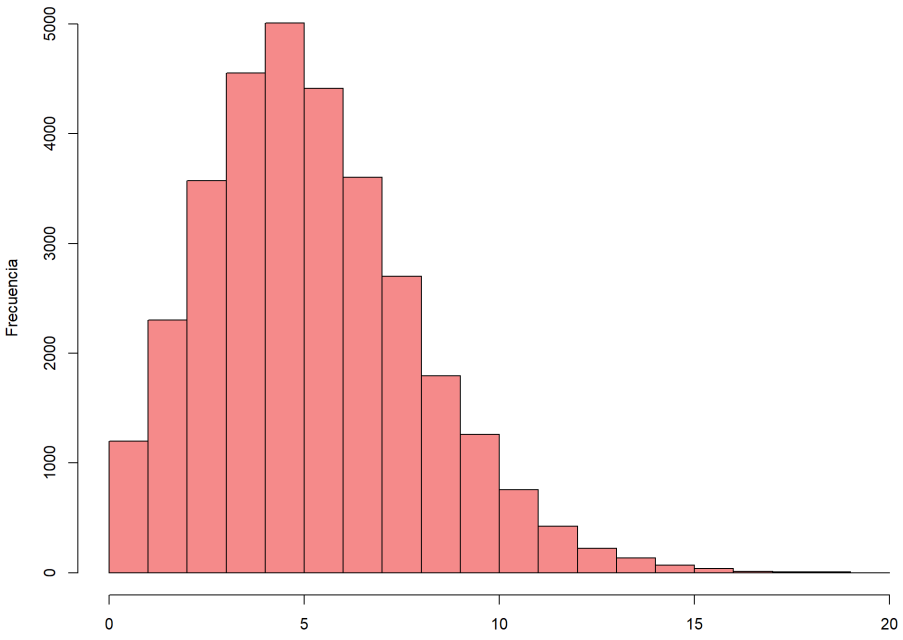
Al libro seleccionado se le realiza un estudio de frecuencia de ocurrencia de la cantidad de letras por palabras, de palabras por oraciones y cantidad de oraciones por párrafo. El análisis será realizado en el programa R versión 4.0.2 [2] en el entorno de desarrollo Rstudio [3].

3.1. Letras por oraciones

De le libro en cuestión se extraen todas las palabras que componen el texto, de las mismas se eliminan la llamadas palabras "vacías", es decir las que no aportan contenido al libro, el resultado del filtro se muestra en el cuadro 1 de la página 3. Una vez realizado este filtrado se procede a contar la cantidad de letras que tiene cada palabra, como se observa en la figura 1(a) de la página 5. En la figura 1(b) de la página 5, se muestra la distribución binomial negativa con parámetros $n =$ cantidad de palabras después del filtro, $k = 18$ y $p = 0{,}75$, a la que se asemeja la distribución de las cantidad de letras por palabras en el texto analizado.



(a) Cantidad de letras por palabras



(b) distribución binomial negativa creada con la función *rnbinom* de R

Figura 1: Histogramas de distribución de cuantiad de letras por palabras en el texto

Cuadro 1: fragmento de *Data frame* resultante del filtrado

	Gutenberg_id	Palabras
2	1661	sherlock
3	1661	holmes
19	1661	mystery
20	1661	orange
167	1661	crime
168	1661	occupied
169	1661	immense
170	1661	faculties
171	1661	extraordinary
172	1661	powers

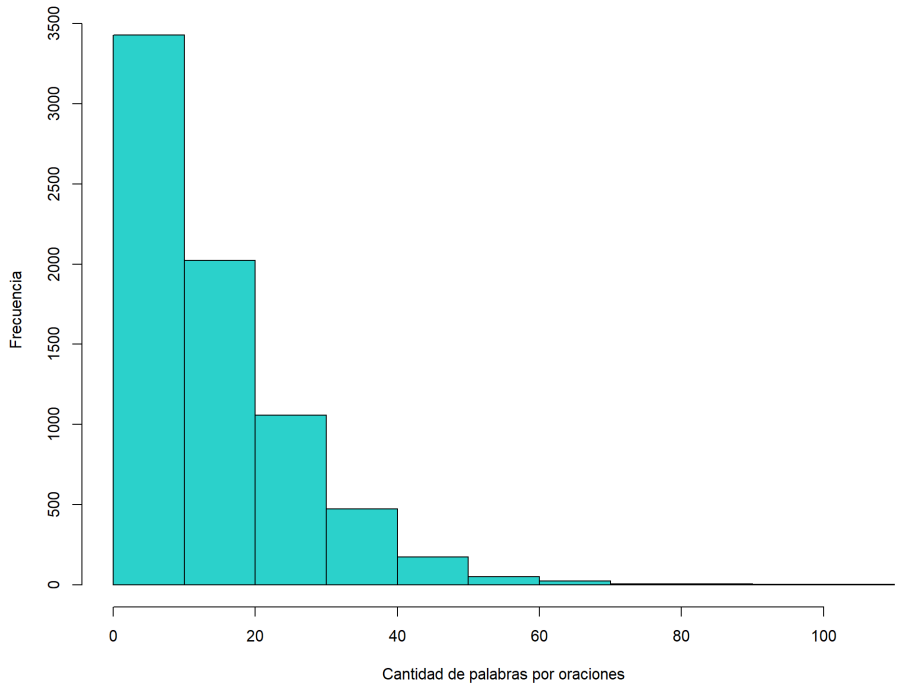
3.2. Palabras por oraciones

Para el trabajo con las oraciones, se extraen todas las presentes en el texto y se cuenta la cantidad de palabras que conforman cada una de estas oraciones. Lo anterior se puede observar en la en la figura 2(a) de la página 4. En la figura 2(b) de la página 4 se muestra una distribución geométrica con parámetros n = cantidad de oraciones del texto y $p = 0{,}055$. Dicha distribución se asemeja a la distribución de la cantidad de palabras en las oraciones del texto.

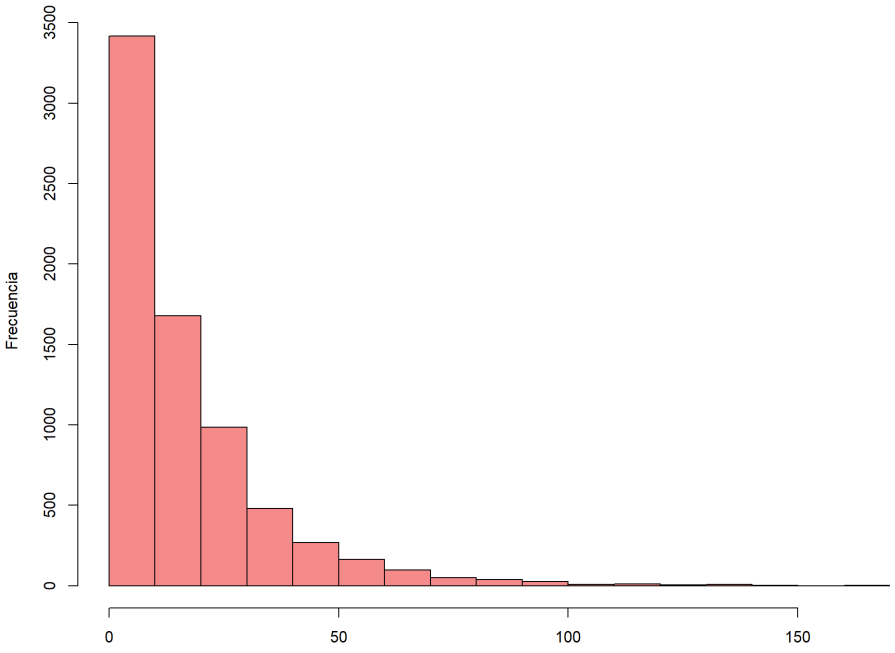
3.3. Oraciones por párrafos

Análogamente a lo realizado en la subsecciones anteriores se procede al análisis de la cantidad de oraciones por párrafo, como se puede observar en la figura 3(a) de la página 5. En la figura 3(b) de la página 5 se muestra una distribución geométrica con parámetros n = cantidad de párrafos del texto y $p = 0{,}29$. La distribución mostrada en figura 3(b) es muy similar a la distribución de la cantidad de oraciones por párrafos en el libro.

El código general se encuentra disponible en el repositorio https://github.com/Albertomnoa/Tareas

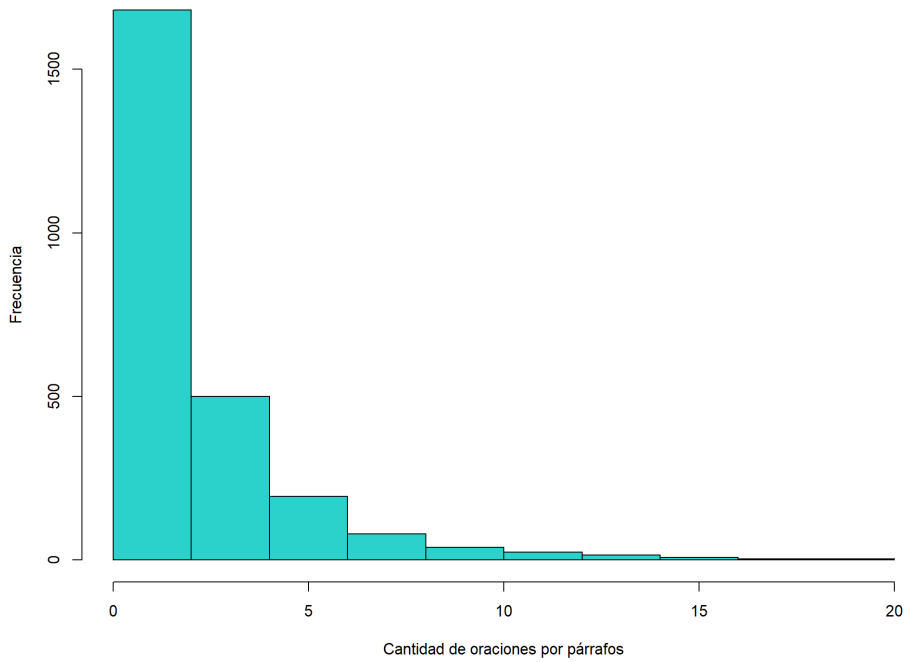


(a) Cantidad de palabras por oraciones

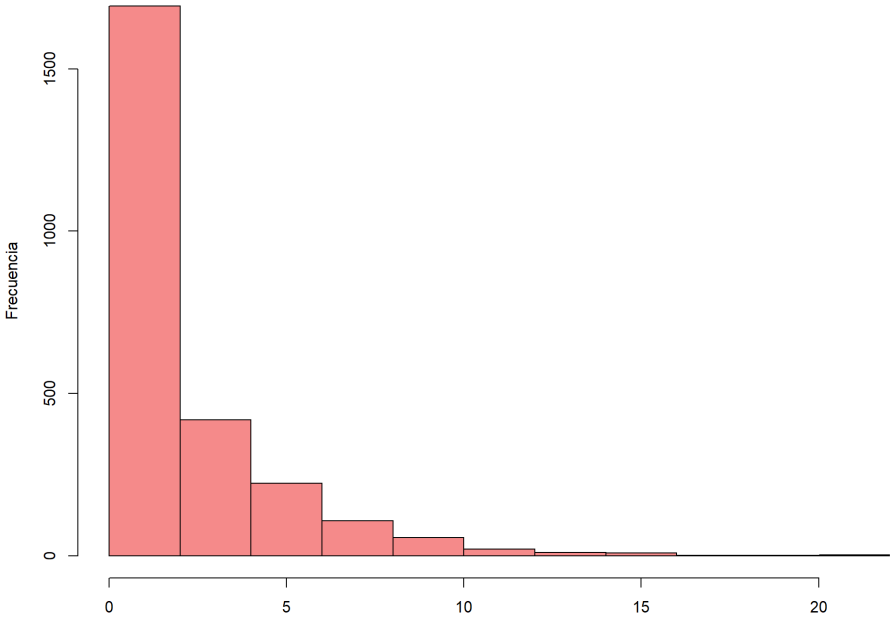


(b) distribución geométrica creada con la función *rgeom* de R

Figura 2: Histogramas de distribución de cuantiad de palabras por oraciones en el texto



(a) Cantidad de oraciones por párrafos



(b) distribución geométrica creada con la función *rgeom* de R

Figura 3: Histogramas de distribución de cuantiad de oraciones por párrafos en el libro

Referencias

[1] Arthur Conan Doyle. *The Adventures of Sherlock Holmes*. George Newnes, United Kingdom, 1892.

[2] R Core Team. R: Un lenguaje y un entorno para la informática estadística, 2020.

[3] RStudio Team. Rstudio: Entorno de desarrollo integrado para R, 2020.

Tarea 4 de Modelos Probabilistas Aplicados

Distribución de Poisson

5271

29 de septiembre de 2020

1. Introducción

En este trabajo se presenta un acercamiento a la simulación de variables aleatorias de Poisson a partir de variables aleatorias de distribución Uniforme, Exponencial y Normal.

2. Proceso de Poisson

La distribución de Poisson es una distribución de probabilidad discreta que expresa, a partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un determinado número de eventos durante cierto período de tiempo o espacio. Se centra en la probabilidad de ocurrencia de eventos con probabilidades muy pequeñas. Se especifica por un parámetro lambda (λ). Este parámetro es igual a la media y la varianza de la distribución.

3. Aproximación mediante el uso de variables aleatorias Exponenciales

Los momentos en que se incrementa el proceso de Poisson se denominan tiempos de llegada o tiempos de ocurrencia , ya que en los modelos estocásticos clásicos representan las llegadas o ocurrencias de algo, como llegadas de clientes a la fila de un banco. Las diferencias entre tiempos consecutivos se denominan tiempos entre llegadas. Los tiempos entre llegadas de un proceso de Poisson homogéneo lo forman variables aleatorias exponenciales independientes , un resultado conocido como el Teorema del intervalo . A partir de esta relación se pueden generar variables aleatorias exponenciales $E_1, E_2, ...N_n$ y N es el número entero más pequeño tal que:

$$\sum_{k=1}^{n} E_k > 1 \tag{1}$$

Entonces N es Poisson (λ), Lema 3.2 Capítulo 10 [1]

A continuación realizaremos una comparación entre las variables aleatorias de Poisson creadas a partir de la suma de las variables aleatorias exponenciales y las creadas a partir de la biblioteca *rpois*.

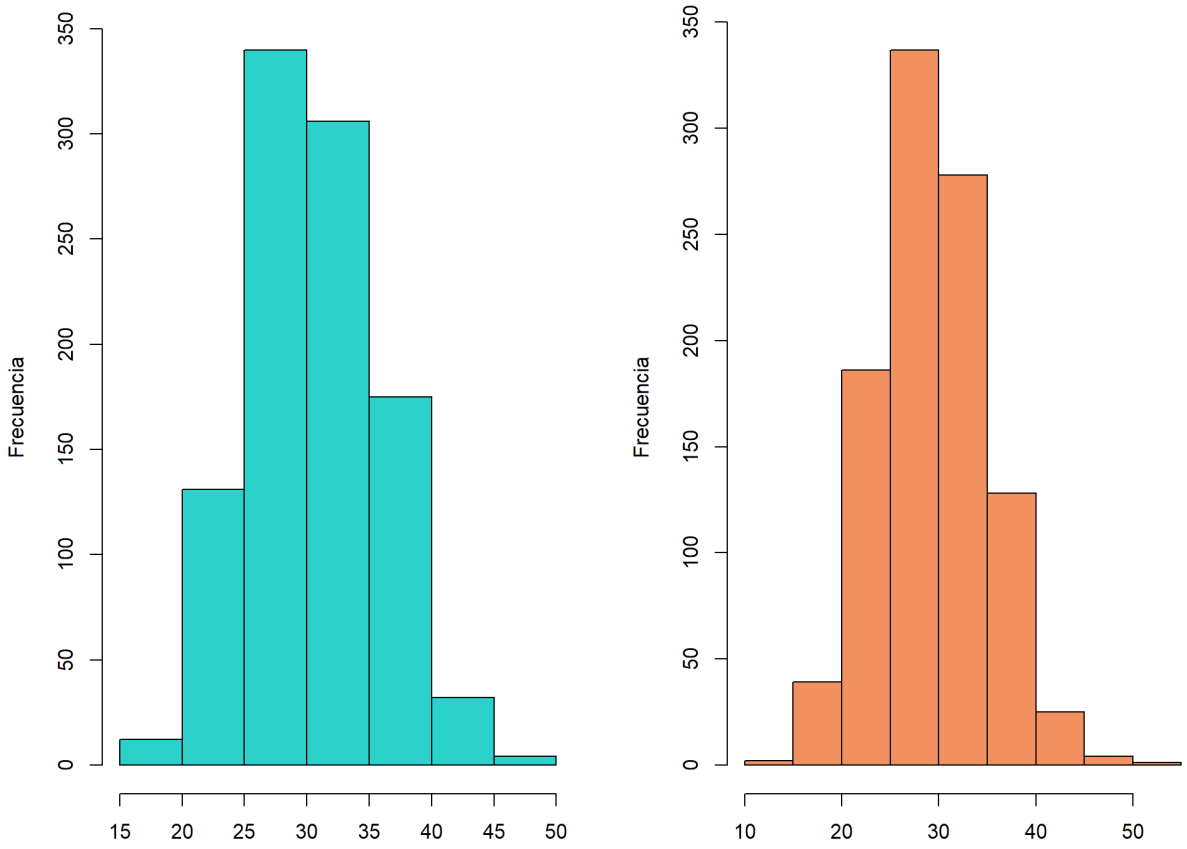


Figura 1: Histogramas de las poblaciones simuladas, a la izquierda la aproximación a partir de la exponencial y a la derecha la creada a partir de la biblioteca *rpois*

En la figura 1 de la página 2 se muestra los histogramas de las dos poblaciones simuladas. En la figura 2 de la página 3, se muestran superpuestos los diagramas de densidad de ambas poblaciones.

Para calcular la diferencia entre las distribuciones se realiza el cálculo de la distancia Kolmogorov—Smirnov, que se define como la distancia vertical máxima entre las funciones de distribución acumulada empíricas de dos muestras, donde el valor de la distancia es 0,11. Esto se puede observar en la figura 3 de la página 4.

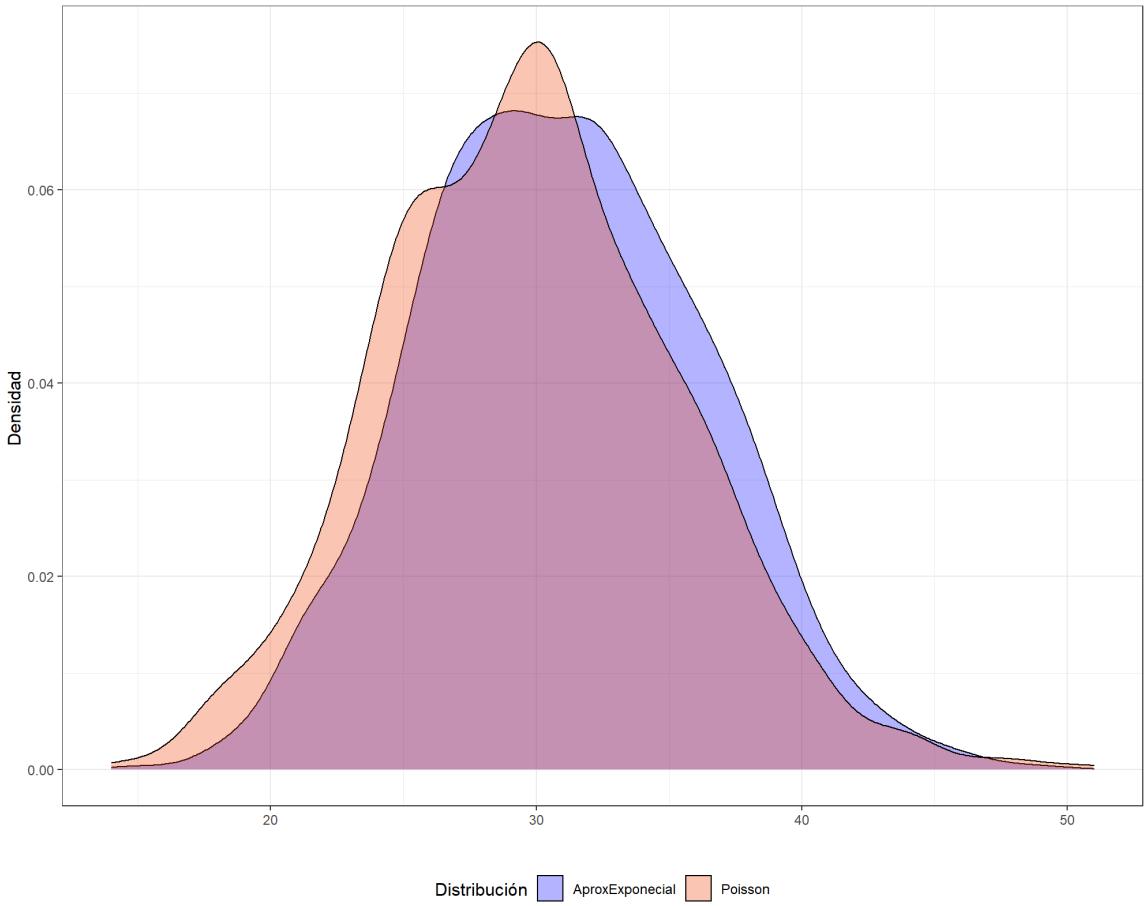


Figura 2: Diagramas de densidad de ambas poblaciones

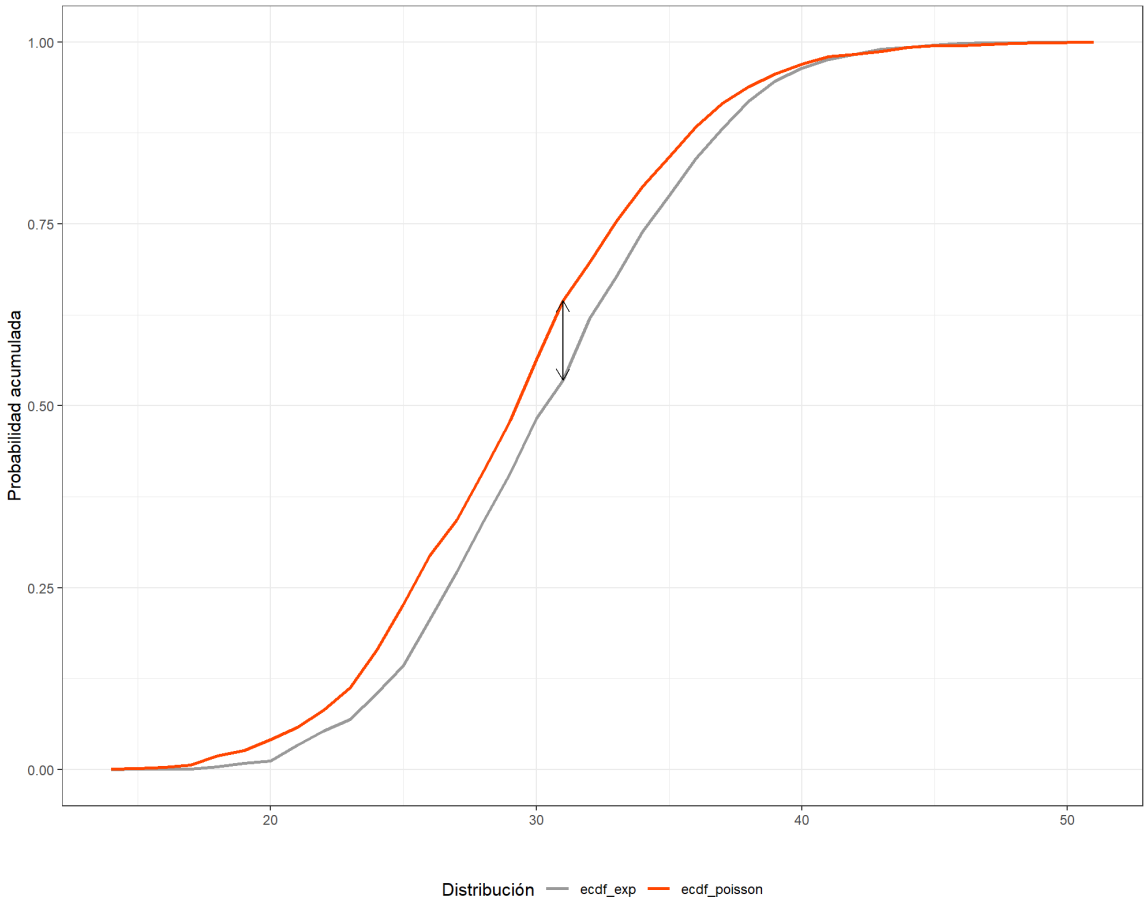


Figura 3: Distancia de Kolmogorov—Smirnov entre ambas poblaciones

Para concluir si ambas poblaciones viene de una misma distribución se procede a aplicar la prueba de Cucconi, que es una prueba no paramétrica para comparar conjuntamente la tendencia central y la variabilidad (detectando cambios de ubicación y escala) en dos muestras. Lo s resultados son para el estadístico $C = 12{,}287$ y $p - valor = 0$, se rechaza la hipótesis nula que ambas poblaciones viene de una misma distribución.

3.1. Aproximación mediante el uso de variables aleatorias Exponenciales

Para reducir los cálculos, se reformula el método que utiliza variables aleatorias exponenciales, por el productos de variables aleatorias uniformes, debido a identidades logarítmicas. Se usan variables aleatorias uniformes estándar $U_1, U_2, ...U_n$ y N es el número entero más pequeño tal que:

$$\prod_{k=1}^{n} U_k > e^{-\lambda} \tag{2}$$

Entonces N es Poisson (λ), Lema 3.3 Capítulo 10 [1]

A continuación realizaremos una comparación entre las variables aleatorias de Poisson creadas a partir de la suma del producto de variables aleatorias uniformes y las creadas a partir de la biblioteca *rpois*.

En la figura 4 de la página 6 se muestra los histogramas de las dos poblaciones simuladas. En la figura 5 de la página 7, se muestran superpuestos los diagramas de densidad de ambas poblaciones.

Para calcular la diferencia entre las distribuciones se realiza el cálculo de la distancia Kolmogorov—Smirnov, que se define como la distancia vertical máxima entre las funciones de distribución acumulada empíricas de dos muestras, donde el valor de la distancia es 0,154. Esto se puede observar en la figura 6 de la página 8.

Para concluir si ambas poblaciones viene de una misma distribución se procede a aplicar la prueba de Cucconi, que es una prueba no paramétrica para comparar conjuntamente la tendencia central y la variabilidad (detectando cambios de ubicación y escala) en dos muestras. Lo s resultados son para el estadístico $C = 12{,}287$ y $p - valor = 0$, se rechaza la hipótesis nula que ambas poblaciones viene de una misma distribución.

3.2. Aproximación mediante distribución normal

A continuación realizaremos una comparación entre las variables aleatorias de Poisson creadas a partir de la biblioteca *rnorm* y las creadas a partir de la biblioteca *rpois*.

En la figura 7 de la página 9 se muestra los histogramas de las dos poblaciones simuladas. En la figura 8 de la página 10, se muestran superpuestos los diagramas de densidad de ambas poblaciones.

