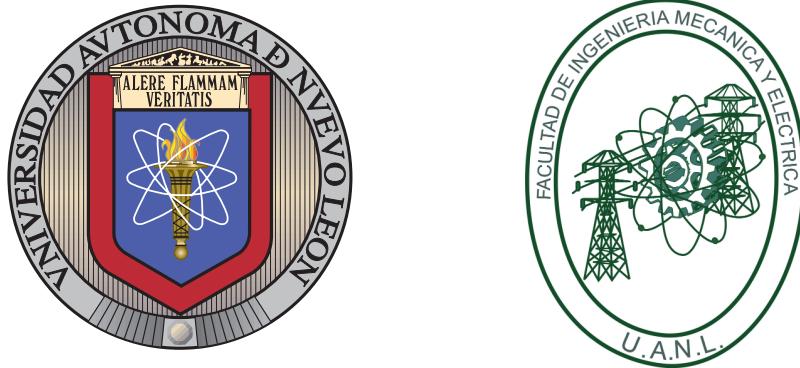


UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE INGENIERÍA MECÁNICA Y ELÉCTRICA
POSGRADO EN INGENIERÍA DE SISTEMAS
DOCTORADO



PORTAFOLIO DE EVIDENCIAS

DE

JOSÉ ALBERTO BENAVIDES VÁZQUEZ

1373079

PARA EL CURSO DE MODELOS PROBABILISTAS APLICADOS,

CON LA PROFESORA ELISA SCHAEFFER.

SEMESTRE AGOSTO 2020 - ENERO 2021.

[HTTPS://GITHUB.COM/JBENAVIDESV87/PROBABILIDAD](https://github.com/jbenavidesv87/probabilidad)

Conceptos básicos de probabilidad

Alberto Benavides
7 de septiembre de 2020

1. SOBRE LOS DATOS

En este reporte se estudia la estadística descriptiva básica de los datos de cantidad de derechohabientes en las instituciones de salud de México durante 2015. Los datos fueron obtenidos del INEGI [1]. Un ejemplo de estos datos se muestra en el cuadro 1.1.

Para el análisis de estos datos, se ha elegido el uso del programa R versión 4.0.2 [4] que se ejecuta en un entorno de Jupyter [2] que se encuentra disponible en <https://tinyurl.com/y52zh15c>.

Los datos obtenidos de la página del INEGI están en formato XLSX, los cuales se editaron con Microsoft Excel [3] para obtener sólo las instituciones y los estados donde se presentan. Los estados son todos los de la república mexicana, mientras que las instituciones son el IMMS, el ISSSTE, las afiliadas a PEMEX, el Seguro Popular y sus derivados actuales y, por último, las instituciones privadas.

Lugar	IMSS	ISSSTE	PEMEX	S. P.	Privada	Otras
Michoacán de Ocampo	954244	265494	13068	2154013	52033	26847
Nuevo León	2973597	197764	21941	905381	424545	137741
Durango	620648	182302	9135	672183	21418	9999

Tabla 1.1: Ejemplo de datos obtenidos del INEGI que contienen la cantidad de derechohabientes por institución de salud en México durante 2015.