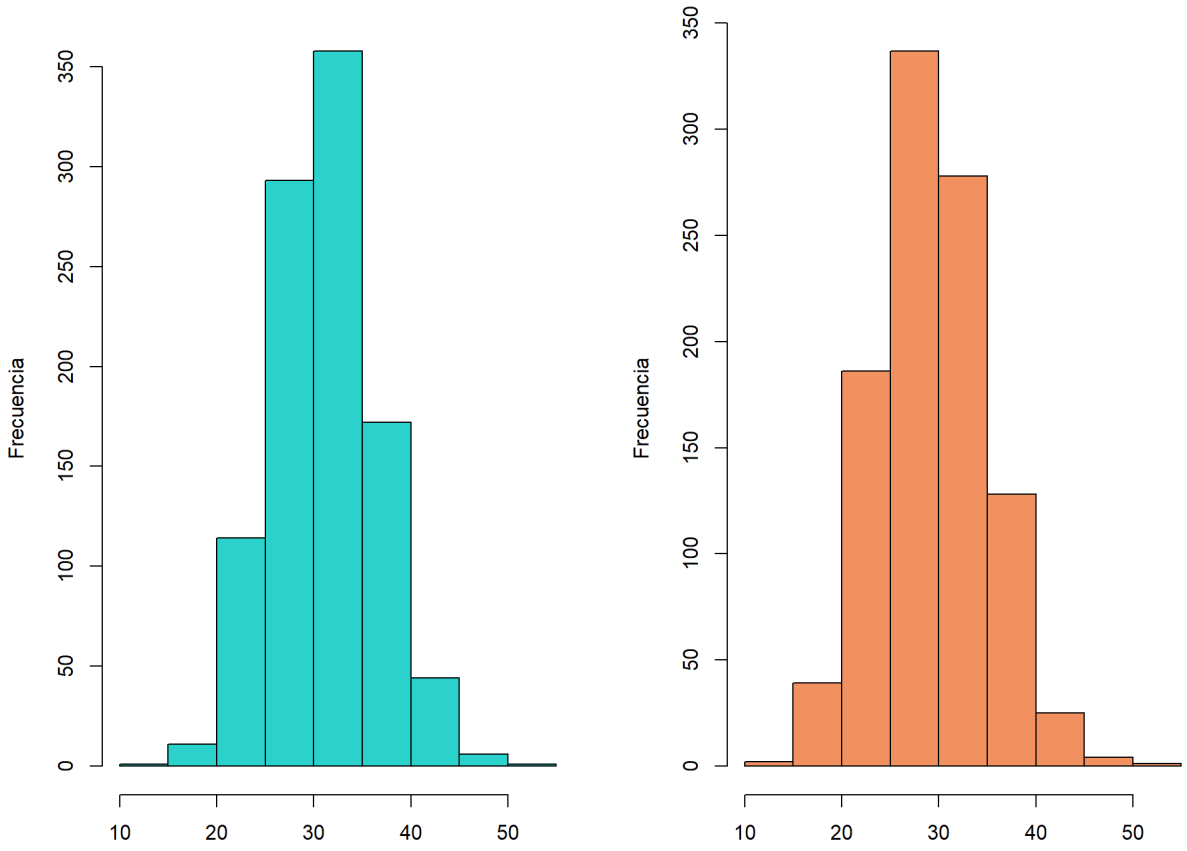


Figura 4: Histogramas de las poblaciones simuladas, a la izquierda la aproximación a partir de la Uniforme y a la derecha la creada a partir de la biblioteca *rpois*

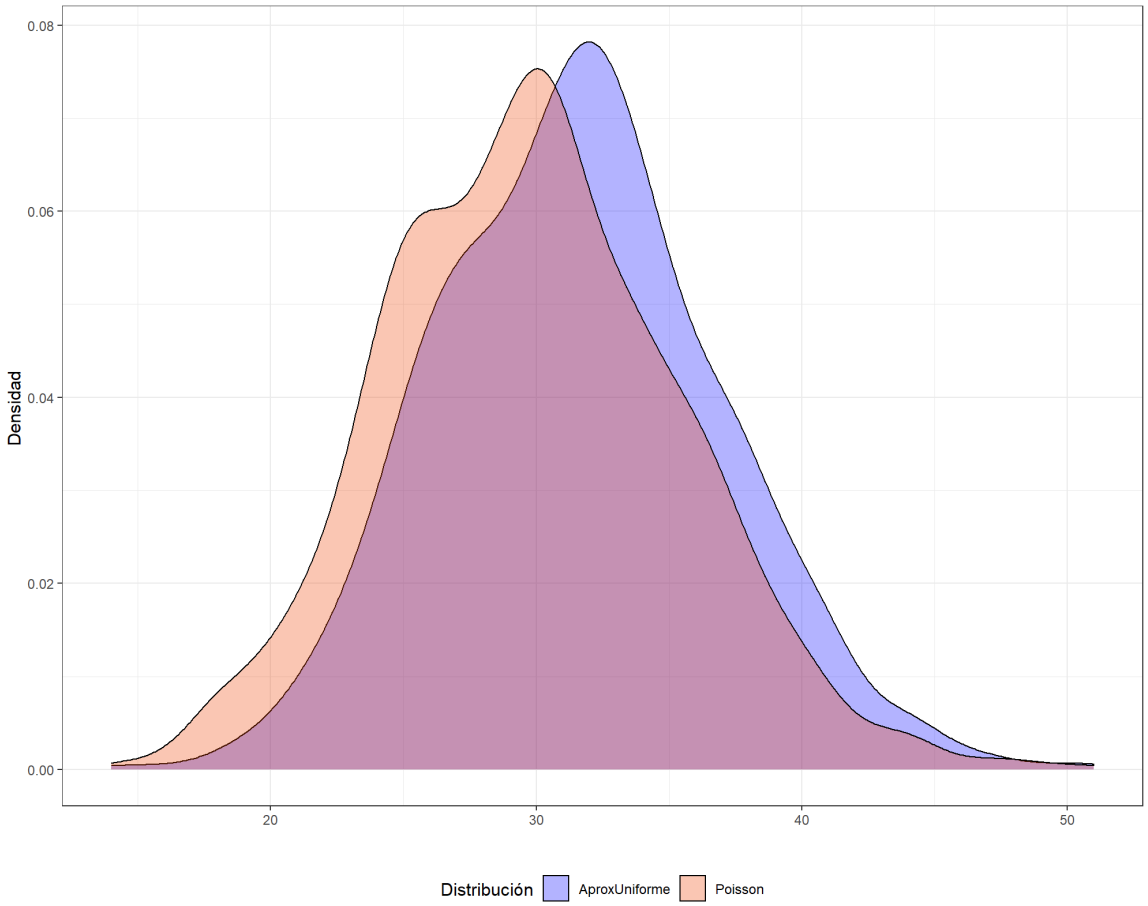


Figura 5: Diagramas de densidad de ambas poblaciones

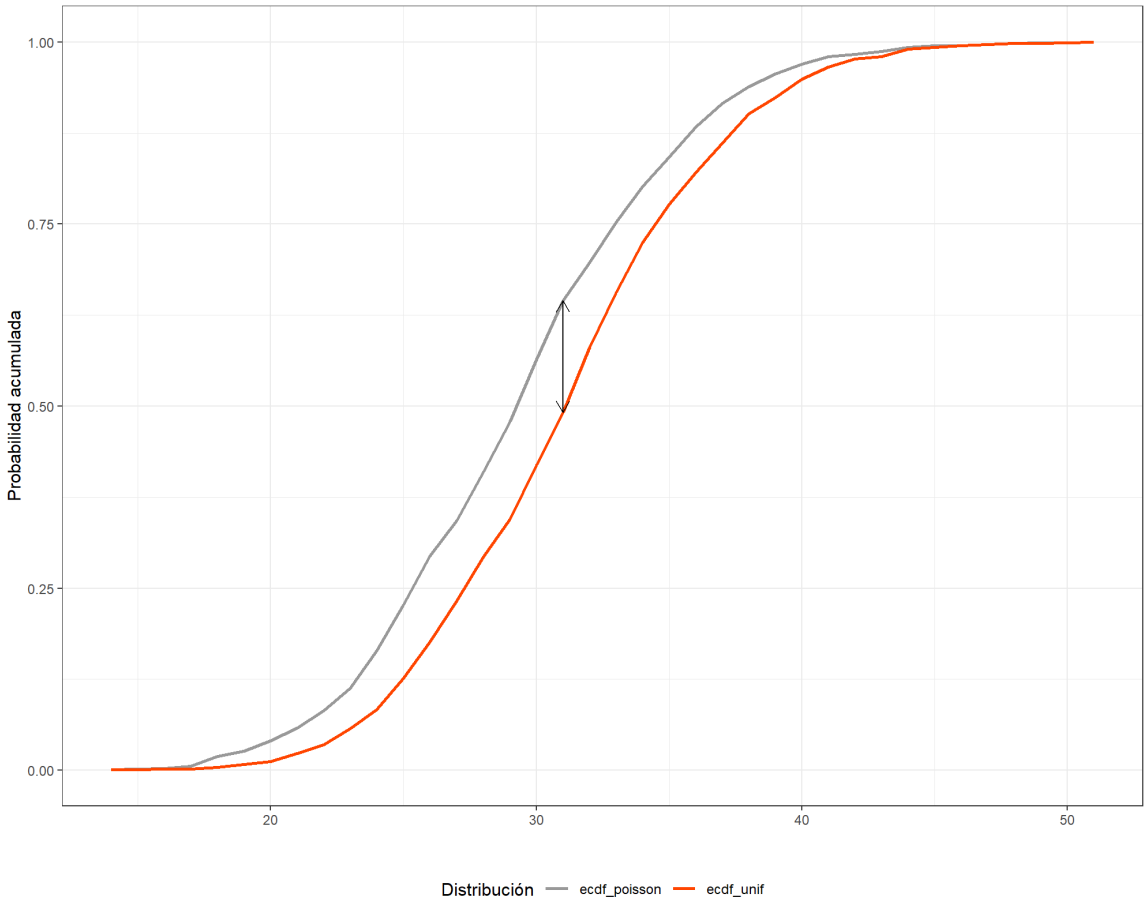


Figura 6: Distancia de Kolmogorov—Smirnov entre ambas poblaciones

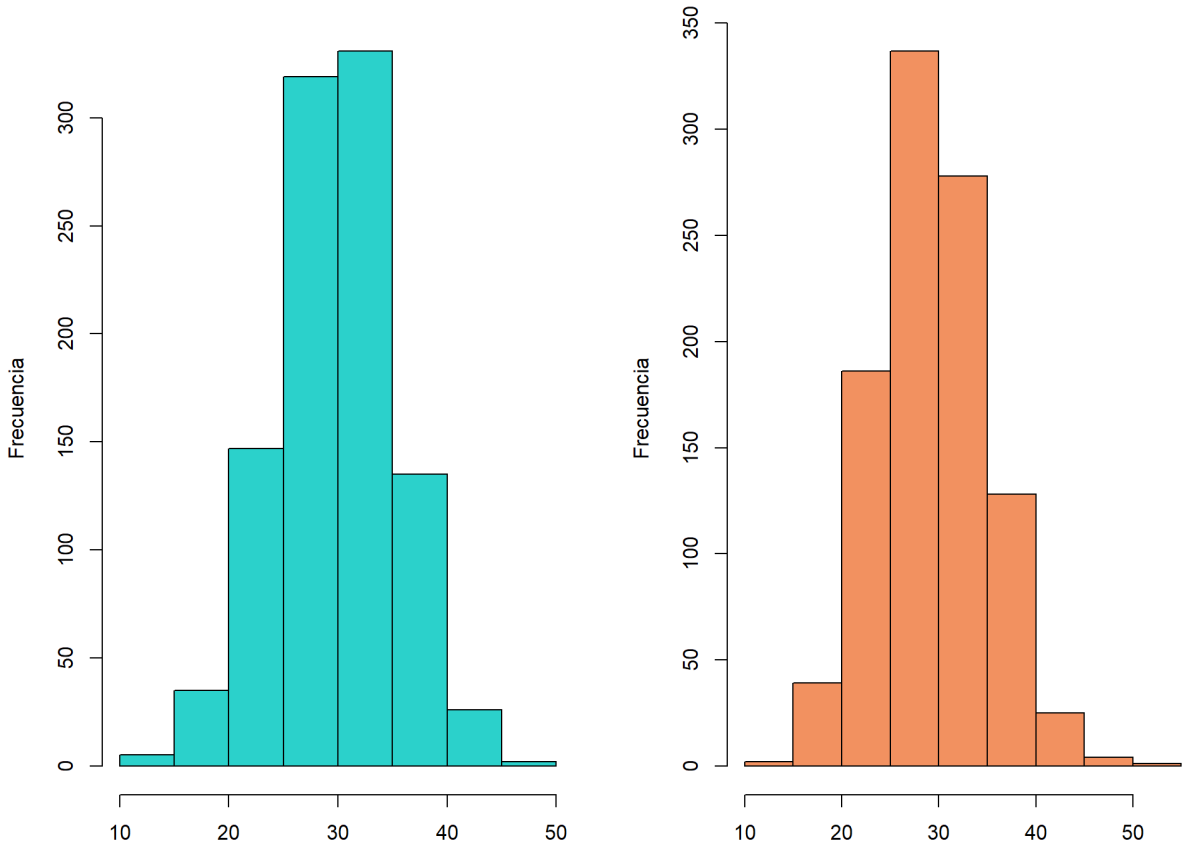


Figura 7: Histogramas de las poblaciones simuladas, a la izquierda la aproximación a partir de la Normal y a la derecha la creada a partir de la biblioteca *rpois*

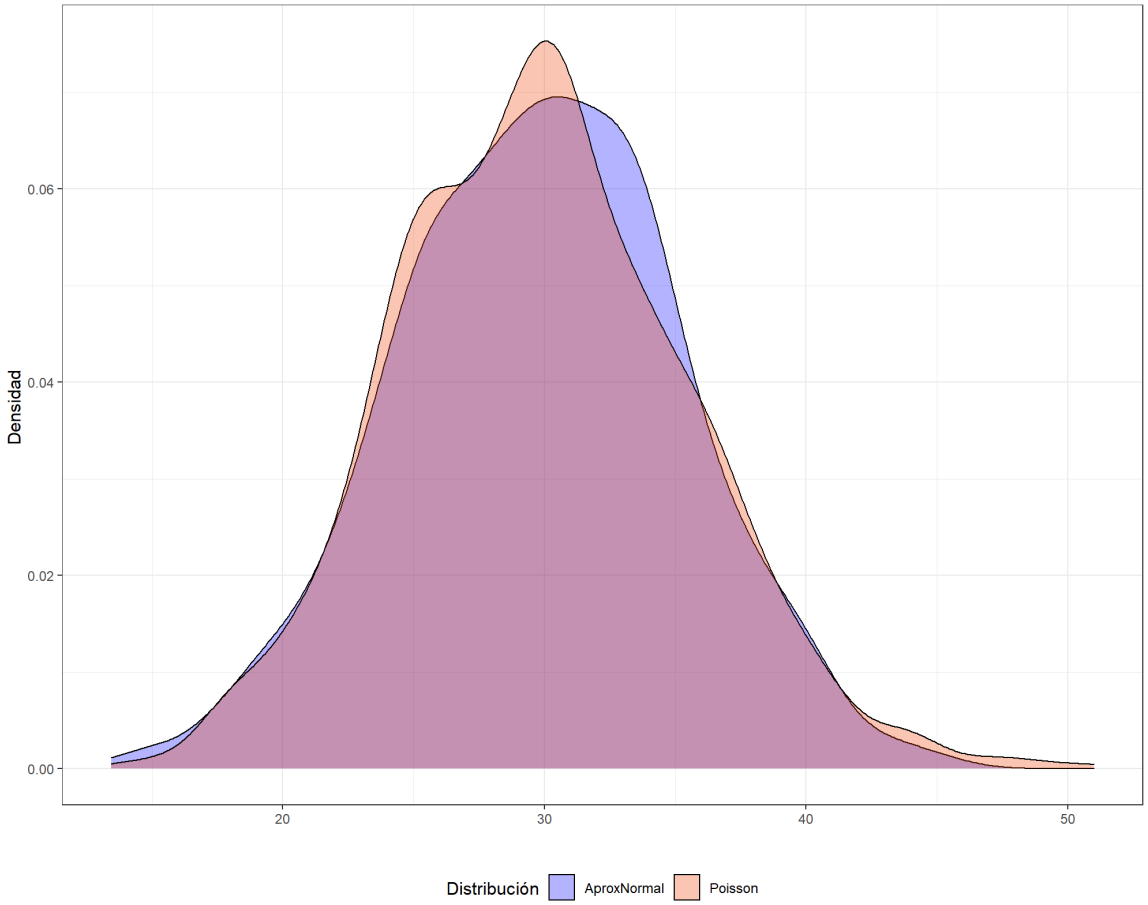


Figura 8: Diagramas de densidad de ambas poblaciones

Para calcular la diferencia entre las distribuciones se realiza el cálculo de la distancia Kolmogorov—Smirnov, que se define como la distancia vertical máxima entre las funciones de distribución acumulada empíricas de dos muestras, donde el valor de la distancia es 0,069. Esto se puede observar en la figura 9 de la página 11.

Para concluir si ambas poblaciones viene de una misma distribución se procede a aplicar la prueba de Cucconi, que es una prueba no paramétrica para comparar conjuntamente la tendencia central y la variabilidad (detectando cambios de ubicación y escala) en dos muestras. Lo s resultados son para el estadístico $C = 0{,}106$ y $p - valor = 0{,}891$, se acepta la hipótesis nula que ambas poblaciones viene de una misma distribución.

El código general se encuentra disponible en el repositorio. https://github.com/Albertomnoa/Tareas

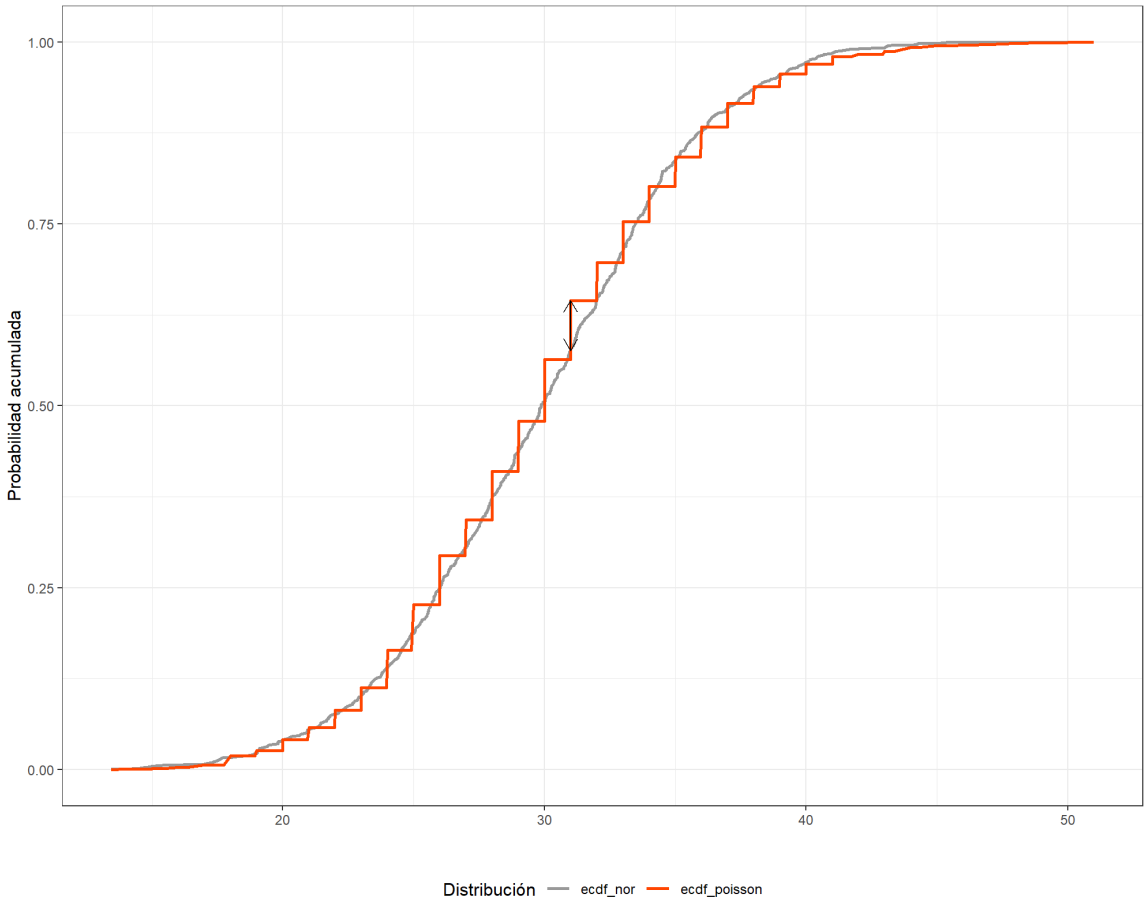


Figura 9: Distancia de Kolmogorov—Smirnov entre ambas poblaciones

Referencias

[1] Luc Devroye. *Non-Uniform Random Variate Generation.* Springer-Verlag, New York, 1986.

Tarea 5 de Modelos Probabilistas Aplicados

Algoritmos generadores de números pseudo-aleatorios con distribución Uniforme y distribución Normal

5271

6 de octubre de 2020

1. Introducción

En este trabajo se presenta el análisis a varios algoritmos de generación de números pseudo-aleatorios con distribución Uniforme y distribución Normal. Así como el impacto de los parámetros de dichos algoritmos en la calidad de los números pseudo-aleatorios. El análisis será realizado en el programa R versión 4.0.2 [2] en el entorno de desarrollo Rstudio [3]

2. Generador congruencial lineal(Mixto)

Entre los principales generadores de números pseudo-aleatorios que se emplean en la actualidad están los llamados generadores congruenciales lineales, introducidos por Lehmer en 1951. Estos métodos comienza con un valor inicial x_0 (semilla), y los sucesivos valores $x_n, n \geq 0$ se obtienen recursivamente con la ecuación 1:

$$X_{n+1} = (aX_n + c) \, mod \, m, \quad n \geq 0. \tag{1}$$

Con parámetros:

$$\begin{array}{lll} m, & \textit{el módulo}; & 0 < m. \\ a, & \textit{el multiplicador}; & 0 \leq a < m. \\ c, & \textit{el incremento}; & 0 \leq c < m. \\ X_0, & \textit{la semilla}; & 0 \leq X_0 < m. \end{array} \tag{2}$$

Se crea una función con la ecuación 1 en R, como muestra el código 1.

```r
dist_uniforme = function(n, semilla) {
    a = 7
    cc = 7
    m = 10
    datos = numeric()
    x = semilla
    while (length(datos) < n) {
        x = (a * x + cc) %%m
        datos = c(datos, x)
    }
return(datos / (m - 1))}
```

Tarea5.R

Pero con esto no es suficiente para garantizar la calidad de los números pseudos-aleatorios, una de estas medidas de calidad es el período de los números generados. El período no es más que cada cuentos números generados se vuelve a repetir la secuencia y se representa $\lambda^*(m)$ es decir que si $\lambda^* < m$ el generador no es de buena calidad. Los generadores de buena calidad deben tener un período completo, es decir $\lambda^*(m) = m$. Para garantizar el período completo se tiene de [1]:

Teorema 1 *La secuencia lineal congruencial definida por* m, a, c, X_0 *tiene período completo sí y solo sí*

I. c *es primo relativo de* m.

II. $b = a - 1$ *es múltiplo de* $p, \forall\, p$ *primo dividiendo* m.

III. b *es múltiplo de cuatro, sí* m *es un múltiplo de cuatro.*

2.1. Selección del módulo

Para la selección del modulo m la siguiente expresión:

$$m = P^e \begin{cases} \text{P } \textit{es la base que utiliza} \\ \\ \text{e } \textit{es el número de bit} \end{cases} \tag{3}$$

La base mayormente usada es dos y el número de bit es 32.Para la selección del valor para el módulo se creo la función 2.1 en R.

```
modu = function (p,e){
 m = p^e
return(m)}
```

Tarea5.R

2.2. Selección del incremento

El incremento o constante aditiva c es un número en el intervalo $0 \leq c < m$ y primo relativo con m, es decir que el Mínimo Común Divisor de $(c, m) = 1$, esto genera una cierta cantidad de posibles valores de c. Para garantizar la elección de un valor de c adecuado se realizo una la función 2.2 en R.

```
aditiva= function(m,po){
lisc=numeric()
for (i in c(1:m)) {
  if(GCD(m,i)==1){
      lisc=c(lisc,i)
  }
}
c = lisc[po]
return(c)
}
```

Tarea5.R

2.3. Selección del multiplicador

Para la elección acertada del multiplicador a se tiene las siguientes expresiones:

$a = 1 + MCM(P_1, P_2, P_3, ..., P_{k-1,P_k,4}) * t, \quad t \in (Z^+ \cup \{0\})$ si cuatro divide a m.

$a = 1 + MCM(P_1, P_2, P_3, ..., P_{k-1,P_k}) * t, \quad t \in (Z^+ \cup \{0\})$ si cuatro no divide a m.

Teniendo en cuenta estas expresiones y con el apoyo en la función LCM de R que calcula el Mínimo Común Múltiplo (MCM), se crea la función 2.3 en R para la correcta selección del parámetro a.

```
multiplicador = function(m,t){
    a = 0
    s = numeric()
    d = numeric()
if ((m %%4 )!=0){
s = primeFactors(m)
d = s[!duplicated(s)]
  if(length(d == 1)){
  a = 1 + d * t

  }else {
  a = 1 + LCM(d) * t}
}else{
    s = primeFactors(m)
    d = s[!duplicated(s)]
    d = c(d,4)
    a = 1 + LCM (d) * t
}
 return(a)}
```

Tarea5.R

2.4. Selección de la semilla

Para la selección de la semilla X_0 solo se debe tener en cuenta que debe encontrarse en el rango $0 \leq X_0 < m$.

2.5. Generador congruencial lineal(Mixto) de período completo

Con las funciones 2.1, 2.2, 2.3, se modifica la función 2, dando lugar a la función 2.5 que garantiza el período completo y la posibilidad de reproducir la secuencia si fuera necesario, dando como resultados números pseudo-aleatorios de calidad. Esto se comprueba en la figura 1 de la página 5 y en los resultados de la prueba estadística de uniformidad e independencia Chi-cuadrado con un valor del estadístico $X^2 = 11{,}424$ y el valor p = 0,9088, no se rechaza la hipótesis nula. Por tanto, los números son independientes y Uniformes.

```
uniforme_comp = function(n,p,e,t,ps,po) {
    m = modu(p,e)
    s = semilla(m,ps)
    a = multiplicador(m,t)
    c = aditiva(m,po)
    datos = numeric()
    x = s
```

```
    while (length(datos) < n) {
        x = (a * x + c) %%m
        datos = c(datos, x)
    }
    return(datos / (m - 1))
}
```

Tarea5.R

3. Transformada de Box-Muller

la Transformada de Box-Muller es método para generar pares de independiente, estándar, normalmente distribuido. Este método fue llevado a un programa de R como se muestra en el código 3. A partir de este código se realizo una experimentación donde se variaron los parámetros y se utilizó en uno de los casos el generador *UniformeGLC* propuesto. Como se muestra en el cuadro 1 de la página 4. En este cuadro muestra como afecta los parámetros a la normalidad de los valores.

Cuadro 1: Resultados de la prueba de Shapiro–Wilk

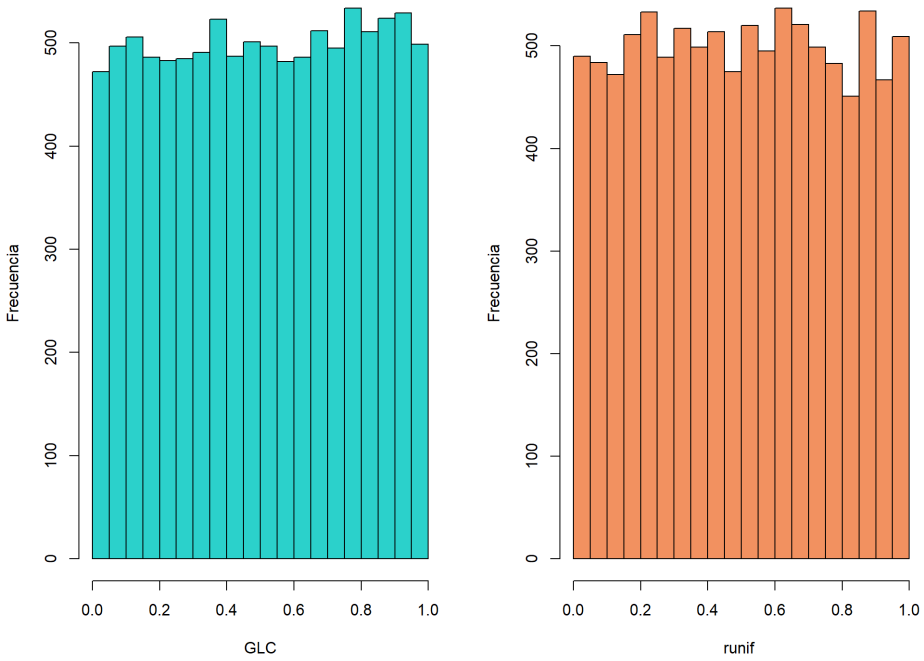
Variante	**Valor p**
rnorm	0.9849
runif	0.6346
UniformeGLC	0.7346
u1/u2	0.0000
Z1	0.0100

```
gaussian = function (mu, sigma) {
 u = runif(2)
 z0 = sqrt(-2*log(u[1])) * cos(2*pi*u[2])
 z1 = sqrt(-2*log(u[1])) * sin(2*pi*u[2])
 datos = c(z0,z1)
 return (sigma * datos + mu)
```

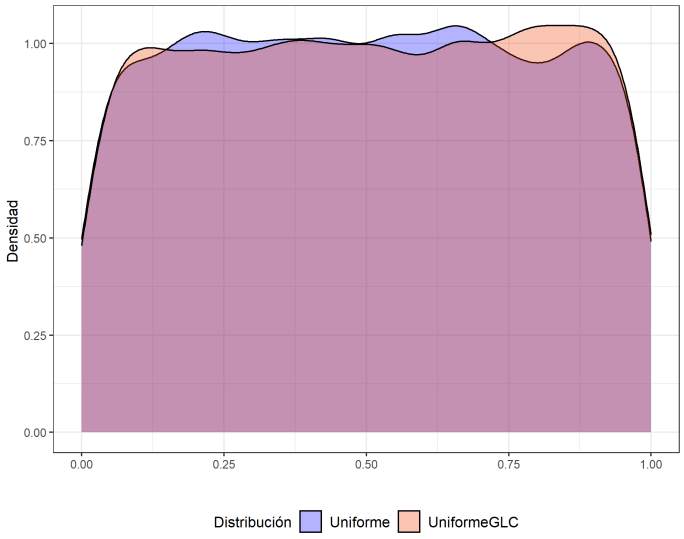
Tarea5.R

En la imagen 2 de la página 6 se muestran los histogramas de las diferentes variantes analizadas, como se puede observar la variante la sdos variantes representan una distribución normal, por la prueba de Shapiro–Wilk con un valor p de 0.98 para la a) y 0.87 para b).

El código general se encuentra disponible en el repositorio. https://github.com/Albertomnoa/Tareas

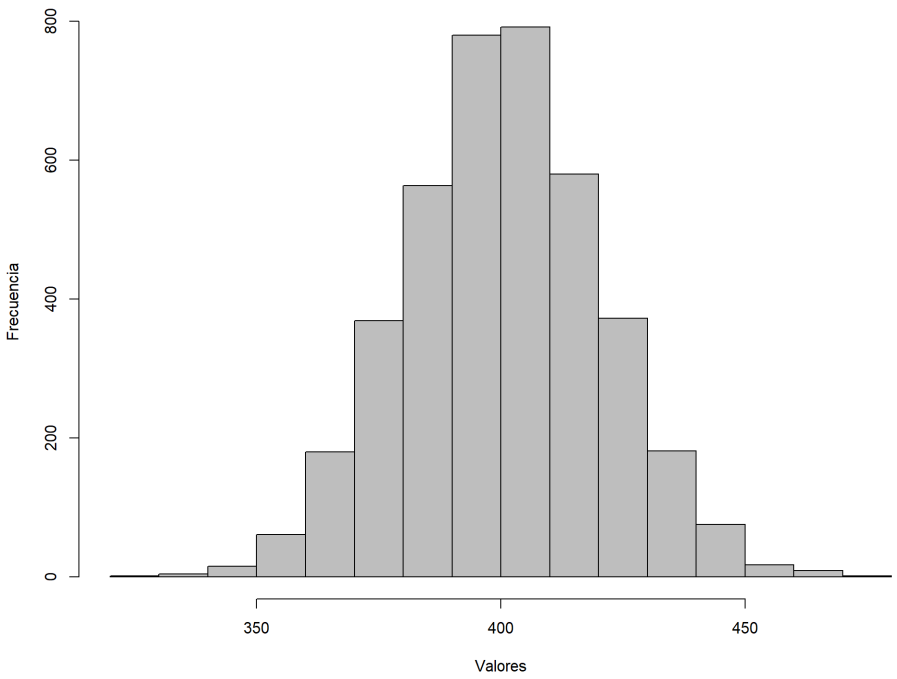


(a) Histogramas de frecuencia de los números pseudo-aleatorios creados por la función *UniformeGCL* presentada y *runif* de R.

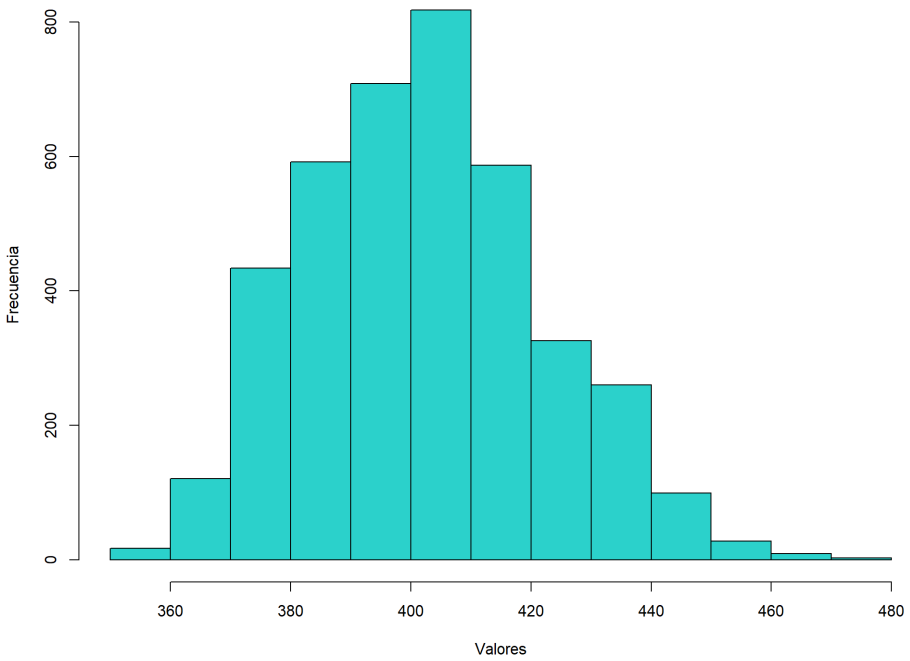


(b) Diagrama de densidad de ambas poblaciones generadas decreciente

Figura 1: Comparación de ambos métodos de generación de números pseudo-aleatorios



(a) Histograma de frecuencia de la distribución normal creada por *runif*



(b) Histograma de frecuencia de la distribución normal creada por *UniformeGLC*

Figura 2: Comparación de ambos métodos de generación de números pseudo-aleatorios

Referencias

[1] Donald E. Knuth. *The Art of Computer Programming, Volume 2 (3rd Ed.): Seminumerical Algorithms*. Addison-Wesley Longman Publishing Co., Inc., USA, 1997.

[2] R Core Team. R: Un lenguaje y un entorno para la informática estadística, 2020.

[3] RStudio Team. Rstudio: Entorno de desarrollo integrado para R, 2020.

Tarea 6 de Modelos Probabilistas Aplicados

Pruebas estadísticas

5271

13 de octubre de 2020

1. Introducción

En este trabajo se presenta un acercamiento al tema pruebas estadísticas, dando respuestas a una serie de preguntas que dan cuerpo a una breve introducción en dicho tema. Además se realizan diversas pruebas estadísticas como ejemplos. Los datos que se utilizan en este trabajo fueros obtenidos en el sitio https://www.inegi.org.mx que pertenece al Instituto Nacional de Estadística y Geografía (INEGI). Se escogió el apartado Construcción donde se muestra información sobre los principales resultados de las Empresas Constructoras, comprende unidades económicas dedicadas principalmente a la edificación; a la construcción de obras de ingeniería civil y a la realización de trabajos especializados de construcción. De este apartado se descargó el tabulado (Valor de producción generado por las empresas constructoras según el tipo de obra) en formato *csv*. Las pruebas se realizan en el programa R versión 4.0.2 [4] en el entorno de desarrollo Rstudio [5]

2. Pruebas estadísticas

En esta sección se realiza un acercamiento a temas importantes relacionados con las pruebas estadísticas, como su características generales, los tipos de pruebas e interpretación de las mismas.

2.1. Relación entre contraste de hipótesis y pruebas estadísticas

Una prueba estadística un procedimiento en el cual se analiza la evidencia proporcionada por los datos con el fin de probar una Hipótesis.La hipótesis estadística es una afirmación sobre los valores del parámetro θ (θ puede ser: μ, p, σ^2, entre otros) de una población o proceso, que es susceptible de probarse a partir de la información contenida en una muestra representativa que es obtenida de la población. Trasladando esto al tema de investigación del autor (Empaquetamiento óptimo), se tiene el siguiente ejemplo, la afirmación "Existe diferencia en el porciento de ocupación de los diferentes tipos de figuras en el contenedor". La veracidad de esta afirmación se obtiene al contrastar las siguientes hipótesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6, \tag{1}$$

$$H_1 : \mu_i \neq \mu_j \quad para \quad alguna \quad i \neq j. \tag{2}$$

Cuadro 1: Influencia del tipo de figura en el % de ocupación del contenedor

	Factor	SS	DF	MS	F	Valor p	np2
0	Tipo de figura	0.0132	5.0000	0.0026	1.0302	0.4147	0.1252
1	Within	0.0920	36.0000	0.0026			

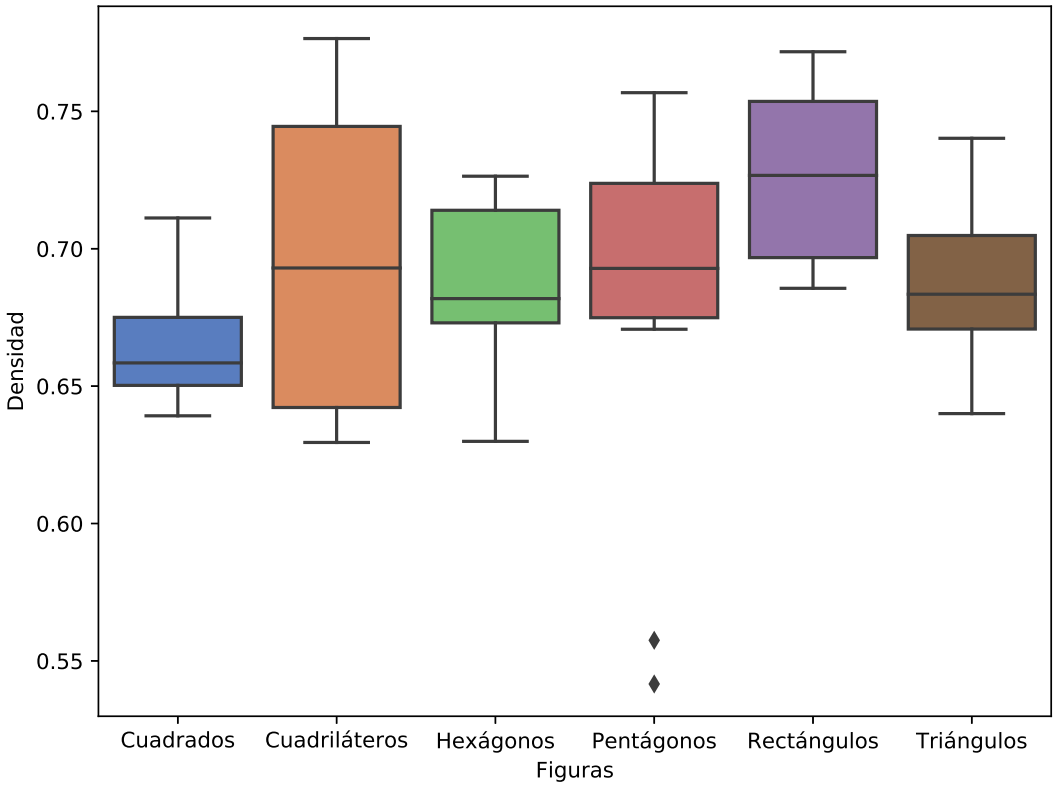


Figura 1: Diagrama de caja y bigotes que relaciona los tipos de figuras y el % de ocupación del contenedor.

A la expresión (13) se le conoce como hipótesis nula y a la expresión (14) se le nombra como hipótesis alternativa. El nombre de hipótesis nula nace de que comúnmente se plantea como una igualdad. Generalmente se trata de probar que la hipótesis nula es verdadera, y que en caso de ser rechazada por la evidencia que aportan los datos, se aceptará la hipótesis alternativa.

En el cuadro 1 de la página 2 se muestra el resultado de la aplicación de un análisis de varianza (ANOVA) al ejemplo antes mencionado. Con los valores del estadístico de prueba $\mathbf{F} = 1.0302$ y el **valor p** $= 0.4117$, se tiene evidencia para aceptar la hipótesis H_0, la cual indica que no existe diferencias estadísticas significativas entre los tratamientos con un intervalo de confianza del 95 %. Esto se puede observar en la figura 1 de la de la página 2.

Cuadro 2: Influencia de la cantidad de figura en el % de ocupación del contenedor

	Factor	SS	DF	MS	F	Valor p	np2
0	Cantidad defiguras	0.0429	5.0000	0.0086	4.9514	0.0015	0.4075
1	Within	0.0623	36.0000	0.0017	–	–	–

2.2. En que casos se rechaza la hipótesis nula

Para probar una hipótesis se trata de corroborar que la afirmación planteada por la hipótesis nula (H_0) es verdad o no. Se parte del supuesto que es verdadera, si la evidencia arrojada por los datos es suficiente para contradecir dicho supuesto se rechaza H_0 y se acepta la Hipótesis alternativa (H_1). En caso de no haber evidencias suficientes que demuestren la falsedad de la H_0 esta no se rechaza, es decir H_0 es verdad hasta que no se demuestre lo contrario.

Para discernir si H_0 se rechaza o no, se utiliza un estadístico de prueba el cual es un es un número calculado a partir de los datos y la hipótesis nula. Al conjunto de posibles valores del estadístico de prueba que llevan a rechazar H_0, se le llama región o intervalo de rechazo para la prueba, y a los posibles valores donde no se rechaza H_0 se les conoce como región o intervalo de confianza.

El estadístico de prueba, construido bajo el supuesto de que H_0 es verdad, es una variable aleatoria con distribución conocida. Si efectivamente H_0 es verdad, el valor del estadístico de prueba debería caer dentro del rango de valores más probables de su distribución asociada, el cual se conoce como región de confianza. Si cae en una de las colas de su distribución asociada, fuera del rango de valores más probables (en la región de rechazo), es evidencia en contra de que este valor pertenece a dicha distribución. De aquí se deduce que debe estar mal el supuesto bajo el cual se construyó, es decir, H0 debe ser falsa [3].

Análogamente al análisis del ejemplo de la sección 2.1, se realiza la misma prueba para el factor de control cantidad de figuras. Obtenemos como resultado el cuadro 2 de la página 3. Con los valores del estadístico de prueba $\mathbf{F} = 4.9514$ y con un **valor p** $= 0.0015$ menor que 0.05, se tiene evidencia para rechazar la hipótesis H_0 ya que existe diferencias estadísticas significativas al menos entre algunos de los grupos con un intervalo de confianza de un **95 %** como se muestra en la figura 2 de la página 4.

2.3. Interpretación de la salida de una prueba estadística

En la sección 2.1 y 2.2 muestran ejemplos de como interpretar la salidas de las pruebas estadísticas a través de los estadísticos de prueba explicados en la sección 2.2 y el **valor p**, representa una probabilidad que mide la evidencia en contra de la hipótesis nula. **valor p** más pequeño proporciona una evidencia más fuerte en contra de la hipótesis nula.

Para determinar si la diferencia entre las desviaciones estándar o las varianzas de las poblaciones es estadísticamente significativa, se compara el **valor p** con el nivel de significancia. Por lo general, el nivel de significancia (denotado como α) indica el riesgo de concluir que existe una diferencia cuando realmente no la hay.

- Sí el **valor p** $> \alpha$, la relación de las desviaciones estándar o las varianzas no es estadísticamente significativa (no puede rechazar H_0).
- Sí el **valor p** $\leq \alpha$, la relación de las desviaciones estándar o las varianzas es estadísticamente significativa (se rechaza H_0).

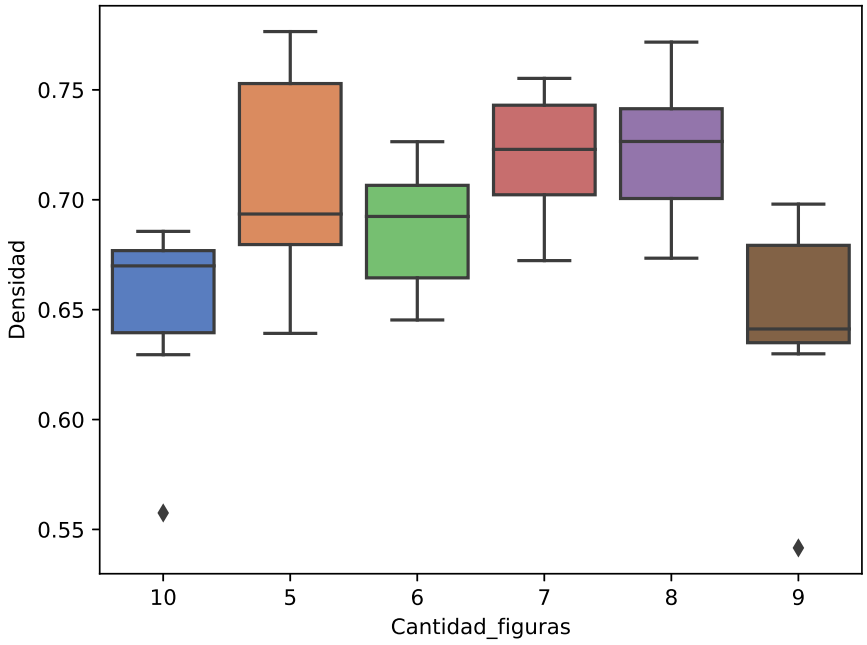


Figura 2: Diagrama de caja y bigotes que relaciona los cantidad de figuras y el % de ocupación del contenedor.

2.4. Selección de nivel de significancia

El valor de α es el máximo nivel de riesgo aceptable de rechazar la hipótesis nula cuando la hipótesis nula es verdadera, como se comento en la sección 2.3. Por lo general, se elige elige el nivel de significancia antes de analizar los datos. El este nivel se calcula como 1 menos el intervalo de confianza. Para los ejemplos anteriores se empleo un intervalo de confianza del 95 %, por lo que α es igual a 0.05.

Los valores de α son elegidos en dependencia de la finalidad de la prueba, es decir si se quiere determinar cualquier diferencia que exista se utiliza un valor de α mas grande (0.10). En el caso que se quiera asegurar que el las diferencias que existen son reales, se emplea un valor de α muy pequeño (0.01). Comúnmente se utiliza el valor de α igual a 0.05.

2.5. Errores frecuentes de interpretación del valor p

Probar una hipótesis estadística es una decisión probabilística, por lo que existe el riesgo de cometer un error **tipo I** o un error **tipo II**. El primero ocurre cuando se rechaza H_0 cuando ésta es verdadera, y el error **tipo II** es cuando se acepta H_0 y ésta es falsa. Con α y β se denotan las probabilidades de los errores **tipo I** y **tipo II**, respectivamente.

Donde α es el nivel de significancia explicado en la sección 2.4. Y β es la llamada potencia de prueba.

Cuadro 3: Prueba Shapiro-Wilk a la variable % de ocupación

	W	**Valor p**
% de ocupación	0.9537	0.0877

2.6. Potencia de prueba, utilidad

A $1 - \beta$ se le llama potencia de la prueba, y es la probabilidad de rechazar H_0 cuando es falsa. Por lo general, en las pruebas de hipótesis se especifica el valor de α y se diseña la prueba de tal forma que el valor de β sea pequeño. Esto es, la probabilidad del error **tipo I** se controla directamente, mientras que la probabilidad de error **tipo II** se controla de manera indirecta con el tamaño de la muestra, ya que a más datos β será menor. Es decir, con una muestra grande es mayor la potencia de la prueba, de manera que se incrementa la probabilidad de rechazar H_0 si ésta es falsa.

2.7. Pruebas paramétricas

Para la aplicación de las pruebas paramétricas, los datos deben cumplir ciertas presunciones [1]:

- Los datos son de escala de intervalo o razón.
- La población de la muestra debe aproximarse a una distribución normal.
- Las varianzas de las muestras deben aproximadamente similares.
- Las observaciones deben ser independientes entre sí.

Entre las pruebas perimétricas más usadas están:

- Prueba de Shapiro-Wilks.
- Prueba de Fisher.
- Prueba t de Student para muestras independientes.
- Coeficiente de correlación de Pearson.
- Regresión lineal.
- Análisis de varianza factorial (ANOVA).
- Análisis de covarianza (ANCOVA).

En ambos ejemplos presentados en la sección 2.1 y 2.2 Antes de realizar el ANOVA, se realiza la prueba de Shapiro-Wilk, que calcula un W estadístico que prueba si una muestra aleatoria $x_1, x_2, ..., x_n$ proviene de una distribución normal. El resultado de esta prueba se muestra en el cuadro 3 de la página 5. En dicho cuadro con un $\mathbf{W} = 0.9537$ y un **valor p** $= 0.0877$ mayor que 0.05, se puede observar que la variable dependiente (% de ocupación) sigue una distribución normal con un intervalo de confianza del 95 %.

Cuadro 4: Prueba Shapiro-Wilk a la variable % de ocupación variando el contenedor

	W	Valor p	Normal
Densidad	0.9527	0.0037	Falso

Cuadro 5: Prueba H de Kruskal-Wallis que relaciona el % de ocupación con el tipo de contenedor

	Factor	ddof1	H	pval
Kruskal	Contenedor	1.0000	13.9167	0.0002

2.8. Pruebas no paramétricas

En la sección 2.7 se plantea características que deben tener los datos para usar las pruebas paramétricas. Cuando los datos no cumplen una de estas características. Se emplean las pruebas no paramétricas, ejemplos de estas son [2]:

- Prueba Chi-cuadrada.
- Coeficientes de correlación e independencia para tabulaciones cruzadas.
- Coeficientes de correlación por rangos ordenados Spearman y Kendall.
- Prueba H de Kruskal-Wallis.

Como en los ejemplos anteriores se realiza una prueba de Shapiro-Wilk para corroborar que la variable dependiente (% de ocupación) variando el tipo de contenedor, sigue una distribución normal. Los resultados obtenidos con un **W** = 0.9528 y **valor p**=0.0038 menor que 0.05, nos muestra que dicha variable no sigue una distribución normal con un intervalo de confianza del **95 %**. Lo antes expuestos se puede observar en la cuadro 4 de la página 6.

Dado que los valores de la variable dependiente no sigue una distribución normal, se aplica la prueba H de Kruskal-Wallis que es una versión no paramétrica de ANOVA. En este caso se plantea la afirmación: "Existen diferencias en el % de ocupación según el tipo de contenedor".

Según los datos obtenidos en la prueba que se muestran en el cuadro 5 de la página 6. Estos resultados con un valor del estadístico de prueba **H** = 13.9167 y **valor p** = 0.0002, se rechaza la hipótesis H_0 aceptando la H_1, la cual indica que existen diferencias estadísticamente significativas entre los tipos de contenedores. En esta ocasión con los resultados de la prueba H de Kruskal-Wallis son suficientes para asegurar lo anterior dado que solo son dos los tipos de contenedores analizados. Gráficamente se puede observar esta conclusión en la figura 3 de la página 7.

3. Aplicación de Pruebas estadísticas

En esta sección se presentaran diversas pruebas estadísticas aplicas a los datos del sector de la construcción obtenidos de INEGI.

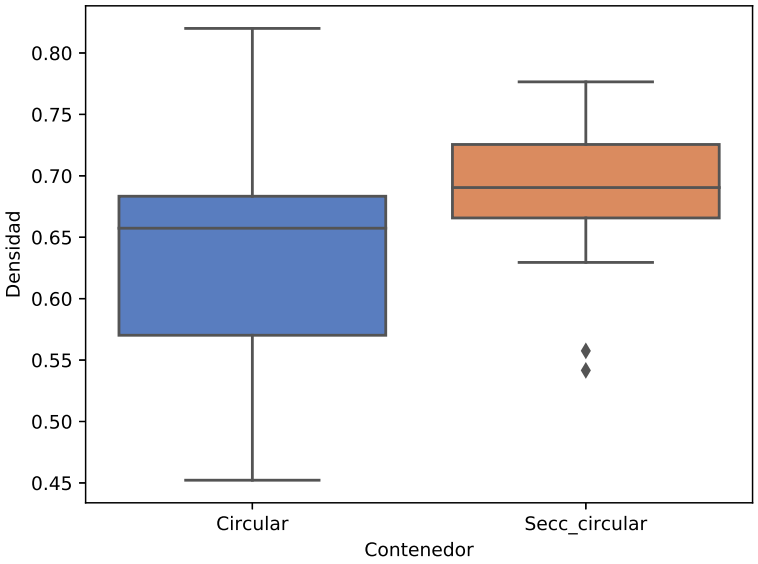


Figura 3: Diagrama de caja y bigotes que relaciona los tipos de contenedor y el % de ocupación del mismo.

Cuadro 6: Resultados de la prueba de Shapiro-Wilk

Tipo de obra	W	Valor p	
Agua, riego y saneamiento	0.9907	0.3223	$> \alpha$
Petróleo y petroquímica	0.9659	0.0002	$< \alpha$
Otras construcciones	0.9451	0.0000	$< \alpha$

3.1. Prueba de Shapiro-Wilk

Como se explica en la sección 2.7 la prueba de Shapiro-Wilk se utiliza para probar si los datos siguen una distribución normal. Por lo cual esta prueba es la primera que se le realiza a los datos para conocer la naturaleza de los mismo y así elegir el tipo de prueba a utilizar. Para el caso de ejemplo se tomara los valores de producción de las constructoras según el tipo de obras. En el cuadro 6 de la página 7 se muestra obtenidos en las pruebas. Además la figura 4 de la página 8 muestra como se comportan los datos según el tipo de obra.

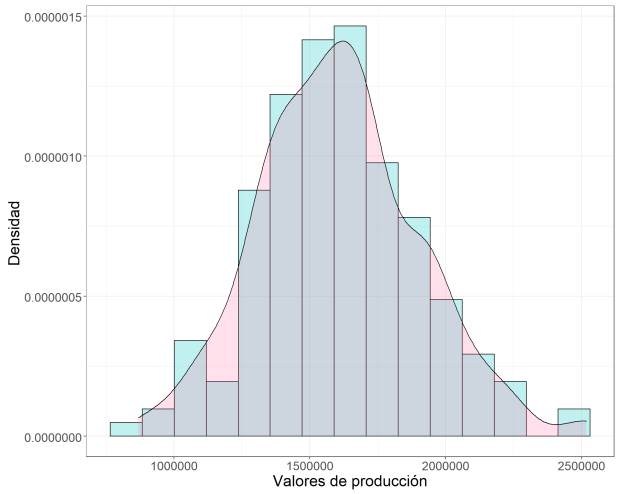
Las pruebas se realizan con un $\alpha = 0.05$, y la hipótesis planteada es la siguiente:

$$H_0 : \text{La muestra sigue un distribución igual a la normal,} \tag{3}$$

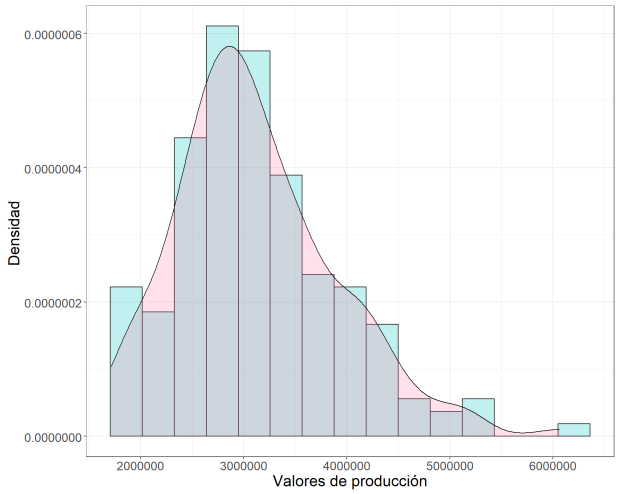
$$H_1 : \text{La muestra sigue un distribución diferente a la normal.} \tag{4}$$

3.2. Prueba t de una muestra

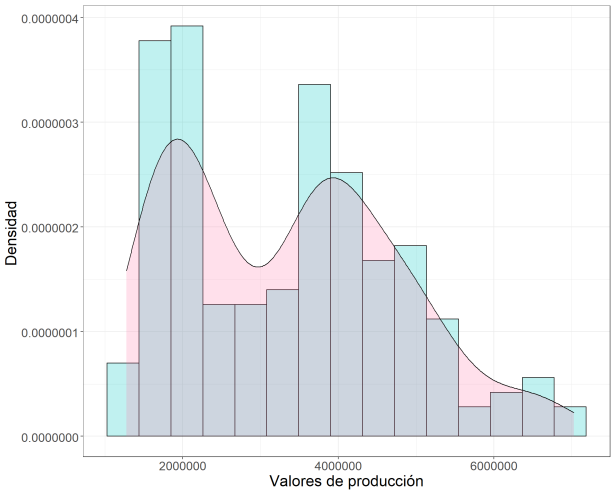
la prueba t de una muestra, es una prueba paramétrica, se utiliza para probar si la media de una muestra puede ser un valor especifico. Al ser paramétrica como premisa necesita que la muestra siga una distribución normal. De los resultados en el cuadro 6 de la página 7, podemos concluir que solo



(a) Agua, riego y saneamiento



(b) Petróleo y petroquímica



(c) Otras construcciones

Figura 4: Histogramas de distribución de los datos por tipo de obras

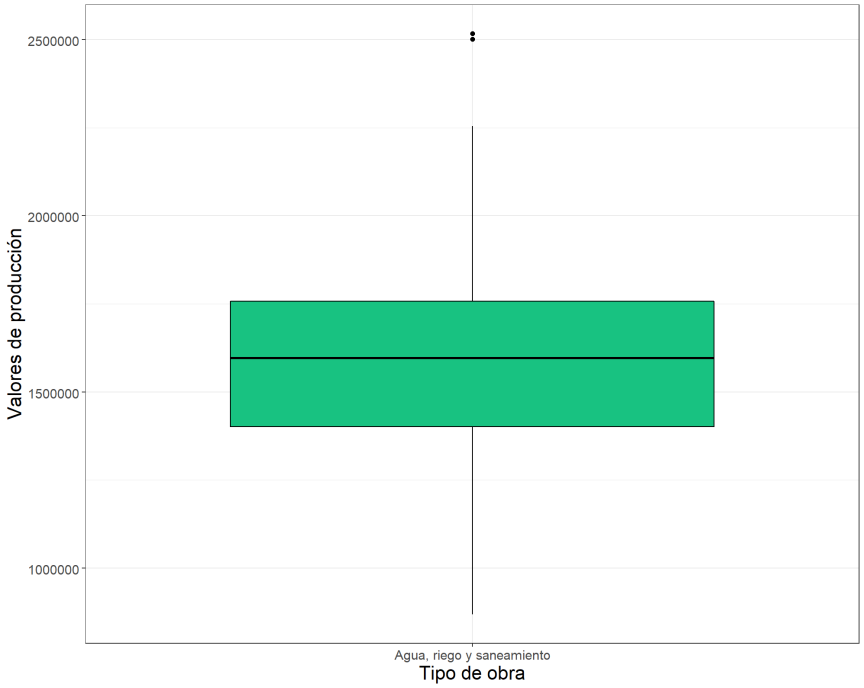


Figura 5: Diagrama de caja y bigotes del tipo de obra Agua, riego y saneamiento

podemos aplicar la prueba t a los datos del tipo de obras Agua, riego y saneamiento, por ser los únicos que siguen una distribución normal.

La prueba se realiza con un $\alpha = 0.05$, y la hipótesis planteada es la siguiente:

$$H_0 : \text{La } \mu \text{ de los de los valores de producción es } 1600000, \tag{5}$$

$$H_1 : \text{La } \mu \text{ de los de los valores de producción diferente } 1600000. \tag{6}$$

Los resultados obtenidos en la prueba, con un estadístico $\mathbf{t} = 0.3014$ y un **valor p** $= 0.7635$ mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo tanto la media de los valores de producción es 1600000 con un intervalo de confianza de 95 %. Esto se puede observar en la figura 5 de la página 9.