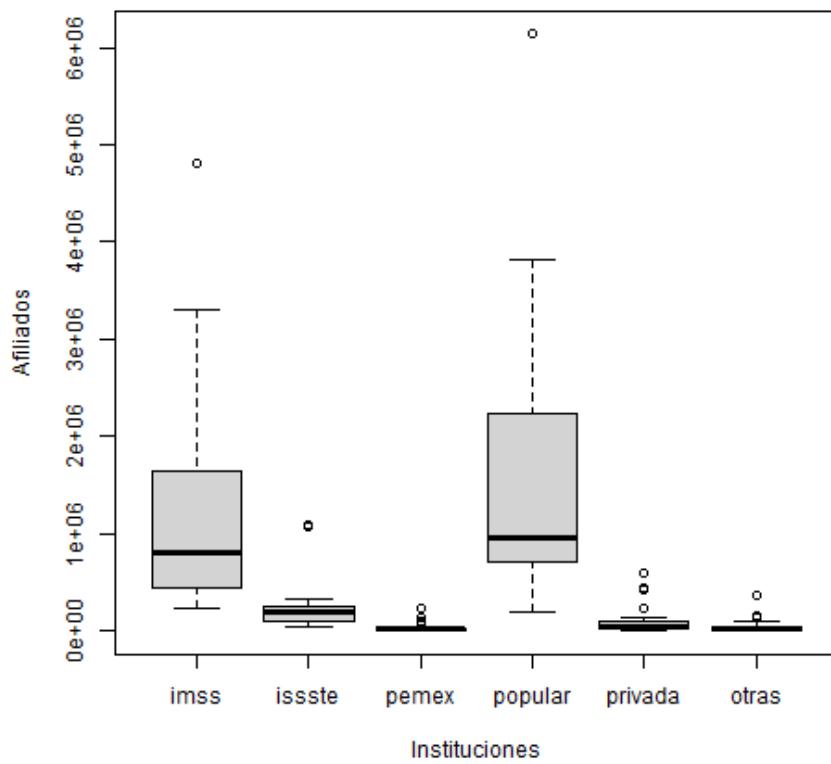


Figura 2.1: Diagramas de cajas y bigotes de la cantidad de derechohabientes por institución de salud en México durante 2015.

2. REPRESENTACIÓN GRÁFICA

Una manera visualmente directa de conocer los datos consiste en crear un diagrama de cajas y bigotes. Estos diagramas resumen la información al graficar el mínimo, máximo, media, primer cuartil y tercer cuartil de los datos. En la imagen 2.1 (p. 2) puede verse este diagrama de la derechohabiencia por institución en México. Cuando las cajas están muy cercanas entre sí, suele convenir hacer un diagrama de cajas y bigotes en escala logarítmica, como se muestra en la imagen 2.2 (p. 3). A partir de estos diagramas se puede ver que el IMSS y el seguro popular son las instituciones con más afiliados. Cabe mencionar que es posible que los afiliados de una institución también lo sean de otras.

Finalmente, también puede obtenerse el diagrama de cajas y bigotes por estado. Éstos se presentan en la figura 2.3 (p. 4). En ésta se puede constatar que entre la Ciudad de México y el estado de México existe una gran cantidad de afiliados.