3.3. Prueba de rango con signo de Wilcoxon

La prueba de rango con signo de Wilcoxon puede verse como una alternativa a la prueba t, dado que su objetivo es determinar si la media de una muestra es un valor especifico sin tener en cuenta el supuesto que la muestra sigue una distribución normal.

Para la prueba se utilizan los datos pertenecientes al tipo de obra Petróleo y petroquímica, debido a que estos no siguen una distribución normal. La prueba se realiza con un $\alpha = 0.05$, y la hipótesis planteada es la siguiente:

$$H_0 : \text{La } \mu \text{ de los de los valores de producción es } 300000, \tag{7}$$

$$H_1 : \text{La } \mu \text{ de los de los valores de producción diferente } 3000000. \tag{8}$$

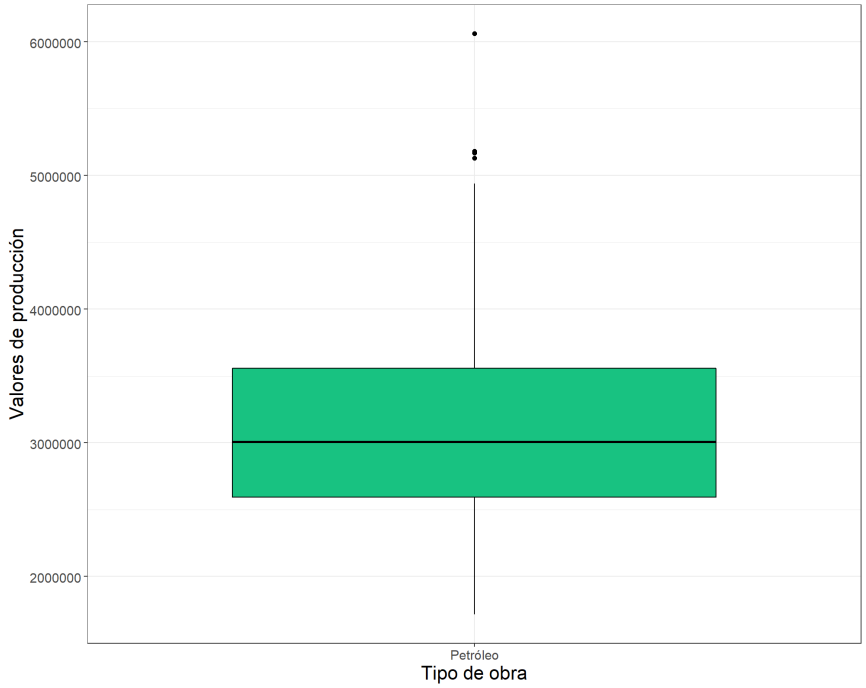


Figura 6: Diagrama de caja y bigotes del tipo de obra Petróleo y petroquímica

Los resultados obtenidos en la prueba, con un estadístico **V** = 8399 y un **valor p** = 0.2375 mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo tanto la media de los valores de producción de las obras Petróleo y petroquímica es 3000000 con un intervalo de confianza de 95 %. Esto se puede observar en la figura 6 de la página 10.

3.4. Prueba t de dos muestras de rangos de Wilcoxon

La Prueba t de dos muestras de rangos de Wilcoxon tiene como objetivo comparar las medias de dos muestras que no siguen una distribución normal.

Para esta prueba se utilizan los datos pertenecientes al tipo de obra Petróleo y petroquímica y Otras construcciones, debido a que estos no siguen una distribución normal. La prueba se realiza con un α = 0.05, y la hipótesis planteada es la siguiente:

$$H_0 : \mu\,(\text{Petróleo y petroquímica}) = \mu\,(\text{Otras construcciones}), \tag{9}$$

$$H_1 : \mu\,(\text{Petróleo y petroquímica}) \neq \mu\,(\text{Otras construcciones}). \tag{10}$$

Los resultados obtenidos en la prueba, con un estadístico **W** = 14364 y un **valor p** = 0.7954 mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo tanto la media de los valores de producción de las obras Petróleo y Otras construcciones tienen la misma media con un intervalo de confianza de 95 %. Esto se puede observar en la figura 7 de la página 11.

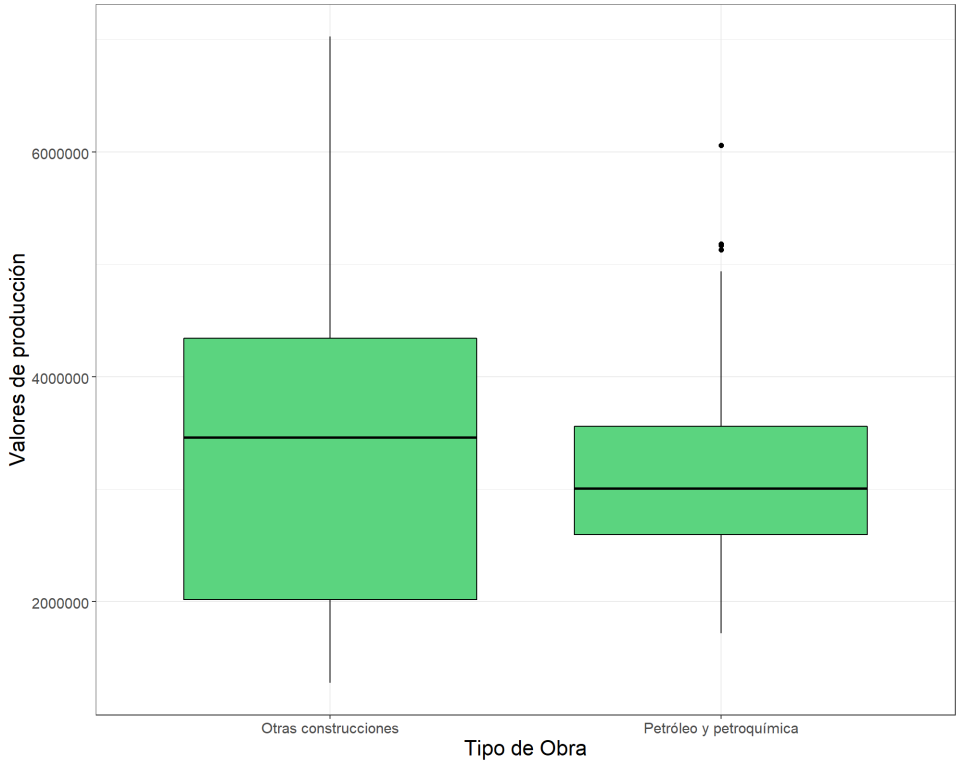


Figura 7: Diagrama de caja y bigotes del tipo de obra Petróleo y petroquímica y Otras construcciones

3.5. Prueba de Kolmogorov y Smirnov

La prueba de Kolmogorov-Smirnov se utiliza para comprobar si dos muestras siguen la misma distribución.

Para esta prueba se utilizan los datos pertenecientes al tipo de obra Agua, riego y saneamiento y datos creados con la función *rnorm* de R. Con el objetivo de verificar los resultados de la prueba Shapiro-Wilk realizada en la 3.2. La prueba se realiza con un $\alpha = 0.05$, y la hipótesis planteada es la siguiente:

$$H_0 : (\text{la distribución de Agua, riego y saneamiento}) = (\text{la distribución } rnorm), \tag{11}$$

$$H_1 : (\text{la distribución de Agua, riego y saneamiento}) \neq (\text{la distribución } rnorm). \tag{12}$$

Los resultados obtenidos en la prueba, con una distancia de Kolmogorov-Smirnov $\mathbf{D} = 0.086207$ y un **valor p** $= 0.5375$ mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo tanto los valores de producción de tipo de obra Agua, riego y saneamiento siguen una distribución normal con un intervalo de confianza de 95 %. Esto se puede observar en la figura 8 de la página 13.

3.6. Prueba F de Fisher

La prueba F de Fisher se puede utilizar para comprobar que dos muestras tienen la misma varianza.

Para esta prueba se utilizan los datos pertenecientes al tipo de obra Agua, riego y saneamiento y datos creados con la función *rnorm* de R. Con el objetivo de verificar si ambos poseen la misma varianza. La prueba se realiza con un $\alpha = 0.05$, y la hipótesis planteada es la siguiente:

$$H_0 : (\text{la varianza de Agua, riego y saneamiento}) = (\text{la varianza } rnorm), \tag{13}$$

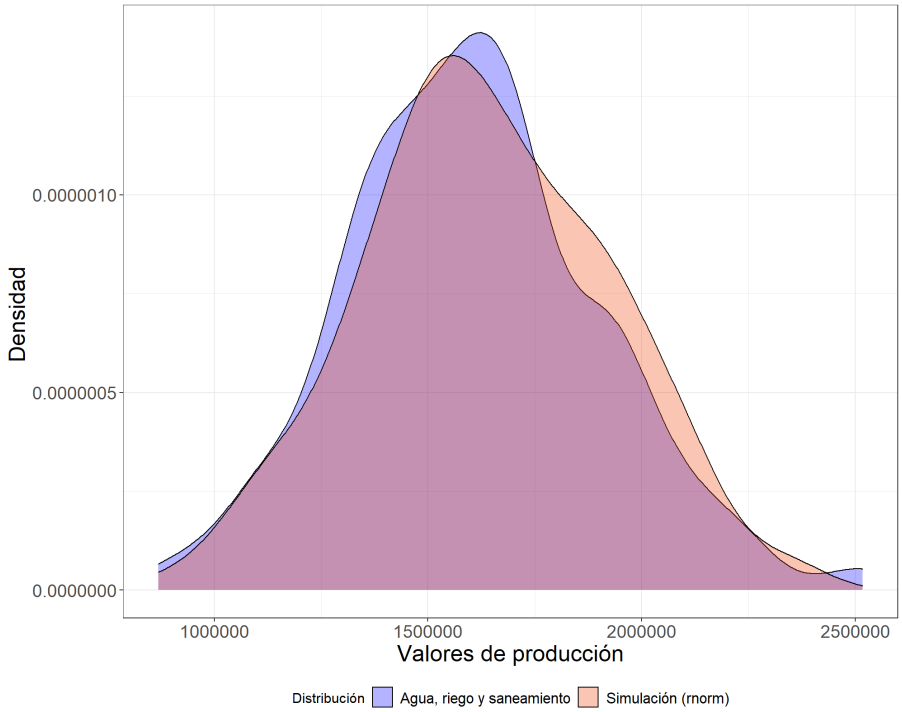
$$H_1 : (\text{la varianza de Agua, riego y saneamiento}) \neq (\text{la varianza } rnorm). \tag{14}$$

Los resultados obtenidos en la prueba, con un estadístico $\mathbf{f} = 1.0442$ y un **valor p** $= 0.7766$ mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo tanto los valores de producción de tipo de obra Agua, riego y saneamiento y los valores creados con *rnorm* tiene la misma varianza como era de esperarse, con un intervalo de confianza de 95 %.

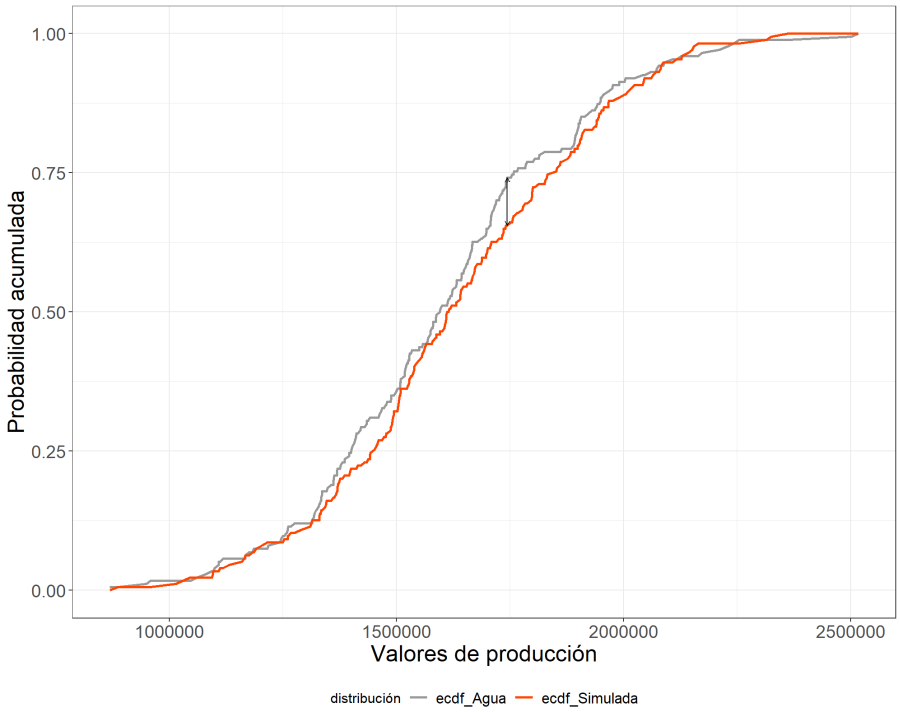
3.7. Prueba de Chi-cuadrada

La prueba de chi-cuadrado en R se puede utilizar para probar si dos variables son dependientes.

Para esta prueba se utilizan los datos pertenecientes al tipo de obra Agua, riego y saneamiento. Con el objetivo de verificar estos son independientes con respecto al periodo en que fueron registrados. La prueba se realiza con un $\alpha = 0.05$. Los resultados obtenidos en la prueba, con un estadístico $X^2 = 2088$ y un **valor p** $= 0.4222$ mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo tanto los valores de producción de tipo de obra Agua, riego y saneamiento son valores independientes con respecto al periodo en que fueron registrados,con un intervalo de confianza de 95 %.



(a) Diagrama de densidad de Agua, riego y saneamiento superpuesto al simulado con *rnorm*



(b) [Diferencia de Kolmogorov-Smirnov entre Agua, riego y saneamiento superpuesto y la simulación con *rnorm*

Figura 8: Prueba de Kolmogorov-Smirnov

3.8. Correlación

La correlación nos indica la relación lineal de dos variables continuas.

Para esta prueba se utilizan los datos pertenecientes al tipo de obra Petróleo y petroquímica y Otras construcciones. La prueba se realiza con un $\alpha = 0.05$.Los resultados obtenidos en la prueba, con un estadístico **t** = 2.4082 y un **valor p** = 0.05709 mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo la correlación es cero,con un intervalo de confianza de 95 %.

El código general se encuentra disponible en el repositorio. https://github.com/Albertomnoa/Tareas

Referencias

[1] Ecured. Pruebas estadísticas. https://www.ecured.cu/Pruebas_estad%C3%ADsticas.

[2] Rosana Ferrero. Guía definitiva para encontrar las pruebas estadísticas que buscas. https://www.maximaformacion.es/blog-dat/guia-para-encontrar-tu-prueba-estadistica/.

[3] Gutiérrez Pulido Humberto. *Análisis y diseño de experimentos*. McGRAW-HILL/INTERAMERICANA EDITORES, S.A. de C.V., México, D.F, 2008.

[4] R Core Team. R: Un lenguaje y un entorno para la informática estadística, 2020.

[5] RStudio Team. Rstudio: Entorno de desarrollo integrado para R, 2020.

Tarea 7 de Modelos Probabilistas Aplicados

Transformaciones mediante Escalera de poderes de Tukey

5271

20 de octubre de 2020

1. Introducción

En este trabajo se realiza un breve acercamiento al tema Transformaciones mediante Escalera de poderes de Tukey y regresión lineal. Además presenta un algoritmo desarrollado en lenguaje R para extraer la función que genera la variable dependiente así como sus coeficientes. La experimentación se realiza en el programa R versión 4.0.2 [1] en el entorno de desarrollo Rstudio [2]

2. Transformaciones mediante Escalera de poderes de Tukey

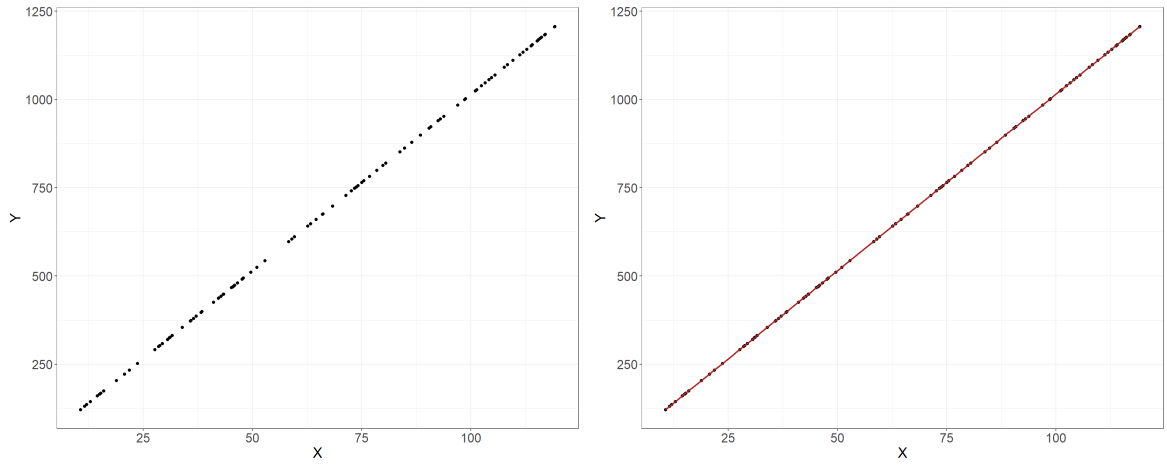
Para entender mejor como funciona las transformaciones mediante Escalera de poderes de Tukey creamos un *data.frame* con datos bivariados (x_1, y) de manera que x_1 se distribuye uniformemente e $y = ax_1 + b$. Lo anterior se realiza con el código 2 en R.

```r
x_1 = runif(100,10,120)
a = 8
b = 15
y = a*(x_1) + b
datas = as.data.frame(y)
datas = cbind(datas,x_1)
```

Tarea7n.R

Como primer paso se trazan los datos en un diagrama de dispersión. En la figura 1 de la página 2 se puede observar como se relaciona las variables x_1 e y. Como se aprecia en dicha figura la relación entre x_1 e y es una dependencia lineal y si aplicamos una regresión lineal obtendremos exactamente los coeficientes y formulación con la cual se crea la variable dependiente y. A continuación se metra el resultado de aplicar la regresión lineal. De donde se pueden extraer los coeficientes a y b de la ecuación, quedando $y = 15x_1 + 8$.

```
Call:
lm(formula = datas$y ~ datas$x_1)

Residuals:
```



(a) $y = ax_1 + b$ (b) regresión lineal

Figura 1: Diagramas de dispersión de los datos bivariados (x_1, y)

```
       Min         1Q     Median         3Q        Max
-8.102e-13 -3.110e-15  8.040e-15  1.811e-14  9.951e-14

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 1.500e+01  2.016e-14 7.442e+14   <2e-16 ***
datas$x_1   8.000e+00  2.744e-16 2.915e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.501e-14 on 98 degrees of freedom
Multiple R-squared:      1,Adjusted R-squared:      1
F-statistic: 8.498e+32 on 1 and 98 DF,  p-value: < 2.2e-16

Warning message:
In summary.lm(modelo_lineal) :
  essentially perfect fit: summary may be unreliable
```

El ejemplo anterior es sencillo pues con solo una regresión lineal se pueden obtener los coeficientes y la función que crea a y. En cambio si creamos representamos los datos bivariados creados con $y = a * log(x_1) + b$, se puede observar en la figura 2 de la página 3 que la relación de dependencia no es lineal, por lo que una simple regresión no arrogará los resultados deseados, por lo cual antes de aplicar la regresión se debe hacer una transformación a una de las variable, es decir tratar de linealizar la relación de dependencia.

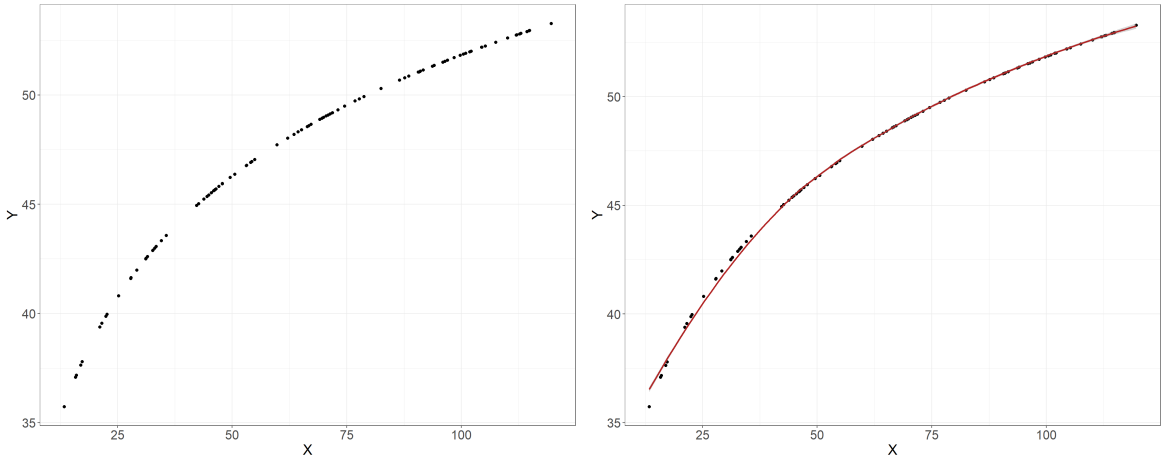
No hay ninguna restricción sobre los valores de λ que podamos considerar. Obviamente, elegir $\lambda = 1$ deja los datos sin cambios. Los valores negativos de λ también son razonables. Lo que da lugar a la ecuación 2. El cuadro 1 de la página 3 se muestran ejemplos de la escalera de transformaciones de Tukey.

Tukey sugiere explorar relaciones como:

$$y = ax_1^{\lambda} + b \tag{1}$$

Cuadro 1: Escalera de transformaciones de Tukey

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



(a) $y = a * log(x_1) + b$ (b) regresión lineal

Figura 2: Diagramas de dispersión de los datos bivariados (x_1, y) con dependencia no lineal

donde λ es una parámetro que elegido con el fin de linealizar la relación.

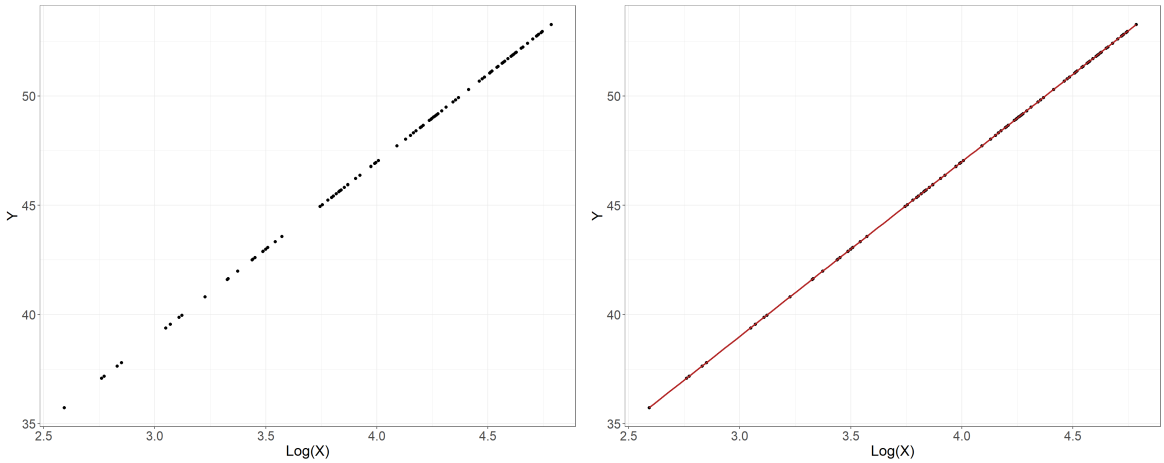
$$y = \begin{cases} x^{\lambda} & \text{si} \lambda > 0 \\ \log \text{ x} & \text{si} \lambda = 0 \\ -(x^{\lambda}) & \text{si} \lambda < 0 \end{cases} \tag{2}$$

Se Aplica transformaciones mediante la escalera de Tukey a los datos, en forma que se calcula el logaritmo de la variable x_1, y se vuelve a representar los datos y se realiza la regresión lineal. En la figura 3 de la página 4 se puede observar la transformación de los datos.

3. Algoritmo propuesto

Según lo visto en la sección anterior podemos mediante las transformaciones de la Escalera de Tukey, podemos acercar la relación entre los datos bivariados a una dependencia lineal. A partir de está conclusión se creo una función en R que la pasarle como parámetros un *data.frame* y la cantidad de variables de cuales depende y (máximo tres variables), devuelve un *data.frame* con los valores de λ encontrados para la linealización, los coeficientes de la función y el error estándar de la estimación delos coeficientes. En el código **??** se muestra dicha función.

```
ajuste1 = function(datos,v){
  rsq=numeric()
  landax1=numeric()
  landax2=numeric()
  landax3=numeric()
  coefi_a=numeric()
```



(a) $y = a * log(x_1) + b$ (b) regresión lineal

Figura 3: Diagramas de dispersión de los datos bivariados (x_1, y) con dependencia no lineal

```
coefi_b=numeric()
coefi_c=numeric()
error_a=numeric()
error_c=numeric()
error_b=numeric()
corr = numeric()
diferencia = numeric()
s=seq(-2,2,0.25)
if(v==1){
  for (i in s) {
    x1=numeric()
    if(i<0){
      x1=-1/((datos$x_1)**(i*-1))
    }else if(i>0){
      x1=(datos$x_1)**i
    }else{
      x1 =log(datos$x_1)}

    modelo_lineal = lm(datos$y ~ x1)
    rsq = c(rsq,summary(modelo_lineal)$r.squared)
    diferencia = c(diferencia,(1-summary(modelo_lineal)$r.squared))
    coefi_a=c(coefi_a,modelo_lineal$coefficient[1])
    error_a=c(error_a,round(summary(modelo_lineal)$coefficient[1,2],4))
    coefi_b=c(coefi_b,modelo_lineal$coefficient[2])
    error_b=c(error_b,round(summary(modelo_lineal)$coefficient[2,2],4))
    landax1=c(landax1,i)
  }
  corre= as.data.frame(rsq)
  corrl= cbind(corre,landax1,diferencia,coefi_a,error_a,coefi_b,error_b)
  resultado=filter(corrl,diferencia==min(corrl$diferencia))


}else if(v==2){
  for (i in s) {
    x1=numeric()
    if(i<0){
      x1=-1/((datos$x_1)**(i*-1))
    }else if(i>0){
      x1=(datos$x_1)**i
    }else{
```

```
        x1 =log(datos$x_1)}
      for (j in s) {
        x2=numeric()
        if(j<0){
          x2=-1/((datos$x_2)**(j*-1))
        }else if(j>0){
          x2=(datos$x_2)**j
        }else{
          x2 =log(datos$x_2)}

        modelo_lineal = lm(datos$y ~ x1+x2)
        rsq = c(rsq,summary(modelo_lineal)$r.squared)
        diferencia = c(diferencia,(1-summary(modelo_lineal)$r.squared))
        coefi_a=c(coefi_a,modelo_lineal$coefficient[2])
        error_a=c(error_a,round(summary(modelo_lineal)$coefficient[2,2],4))
        coefi_b=c(coefi_b,modelo_lineal$coefficient[3])
        error_b=c(error_b,round(summary(modelo_lineal)$coefficient[3,2],4))
        landax1=c(landax1,i)
        landax2=c(landax2,j)

      }
    }
    corre= as.data.frame(rsq)
    corrl= cbind(corre,landax1,landax2,diferencia,coefi_a,error_a,coefi_b,error_b)
    resultado=filter(corrl,diferencia==min(corrl$diferencia))
  }else if(v==3){

    for (i in s) {
      x1=numeric()
      if(i<0){
        x1=-1/((datos$x_1)**(i*-1))
      }else if(i>0){
        x1=(datos$x_1)**i
      }else{
        x1 =log(datos$x_1)}
      for (j in s) {
        x2=numeric()
        if(j<0){
          x2=-1/((datos$x_2)**(j*-1))
        }else if(j>0){
          x2=(datos$x_2)**j
        }else{
          x2 =log(datos$x_2)}
        for (k in s) {
          x3=numeric()
          if(k<0){
            x3=-1/((datos$x_3)**(k*-1))
          }else if(k>0){
            x3=(datos$x_3)**k
          }else{
            x3 =log(datos$x_3)}

          modelo_lineal = lm(datos$y ~ x1+x2+x3)
          rsq = c(rsq,summary(modelo_lineal)$r.squared)
          diferencia = c(diferencia,(1-summary(modelo_lineal)$r.squared))
          coefi_a=c(coefi_a,modelo_lineal$coefficient[2])
          error_a=c(error_a,round(summary(modelo_lineal)$coefficient[2,2],4))
          coefi_b=c(coefi_b,modelo_lineal$coefficient[3])
          error_b=c(error_b,round(summary(modelo_lineal)$coefficient[3,2],4))
          coefi_c=c(coefi_c,modelo_lineal$coefficient[4])
          error_c=c(error_c,round(summary(modelo_lineal)$coefficient[4,2],4))
          landax1=c(landax1,i)
          landax2=c(landax2,j)
          landax3=c(landax3,k)
        }
```

```
      }
    }
    corre= as.data.frame(rsq)
    corrl= cbind(corre,landax1,landax2,landax3,diferencia,coefi_a,error_a,coefi_b,
    error_b,coefi_c,error_c)
    resultado=filter(corrl,diferencia == min(corrl$diferencia))

  }

  return(resultado)
}
```

Tarea7n.R

Resultados El código general se encuentra disponible en el repositorio. https://github.com/Albertomnoa/Tareas

Referencias

[1] R Core Team. R: Un lenguaje y un entorno para la informática estadística, 2020.

[2] RStudio Team. Rstudio: Entorno de desarrollo integrado para R, 2020.

Tarea 8 de Modelos Probabilistas Aplicados

Aplicación del teorema de Bayes en datos relacionados con el Covid-19

5271

27 de octubre de 2020

1. Introducción

En este trabajo se realiza un acercamiento a uno de los temas más relevantes de la actualidad, la pandemia del Covid-19. En el mismo se realiza un análisis sobre artículos que emplean el teorema de Bayes en datos de las pruebas diagnostico realizadas en la detención de la enfermedad, teniendo en cuenta la sensibilidad, especificidad y valores predictivos (positivos, negativos) de las pruebas. Además se presenta el cálculo de los valores predictivos (positivos, negativos) con datos abiertos sobre la aplicación de pruebas de Covid-19 en México.

2. Teorema de Bayes

Sea $B_1, B_2, ..., B_n$ una partición del espacio muestral (Ω) tal que $P(B_i) \neq 0$ para $i = 1, 2, ..., n$ y sea A un evento tal que $P(A) \neq 0$. Entonces para cada $j = 1, 2, ..., n$ se tiene:

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}. \tag{1}$$

3. Sensibilidad y especificidad de una prueba diagnostico

En esta sección se explicarán dos conceptos básicos a tener en cuenta en el análisis de una prueba diagnóstico, como son la sensibilidad y la especificidad.

La sensibilidad representa la probabilidad de que una persona enferma obtenga un resultado positivo. Esta se define como:

- $Sensibilidad := \frac{VP}{VP+FN}$
  VP (verdaderos positivos) y FN (falsos negativos).

La especificidad representa la probabilidad de que una persona sana obtenga un resultado negativo. Esta se define como:

Cuadro 1: Matriz de confusión con posibles resultados de una prueba diagnostico

		Valor real		
		p	***n***	**Total**
Predicción	***p'***	Verdaderos Positivos	Falsos Positivos	P'
	n'	Falsos Negativos	Verdaderos Negativos	N'
	Total	P	N	

- $Especificidad := \frac{VN}{VN+FP}$
  VP (verdaderos negativos) y FN (falsos positivos).

A partir de estos conceptos se puede crear una matriz de confusión donde se reflejen los cuatro posibles resultados de un experimento a partir de P instancias positivas y N instancias negativas. Esta matriz se puede observar en el cuadro 1 de la página 2.

De la matriz en el cuadro 1 se derivan los conceptos, valores predictivos (positivo y negativo) que son hallados por el teorema de Bayes. Estos valores miden la eficacia real de una prueba diagnóstica, los mismos dan la probabilidad de padecer o no una enfermedad una vez conocido el resultado de la prueba. A continuación, se muestra cómo se definen dichos valores.

- Valor predictivo positivo ($PV+$): es la probabilidad de tener la enfermedad si el resultado de la prueba es positivo.
  $PV+ := \frac{VP}{FP+VP}$

- Valor predictivo negativo ($PV-$): es la probabilidad de no tener la enfermedad si el resultado de la prueba es negativo.
  $PV- := \frac{VN}{VN+FN}$

4. Análisis de artículos donde se aplica el teorema de Bayes

En esta sección se presenta un resumen de lo propuesto de varios artículos que utilizan el teorema de Bayes con datos de pruebas diagnostico que para detectan el Covid-19. El objetivo de estos documentos es poder hallar con que probabilidad una persona que da positivo en un examen está realmente enferma y con los pacientes que resultan negativos que probabilidad existe que estas personas realmente no posean la enfermedad.

4.1. Suplemento de precisión de la prueba Covid-19: las matemáticas del teorema de Bayes

En el artículo "Suplemento de precisión de la prueba Covid-19: las matemáticas del teorema de Bayes"disponible en la liga: (https://www.statnews.com/2020/08/20/covid-19-test-accuracy-supplement-the-math-of-bayes-th- eorem/), los autores aplican los valores predictivos (positivo y negativo), variando la prevalencia (P) de la enfermedad.

En el primer caso toman como prevalencia o probabilidad de estar enfermo bajo, con el valor de $P = 1\,\%$ y valores de sensibilidad (SE) $= 80\,\%$ y especificidad (SP) $= 100\,\%$. Dando como resultado $PV+ = 100\,\%$ y un $PV- = 99,8\,\%$. En el segundo caso se mantienen los valores de SE y SP y se varia la prevalencia con $P = 30\,\%$ esto significa que la probabilidad de dar positivo es mayor. Con estos valores se obtiene $PV+ = 100\,\%$ y un $PV- = 92\,\%$. Aquí se puede observar cómo afecta la prevalencia en los valores predictivos.

4.2. Teorema de Bayes y pruebas de Covid-19

En el artículo "Teorema de Bayes y pruebas de Covid-19"disponible en la liga: (https://www.significance-magazine.com/science/660-bayes-theorem-and-covid-19-testing), el autor calcula al igual que en artículo anterior el valor $PV+$ mediante la aplicación del teorema de Bayes para un $SE = 99\,\%$ y un $SP = 99\,\%$. Con valores de $P = (0{,}1\,\%, 1\,\%, 10\,\%)$ en cada caso. Con lo cual se muestra que incluso cuando se utiliza una prueba muy sensible $SE = 99\,\%$ cuanto menor sea la prevalencia, más probabilidades tenemos de obtener falsos positivos. También se puede demostrar que cuanto mayor sea prevalencia, más probabilidades tenemos de obtener falsos negativos. Esto significa que la calidad de las pruebas de Covid-19 depende de la magnitud del brote.

4.3. COVID-19, teorema de Bayes y toma de decisiones probabilísticas

Este articulo disponible en la liga: (https://towardsdatascience.com/covid-19-bayes-theorem-and-taking-data-driven-decisions-part-1-b61e2c2b3bea), plantea la realización de pruebas de personas al azar en lugares con prevalencia de la enfermedad muy pequeña (0.0001) no resulta útil en la detección del Covid-19. Además, plantea que en estos casos sería mejor hacerlo por grupos de muestras.

4.4. Interpretación de los resultados de la prueba COVID-19: un enfoque bayesiano y Teorema de Bayes, COVID19 y pruebas de detección

En los artículos "Interpretación de los resultados de la prueba COVID-19: un enfoque bayesianoz "Teorema de Bayes, COVID19 y pruebas de detección"disponibles en las ligas:(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7269418/) y (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7315940/) respectivamente, llegan a una conclusión similar planteando que las acciones y recomendaciones posteriores al diagnóstico se basan en gran medida en las probabilidades posteriores a la prueba, no en los resultados categóricos de "positivo.º "negativo". Por lo que un resultado negativo de una prueba sin conocer el $VP-$ de un individuo puede limitar la capacidad de los médicos para realizar las siguientes acciones y disposiciones apropiadas. Para todas las pruebas de detección, ya sea para COVID19 u otros diagnósticos, la comprensión de los valores predictivos y las razones de probabilidad con la ayuda del teorema de Bayes garantizará una interpretación sólida y las recomendaciones y acciones resultantes por parte del personal de salud y las partes interesadas. Un resultado de prueba negativo, en este paradigma, nunca es absolutamente negativo. Más bien, ajusta la probabilidad previa a la prueba de tener una enfermedad más baja.

5. Aplicación del teorema de Bayes a datos de pruebas de Covid-19 en México

Como se pudo apreciar en las secciones anteriores en lo que a las pruebas diagnósticos se refiere, no solo interesa si el resultado es "positivo.º "negativo", además hay que tener en cuenta los valores predictivos (positivo y negativo). Estos últimos nos van a dar la probabilidad de un diagnóstico certero y los pasos a seguir del personal de salud a la hora de descartar o no la enfermedad en los pacientes.

5.1. PCR

En México durante el enfrentamiento a la pandemia del Covid-19 la prueba diagnostico más utilizada es la prueba de detección de ácidos nucleicos: reacción en cadena de la polimerasa (PCR). El PCR tiene las siguientes características [2]:

- Especificidad, próxima al 100 %.
- Sensibilidad variable dependiendo del momento del proceso infeccioso, es decir, de la carga viral, y del lugar de toma de la muestra. Entre cero y siete días tras el comienzo de la enfermedad, las sensibilidades para pacientes leves como severos fueron:
  - Esputo: 89 %
  - Nasal: 73 %
  - Oro-faringe: 60 %

5.2. Datos utilizados

Para el cálculo de $VP+$ y $VP-$ de pruebas diagnósticos PCR en Mexico, se utilizan datos abiertos proporcionados por el gobierno de México y disponibles en [1], en la utilización del 16 de octubre del 2020. De dichos datos se extraen los siguientes valores:

- Total de pruebas realizadas $(TP) = 1{,}903{,}285$.
- Total de pruebas positivas $(T+) = 837{,}445$.
- Total de pruebas negativas $(T-) = 1{,}065{,}840$.
- Tasa positiva $(P) = (0.30\text{–}0.50)$

Con los valores de $SE = (60\,\%, 73\,\%, 89\,\%)$, $SP = 99,9\,\%$ del PCR y los valores TP, $T+$, $T-$ se crean matrices de confusión, para la obtención de los valores a emplear en teorema de Bayes, en el cuadro 2 de la página 5 se muestra un ejemplo de matriz de confección para $SE = 60\,\%$. En el cuadro 3 de la página 5 se puede observar los resultados obtenidos del cálculo de los valores predictivos para cada una de las variantes. Como prevalencia de la enfermedad se utiliza el valor la tasa positiva que es una buena aproximación a la probabilidad previa a la prueba por lo que se utilizan valores de $P = (0{,}30, 0{,}50)$ que son los valores por lo que se mueve la tasa positiva en México.

De los resultados obtenidos se puede concluir que el personal médico debe tener mayor cuidado con los valores negativos de las pruebas de PCR, dado que los $VP-$ nos muestran que aumentan las

Cuadro 2: Matriz de confusión para $SE = 60\,\%$

		Valor real		
		p	***n***	**Total**
Predicción	***p'***	502,467.00	1,065.00	P'
	n'	334,978.00	1,064,774.00	N'
	Total	P	N	

Cuadro 3: Resultados de la aplicación del teorema de Bayes

SE(%)	***SP(%)***	***P***	***VP+(%)***	***VP-(%)***
60.00	99.90	0.3	99.61	85.35
		0.5	99.83	71.41
73.00	99.90	0.3	99.68	89.61
		0.5	99.86	78.72
89.00	99.90	0.3	99.73	95.49
		0.5	99.88	90.08

probabilidades de falsos negativos cuando la sensibilidad es baja y la prevalencia aumenta, estos son parámetros importantes a tener en cuenta a la hora de emitir el diagnostico final.

Referencias

[1] Esteban Ortiz-Ospina Max Roser, Hannah Ritchie and Joe Hasell. Coronavirus pandemic (covid-19). *Our World in Data*, 2020. https://ourworldindata.org/coronavirus.

[2] Yang Yang, Minghui Yang, Chenguang Shen, Fuxiang Wang, Jing Yuan, Jinxiu Li, Mingxia Zhang, Zhaoqin Wang, Li Xing, Jinli Wei, Ling Peng, Gary Wong, Haixia Zheng, Mingfeng Liao, Kai Feng, Jianming Li, Qianting Yang, Juanjuan Zhao, Zheng Zhang, Lei Liu, and Yingxia Liu. Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-ncov infections. *medRxiv*, 2020.

Tarea 9 de Modelos Probabilistas Aplicados

Ejercicios

5271

3 de noviembre de 2020

1. Introducción

En este documento se presentan los resultados de varios ejercicios del libro "*Introduction to Probability*"[1].

2. Ejercicio 1 de la 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?*

Respuesta: Definimos la variable aleatoria X como, dólar que se pierde o gana el jugador. Por lo que:

$$P(X = 1) = P(3, 5, 7, 9) = \frac{4}{9}. \tag{1}$$

$$P(X = -1) = 1 - P(3, 5, 7, 9) = 1 - \frac{4}{9} = \frac{5}{9}. \tag{2}$$

Ahora para variables aleatorias discretas el valor esperado es:

$$E[X] = \sum_{x \in \Omega} x \times P(X = x). \tag{3}$$

Sustituyendo los resultados de la ecuación 1 y 2 en la ecuación 3 se tiene:

$$\begin{aligned} E[X] &= 1 \times \tfrac{4}{9} - 1 \times \tfrac{5}{9} \\ E[X] &= -\tfrac{1}{9}. \end{aligned} \tag{4}$$

Por tanto, el valor esperado de las ganancias para el jugador es -1/9 de dólar.

3. Ejercicio 6 de la página 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?

Respuesta: Sea $X = a + b$ donde a es el valor del primer tiro y b el segundo, se tiene: $X = (2, 3, 4, 5, 6, 7, 3, 4, 5, 6, 7, 8, 4, 5, 6, 7, 8, 9, 5, 6, 7, 8, 9, 10, 6, 7, 8, 9, 10, 11, 7, 8, 9, 10, 11, 12)$. La frecuencia de los valores de X se muestran en el cuadro 1 de la página 2.

Cuadro 1: Frecuencia de valores de X

Posibles valores de x	**frecuencia** (x)	$P(X = x)$
2	1	0.028
3	2	0.056
4	3	0.083
5	4	0.111
6	5	0.139
7	6	0.167
8	5	0.139
9	4	0.111
10	3	0.083
11	2	0.056
12	1	0.028

Del cuadro 1 podemos obtener que:

$$P(X = x_i) = \frac{\text{frecuencia de } x_i}{36}. \tag{5}$$

Sustituyendo en la ecuación 3

$$\begin{aligned} E[X] &= \textstyle\sum_{i=1}^{11} x_i \times P(X = x_i) \\ E[X] &= 7 \end{aligned} \tag{6}$$

Análogamente, sea $Y = a - b$. De Y se obtiene los valores que muestra el cuadro 2 de la página 3.

Cuadro 2: Frecuencia de valores de X

Posibles valores de *x*	**frecuencia *(x)***	***P(X = x)***
-5	1	0.028
-4	2	0.056
-3	3	0.083
-2	4	0.111
-1	5	0.139
0	6	0.167
1	5	0.139
2	4	0.111
3	3	0.083
4	2	0.056
5	1	0.028

Sustituyendo en la ecuación 3

$$\begin{aligned} E[X] &= \sum_{i=1}^{11} x_i \times P(X = x_i) \\ E[X] &= 0 \end{aligned} \tag{7}$$

Dado que $E(XY) = E((a+b)(a-b))$, tenemos para XY los valores que se muestran en el cuadro 3 de la página 3.

Cuadro 3: Fragmento de valores de frecuencia de XY

Posibles valores de *xy*	**frecuencia *(xy)***	***P(XY = xy)***
-35	1	0.028
-32	1	0.028
-9	1	0.028
-8	1	0.028
-7	1	0.028
-5	1	0.028
-3	1	0.028
0	6	0.167
3	1	0.028
5	1	0.028
7	1	0.028
8	1	0.028
9	1	0.028
32	1	0.028
35	1	0.028

Sustituyendo en la ecuación 3

$$\begin{aligned} E[XY] &= \sum_{i=1}^{31} xy_i \times P(XY = xy_i) \\ E[XY] &= 0 \end{aligned} \tag{8}$$

Ahora para comprobar si X e Y son variables aleatorias independientes, hay que probar que se cumple para todos los casos se cumpla que $P(X, Y) = P(X)P(Y)$. Por lo anteriormente mencionado, con encontrar un caso donde no se cumpla la propiedad para decir que las variables aleatorias no son independientes. Como es el caso:

$$P(X = 12, Y = 0) = P(a = 6, b = 6) = \frac{1}{36}. \tag{9}$$

$$P(X = 12)P(Y = 0) = \frac{1}{36} \times \frac{1}{6} = \frac{1}{216}. \tag{10}$$

Por tanto $P(X = 12, Y = 0) \neq P(X = 12)P(Y = 0)$, lo que implica que X e Y no son variables aleatorias independientes.

4. Ejercicio 15 de la 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.

Respuesta: Sea la bola dorada (G) y la bola plateada (S), dado que el juego se detiene cuando se está un dólar arriba o se acaban las bolas doradas. La variable aleatoria X viene dada por las posibles maneras de jugar (ver cuadro 4).

Cuadro 4: Orden de selección de la bolas, ganancias y probabilidades respectivas

Orden Selección	**Ganancias**	$\boldsymbol{P(X = x)}$
G	1	2/5
SGG	1	1/10
SGSG	0	1/10
SSGG	0	1/10
SSSGG	-1	1/10
SSGSG	-1	1/10
SGSSG	-1	1/10

Ahora sustituyendo en la ecuación 3 se tiene:

$$\begin{aligned} E[X] &= 1 \times P(G) + 1 \times P(SSG) + 0 \times P(SGSG) + 0 \times P(SSGG) \\ &\quad -1 \times P(SSSGG) - 1 \times P(SSGSG) - 1 \times P(SGSSG) \\ &= 1 \times (\tfrac{2}{5}) + 1 \times (\tfrac{1}{10}) - 1 \times (\tfrac{1}{10}) - 1 \times (\tfrac{1}{10}) - 1 \times (\tfrac{1}{10}) \\ E[X] &= \tfrac{1}{5}. \end{aligned} \tag{11}$$

Dado que la $E[X]$, es mayor que cero se puede concluir que es un juego favorable.

5. Ejercicio 18 de la 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?

Respuesta: La variable aleatoria X viene dada por el número de intentos fallidos antes de un éxito. Por lo que se tienen los valores en el cuadro 5 de la página 5.

Cuadro 5: Probabilidades de ocurrencia de x

# de intentos (x)	$\boldsymbol{P(X = x)}$
0	$1/6$
1	$(5/6) \times (1/5) = 1/6$
2	$(5/6) \times (4/5) \times (1/4) = 1/6$
3	$(5/6) \times (4/5) \times (3/4) \times (1/3) = 1/6$
4	$(5/6) \times (4/5) \times (3/4) \times (2/3) \times (1/2) = 1/6$
5	$(5/6) \times (4/5) \times (3/4) \times (2/3) \times (1/2) \times 1 = 1/6$

Entonces se sustituye en la ecuación 3:

$$\begin{aligned} E[X] &= 0 \times \left(\tfrac{1}{6}\right) + 1 \times \left(\tfrac{1}{6}\right) + 2 \times \left(\tfrac{1}{6}\right) + 3 \times \left(\tfrac{1}{6}\right) + 4 \times \left(\tfrac{1}{6}\right) + 5 \times \left(\tfrac{1}{6}\right) \\ E[X] &= \left(\tfrac{5}{2}\right). \end{aligned} \tag{12}$$

Por lo que el número esperado de intentos antes de abrir la puerta es 2.5.

6. Ejercicio 19 de la página 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.

Respuesta: Sea la puntuación obtenida en la elección del subconjunto de posibles respuestas una variable aleatoria X, se tiene que el espacio muestral está conformado por dieciséis formas de escoger un subconjunto. En el cuadro 6 de la página 6 se muestra la cardinalidad de los posibles subconjuntos, así como los posibles valores de x y las probabilidades asociadas a dichos valores.

Cuadro 6: Cardinalidad de los subconjunto con sus posibles valores y probabilidades

Elementos de Subconjuntos	**Posibles Valores de x**	$P(X = x)$
0	0	1/16
1	3	1/16
	-1	3/16
2	2	3/16
	-2	3/16
3	1	3/16
	-3	1/16
4	0	1/16

Sustituyendo en la ecuación 3:

$$\begin{aligned} E[X] &= 3 \times \left(\tfrac{1}{16}\right) - 1 \times \left(\tfrac{3}{16}\right) + 2 \times \left(\tfrac{3}{16}\right) - 2 \times \left(\tfrac{3}{16}\right) + 1 \times \left(\tfrac{3}{16}\right) - 3 \times \left(\tfrac{1}{16}\right) \\ E[X] &= 0. \end{aligned} \tag{13}$$

Entonces si solo adivina un subconjunto de manera uniforme y aleatoria, su puntuación esperada es cero.

7. Ejercicio 1 de la página 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.

Respuesta: Sustituyendo en la ecuación 3 se tiene:

$$\begin{aligned} E[X] &= -1 \times \left(\tfrac{1}{3}\right) + 0 \times \left(\tfrac{1}{3}\right) + 1 \times \left(\tfrac{1}{3}\right) \\ E[X] &= 0. \end{aligned} \tag{14}$$

Para calcular la varianza se tiene la expresión $\sigma^2(X) = E(X^2) - E(X)^2$ por lo que:

$$\begin{aligned} \sigma^2 &= (-1)^2 \times \left(\tfrac{1}{3}\right) + (0)^2 \times \left(\tfrac{1}{3}\right) + (1)^2 \times \left(\tfrac{1}{3}\right) - (0)^2 \\ \sigma^2 &= \tfrac{2}{3}. \end{aligned} \tag{15}$$

La desviación estándar no se calcula hallando $\sqrt{\sigma^2}$ por lo que $\sigma = \sqrt{\frac{2}{3}}$.

8. Ejercicio 9 de la página 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 $D(X)$.

Respuesta: Ya que el dado está cargado y la probabilidad de cada cara es proporcional al valor de la cara correspondiente. Es decir que la $P(X = x)$ es proporcional a x, y se define un factor de proporcionalidad c, así que $P(X = x) = xc$. Por lo que:

$$\begin{aligned} 1 &= c + 2c + 3c + 4c + 5c + 6c \\ c &= \tfrac{1}{21}. \end{aligned} \tag{16}$$

En el cuadro 7 de la página 7 se muestran las probabilidades de cada cara para la constante de proporcionalidad $c = \frac{1}{21}$

Cuadro 7: Probabilidades de cada cara del dado cargado

Valores de x	***P(X = x)***	***P(X =x) para c = 1/21***
1	1c	1/21
2	2c	2/21
3	3c	3/21
4	4c	4/21
5	5c	5/21
6	6c	6/21

Con los valores de probabilidad de cada se procede a calcular $E(X)$:

$$\begin{aligned} E[X] &= 1 \times \left(\tfrac{1}{21}\right) + 2 \times \left(\tfrac{2}{21}\right) + 3 \times \left(\tfrac{3}{21}\right) + 4 \times \left(\tfrac{4}{21}\right) + 5 \times \left(\tfrac{5}{21}\right) + 6 \times \left(\tfrac{6}{21}\right) \\ E[X] &= \tfrac{91}{21} = \tfrac{13}{3}. \end{aligned} \tag{17}$$

Calculando σ^2

$$\begin{aligned} \sigma^2 &= (1)^2 \times \left(\tfrac{1}{21}\right) + (2)^2 \times \left(\tfrac{2}{21}\right) + (3)^2 \times \left(\tfrac{3}{21}\right) + (4)^2 \times \left(\tfrac{4}{21}\right) + (5)^2 \times \left(\tfrac{5}{21}\right) + (6)^2 \times \left(\tfrac{6}{21}\right) - (\tfrac{13}{3})^2 \\ \sigma^2 &= \tfrac{20}{9}. \end{aligned} \tag{18}$$

Por tanto $\sigma = \sqrt{\frac{20}{9}}$.

9. Ejercicio 12 de la página 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.

Respuesta: Teniendo en cuenta las propiedades siguientes de la esperanza (ecuación 19 y 20 y la definición de σ^2(ecuación 21) :

$$E(cX) = cE(X). \tag{19}$$

$$E(X + Y) = E(X) + E(Y). \tag{20}$$

$$\sigma^2 = E(X - \mu^2). \tag{21}$$

Se calcula:

$$\begin{aligned} E[X^*] &= E\left(\frac{X-\mu}{\sigma}\right) \\ &= \frac{1}{\sigma}E(X-\mu) \\ &= \frac{1}{\sigma}\left[E(X)-E(\mu)\right] \\ &= \frac{1}{\sigma}\times 0 \\ E[X^*] &= 0. \end{aligned} \tag{22}$$

Ya con el valor de $E(X^*)$ se cálcala el valor σ^2

$$\begin{aligned} \sigma^2 &= E(X^{*2}) - 0^2 \\ &= E\left(\frac{X-\mu}{\sigma}\right)^2 \\ &= \frac{1}{\sigma^2}E(X-\mu)^2 \\ &= \frac{1}{\sigma^2}\times\sigma^2 \\ \sigma^2 &= 1. \end{aligned} \tag{23}$$

10. Ejercicio 3 de la página 278

The lifetime, measure in hours, of the ACME super light bulb is a random variable T with density function $f_T(t) = \lambda^2 te^{-\lambda t}$, where $\lambda = 0{,}05$. What is the expected lifetime of this light bulb? What is its variance?

Respuesta: la expresión para la esperanza de vida útil de la bombilla esta dada por:

$$\begin{aligned} E(T) &= \int_0^\infty t\lambda^2 te^{-\lambda t}dt \\ &= \lambda^2\left(\frac{-t^2e^{-\lambda t}}{\lambda} - \int\frac{-2te^{-\lambda t}}{\lambda}dt\right) \\ &= \lambda^2\left[\frac{-t^2e^{-\lambda t}}{\lambda} + \frac{2}{\lambda}\left(-\frac{te^{-\lambda t}}{\lambda} + \frac{1}{\lambda}\left[\frac{-e^{-\lambda t}}{\lambda}\right]\right)\right] \\ &= \left[\frac{\left(-t^2\lambda^2-2t\lambda-2\right)e^{-\lambda t}}{\lambda}\right]_0^\infty \\ &= \frac{2}{\lambda} = \frac{2}{0{,}05} \\ E(T) &= 40. \end{aligned} \tag{24}$$

Calculando $\sigma^2(T)$:

$$
\begin{aligned}
\sigma^2(T) &= E(T^2 - E(T)^2) \\
\sigma^2(T) &= (\int_0^\infty t^3\lambda^2 e^{-\lambda t}dt) - 40^2 \\
&= \lambda^2\left[\frac{-t^3e^{-\lambda t}}{\lambda} + \frac{3}{\lambda}\left(-\frac{t^2e^{-\lambda t}}{\lambda} - \frac{2te^{-\lambda t}}{\lambda^2} - \frac{2e^{\lambda t}}{\lambda^3}\right)\right] - 1600 \\
&= \left[\frac{\left(-t^3\lambda^3 - 3t^2\lambda^2 - 6t\lambda - 6\right)e^{-\lambda t}}{\lambda^2}\right]_0^\infty - 1600 \\
&= \frac{6}{\lambda^2} - 1600 \\
&= 2400 - 1600 \\
\sigma^2(T) &= 800.
\end{aligned}
\tag{25}
$$

Referencias

[1] Grinstead, Charles M., Snell, J. Laurie. *Introduction to Probability.* American Mathematical Society, 2006.

Tarea 10 de Modelos Probabilistas Aplicados

Ejercicios

5271

10 de noviembre de 2020

1. Introducción

En este documento se presentan los resultados de varios ejercicios del libro "*Introduction to Probability*"[1] encontrados de forma analítica, así como los resultados alcanzados numéricamente mediante simulación.

2. Valor esperado

Ahora para variables aleatorias discretas el valor esperado es:

$$E[X] = \sum_{x \in \Omega} x \times P(X = x). \tag{1}$$

3. Ejercicio 6 de la página 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?

3.1. $X = a + b$

Sea $X = a + b$ donde a es el valor del primer tiro y b el segundo, se tiene: $X = 2$–12. La frecuencia de los valores de X se muestran en el cuadro 1 de la página 2.

Cuadro 1: Frecuencia de valores de X

Posibles valores de *x*	**frecuencia *(x)***	***P(X = x)***
2	1	0.028
3	2	0.056
4	3	0.083
5	4	0.111
6	5	0.139
7	6	0.167
8	5	0.139
9	4	0.111
10	3	0.083
11	2	0.056
12	1	0.028

Del cuadro 1 podemos obtener que:

$$P(X = x_i) = \frac{\text{frecuencia de } x_i}{36}. \tag{2}$$

Sustituyendo en la ecuación 1

$$\begin{aligned} E[X] &= \textstyle\sum_{i=1}^{11} x_i \times P(X = x_i) \\ E[X] &= 7. \end{aligned} \tag{3}$$

Para la comprobación y mejor entendimiento del resultado obtenido se realiza una simulación de la variable aleatoria $X = a + b$, la misma es realizada en R [2] como se muestra en el código 3.1 a continuación.

```
restas = numeric()
sumas = numeric()
mult = numeric()
val = c(1:6)
a = 0
b = 0

for (i in c(1:1000)){
  a = sample(val,1)
  b = sample(val,1)
  r = a-b
  s = a+b
  restas= c(restas,r)
  sumas= c(sumas,s)
  mult= c(mult,r*s)
}
```

Tarea10.R

Los valores para X son almacenados en un *dataframe*, como se muestra en el cuadro 2 de la página 3, al aplicarle la función *summary* de R al *dataframe* se obtiene los siguientes valores.