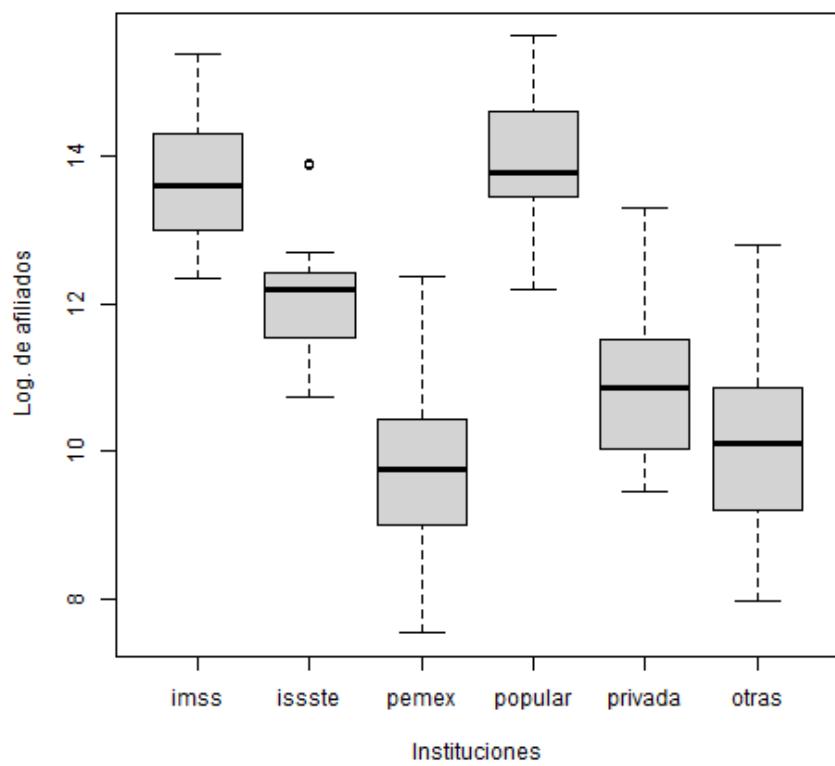


Figura 2.2: Diagramas de cajas y bigotes del logaritmo de la cantidad de derechohabientes por institución de salud en México durante 2015.

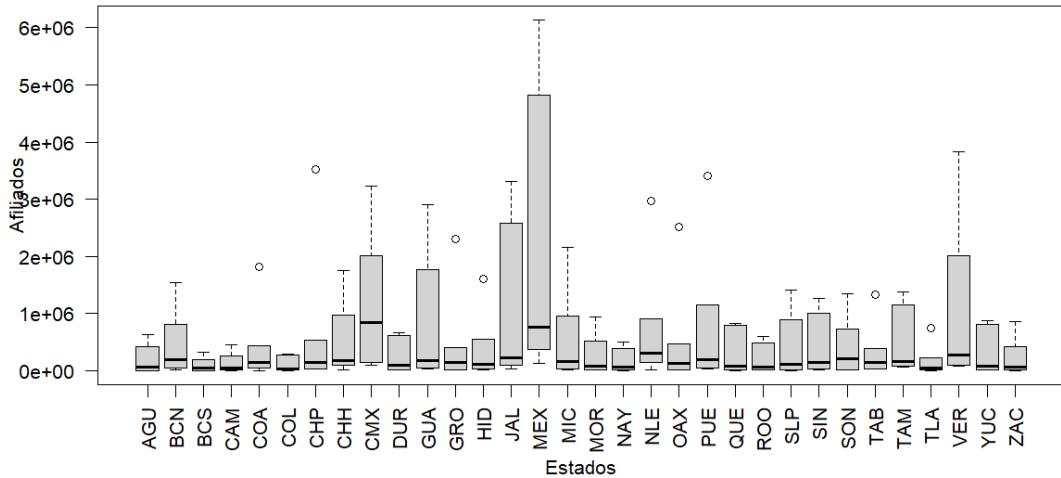


Figura 2.3: Diagramas de cajas y bigotes de la cantidad de derechohabientes por estado de México durante 2015. Los nombres de los estados aparecen abreviados conforme al código ISO 3166-2:MX [5].

REFERENCIAS

- [1] INEGI. Población derechohabiiente por entidad federativa según institución de afiliación, 2015. https://www.inegi.org.mx/app/tabulados/interactivos/?px=Derechohabiencia_02&bd=Derechohabiencia, 2020. [Accedido 2/septiembre/2020].
- [2] Jupyter. Project Jupyter. <https://jupyter.org/>, 2020.
- [3] Microsoft Office. Microsoft excel. <https://www.microsoft.com/es-mx/microsoft-365/excel>, 2020.
- [4] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.
- [5] Wikipedia. Iso 3166-2:mx. https://es.wikipedia.org/wiki/ISO_3166-2:MX, 2019.

Frecuencias

Alberto Benavides
14 de septiembre de 2020

1. INTRODUCCIÓN

Un análisis estadístico básico de un libro consiste en analizar la frecuencia de las palabras con que está escrito. Una guía bastante completa para lograrlo en R [5] puede consultarse en <https://tinyurl.com/y64j1d27>.

2. RECURSOS

El libro de Dracula [4] se encuentra disponible de manera gratuita en el sitio de Project Gutenberg que alberga libros cuya propiedad intelectual ha caducado y, por lo tanto, se trata de obras que forman parte del dominio público. Se puede acceder a los textos planos de estas obras mediante la librería `gutenbergr` [2] de R y la descomposición en palabras de la obra se facilita con el uso de `tidytext` [3] que, a su vez, utiliza para su funcionamiento la librería `dplyr` [6]. Existe un cuaderno de Jupyter [1] donde se trabajaron los datos disponible en <https://tinyurl.com/y3dp3ng1>.