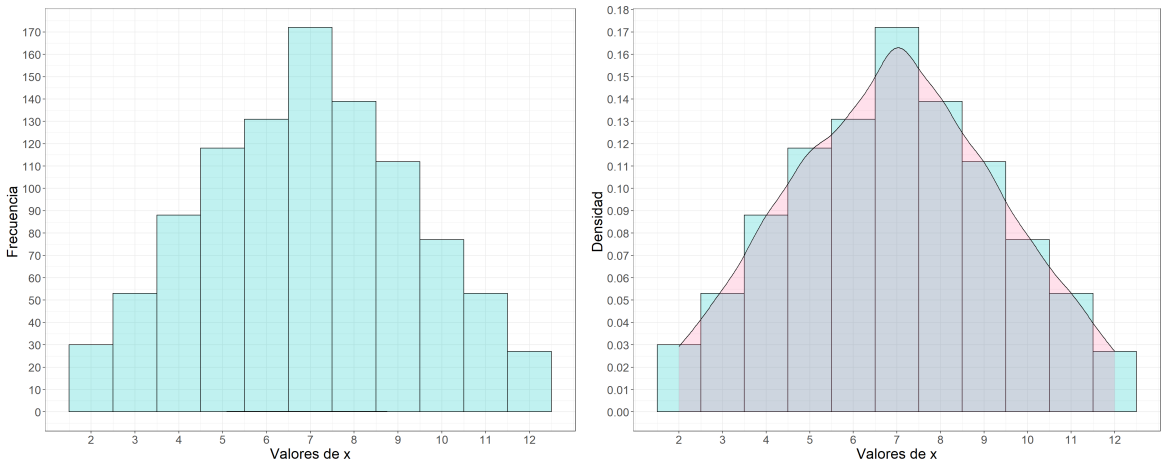
summary.txt

```
> summarydfs$x
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   5.000   7.000   6.948   9.000  12.000
```

En los valores obtenidos se puede observar que la media es igual a 6.848 por lo que se puede decir que el valor promedio obtenido es igual a la $E[X]$, además coincide con la mediana que es el valor con mayor frecuencia de ocurrencia de X como se muestra en la figura 4 de la página 9. Para mejor visualización del experimento se puede ver la figura $X = (a + b).gif$, donde se observa el progreso de los lanzamientos.

Cuadro 2: fragmento de *dataframe* de la variable aleatoria $X = a + b$

Lanzamientos	x
1	6
2	11
3	6
4	9
5	4
161	7
162	3
163	6
164	7
165	8
996	6
997	9
998	11
999	11
1000	5



(a) Histograma de frecuencia de la variable aleatoria X (b) Histograma de densidad de la variable aleatoria X

Figura 1: Histogramas de la variable aleatoria $X = a + b$

3.2. $Y = a - b$

Análogamente, sea $Y = a - b$. De Y se obtiene los valores que muestra el cuadro 3 de la página 4.

Cuadro 3: Frecuencia de valores de X

Posibles valores de x	**frecuencia** (x)	$P(X = x)$
-5	1	0.028
-4	2	0.056
-3	3	0.083
-2	4	0.111
-1	5	0.139
0	6	0.167
1	5	0.139
2	4	0.111
3	3	0.083
4	2	0.056
5	1	0.028

Sustituyendo en la ecuación 1

$$\begin{aligned} E[X] &= \sum_{i=1}^{11} x_i \times P(X = x_i) \\ E[X] &= 0. \end{aligned} \tag{4}$$

Al comprobar con la simulación de la variable aleatoria $X = a - b$, los valores para X son almacenados en un *dataframe*, como se muestra en el cuadro 4 de la página 5, al aplicarle la función *summary* de R al *dataframe* se obtiene los siguientes valores.

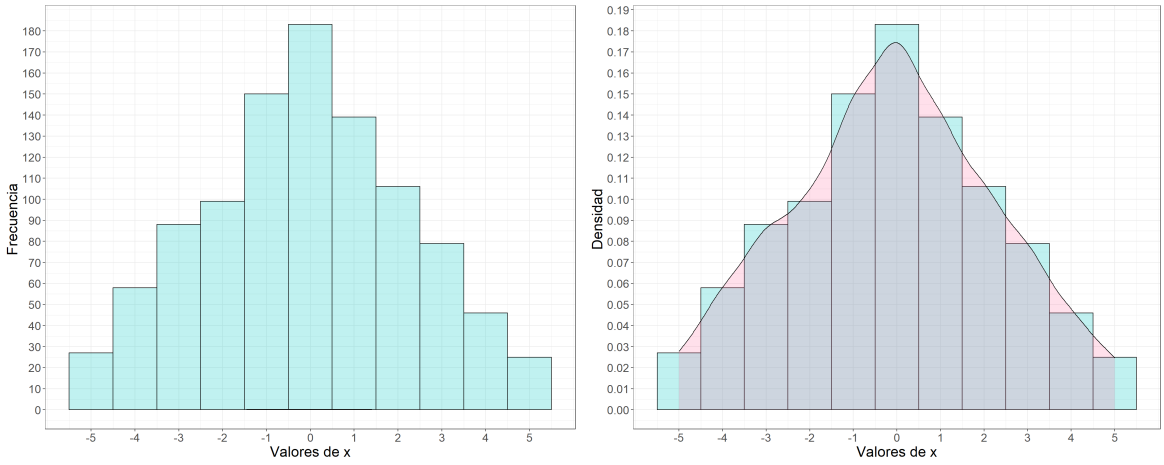
summary.txt

```
> summarydfr$x
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -5.000  -2.000   0.000  -0.082   2.000   5.000
```

Cuadro 4: Fragmento de *dataframe* de la variable aleatoria Y = a - b

Lanzamientos	y
1	-4
2	-1
3	-2
4	-3
5	2
201	-3
202	0
203	1
204	0
205	5
996	2
997	-1
998	1
999	-1
1000	-1

En los valores obtenidos se puede observar que la media es igual a -0.082 por lo que se puede decir que el valor promedio obtenido es igual a la $E[Y]$, además coincide con la mediana que es el valor con mayor frecuencia de ocurrencia de Y como se muestra en la figura 3 de la página 7. Para mejor visualización del experimento se puede ver la figura $Y = (a - b)$.gif, donde se observa el progreso de los lanzamientos.



(a) Histograma de frecuencia de la variable aleatoria Y (b) Histograma de densidad de la variable aleatoria Y

Figura 2: Histogramas de la variable aleatoria $Y = a - b$

3.3. $E(XY) = E((a + b)(a - b))$

Dado que $E(XY) = E((a + b)(a - b))$, tenemos para XY los valores que se muestran en el cuadro 5 de la página 6.

Cuadro 5: Fragmento de valores de frecuencia de XY

Posibles valores de *xy*	**frecuencia *(xy)***	***P(XY = xy)***
-35	1	0.028
-32	1	0.028
-9	1	0.028
-8	1	0.028
-7	1	0.028
-5	1	0.028
-3	1	0.028
0	6	0.167
3	1	0.028
5	1	0.028
7	1	0.028
8	1	0.028
9	1	0.028
32	1	0.028
35	1	0.028

Sustituyendo en la ecuación 1

$$\begin{aligned} E[XY] &= \sum_{i=1}^{31} xy_i \times P(XY = xy_i) \\ E[XY] &= 0 \end{aligned} \tag{5}$$

Ahora para comprobar si X e Y son variables aleatorias independientes, hay que probar que se cumple para todos los casos se cumpla que $P(X, Y) = P(X)P(Y)$. Por lo anteriormente mencionado, con encontrar un caso donde no se cumpla la propiedad para decir que las variables aleatorias no son independientes. Como es el caso:

$$P(X = 12, Y = 0) = P(a = 6, b = 6) = \frac{1}{36}. \tag{6}$$

$$P(X = 12)P(Y = 0) = \frac{1}{36} \times \frac{1}{6} = \frac{1}{216}. \tag{7}$$

Por tanto $P(X = 12, Y = 0) \neq P(X = 12)P(Y = 0)$, lo que implica que X e Y no son variables aleatorias independientes.

Al realizar la simulación, los valores para la variable aleatoria $XY = (a + b)(a - b)$, los valores para XY son almacenados en un *dataframe*, como se muestra en el cuadro 6 de la página 7, al aplicarle la función *summary* de R al *dataframe* se obtiene los siguientes valores.

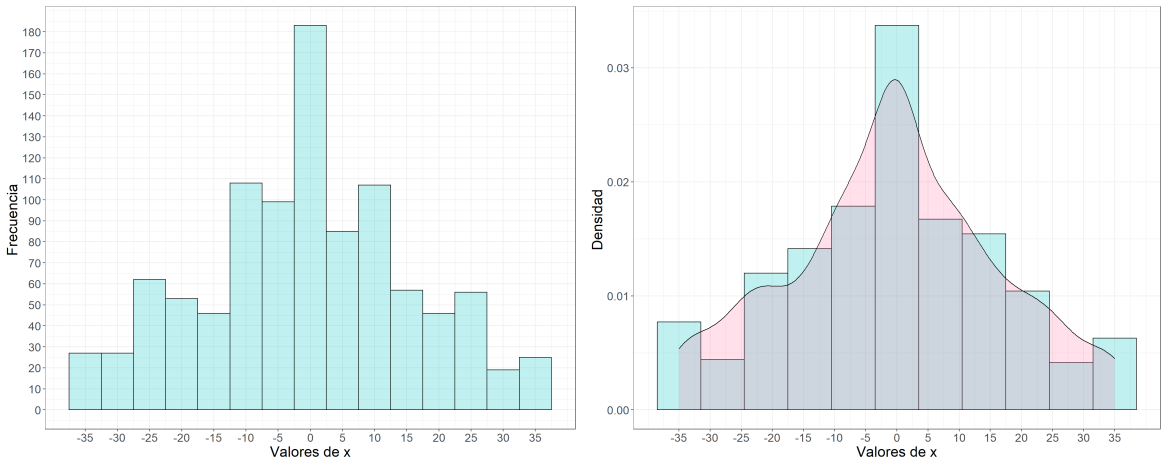
summary.txt

```
summarydfm$x
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-35.000 -11.000   0.000  -0.504  11.000  35.000
```

Cuadro 6: Fragmento de *dataframe* de la variable aleatoria $XY = (a+b)(a-b)$

Lanzamientos	xy
1	-24
2	-11
3	-12
4	-27
5	8
614	0
615	-15
616	0
617	-24
618	9
996	12
997	-9
998	11
999	-11
1000	-5

En los valores obtenidos se puede observar que la media es igual a -0.504 por lo que se puede decir que el valor promedio obtenido es similar a la $E[XY]$, además la $E[XY]$ coincide con la mediana que es el valor con mayor frecuencia de ocurrencia de XY como se muestra en la figura 3 de la página 7. Para mejor visualización del experimento se puede ver la figura $XY = (a+b)(a-b)$.gif, donde se observa el progreso de los lanzamientos.



(a) Histograma de frecuencia de la variable aleatoria XY (b) Histograma de densidad de la variable aleatoria XY

Figura 3: Histogramas de la variable aleatoria $XY = (a+b)(a-b)$

Mediante la simulación se pudo constatar que los los resultados para la esperanza

4. Ejercicio 18 de la 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?

Respuesta: La variable aleatoria X viene dada por el número de intentos fallidos antes de un éxito. Por lo que se tienen los valores en el cuadro 7 de la página 8.

Cuadro 7: Probabilidades de ocurrencia de x

# de intentos (x)	$\boldsymbol{P(X = x)}$
0	$1/6$
1	$(5/6) \times (1/5) = 1/6$
2	$(5/6) \times (4/5) \times (1/4) = 1/6$
3	$(5/6) \times (4/5) \times (3/4) \times (1/3) = 1/6$
4	$(5/6) \times (4/5) \times (3/4) \times (2/3) \times (1/2) = 1/6$
5	$(5/6) \times (4/5) \times (3/4) \times (2/3) \times (1/2) \times 1 = 1/6$

Entonces se sustituye en la ecuación 1:

$$\begin{aligned} E[X] &= 0 \times \left(\tfrac{1}{6}\right) + 1 \times \left(\tfrac{1}{6}\right) + 2 \times \left(\tfrac{1}{6}\right) + 3 \times \left(\tfrac{1}{6}\right) + 4 \times \left(\tfrac{1}{6}\right) + 5 \times \left(\tfrac{1}{6}\right) \\ E[X] &= \left(\tfrac{5}{2}\right). \end{aligned} \tag{8}$$

Por lo que el número esperado de intentos antes de abrir la puerta es 2.5.

Para la comprobación y mejor entendimiento del resultado obtenido se realiza una simulación de la variable aleatoria X, la misma es realizada por el código 5 a continuación.

```
llaves = c(1:6)
correcta= 2
inten = 0
intentos= numeric()
for (k in c(1:1000)){
  for (j in c(1:6)){
    int = sample(llaves,1)
    if(int ==correcta){
    inten= j-1
      }
  }
  intentos= c(intentos,inten)
}
```

Tarea10.R

Los valores para X son almacenados en un *dataframe*, como se muestra en el cuadro **??** de la página **??**, al aplicarle la función *summary* de R al *dataframe* se obtiene los siguientes valores.

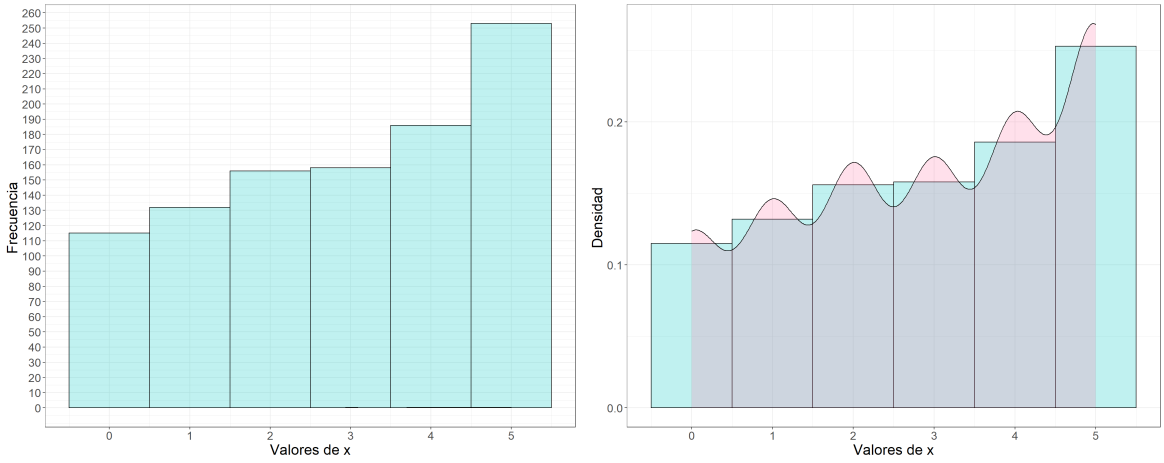
summary.txt

```
> summaryintentos
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.000   3.000   2.927   5.000   5.000
```

Cuadro 8: Fragmento de *dataframe* con valores de X obtenido en los experimentos

Experimentos	X
1	3
2	5
3	3
4	3
5	5
177	5
178	0
179	0
180	1
181	1
996	3
997	3
998	2
999	4
1000	5

En los valores obtenidos se puede observar que la media es igual a 2.927 por lo que se puede decir que el valor promedio obtenido no es igual a la $E[X]$, además no coincide con la mediana ni con el valor de valor con mayor frecuencia de ocurrencia de X que es cinco intentos. Lo anterior se muestra en la figura 4 de la página 9.



(a) Histograma de frecuencia de la variable aleatoria X (b) Histograma de densidad de la variable aleatoria X

Figura 4: Histogramas de la variable aleatoria X

Con los valores numéricos obtenidos en la simulación se llega a una contradicción a los resultados obtenidos de forma analítica.

5. Ejercicio 19 de la página 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.

Respuesta: Sea la puntuación obtenida en la elección del subconjunto de posibles respuestas una variable aleatoria X, se tiene que el espacio muestral está conformado por dieciséis formas de escoger un subconjunto. En el cuadro 9 de la página 10 se muestra la cardinalidad de los posibles subconjuntos, así como los posibles valores de x y las probabilidades asociadas a dichos valores.

Cuadro 9: Cardinalidad de los subconjunto con sus posibles valores y probabilidades

Elementos de Subconjuntos	**Posibles Valores de x**	$P(X = x)$
0	0	1/16
1	3	1/16
	-1	3/16
2	2	3/16
	-2	3/16
3	1	3/16
	-3	1/16
4	0	1/16

Sustituyendo en la ecuación 1:

$$\begin{aligned} E[X] &= 3 \times \left(\tfrac{1}{16}\right) - 1 \times \left(\tfrac{3}{16}\right) + 2 \times \left(\tfrac{3}{16}\right) - 2 \times \left(\tfrac{3}{16}\right) + 1 \times \left(\tfrac{3}{16}\right) - 3 \times \left(\tfrac{1}{16}\right) \\ E[X] &= 0. \end{aligned} \tag{9}$$

Entonces si solo adivina un subconjunto de manera uniforme y aleatoria, su puntuación esperada es cero.

Para la comprobación y mejor entendimiento del resultado obtenido se realiza una simulación de la variable aleatoria X, la misma es realizada por el código 5 a continuación.

```
resp = c(0,0,1,0)
dos =numeric()
esperado = numeric()
for (k in c(1:1000)){
  sub = numeric()
  uno=sample(resp,1)
  if(uno == 1){
    sub=c(sub, 3)
  }else {
    sub=c(sub, -1)
  }
```

```
  d=sample(resp,2)
  dos=c(d)
  for (i in c(1:2)) {
    if(dos[i] == 1){
      sub=c(sub, 2)
    }else {
      sub=c(sub, -2)
    }
  }
  t=sample(resp,3)
  tres=c(t)
  for (i in c(1:3)) {
    if(tres[i] == 1){
      sub=c(sub, 1)
    }else {
      sub=c(sub, -3)
    }
  }

  esperado= c(esperado,sub)
}
```

Tarea10.R

Los valores para X son almacenados en un *dataframe*, como se muestra en el cuadro **??** de la página **??**, al aplicarle la función *summary* de R al *dataframe* se obtiene los siguientes valores.

summary.txt

```
summaryesperado
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -3.000  -3.000  -2.000  -1.335  -1.000   3.000
```

Cuadro 10: Fragmento

Experimentos	x
1	3
2	-2
3	-2
4	-3
5	-3
6	-3
7	3
8	-2
9	-2
10	1
5996	-2
5997	-2
5998	-3
5999	-3
6000	1

En los valores obtenidos se puede observar que la media es igual a -2 por lo que se puede decir que el

valor promedio es negativo por lo que se puede decir que va a obtener cero en el examen. El código general se encuentra disponible en el repositorio. https://github.com/Albertomnoa/Tareas

Referencias

[1] Grinstead, Charles M., Snell, J. Laurie. *Introduction to Probability.* American Mathematical Society, 2006.

[2] R Core Team. R: Un lenguaje y un entorno para la informática estadística, 2020.

Tarea 11 de Modelos Probabilistas Aplicados

Convolución, χ^2, covarianza

5271

17 de noviembre de 2020

1. Introducción

En este documento se presentan los resultados del análisis de varias propiedades y conceptos de la convolución, χ^2 y covarianza de forma analítica, así como los resultados alcanzados numéricamente mediante simulación. Para la simulación se utiliza lenguaje R [1].

2. Convolución

Una convolución es un operador matemático que transforma dos funciones f_1 y f_2 en una tercera función f_c que representa la magnitud en la que se superponen f_1 y una versión trasladada e invertida de f_2. Para el caso discreto se tiene la expresión:

$$\begin{aligned} f_c &= f_1 * f_2 \\ f_c(i) &= \sum_j f_1(j) \times f_2(i-j) \end{aligned}.$$

Y para el caso continuo se cuenta con la siguiente ecuación:

$$(f * g)(z) = \int_{-\infty}^{\infty} f(z-y) \times g(y) \mathrm{d}y = \int_{-\infty}^{\infty} f(z-x) \times g(x) \mathrm{d}x.$$

2.1. Aplicaciones

La convolución tiene muchas aplicaciones prácticas como son:

- Procesamiento de imágenes.
- Filtrado de señales.
- Multiplicación de polinomios.
- Procesamiento de audio.

- Inteligencia artificial.
- Óptica.
- Teoría de la probabilidad.
- Mercado financiero.

Teoría de la probabilidad: Si se considera una situación en la que se tienen dos variables aleatorias independientes, X e Y, con funciones de densidad de probabilidad f y g respectivamente. Y se desea calcular la función de densidad $(X + Y)$, podemos usar la convolución de f y g. Por lo que se puede calcular la suma de cualquier número de variables independientes. Esto es importante porque muchas de las distribuciones estándar se caracterizan por patrones de convolución simples, lo que significa que podemos encontrar sus funciones de densidad de probabilidad mediante convolución.

Para realizar una prueba numérica de lo antes mencionado se utilizan los valores de rendimiento del oro y del platino obtenidos del sitio *Yahoo!finanzas* en la liga: https://es.finance.yahoo.com/materias-primas. Los valores de rendimientos fueron calculados a partir de los datos obtenidos como el valor de cierre menos el valor de apertura, el mismo es una variable aleatoria. Se pretende calcular la probabilidad conjunta de las variables X (oro) e Y (platino), para lo cual se utiliza la función *convolve* de la librería de R que utiliza la transformada rápida de Fourier para calcular la convolución de dos secuencias. En la figura 1 de la página 3 se puede observar los resultados obtenidos.

En la figura 2.1 y 2.1 de la página 3, muestran un comportamiento similar a la distribución normal, aunque no es tan evidente como en la 2.1 que muestra que la distribución de la convolución de las variables se distribuye normalmente. Lo anterior se corrobora aplicando la prueba de Shapiro-Wilk, que calcula un W estadístico que prueba si una muestra aleatoria $x1, x_2, ..., x_n$ proviene de una distribución normal. Al aplicar la prueba con un valor del estadístico $W = 0{,}99112$ y un valor $p = 0{,}0643 > 0{,}05$ no se tiene suficiente evidencia para rechazar la hipótesis nula, que plantea que la variable sigue una distribución normal, esto se puede afirmar con un intervalo de confianza del 95 %.

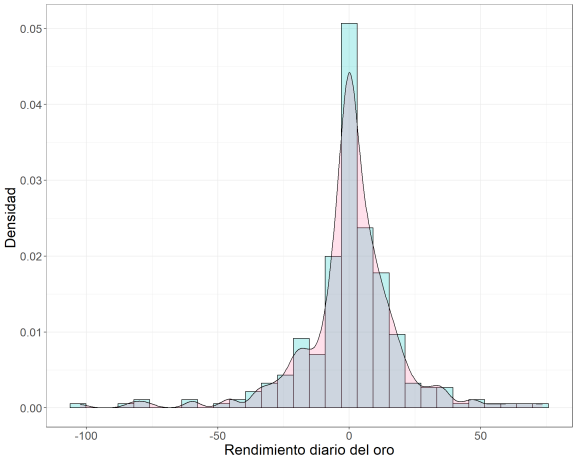
`prueba.txt`

```
        Shapiro-Wilk normality test

data:  re$resul
W = 0.9912, p-value = 0.06439
```

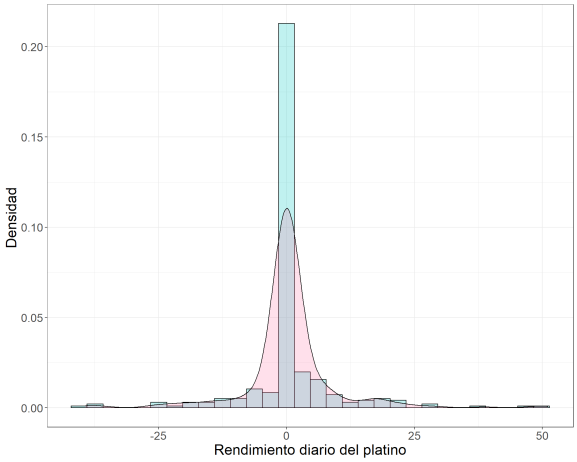
3. Chi cuadrada (χ^2)

La distribución χ^2 se utiliza para examinar tiene si un conjunto de datos presenta una diferencia estadísticamente significativa de lo que se espera. Para la misma necesitamos conocer la distribución que se espera ver y las frecuencias observadas de cada valor posible. Para el caso de estudio en este trabajo, se tiene el valor de densidad de empaquetamiento de seis tipos de figuras (triángulos,rectángulos, cuadrados,pentágonos, cuadriláteros mixtos y hexágonos) en un contenedor circular. Los valores de densidad se clasifican en tres categorías:

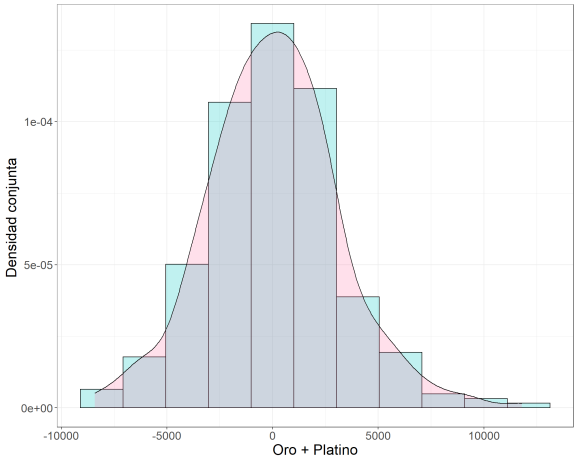
- Malos (densidad $\leq 0{,}40$).



(a) Histograma de densidad de la variable aleatoria X



(b) Histograma de densidad de la variable aleatoria Y



(c) Histograma de densidad conjunta de las variable aleatoria X e Y

Figura 1: Histogramas de la variables aleatorias X e Y y su convolución

- Buenos (0,40 > densidad $\leq$ 0,70).
- Muy buenos (densidad > 0,70).

Esta clasificación se puede ver en el cuadro 1 de la página 4, teniendo esto se plantea la hipótesis nula (H_0) que el tipo de figura no afecta en que la densidad de empaquetamiento sea mala, buena o muy buena. Para probar esta H_0 se realiza la prueba de *chisq.test* de la librería de R como se muestra a continuación.

prueba.txt

```
        Chi-squared test for given probabilities

data:  tabla
X-squared = 16.171, df = 2, p-value = 0.0003079
```

Con el estadístico $\chi^2 = 16{,}171$ y un valor $p = 0{,}0003 < 0{,}05$, se tiene suficiente evidencia para rechazar la H_0, por lo que se puede decir que el tipo de figura si influye en la calidad de la densidad de empaque, con un intervalo de confianza del 95 %.

Cuadro 1: Clasificación de los resultados de densidad para cada figura

Figuras	Malos	Buenos	Muy buenos	**Total**
Triángulo	1	20	14	**35**
Rectángulo	2	22	11	**35**
Cuadrados	5	20	10	**35**
Pentágono	7	16	7	**30**
Cuadrilátero mix	7	23	5	**35**
Hexágono	5	24	6	**35**
Total	**27**	**125**	**53**	**205**

4. Covarianza

En probabilidad y la estadística, la covarianza es una medida del grado en que dos variables aleatorias (X, Y) varían conjuntamente entorno a los valores de sus medias. Si las variables tienden a mostrar un comportamiento similar, la covarianza es positiva. En el caso contrario, cuando son inversamente proporcionales, la covarianza es negativa. Por lo que se puede carcular como se muestra en la ecuación 1 de la página 4.

$$\mathrm{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)]. \tag{1}$$

4.1. Propiedades de la covarianza

En esta sección se comprobará numéricamente y analíticamente dos de las propiedades de la covarianza.

Primera: Se tiene:

$$\text{Cov}[aX + b, cY + d] = ac \cdot \text{Cov}[X, Y] \tag{2}$$

Para demostrar numéricamente la igualdad planteada en la ecuación 2 de la página 5, se crean diez pares de variables aleatorias X con distribución normal e $Y = \left(\frac{X}{2}\right) * \frac{d}{a}$ como se puede ver en el cuadro 2 de la página 5. Posteriormente se les asignan valores a las constantes a, b, c, d. Lo anterior se realiza con el código 4.1 que se muestra a continuación.

```
mi=numeric()
md=numeric()
normal= seq(100,190,10)
for (i in c(1:10)) {

  X = rnorm(sample(normal,1))
  Y = (X/2)+d/a
  a = sample(2:6,1)
  b = sample(1:8,1)
  c = sample(2:10,1)
  d = sample(3:6,1)

  mi =c(mi,a*c*cov(X,Y))
  md = c(md,cov(((a*X)+b), ((c*Y)+d)))
}
```

R/Tarea11.R

Cuadro 2: Fragmento de uno de los pares de variables aleatorias X e Y creadas

	X	Y
1	1.41293	1.20647
2	-0.44490	0.27755
3	1.02916	1.01458
4	0.54957	0.77478
5	-1.35893	-0.17946
54	-0.13546	0.43227
55	1.06661	1.03331
56	0.65222	0.82611
57	1.04822	1.02411
58	-1.37364	-0.18682

Cuadro 3: Resultados de ambos miembros de la ecuación 2

	M. izquierdo(M.i)	M. derecho(M.d)	M.i = M.d
1	10.28066	10.28066	Si
2	6.25041	6.25041	Si
3	14.60058	14.60058	Si
4	15.01817	15.01817	Si
5	6.16484	6.16484	Si
6	16.68462	16.68462	Si
7	14.00431	14.00431	Si
8	16.71537	16.71537	Si
9	7.37975	7.37975	Si
10	3.11481	3.11481	Si

Como se observa en el cuadro 3 de la página 5, para todas las variables aleatorias X e Y creadas los valores del miembro izquierdo de la ecuación 2 son iguales al miembro derecho de dicha ecuación. De esta manera queda demostrado numéricamente la igualdad planteada. Para esta comprobar la igualdad de manera analítica es conveniente recordar las dos propiedades siguientes.