3. PREPROCESAMIENTO

Después de obtenido el libro mediante la función `gutenberg_download`, se analiza su contenido inicial y final con las funciones `head` y `tail`. Esto muestra la existencia de apartados introductorios y anexos que se escapan del contenido del libro en sí, por lo que se procede a remover del contenido esos elementos. Destaca aquí el uso de la función `grep` que permite identificar expresiones regulares y, en este caso, la ubicación del final del libro que coincide con las palabras “THE END”.

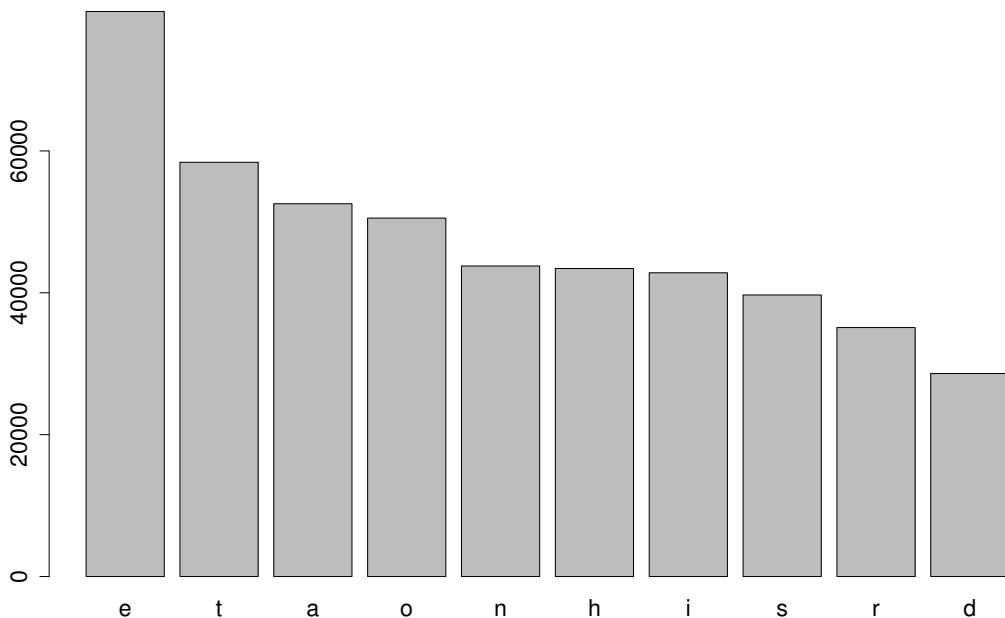


Figura 4.1: Diagrama de barras de las frecuencias de aparición de los caracteres más usados en el libro de *Dracula*.

4. RESULTADOS Y ANÁLISIS

Una vez identificado y aislado el contenido del libro, se extraen las letras y palabras que contiene con la función `unnest_tokens` de la librería `tidytext`. Luego se genera una tabla de frecuencias en orden descendiente por conteo de aparición de las diez letras y palabras más usadas mediante la función `order`. Dichas tablas se grafican en los diagramas de barras de las figuras 4.1 (p. 2) y 4.2 (p. 3).

En cuanto al contexto del libro, esta información no es muy relevante puesto que las palabras que aparecen se denominan palabras vacías (o stop words en inglés) que se refiere a palabras que en un idioma carecen de significado fuera de contexto (más información en <https://tinyurl.com/y2pr7kcb>). En español se incluyen todos las preposiciones, artículos y adverbios. Al eliminar estas palabras, se obtiene una lista de palabras más significativas cuya representación gráfica en una nube de palabras se puede consultar en la figura 4.3 (p. 4).

Asimismo, se hizo una animación de la frecuencia acumulada de las cincuenta palabras más significativas que aparecen en la obra a partir del porcentaje de avance de la misma, de uno en uno por ciento. Esta animación puede verse en <https://tinyurl.com/y2dnxrv8>.

Por último, se analizaron los **bigramas**, o sea las palabras que más comúnmente se encuentran seguidas a lo largo de la obra. Éstos se pueden obtener también con

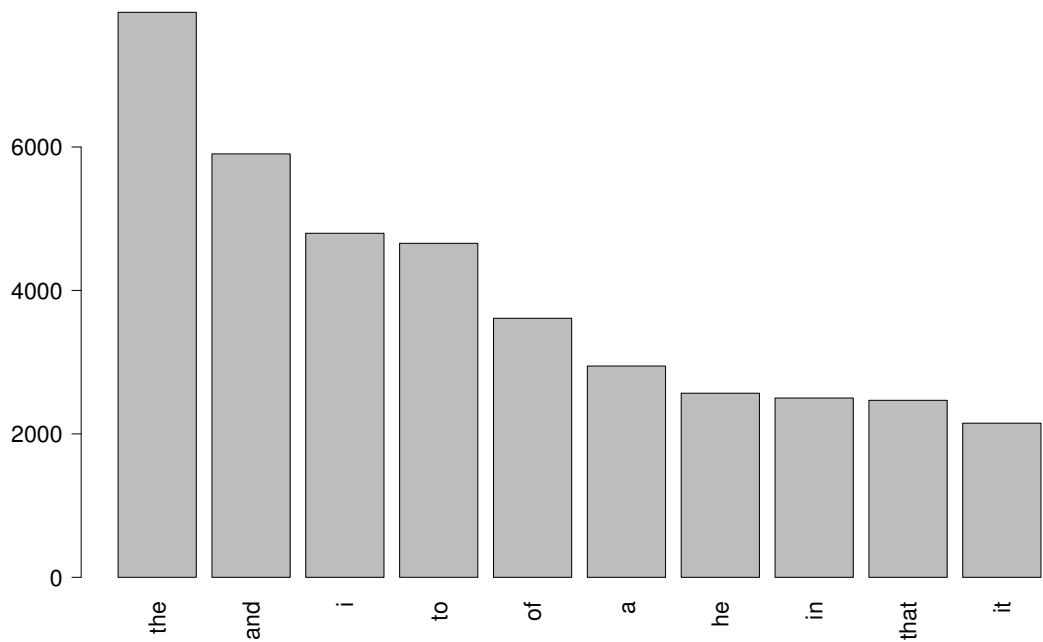


Figura 4.2: Diagrama de barras de las frecuencias de aparición de las palabras más usadas en el libro de *Dracula*.



Figura 4.3: Nube de palabras de las palabras más representativas de la obra *Dracula*.

la función `unnest_tokens(bigram, ...)` (existen dos tutoriales muy completos en <https://tinyurl.com/y5yzdxdm> y <https://tinyurl.com/y2qcvrcq>). El procedimiento consiste en encontrar los bigramas, eliminar aquéllos que contienen palabras vacías y realizar un conteo de los más significativos. Un grafo de estos bigramas con al menos diez apariciones puede observarse en la figura 4.4 (p. 6). Como puede observarse, la mayoría de los bigramas formados pertenecen a nombres de personas o a títulos seguidos de nombres, por ejemplo “Dr. Seward”. Aparte de estos bigramas, destacan los “poor Lucy” o “dear Lucy” personaje por el que podría arrojarse la suposición de que es tan querida como commiserada durante la obra.

REFERENCIAS

- [1] Jupyter. Project Jupyter. <https://jupyter.org/>, 2020.
- [2] David Robinson. gutenbergr. <https://www.rdocumentation.org/packages/gutenbergr/versions/0.1.5>, 2019.
- [3] Julia Silge. tidytext. <https://www.rdocumentation.org/packages/tidytext/versions/0.2.5>, 2020.
- [4] Bram Stoker. Dracula by Bram Stoker. <https://www.gutenberg.org/ebooks/345>, 2020.
- [5] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.
- [6] Hadley Wickham. dplyr. <https://www.rdocumentation.org/packages/dplyr/versions/0.7.8>, 2018.