Sea a y c dos constantes,

$$\text{Cov}(aX, Y) = a \cdot \text{Cov}(X, Y) \tag{3}$$

$$\text{Cov}(X + c, Y) = \text{Cov}(X, Y). \tag{4}$$

Desarrollando el lado izquierdo de la ecuación 2 se tiene:

$$\begin{aligned} \text{Cov}[aX + b, cY + d] &= ac \cdot \text{Cov}[X + b, Y + d], && \text{por la propiedad 3,} \\ &= ac \cdot \text{Cov}[X, Y], && \text{por propiedad 4.} \end{aligned} \tag{5}$$

Segunda: Se tiene:

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2 \cdot \text{Cov}[X, Y] \tag{6}$$

Para la comprobación numérica de la igualdad 6 de la página 6, se utiliza el código 4.1, los valores de las variables aleatorias y las constantes tienen las mismas características que en el caso anterior.

```r
mi=numeric()
md=numeric()
normal= seq(100,190,10)
for (i in c(1:10)) {

  X = rnorm(sample(normal,1))
  Y = (X/2)+d/a
  a = sample(2:6,1)
  b = sample(1:8,1)
  c = sample(2:10,1)
  d = sample(3:6,1)

  mi =c(mi,var(X+Y))
  md = c(md,var(X)+ var(Y)+2*cov(X,Y))
}
```

R/Tarea11.R

Cuadro 4: Resultados de ambos miembros de la ecuación 6

	M. izquierdo(M.i)	M. derecho(M.d)	M.i = M.d
1	2.49469	2.49469	Si
2	2.23400	2.23400	Si
3	2.44015	2.44015	Si
4	2.20168	2.20168	Si
5	2.37420	2.37420	Si
6	2.40887	2.40887	Si
7	2.17660	2.17660	Si
8	1.95016	1.95016	Si
9	2.60644	2.60644	Si
10	1.98039	1.98039	Si

En los resultados que se muestran en el cuadro 4 de la página 6, queda comprobado de forma numérica la igualad planteada en la ecuación 6. Para la corporación analítica de ecuación en cuestión, no apoyamos de las expresiones siguientes:

$$\text{Var}[X] = E[X^2] - E[X]^2. \tag{7}$$

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]. \tag{8}$$

Desarrollando el lado izquierdo de la ecuación 6 se tiene:

$$\begin{aligned}
\text{Var}[X+Y] &= E[(X+Y)^2] - E[X+Y]^2 \\
&= E[(X+Y)^2] - E[X+Y]^2 \\
&= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 + 2 \cdot E[XY] - 2 \cdot E[X]E[Y] \\
&= \text{Var}[X] + \text{Var}[Y] + 2(E[XY] - E[X]E[Y]) \\
&= \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}(X, Y).
\end{aligned} \tag{9}$$

El código general se encuentra disponible en el repositorio. https://github.com/Albertomnoa/Tareas

Referencias

[1] R Core Team. R: Un lenguaje y un entorno para la informática estadística, 2020.

Tarea 12 de Modelos Probabilistas Aplicados

Ejercicios

5271

24 de noviembre de 2020

1. Introducción

En este documento se presentan los resultados de varios ejercicios del libro "*Introduction to Probability*"[1].

2. Ejercicio 1 de la página 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) $p_0 = \frac{1}{2}, p_1 = \frac{1}{4}, p_2 = \frac{1}{4}$.

b) $p_0 = \frac{1}{3}, p_1 = \frac{1}{3}, p_2 = \frac{1}{3}$.

c) $p_0 = \frac{1}{3}, p_1 = 0, p_2 = \frac{2}{3}$.

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

e) $p_j = (\frac{1}{3})(\frac{2}{3})^j$, for $j = 0, 1, 2, ...$

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

De acuerdo al teorema 10.3 si el número medio m de descendientes producidos por un solo padre es ≤ 1, entonces d = 1 y el proceso se extingue con probabilidad uno. Si m > 1 entonces d < 1 y el proceso se extingue con probabilidad d. Para la resolución de los incisos a), b), c) se utiliza la expresión siguiente:

$$\mathrm{m} = p_1 + 2p_2 = 1 - p_0 + p_2 \tag{1}$$

a) $p_0 = \frac{1}{2}, p_1 = \frac{1}{4}, p_2 = \frac{1}{4}$.

$$\begin{aligned} \mathrm{m} &= p_1 + 2p_2 && (2)\\ &= \frac{1}{4} + 2\left(\frac{1}{4}\right) && (3)\\ &= \frac{3}{4} && (4) \end{aligned}$$

Por tanto m es menor que uno y la probabilidad de decadencia es uno.

b) $p_0 = \frac{1}{3}, p_1 = \frac{1}{3}, p_2 = \frac{1}{3}$.

$$\begin{aligned} \mathrm{m} &= p_1 + 2p_2 && (5)\\ &= \frac{1}{3} + 2\left(\frac{1}{3}\right) && (6)\\ &= 1 && (7) \end{aligned}$$

Por tanto m es igual a uno y la probabilidad de decadencia es uno.

c) $p_0 = \frac{1}{3}, p_1 = 0, p_2 = \frac{2}{3}$.

$$\begin{aligned} \mathrm{m} &= p_1 + 2p_2 && (8)\\ &= 0 + 2\left(\frac{2}{3}\right) && (9)\\ &= \frac{4}{3} && (10) \end{aligned}$$

Por tanto m es mayor que uno y la probabilidad de decadencia es igual a d. A contención pasamos a calcular la probabilidad de decadencia (d)

$$\begin{aligned} d &= \frac{p_0}{p_2} && (11)\\ &= \frac{\frac{1}{3}}{\frac{2}{3}} && (12)\\ d &= \frac{1}{2} && (13)\\ & && (14) \end{aligned}$$

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

$$h(z) = p_0 + p_1 z + p_2 z^2 + p_3 z^3 + p_4 z^4 ... \tag{15}$$
$$= \frac{1}{2}^{0+1} + \frac{1}{2}^{1+1} z + \frac{1}{2}^{2+1} z^2 + \frac{1}{2}^{3+1} z^3 + \frac{1}{2}^{4+1} z^4 + ... \tag{16}$$
$$= \frac{1}{2}^{1} + \frac{1}{2}^{2} z + \frac{1}{2}^{3} z^2 + \frac{1}{2}^{4} z^3 + \frac{1}{2}^{5} z^4 .. \tag{17}$$
$$= \frac{1}{2}(1 + 1/2^1 z + 1/2^2 z^2 + 1/2^3 z^3 + ...) \tag{18}$$
$$= \frac{1}{2}\left(\frac{1}{1 - \frac{1}{2}z}\right) \tag{19}$$
$$= \frac{1}{2 - z}. \tag{20}$$

Calculando h'(z) tenemos:

$$h'(z) = \frac{d}{dz}\left(\frac{1}{2 - z}\right) \tag{21}$$
$$= \frac{-\frac{d}{dz}(2 - z)}{(2 - z)^2} \tag{22}$$
$$= \frac{1}{(2 - z)^2} \tag{23}$$
$$\tag{24}$$

ahora calculamos h'(1): $m = h'(1) = \frac{1}{(2-1)^2} = 1$ Como m es igual a 1 entones la probabilidad de decadencia es uno.

3. Ejercicio 3 de la página 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.

Para la resolver este ejercicio se se emplea la siguiente expresión:

$$50\mathrm{m} + 50\mathrm{m}^{12}, \quad \text{con} \quad \mathrm{m} = p_1 + 2p_2. \tag{25}$$

a) $p_0 = 1/2$, $p_1 = 0$, $p_2 = 1/2$. Sustituyendo en la expresión anterior se tiene:

$$\mathrm{m} = 0 + 2\left(\frac{1}{2}\right) \tag{26}$$
$$= 1, \tag{27}$$

por lo que:

$$50(1) + 50(1^{12}) - 100 = 0 \tag{28}$$

Por tanto el la ganancia esperada es cero.

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

$$\mathrm{m} = \frac{1}{2} + 2\left(\frac{1}{3}\right) \tag{29}$$

$$= \frac{7}{6}, \tag{30}$$

por lo que:

$$50\left(\frac{7}{6}\right) + 50\left(\frac{7}{6}\right)^{12} - 100 \approx 276{,}3 \tag{31}$$

Por tanto el la ganancia esperada es 276.3.

Para todos los posibles valores de $p_0 > 1/2$, $p_2 < p_0$ por lo tanto $\mathrm{m} < 1$ y $50\mathrm{m} + 50\mathrm{m}^{12} < 100$, por lo que el juego no va a ser favorable.

4. Ejercicio 1 de la página 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) $f_X(x) = \frac{3}{8}x^2$.

Para la realización de este ejercicio se utiliza la ecuación siguiente:

$$g(t) = \mathbb{E}[e^{tx}] \tag{32}$$

$$= \int_{-\infty}^{\infty} e^{tx} fX(x)\mathrm{d}x \tag{33}$$

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

$$g(t) = \int_0^2 e^{tx} \cdot \frac{1}{2} \mathrm{d}x \tag{34}$$
$$= \int \frac{\mathrm{e}^{tx}}{2} \, \mathrm{d}x \tag{35}$$
$$= \frac{1}{2t} \int \mathrm{e}^{u} \, \mathrm{d}u \tag{36}$$
$$= \frac{\mathrm{e}^{tx}}{2t} \tag{37}$$
$$= \frac{\frac{\mathrm{e}^{2t}}{t} - \frac{1}{t}}{2} \tag{38}$$
$$= \frac{\mathrm{e}^{2t} - 1}{2t}. \tag{39}$$

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

$$g(t) = \int_0^2 e^{tx} \cdot \frac{1}{2} x \mathrm{d}x \tag{40}$$
$$= \frac{1}{2} \int x \mathrm{e}^{tx} \, \mathrm{d}x. \tag{41}$$

Resolviendo:

$$\int x \mathrm{e}^{tx} \, \mathrm{d}x,$$

se tiene:

$$= \frac{x\mathrm{e}^{tx}}{t} - \int \frac{\mathrm{e}^{tx}}{t} \, \mathrm{d}x \tag{42}$$

Ahora se resuelve:

$$\int \frac{\mathrm{e}^{tx}}{t} \, \mathrm{d}x,$$

y se tiene:

$$= \frac{1}{t^2} \int \mathrm{e}^{u} \, \mathrm{d}u \tag{43}$$
$$= \frac{\mathrm{e}^{u}}{t^2} \tag{44}$$
$$= \frac{\mathrm{e}^{tx}}{t^2}. \tag{45}$$

Remplazando las integrales resuelta se tiene:

$$= \frac{1}{2} \left[\frac{x\mathrm{e}^{tx}}{t} - \frac{\mathrm{e}^{tx}}{t^2} \right] \tag{46}$$
$$= \frac{x\mathrm{e}^{tx}}{2t} - \frac{\mathrm{e}^{tx}}{2t^2} \tag{47}$$
$$= \frac{(tx - 1)\,\mathrm{e}^{tx}}{2t^2} \tag{48}$$
$$= \frac{(2t - 1)\,\mathrm{e}^{2t} + 1}{2t^2}. \tag{49}$$

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

$$g(t) = \int_0^2 e^{tx} \cdot \left(1 - \frac{1}{2}x\right) \mathrm{d}x \tag{50}$$

$$= -\frac{1}{2}\int (x-2)\,\mathrm{e}^{tx}\,\mathrm{d}x. \tag{51}$$

Resolviendo:

$$\int (x-2)\,\mathrm{e}^{tx}\,\mathrm{d}x,$$

se tiene:

$$= \frac{(x-2)\,\mathrm{e}^{tx}}{t} - \int \frac{\mathrm{e}^{tx}}{t}\,\mathrm{d}x \tag{52}$$

Ahora se resuelve:

$$\int \frac{\mathrm{e}^{tx}}{t}\,\mathrm{d}x,$$

y se tiene:

$$= \frac{1}{t^2}\int \mathrm{e}^{u}\,\mathrm{d}u \tag{53}$$

$$= \frac{\mathrm{e}^{u}}{t^2} \tag{54}$$

$$= \frac{\mathrm{e}^{tx}}{t^2}. \tag{55}$$

Remplazando las integrales resuelta se tiene:

$$= -\frac{1}{2}\left[\frac{(x-2)\,\mathrm{e}^{tx}}{t} - \frac{\mathrm{e}^{tx}}{t^2}\right] \tag{56}$$

$$= \frac{\mathrm{e}^{tx}}{2t^2} - \frac{(x-2)\,\mathrm{e}^{tx}}{2t} \tag{57}$$

$$= \frac{\mathrm{e}^{2t}}{2t^2} - \frac{2t+1}{2t^2} \tag{58}$$

$$= \frac{\mathrm{e}^{2t} - 2t - 1}{2t^2}. \tag{59}$$

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

$$g(t) = \int_0^2 e^{tx} \cdot |x-1|\,\mathrm{d}x \tag{60}$$

$$= \frac{|x-1|\,\mathrm{e}^{tx}}{t} - \int \frac{(x-1)\,\mathrm{e}^{tx}}{t\,|x-1|}\,\mathrm{d}x. \tag{61}$$

Resolviendo:

$$\int \frac{(x-1)\,\mathrm{e}^{tx}}{t\,|x-1|}\,\mathrm{d}x,$$

se tiene:

$$= \frac{1}{t^2}\int 1 \,\mathrm{d}u, \tag{62}$$

ahora se resuelve:

$$\int 1 \,\mathrm{d}u,$$

y se tiene:

$$= u \tag{63}$$

$$= \frac{u}{t^2}, \tag{64}$$

desciendo la sustitución:

$$= \frac{(x-1)\,\mathrm{e}^{tx}}{t^2\,|x-1|}. \tag{65}$$

Remplazando las integrales resuelta se tiene:

$$= \frac{|x-1|\,\mathrm{e}^{tx}}{t} - \frac{(x-1)\,\mathrm{e}^{tx}}{t^2\,|x-1|} \tag{66}$$

$$= \frac{(t-1)\,\mathrm{e}^{2t}}{t^2} + \frac{2\mathrm{e}^{t}}{t^2} - \frac{t+1}{t^2} \tag{67}$$

$$= \frac{(t-1)\,\mathrm{e}^{2t} + 2\mathrm{e}^{t} - t - 1}{t^2}. \tag{68}$$

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

$$g(t) = \int_0^2 e^{tx} \cdot \frac{3}{8}x^2 \mathrm{d}x \tag{69}$$

$$= \frac{3}{8}\int x^2 \mathrm{e}^{tx} \,\mathrm{d}x. \tag{70}$$

Resolviendo:

$$\int x^2 \mathrm{e}^{tx} \,\mathrm{d}x,$$

se tiene:

$$= \frac{x^2 \mathrm{e}^{tx}}{t} - \int \frac{2x\mathrm{e}^{tx}}{t} \,\mathrm{d}x, \tag{71}$$

ahora se resuelve:

$$\int \frac{2x\mathrm{e}^{tx}}{t} \,\mathrm{d}x,$$

y se tiene:

$$= \frac{2}{t}\int x\mathrm{e}^{tx} \,\mathrm{d}x, \tag{72}$$

resolviendo:

$$\int x\mathrm{e}^{tx} \,\mathrm{d}x,$$

se tiene:

$$= \frac{1}{t^2}\int \mathrm{e}^u \,\mathrm{d}u \tag{73}$$

$$= \frac{\mathrm{e}^u}{t^2} \tag{74}$$

$$== \frac{\mathrm{e}^{tx}}{t^2} \tag{75}$$

Remplazando las integrales resuelta se tiene:

$$= \frac{3}{8}\left[\frac{x^2\mathrm{e}^{tx}}{t} - \frac{2x\mathrm{e}^{tx}}{t^2} + \frac{2\mathrm{e}^{tx}}{t^3}\right] \tag{76}$$

$$= \frac{3x^2\mathrm{e}^{tx}}{8t} - \frac{3x\mathrm{e}^{tx}}{4t^2} + \frac{3\mathrm{e}^{tx}}{4t^3} \tag{77}$$

$$= \frac{\left(6t^2 - 6t + 3\right)\mathrm{e}^{2t} - 3}{4t^3}. \tag{78}$$

5. Ejercicio 6 de la página 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

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

Para la solución de este ejercicio se utilizara la formula que se muestra a continuación:

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\tau x}\cdot k_X(\tau)\mathrm{d}\tau. \tag{79}$$

Respuesta sustituyendo en la ecuación anterior se tiene:

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\tau x}\cdot e^{-|\tau|}\mathrm{d}\tau \tag{80}$$

$$= \frac{1}{2\pi}\left(\int_{-\infty}^{0} e^{-ix\tau}e^{\tau}\,d\tau\right) + \frac{1}{2\pi}\left(\int_{0}^{\infty} e^{-ix\tau}e^{-\tau}\,d\tau\right) \tag{81}$$

$$= \frac{1}{2\pi}\left(\frac{\mathrm{i}x+1}{x^2+1}\right) + \frac{1}{2\pi}\left[-\left(\frac{\mathrm{i}x-1}{x^2+1}\right)\right] \tag{82}$$

$$= \frac{1}{2\pi}\left(\frac{\mathrm{i}x+1-\mathrm{i}x+1}{x^2+1}\right) \tag{83}$$

$$= \frac{1}{2\pi}\left(\frac{2}{x^2+1}\right) \tag{84}$$

$$= \frac{1}{\pi(1+x^2)}. \tag{85}$$

6. ejercicio 10 de la página 402

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

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

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 $n \to \infty$.

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

$$\begin{aligned}
\mathbb{E}(X) &= \int_{-\infty}^{\infty} x\frac{e^{-|x|}}{2}dx && (86)\\
&= \frac{1}{2}\int_{-\infty}^{\infty} \frac{x}{|x|}e^{-|x|}|x|dx && (87)\\
&= -\frac{e^{-|x|}|x|}{2} - \frac{e^{-|x|}}{2} && (88)\\
&= \left[\frac{e^{-|x|}(-|x|-1)}{2}\right]_{-\infty}^{\infty} && (89)\\
&= 0. && (90)
\end{aligned}$$

Ya se tiene el valor de $\mathbb{E}[X]$, pasamos a calcular $\mathbb{V}(X)$ con la expresión:

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2) \quad (91)$$

$$\begin{aligned}
\mathbb{E}[X]^2) &= \int_{-\infty}^{\infty} x^2\frac{1}{2}e^{-|x|}\,dx && (92)\\
&= \frac{1}{2}\left[\int_{-\infty}^{0} x^2e^{-(-x)}\,dx + \int_{0}^{\infty} x^2e^{-(x)}\,dx\right] && (93)\\
&= \frac{1}{2}\left[\int_{-\infty}^{0} x^2e^{(x)}\,dx + \int_{0}^{\infty} x^2e^{-(x)}\,dx\right] && (94)\\
&= \frac{1}{2}\left[2+2\right] && \\
&= 2. && (95)
\end{aligned}$$

Entonces:

$$\begin{aligned}
\mathbb{V}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2) && (96)\\
&= 2 - (0)^2 && (97)\\
&= 2 - 0 && (98)\\
&= 2. && (99)
\end{aligned}$$

Referencias

[1] Grinstead, Charles M., Snell, J. Laurie. *Introduction to Probability.* American Mathematical Society, 2006.

Tarea 13 de Modelos Probabilistas Aplicados

Ley de los grandes números

5271

1 de diciembre de 2020

1. Introducción

En este documento se presenta las nociones básicas sobre la ley de los grandes números, ejemplos y aplicaciones.

2. Ley de los grandes números

LA ley de los grandes números plantea formalmente que con una sucesión de variables aleatorias independientes e idénticamente distribuidas con varianza finita se asegura que el promedio de las n primeras observaciones (variables aleatorias) se acerca a la media teórica cuando el número n de repeticiones tiende hacia infinito. Lo anterior se apoya en el teorema 8.2 del libro "*Introduction to Probability*"[1]:

Sean $X_1, X_2, ..., X_n$ un proceso de pruebas independientes e igualmente distribuidos con un valor esperado finito $\mu = E[X]$ y una y una varianza finita $\sigma^2 = \text{Var}[X]$. Sea $S_n = X_1 + X_2 + \ldots + X_n$. Entonces para cualquier valor de $\epsilon > 0$ y n que tiende al infinito se tiene:

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0 \tag{1}$$

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) = 1 \tag{2}$$

Un caso particular de esta ley es el principio de estabilidad de las frecuencias, o teorema de Bernoulli, que ya hemos visto. Efectivamente, recordemos que una variable de Bernoulli es aquella que toma solo el valor 0 o 1 cuando no ocurre (u ocurre, respectivamente) un suceso A con probabilidades $1 - p$ y p. Sumar n variables de Bernoulli es contar el número de veces que se repite el suceso A en n pruebas. Una variable de Bernoulli tiene media p. Luego la media de n medias sera también p.

Para una mejor comprensión de lo anteriormente expuesto se tomará como ejemplo la resolución del ejercicio 3 de la página 312 del mismo libro, que dice:

Write a program to toss a coin 10,000 times. Let S_n be the number of heads in the first n tosses. Have your program print out, after every 1000 tosses, $S_n - \frac{n}{2}$. On the basis of this simulation, is it correct to say that you can expect heads about half of the time when you toss a coin a large number of times?

Para resolver este ejercicio se realiza el programa 2 en lenguaje R [3] que se muestra a continuación.

```
n = 1000 # N mero de lanzamientos igual a 1000
Sn = c() # Numero de caras obtenidos
suma_caras = 0 # suma de los numeros de cara
n = c() # cantidad acomulada de lanzamientos
for(i in 1:10){
  simul = sample(0:1, 1000, rep = T) # Lanza una moneda 1000 veces
  suma_caras = suma_caras + sum(simul == 1) # Suma de las caras
  Sn = c(Sn, suma_caras / (1000 * i)) # N mero de caras
  n = c(n, 1000 * i) # N mero de lanzamientos.
}
data_frame = data.frame(n, Sn)
data_frame
```

Tarea13.R

El cuadro 1 de la página 2 se puede observar claramente, que para el programa 2 donde se ejecuta diez simulaciones de 1000 lanzamientos de una moneda donde la cara vale uno y la cruz 0. Es correcto decir que puede esperar cara la mitad de las veces cuando lanza una moneda una gran cantidad de veces. En la figura 1 de la página 3 se muestra los resultados para 100 repeticiones.

Cuadro 1: Resultados de la simulación de sacar cara en el lanzamiento de una moneda

	n	Sn
1	1,000	0.513
2	2,000	0.511
3	3,000	0.510
4	4,000	0.506
5	5,000	0.504
6	6,000	0.505
7	7,000	0.503
8	8,000	0.502
9	9,000	0.502
10	10,000	0.502

3. Aplicación en la estimación de subconjuntos en un método estocástico de ramificación y poda

En [2], se propone un método estocástico de ramificación y poda para resolver problemas de optimización global estocástica. Como en el caso determinista, el conjunto factible se divide en subconjuntos compactos. Para guiar el proceso de partición, el método utiliza estimaciones estocásticas superior e inferior del valor óptimo de la función objetivo en cada subconjunto. Y se demuestra la convergencia del método y se obtienen estimaciones de precisión aleatorias. Se discuten los métodos para construir límites estocásticos superior e inferior. Las consideraciones teóricas la ilustran con un ejemplo de un problema de ubicación de instalaciones.

En el método estocástico de ramificación y poda [2] se construye una secuencia de conjuntos $X^k(\omega) \subset X^{k-1}(\omega)$, y se tiene que estimar el valor de cota inferior $L(\cdot)$ en el límite que establece $X^* = \lim_k X(\omega)$,

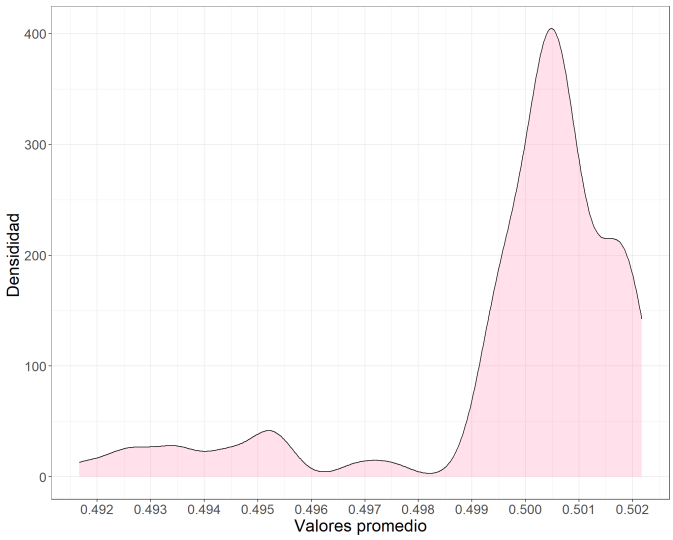


Figura 1: Diagrama de densidad de los resultados para 100 repeticiones del experimento.

usando observaciones independientes de variables aleatorias $\xi(X^k)$ tales que $\mathbb{E}[\xi(X^k)] = L(X^k)$. A tal efecto, en[2] se utiliza la siguiente estimación:

$$L_k(X^k) = \frac{1}{k}\sum_{s=1}^{k} \xi(X^k) \to L(X^*). \tag{3}$$

Referencias

[1] Grinstead, Charles M., Snell, J. Laurie. *Introduction to Probability.* American Mathematical Society, 2006.

[2] Vladimir I Norkin, Georg Ch Pflug, and Andrzej Ruszczyński. A branch and bound method for stochastic global optimization. *Mathematical programming*, 83(1-3):425–450, 1998.

[3] R Core Team. R: Un lenguaje y un entorno para la informática estadística, 2020.

Tarea 13 de Modelos Probabilistas Aplicados

Teorema de límite central

5271

8 de diciembre de 2020

1. Introducción

En este documento se presenta las nociones básicas sobre el teorema del limite central, ejemplos y aplicaciones.

2. Teorema del limite central

El teorema del límite central proporciona una aproximación al comportamiento de las sumas de variables aleatorias. El teorema establece que a medida que aumenta el número de variables aleatorias independientes e idénticamente distribuidas con media finita y varianza finita, la distribución de su suma se vuelve cada vez más normal independientemente de la forma de distribución de las variables aleatorias. Es decir, sea $X_1, X_2, ..., X_n$ una secuencia de variables aleatorias mutuamente independientes e idénticamente distribuidas, cada una de las cuales tiene una media finita μ_x y una varianza finita σ_x^2. Sea S_n definido como sigue:

$$S_n = X_1 + X_2 + \ldots + X_n. \tag{1}$$

Ahora, $\mathbb{E}[S_n] = n\mu_x$ y $\sigma_{S_n}^2 = n\sigma_x^2$. Al convertir S_n en una variable aleatoria normal estándar (es decir, media cero y varianza unitaria), se obtiene.

$$Y_n = \frac{S_n - \mathbb{E}[S_n]}{\sigma_{S_n}^2} = \frac{S_n - n\mu_X}{\sigma_X\sqrt{n}}. \tag{2}$$

El teorema del límite central establece que si $F_{Y_n}(y)$ es la función de densidad de γ_n , entonces:

$$\lim_{n\to\infty} F_{Y_n}(y) = \lim_{n\to\infty} \mathbb{P}[Y_n \leq y] = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du = \Phi(y); \tag{3}$$

Esto significa que $\lim_{n\to\infty} \gamma_n$ se distribuye $\sim N(0,1)$. Por tanto, uno de los roles importantes que juega la distribución normal en estadística es su utilidad como aproximación de otras funciones de distribución de probabilidad. Lo anterior se apoya en el teorema 9.5 del libro "*Introduction to Probability*"[1]:

Para una mejor comprensión de lo anteriormente expuesto se tomará como ejemplo la resolución del ejercicio 11 de la página 355 del mismo libro, que dice:

A tourist in Las Vegas was attracted by a certain gambling game in which the customer stakes 1 dollar on each play; a win then pays the customer 2 dollars plus the return of her stake, although a loss costs her only her stake.Las Vegas insiders, and alert students of probability theory, know that the probability of winning at this game is 1/4. When driven from the tables by hunger, the tourist had played this game 240 times. Assuming that no near miracles happened, about how much poorer was the tourist upon leaving the casino? What is the probability that she lost no money?

Para resolver este ejercicio se realiza el programa 2 en lenguaje R [2] que se muestra a continuación.

```
# Assuming that no near miracles happened, about how much poorer was the tourist upon
    leaving the casino?
valor_e = 0.25*2 + 0.75*(-1)
valor_e_perder = valor_e*240
valor_e_perder

# What is the probability that she lost no money?
n <- 240
p <- 0.25
q <- 1-p
sd <- sqrt(n*p*q)
mean <- -60
normal=as.data.frame(rnorm(n, mean, sd))

png(filename = "ejercicio11.png",width = 2000, height = 1600, res =200)
ggplot(normal, aes(x=`rnorm(n, mean, sd)`)) + scale_x_continuous(breaks=seq(-80, -30,
    5)) + geom_density(alpha=.2, fill="#FF6699")+ theme_bw()+ xlab("Valores simulados")
    +
  ylab("Densididad ")+theme(axis.text = element_text(size = 14))+theme(axis.title =
    element_text(size = 18))
dev.off()
```

Tarea14.R

Como respuesta a la primera pregunta se tiene que $\mathbb{E}[S_n] = n\mu_x = -60$, por lo que si el turista juega 240 veces se espera que pierda 60 dolares. Para la segunda respuesta tenemos la figura 1 de la página 3, que muestra una simulación de 240 veces jugadas con una distribución normal con media $\mu = -60$ y $\sigma^2 = 6{,}708$. en la misma se puede observar que existe cero probabilidad que el jugador gane algún dolar en un total de 240 juegos.

3. Aplicación en prueba de hipótesis

En esta sección un ejemplo de como se utiliza el teorema del limite central para probar hipótesis. Comúnmente para esta fin se utiliza una prueba de χ^2 para determinar si rechazar una distribución de población hipotética (con un número finito de clases) como falso. Aquí se hará esto para cuando la población se descomponga en dos clases (fumadores y no fumadores).

Para el ejemplo plantearemos la hipótesis H_0 que nos dice que el 20 % de los jóvenes en Monterrey fuman, para probar esta hipótesis tomamos datos de una encuesta realizada a 1,888 estudiantes en Monterrey [3] la cual arroja que 18.7 % de los encuestados eran fumadores al momento de la realización de la encuesta.