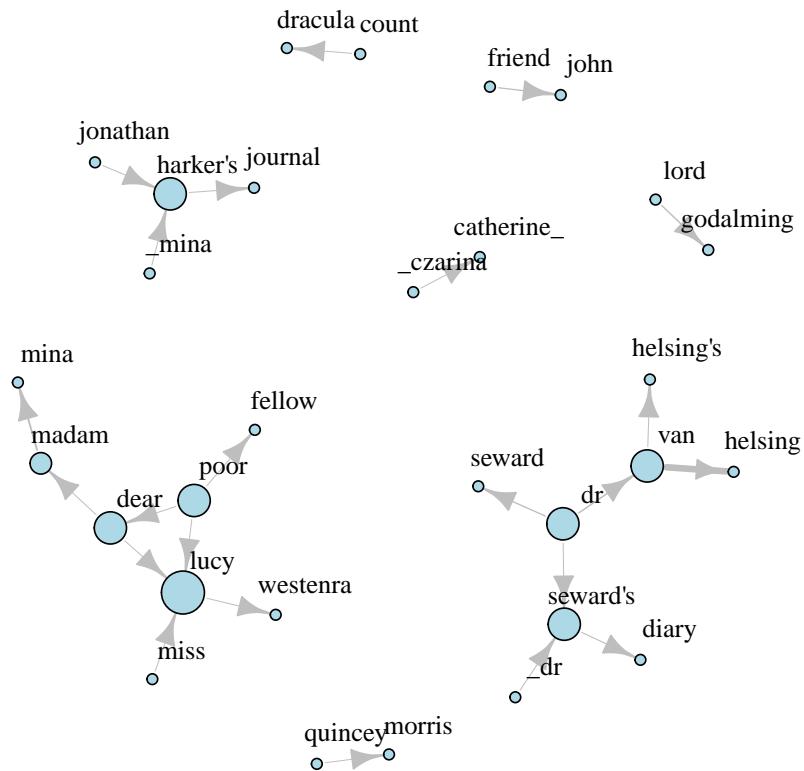


Figura 4.4: Grafo de bigramas más representativos en la obra *Dracula*. Los vértices dirigidos representan la secuencia de los bigramas y el tamaño de los nodos el conteo de cada palabra utilizada en los bigramas.

Distribuciones en oraciones de Dracula

Alberto Benavides

22 de septiembre de 2020

1. INTRODUCCIÓN

En este trabajo me apoyé con Gabriela Sánchez Yépez con repositorio ubicado en <https://github.com/Saphira3000/MPA> para realizar las distribuciones binomial negativa y geométrica del final.

Existen diversos métodos para determinar la autoría de libros. De entre ellos destaca el que utiliza minería de textos combinado con técnicas de agrupamientos como K -medias. Estos métodos utilizan como sus características la cantidad de palabras y comas de las oraciones presentes en las obras, entre otras. Un breve ejemplo sobre esta técnica se halla en <http://www.aicbt.com/authorship-attribution/>.

2. RECURSOS

El libro de *Dracula* [3] escrito por Stoker en [1897] se encuentra disponible de manera gratuita en el sitio de Project Gutenberg [4]. Éste se descarga por la librería `gutenbergr` [2] de R. La separación en oraciones se realiza mediante `corpus` [1].

3. PREPROCESAMIENTO

Después de obtenido el libro mediante la función `gutenberg_download`, se analiza su contenido inicial y final con las funciones `head` y `tail`. Esto muestra la existencia de apartados introductorios y anexos que se escapan del contenido del libro en sí, por lo que se procede a remover del contenido esos elementos. Destaca aquí el uso de la función

`grep` que permite identificar expresiones regulares y, en este caso, la ubicación del final del libro que coindice con las palabras “THE END”.

4. RESULTADOS Y ANÁLISIS

Con el contenido del libro se pueden extraer las oraciones que contiene. Para ello primero es necesario convertir el libro obtenido en una sola cadena de texto, lo que se hace con la función `paste`, seguido del uso de la función `text_split` de la librería `corpus` que separa las oraciones de dicho texto. Se prefiere el uso de esta librería puesto que obvia abreviaturas que usan puntos para evitar separar oraciones por estos motivos. De estas oraciones se cuentan las palabras, las comas, y los puntos y comas que contienen con `strsplit` para la primera y `str_count` para las últimas. En la figura 4.1 se grafican las distribuciones de densidad de las obras de *Dracula* y *Frankenstein* [5]. La última obra fue escrita por Wollstonecraft Shelley en 1831.

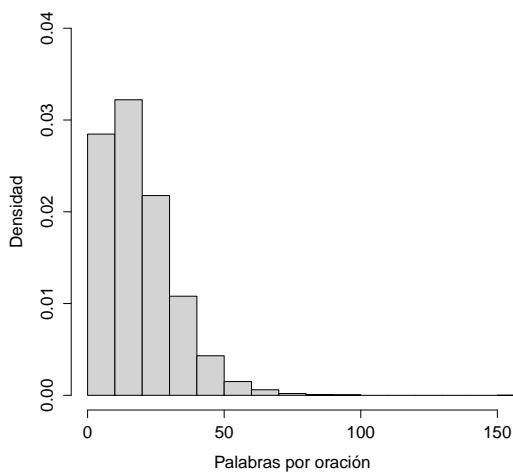
Por último, para la obra de *Dracula* se muestran los histogramas de palabras y comas por oración junto a una función de distribución que se equipara a sus proporciones. Estas gráficas pueden verse en las figuras 4.2 (p. 4) y 4.3 (p. 5), respectivamente. Los procedimientos computacionales y sus códigos pueden revisarse en <https://github.com/jbenavidesv87/probabilidad/blob/master/tema3/tarea.ipynb>

La distribución de palabras por oración que sigue la obra de *Dracula* coincide con una distribución binomial negativa. Este tipo de distribuciones se generan tras obtener, en un transcurso de r_1 repeticiones, las veces que debe repetirse un experimento para que un determinado evento con probabilidad p_1 sea exitoso un total de k veces. En este caso, la distribución binomial negativa generada que equipara los resultados de la distribución del número de palabras por oración, toma los parámetros $r_1 = 90000$, $p_1 = 0.087$ y $k = 2$.

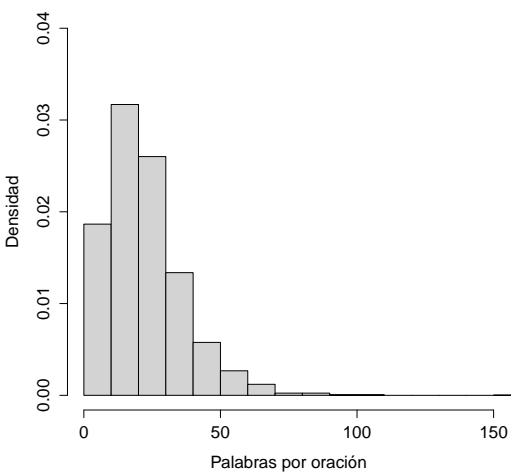
Por su parte, la distribución de comas en una oración de la obra citada coincide con una distribución geométrica que se desarrolla a partir del número de veces que debe hacerse un experimento de probabilidad p_2 (repetido r_2 veces) para que tal experimento sea exitoso una primera vez. Los parámetros que toma esta distribución para ser semejante a la del número de comas por oración en *Dracula* son $p_2 = 0.35$ y $r_2 = 10000$.