Sea $X_i = 1$ si el encuestado i dice que es fumador, y sea $X_i = 0$ si el encuestado i dice que no es fumador. Estas X_i son variables aleatorias de Bernoulli independientes. Tenemos que $S_n = X1 + + X_1, 888$. Si la hipótesis de que el 20 % de los jóvenes de Monterrey fuman es correcta, entonces $\mu = \mathbb{E}[X_i] = 0{,}2$,

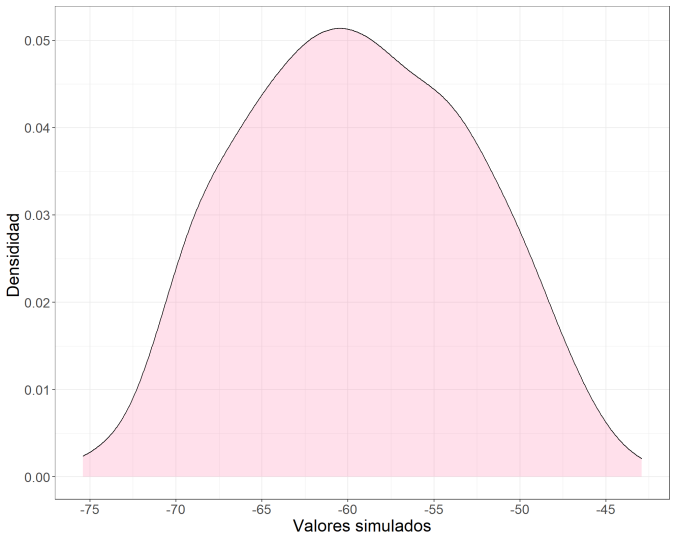


Figura 1: Diagrama de densidad de los resultados para 240 juegos.

y $X_i = (1 -) = 0{,}16$; y entonces, el teorema del límite central nos diría que:

$$\frac{S_{1888} - 378}{\sqrt{1,888 \cdot 0{,}16}}, \tag{4}$$

en el sentido que

$$\mathbb{P}\left(\frac{S_{1888} - 378}{6{,}952} \leq y\right) \approx \Phi(y). \tag{5}$$

Ahora, si $S^*_{1,888}$ es el valor observado $S^1_{1,888}$, y si

$$\gamma = \frac{S_{1888} - 378}{\sqrt{6{,}952}}, \tag{6}$$

entonces, sobre la base del teorema del limite central y la ecuación eq:1 se esperaría que γ es un valor atípico para $\sim N(0,1)$. Específicamente, no se esperaría que γ demasiado grande; es decir no se esperaría que:

$$\mathbb{P}(|N(0,1)||\gamma|) < 0{,}05, \tag{7}$$

si $\mu = 0{,}2$ es la media verdadera. Así se tiene la siguiente prueba estadística: Se fija un $\alpha > 0$, comúnmente se emplea $\alpha = 0{,}05$. Calculando γ como en la ecuación 6, si

$$\mathbb{P}(|N(0,1)||\gamma|) = 2\Phi(-|\gamma|) < \alpha, \tag{8}$$

rechazamos la hipótesis de que el valor medio de X_i es μ, y si esta desigualdad no se satisface, no la rechazamos.

Por lo que se tiene que:

$$\gamma = \frac{353 - 378}{6{,}952} = -3{,}596, \tag{9}$$

y calculando vemos clara mente que

$$2\Phi(-3{,}596) = 2(0{,}49983) = 0{,}9996 > 0{,}05. \tag{10}$$

Por lo tanto no hay suficientes elementos para rechazar H_o, que plantea que el 20 % de los jóvenes en Monterrey fuman. H_0 se acepta con un intervalo de confianza del 95 %.

Referencias

[1] Grinstead, Charles M., Snell, J. Laurie. *Introduction to Probability.* American Mathematical Society, 2006.

[2] R Core Team. R: Un lenguaje y un entorno para la informática estadística, 2020.

[3] Reynales-Shigematsu LM, Valdés-Salgado R, Rodríguez-Bolaños R, Lazcano-Ponce E, Hernández-Ávila M. *Encuesta de Tabaquismo en Jóvenes en México.Análisis descriptivo 2003, 2005, 2006, 2008.* nstituto Nacional de Salud Pública, 2009.

Tarea 16 de Modelos Probabilistas Aplicados

Retroalimentación de propuestas para el proyecto final

5271

15 de diciembre de 2020

1. Propuesta Alberto Benavides

Motivado por el interés de avanzar en mi tema de tesis, esta primera propuesta consiste en encontrar las relaciones entre dos series de tiempo, una de contaminantes del aire y otra de afecciones del sistema respiratorio, ambas de algún rango de tiempo perteneciente a la última década. El estudio de las relaciones entre estas series se abordaría mediante correlaciones y una técnica conocida como Dynamic time warping *a partir de retrasos de tiempo entre las series.*

Retroalimentación: Este tema es bastante interesante debido al impacto social de la contaminación, el método deformación dinámica del tiempo (DTW) lo había visto en algoritmo de reconocimiento de voz y sería muy interesante verlo en tu trabajo esto sería un gran aporte.

2. Propuesta de Gerardo Palafox

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

Retroalimentación: El tema está muy interesante ya que la detección de comunidades ha demostrado ser valiosa en una serie de campos, por ejemplo, biología, ciencias sociales, bibliometría, además esta presentado de manera muy concreta y con sus respectivas referencias teóricas. Me parece una muy buena propuesta.

3. Propuesta de Joaquin Arturo Velarde Moreno

Las empresas necesitan tomar préstamos en el extranjero en dólares, por ello necesitan tener una idea del nivel de depreciación del peso a largo plazo, que les permita valorar el riesgo de tomar estos

préstamos, para ello se va a hacer una regresión lineal con el objetivo de calcular el valor del dólar a 15 años, basándonos en los datos de los últimos 40 años dados por el banco de México.

Retroalimentación Me parece muy interesante, sobre todo en la situación de incertidumbre actual y me parece que sería conveniente mencionar que factores vas a tener en cuenta en el análisis, como podrían ser la inflación interna y externa, la tasa de interés interna y externa, las reservas internacionales, el circulante monetario y la actividad industrial entre otros.

Diseño de experimento para el problema de empaquetamiento óptimo de politopos convexos definidos por sus vérticess

Alberto Martínez-Noa

Facultad de Ingeniería Mecánica y Eléctrica, Universidad Autónoma de Nuevo León

Abstract

En los problemas de C&P contamos con un conjunto de elementos pequeños llamados carga que deben ser dispuestos o asignados a uno o varios objetos de gran tamaño llamados contenedores, y se debe cumplir la no superposición entre elementos pequeños y que dichos conjuntos de elementos no excedan las dimensiones del contenedor al cual han sido asignados. Dichos problemas se clasifican como NP-duros [1] debido a su complejidad computacional. En este Trabajo se presenta un diseño de experimento para un modelo no lineal de empaquetamiento con un enfoque lagrangiano en las condiciones de no intersección que analiza las figuras por sus vértices. En el mismo se analizan como influyen los factores (tipo de figura, cantidad de figura y tipo de contenedor) en la densidad de empaque con el empleo de este modelo.

Keywords: Empaquetamiento óptimo, ANOVA, diseño de experimento, pruebas no perimétricas.

1. Introduction

En los problemas de C&P contamos con un conjunto de elementos pequeños llamados carga que deben ser dispuestos o asignados a uno o varios objetos de gran tamaño llamados contenedores, y se debe cumplir la no superposición entre elementos pequeños y que dichos conjuntos de elementos no excedan las dimensiones del contenedor al cual han sido asignados. Dichos problemas se clasifican como NP-duros [1] debido a su complejidad computacional.

Este trabajo de investigación se centrará en la resolución del problema de empaquetamiento de un surtido arbitrario de elementos pequeños convexos en un único contenedor convexo de dimensiones variables que se puede tratar como el problema de dimensión(es) abierta(as) (ODP, por sus siglas en inglés).

El problema de dimensión abierta es un problema en el cual, el conjunto de elementos pequeños debe ser acomodado completamente dentro de un objeto grande o contenedor. El objeto grande, posee al menos una dimensión puede considerarse variable. En otras palabras, este problema implica una decisión sobre la fijación de las extensiones en las dimensiones variables de los objetos grandes, así como el valor de la entrada u otra medida como longitud, tamaño o volumen se deben minimizar [6].

Este trabajo presenta un diseño de experimento para un modelo no lineal con un enfoque lagrangiano para las condiciones de no intersección que analiza las figuras por sus vértices. La definición de los elementos a empaquetar por sus vértices no ha sido tratada en la literatura revisada [3].

El presente trabajo está estructurado de la siguiente forma. En la Sección 2 se muestra una descripción de las instancias. A continuación, en la Sección 3, se presenta el diseño de experimento. En la sección 4 se realiza un análisis estadísticos de los resultados para el contenedor tipo sección-circular. La sección 5 analiza los resultados del contenedor circular. En la sección 6 se muestra los resultados del análisis por tipo de contenedor, y finalmente en la sección 7 se presentan las conclusiones alcanzadas.

Email address: alberto.martineznn@uanl.edu.mx (Alberto Martínez-Noa)

2. Descripción de las instancias

Las instancias analizadas en esta investigación fueron creadas, con un generador de instancias programado en lenguaje Python [4], con fin de validar el modelo propuesto. En el cuadro 1 se muestra la cantidad de instancias para los dos tipos de contenedores analizados (sección-circular, circular). Las instancias para cada una de las figuras solamente varían en el número de elementos (5, 6, 7, 8, 9, 10, 15, 20, 25, 30) a empacar. En el caso de los pentágonos existen diez instancias para los regulares (pentágonos$_r$) y la misma cantidad de para los irregulares (pentágonos$_i$).

Cuadro 1: Cantidad de instancias por figura

Figuras	Cantidad de instancias
triángulos	10
rectángulos	10
cuadrados	10
pentágonos$_r$	10
pentágonos$_i$	10
cuadriláteros mixtos	10
hexágonos	10
tetraedros	10

3. Diseño de experimento

En esta experimentación se contemplan dos tipos de contenedores y siete tipos de figuras diferentes a empacar con 10 tamaños de instancias diferentes. Por lo que se utiliza un diseño factorial completo con tres factores de control por tratamiento por lo que se tiene 140 tratamientos. Dichos factores de control son:

- Tipo de contenedor de dos niveles (sección-circular, circular),
- Tipos de figuras de siete niveles (triángulos, cuadrados, rectángulos, pentágonos (regulares e irregulares) y hexágonos, cuadriláteros mixtos),
- Cantidad de elementos de la instancia (5, 6, 7, 8, 9, 10, 15, 20, 25, 30).

4. Análisis estadístico contenedor tipo sección-circular

El objetivo de esta es investigación es determinar si los tipos de figuras, la cantidad de estas y el tipo de contenedor influyen en el porciento de ocupación de los elementos en el contenedor. Para cumplir dicho objetivo realizaremos un análisis estadístico de los datos, dividiéndolos por tipos de contenedor.

Para el analizar de los resultado se pretende usar un análisis de varianza (ANOVA) unidireccional, donde se toma como variable dependiente el porciento de ocupación y como factores el tipo de figura, la cantidad de figuras a empaquetar en cada caso.

Antes de realizar el ANOVA, se debe comprobar tres supuestos: las poblaciones (distribuciones de probabilidad de la variable dependiente correspondiente a cada factor) son normales, las K muestras sobre las que se aplican los tratamientos son independientes y que las poblaciones tienen igual varianza (homoscedasticidad) [2].

Para el primer supuesto se realiza la prueba de Shapiro–Wilk, que calcula un W estadístico que prueba si una muestra aleatoria $x_1, x_2, ..., x_n$ proviene de una distribución normal. El resultado de esta prueba con un **W** = 0,966 y un **valor** p = 0.055 mayor que α = 0,050 muestra que no existe suficiente evidencia para rechazar la hipótesis nula, la cual plantea que la variable dependiente (densidad) sigue una distribución normal. Esto se puede afirmar con un intervalo de confianza del 95 %. En la figura 1 se muestra que todos los puntos caen aproximadamente a lo largo de la línea de referencia, podemos asumir la normalidad.

Dado que los datos se encuentran en el límite para aceptar que se distribuye de forma normal, la prueba de Fisher y el de Bartlett no son recomendables. En su lugar es mejor emplear una prueba basada en la mediana prueba de Levene o prueba de Fligner-Killeen. En dicha prueba con χ^2 = 5.319 y un **valor** p = 0.378 mayor que α = 0,050 arroja que no existe suficiente evidencia para rechazar la hipótesis nula, por lo que variable dependiente (densidad) tiene homoscedasticidad con un intervalo de confianza del 95 %.

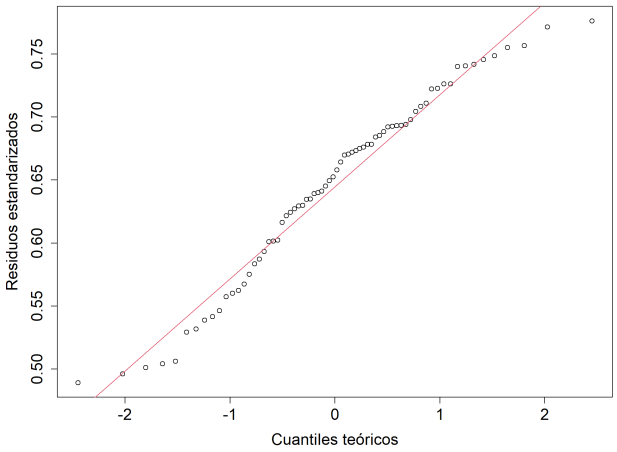


Figura 1: Gráfica Q-Q normal para la variable densidad

Divido a los resultados obtenidos anteriormente se puede utilizar un ANOVA unidireccional, donde se toma como variable dependiente la densidad y como factores el tipo de figura, la cantidad de figuras a empaquetar en cada caso.

Para el primer factor (Tipo de figura) se plantea la siguiente pregunta: ¿Existe diferencia entre el promedio de ocupación de los diferentes tipos de figuras? La respuesta a esta pregunta se obtiene al contrastar las siguientes hipótesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 \quad (1)$$

$$H_1 : \mu_i \neq \mu_j \quad \text{para alguna} \quad i \neq j \quad (2)$$

Con los valores del estadístico de prueba **F** = 1.030 y **valor** p = 0.305 mayor que α = 0,050 no se tiene evidencia suficiente para rechazar la hipótesis H_0 que nos indica que la no existencia diferencias estadísticas significativas entre los grupos con un intervalo de confianza del 95 %. Esto se puede observar en la figura 2.

Análogamente al análisis del primer factor, se realiza la misma prueba para el factor de control cantidad de figuras. Con los valores del estadístico de prueba **F** = 15,380 y **valor** p = 0.000 menor que α = 0,050, se tiene evidencia suficiente para rechazar la hipótesis H_0 ya que existe diferencias estadísticas significativas al menos entre algunos de los grupos con un intervalo de confianza de un 95 %, esto se aprecia en la figura 3.

El ANOVA no ofrece información suficiente para señalar entre que grupos existe la diferencia. Por lo que se aplica una prueba Tukey HSD (*del inglés honestly significant difference*) que muestra las diferencias entre las medias de los grupos. los resultados de la prueba arrogan que se rechaza la hipótesis en los casos de pareo de la instancia de tamaño (10–30, 15–30, 15–5, 15–7, 15–8, 20–5, 20–6, 20–7, 20–8, 25–5, 25–6, 25–7, 25–8, 30–5, 30–6, 30–7, 30–8, 30–9 y 9–8), dado que los **valores** p son menores que α = 0,050, como se aprecia gráficamente en la figura 4.

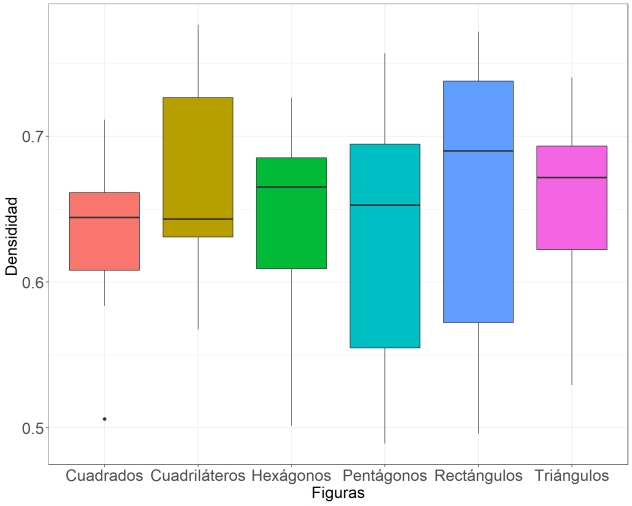


Figura 2: Diagrama de caja y bigotes que relaciona los tipos de figuras y el % de ocupación del contenedor

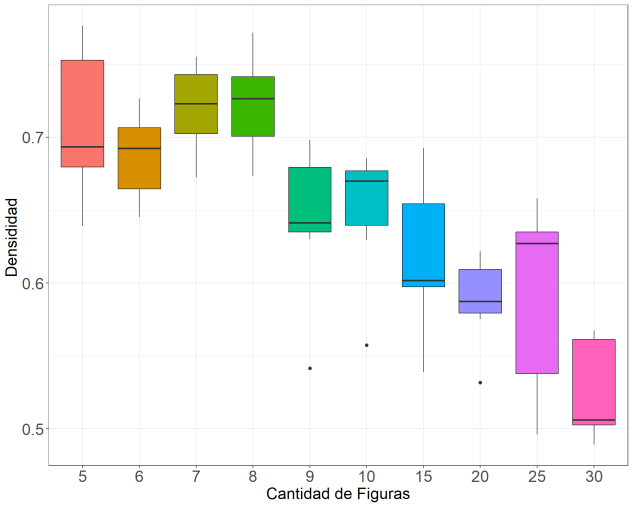


Figura 3: Diagrama de caja y bigotes que relaciona la cantidad de figuras y el % de ocupación del contenedor

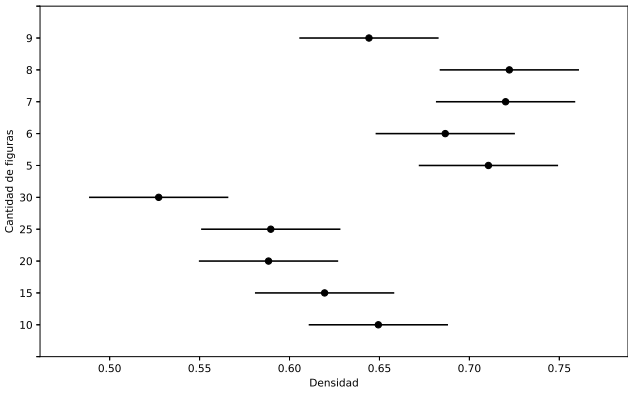


Figura 4: Diagrama simultaneo que relaciona el % de ocupación del contenedor, con los grupos del factor cantidad de figuras

Los resultados obtenidos en esta sección nos indican que para el caso del contenedor tipo sección-circular el tipo de figura a empacar no influye significativamente en el porciento de ocupación del contenedor, por otro lado la cantidad de elementos a empacar si influye en la ocupación del contenedor y se observa que a mayor número de elementos la densidad disminuye.

5. Análisis estadístico contenedor tipo circular

Para el análisis del los resultados obtenidos con el contenedor tipo circular se realiza el mismo procedimiento que en la sección 4.

Al realizar la prueba de Shapiro–Wilk, se obtiene un **W** = 0,928 y un **valor** p = 0.001 mayor que α = 0,050 por lo que existe suficiente evidencia para rechazar la hipótesis que la variable dependiente (densidad) sigue una distribución normal con un intervalo de confianza del 95 %. En la figura 5 se muestra que los puntos no se ajustan a lo largo de la línea de referencia, podemos asumir la falta de normalidad.

Dado que los valores de la variable dependiente no siguen una distribución normal, se procede a aplicar la prueba H de Kruskal-Wallis que es una versión no paramétrica de ANOVA. En este caso se plantea la interrogante: ¿Existe diferencias en el % de ocupación según el tipo de figuras?

Según los datos obtenidos en la prueba, con un valor del estadístico de prueba **H** = 9,639 y **valor** p = 0.086 mayor que α = 0,050, no existe evidencia suficiente para rechazar la nula, la cual indica que no existen diferencias estadísticamente significativas con un intervalo de confianza de un 95 %. Lo anterior se puede observar en la figura 6.

Para el factor cantidad de figuras se tiene como resultado de la prueba de Kruskal-Wallis un **H** = 43,680 y **valor** p = 0.000 menor que α = 0,050, se tiene evidencia suficiente para rechazar la hipótesis H_0 ya que existe diferencias estadísticas significativas al menos entre dos de los grupos con un intervalo de confianza de un 95 %, como se observa en la figura 7.

Al igual que el ANOVA la prueba de Kruskal-Wallis no muestra entre que grupos existen las diferencias. Para saberlo

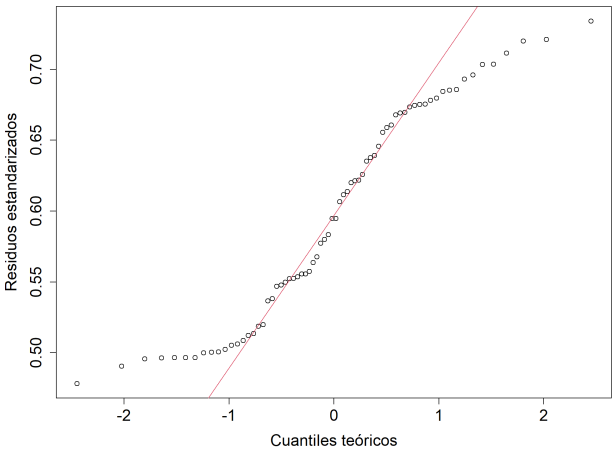


Figura 5: Gráfica Q-Q normal para la variable densidad

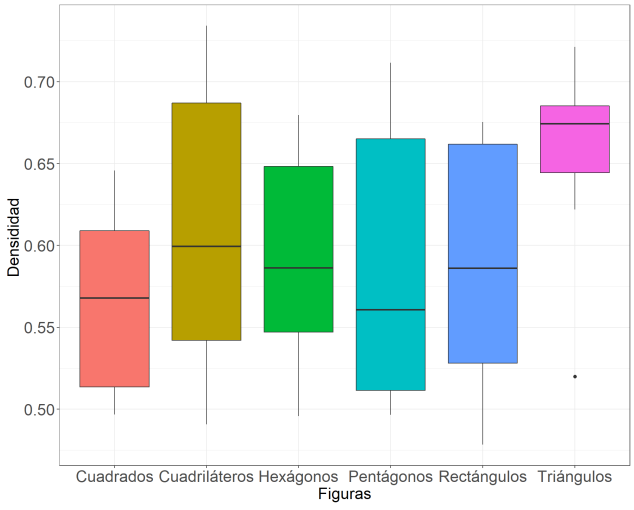


Figura 6: Diagrama de caja y bigotes que relaciona los tipos de figuras y el % de ocupación del contenedor

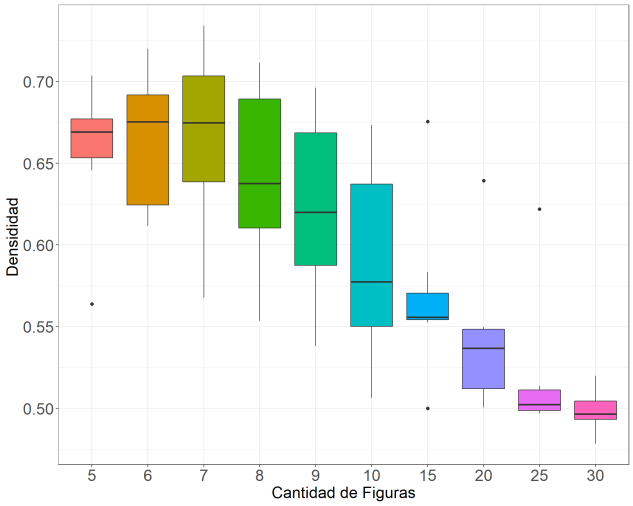


Figura 7: Diagrama de caja y bigotes que relaciona los cantidad de figuras y el % de ocupación del contenedor

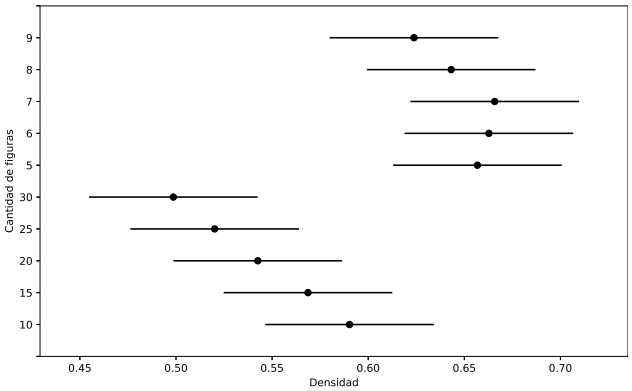


Figura 8: Diagrama simultaneo que relaciona el % de ocupación del contenedor, con los grupos del factor cantidad de figuras

es necesario compararlos todos con todos. Esto implica realizar una corrección del nivel de significancia para evitar incrementar el error de tipo I. El método de comparación *post-hoc* que se utiliza para este caso es la prueba de rango de Tukey con la ficción *kruskalmc()* de R [5], la cual arroja diferencias entre los grupos 5–25, 5–30, 6–25, 6–30, 7–25, 7–30, 8–30 y 9–30. Lo anterior viene dado por los **valores** p del pareo de esos grupos menores que $\alpha = 0{,}050$, por lo que se puede afirmar que existen diferencias estadísticamente significativas entre dichos grupos con un intervalo de confianza del 95 %. Gráficamente se puede ver en la figura 8.

Se puede concluir de la aplicación de estas pruebas que el tipo de figura no influye significativamente en el porciento de ocupación del contenedor tipo circular. Además, se constató que el número de elementos a empacar si influye en dicho porciento.

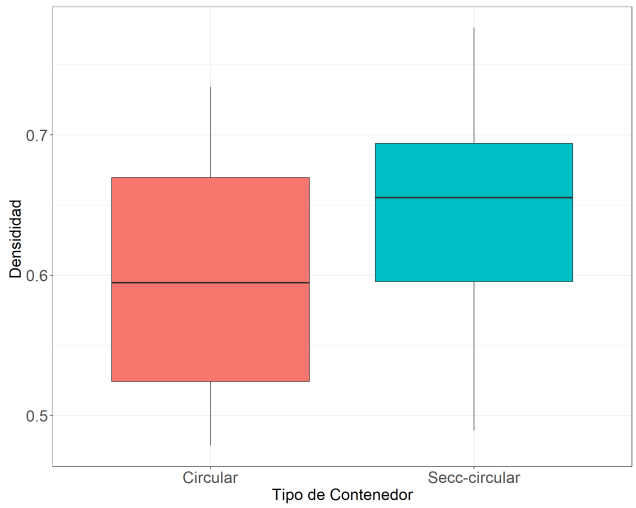


Figura 9: Diagrama de caja y bigotes que relaciona los tipos de contenedores y el % de ocupación del contenedor

6. Análisis estadístico según el tipo de contenedor

Con la aplicación de la prueba de Shapiro–Wilk a la variable dependiente, se obtiene un **W** = 0,955 y un **valor** p = 0.000 menor que α = 0,050 por lo que existe suficiente evidencia para rechazar la hipótesis que la variable sigue una distribución normal con un intervalo de confianza del 95 %.

Por lo que se aplica la prueba no perimétrica H de Kruskal-Wallis con el objetivo de conocer si el tipo de contenedor influye en la densidad del empaquetado. Con un estadístico de pruebas **H** = 12,727 y un **valor** p = 0.000 menor que α = 0,050 se tiene evidencia para decir que el tipo de contenedor si influye en el porciento de ocupación con un intervalo de confianza del 95 %, ya que al solamente haber dos grupos no es necesario realizar la prueba *post-hoc*. En la figura 9 se puede observar lo antes mencionado.

7. Conclusiones

Los métodos estadísticos son muy útiles para poder comprender el comportamiento de los modelos, así como los factores que influyen en su desempeño. Con la aplicación de estos métodos en los resultados obtenidos mediante la experimentación computacional, se pudo identificar que el tipo de figuras a empaquetar no influye significativamente en la densidad de empaque para los dos tipos de contenedores analizados.

En el caso del factor cantidad de figuras se puede concluir que, para los dos tipos de contenedores dicho factor si influye en la densidad del empaque, comprobándose que para instancias más grandes las densidades disminuyen.

Además, se identificó que el tipo de contenedor también influye en el porciento de ocupación del contenedor para el modelo en cuestión.

Referencias

[1] Fowler, R.J., Paterson, M.S., Tanimoto, S.L., 1981. Optimal packing and covering in the plane are np-complete. Information Processing Letters 12, 133 – 137. doi:https://doi.org/10.1016/0020-0190(81)90111-3.

[2] Humberto, G.P., 2008. Análisis y diseño de experimentos. McGraw-Hill Interamericana Editores, S.A. de C.V.

[3] Litvinchev, I., Romanova, T., Corrales-Diaz, R., Esquerra-Arguelles, A., Martínez-Noa, A., 2020. Lagrangian approach to modeling placement conditions in optimized packing problems. Mobile Networks and Applications doi:10.1007/s11036-020-01556-w.

[4] Python Core Team , 2020. Python: A dynamic, open source programming. URL: URLhttps://www.python.org/.

[5] R Core Team, 2020. R: Un lenguaje y un entorno para la informática estadística. URL: https://www.R-project.org/.

[6] Wascher, G., Haubner, H., Schumann, H., 2007. An improved typology of cutting and packing problems. European Journal of Operational Research 183, 1109 – 1130. doi:https://doi.org/10.1016/j.ejor.2005.12.047.