REFERENCIAS

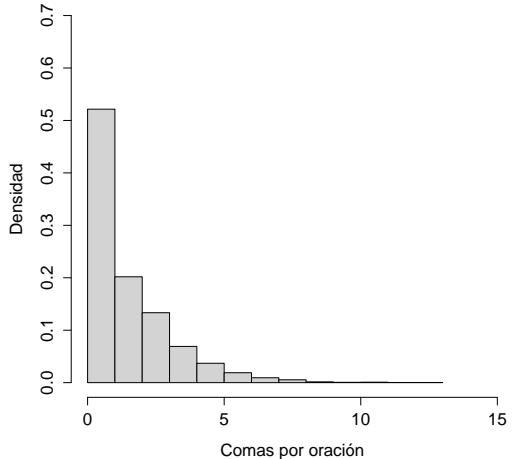
- [1] Leslie Huang. corpus: Text Corpus Analysis. <https://www.rdocumentation.org/packages/tidytext/versions/0.2.5>, 2020.
- [2] David Robinson. gutenbergr. <https://cran.r-project.org/web/packages/corpus/index.html>, 2019.
- [3] Bram Stoker. *Dracula*. Oxford University Press, New York, 1897.
- [4] Bram Stoker. Dracula by Bram Stoker. <https://www.gutenberg.org/ebooks/345>, 2020.
- [5] Mary Wollstonecraft Shelley. *Frankenstein*. Colburn & Bentley, London, 1831.



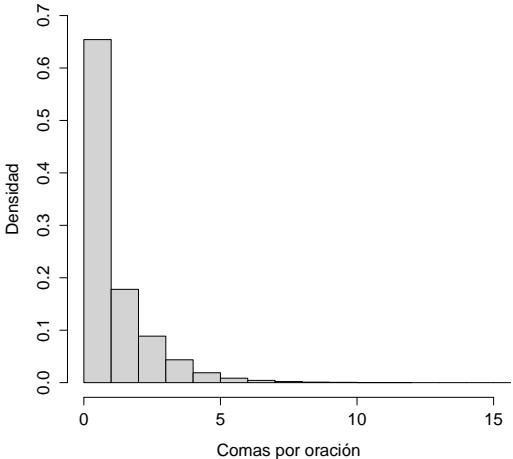
(a) Palabras por oración *Dracula*



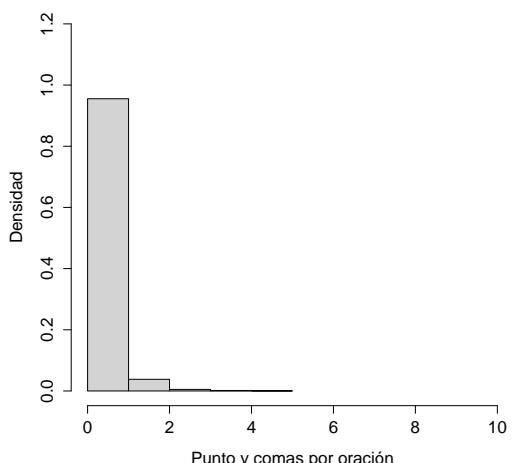
(b) Palabras por oración *Frankenstein*



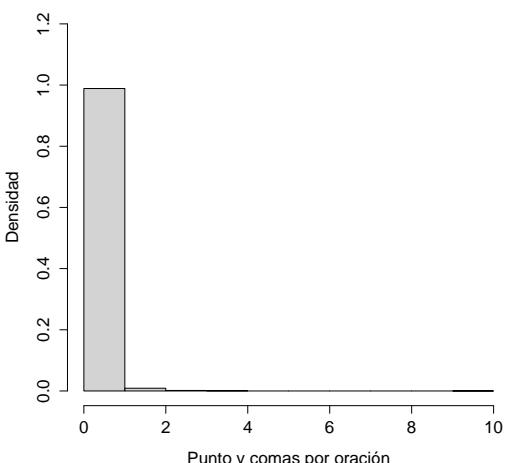
(c) Comas por oración *Dracula*



(d) Comas por oración *Frankenstein*



(e) Punto y comas por oración en *Dracula*



3

(f) Punto y comas por oración en *Frankenstein*

Figura 4.1: Funciones de densidad para las obras de *Dracula* y *Frankenstein*.

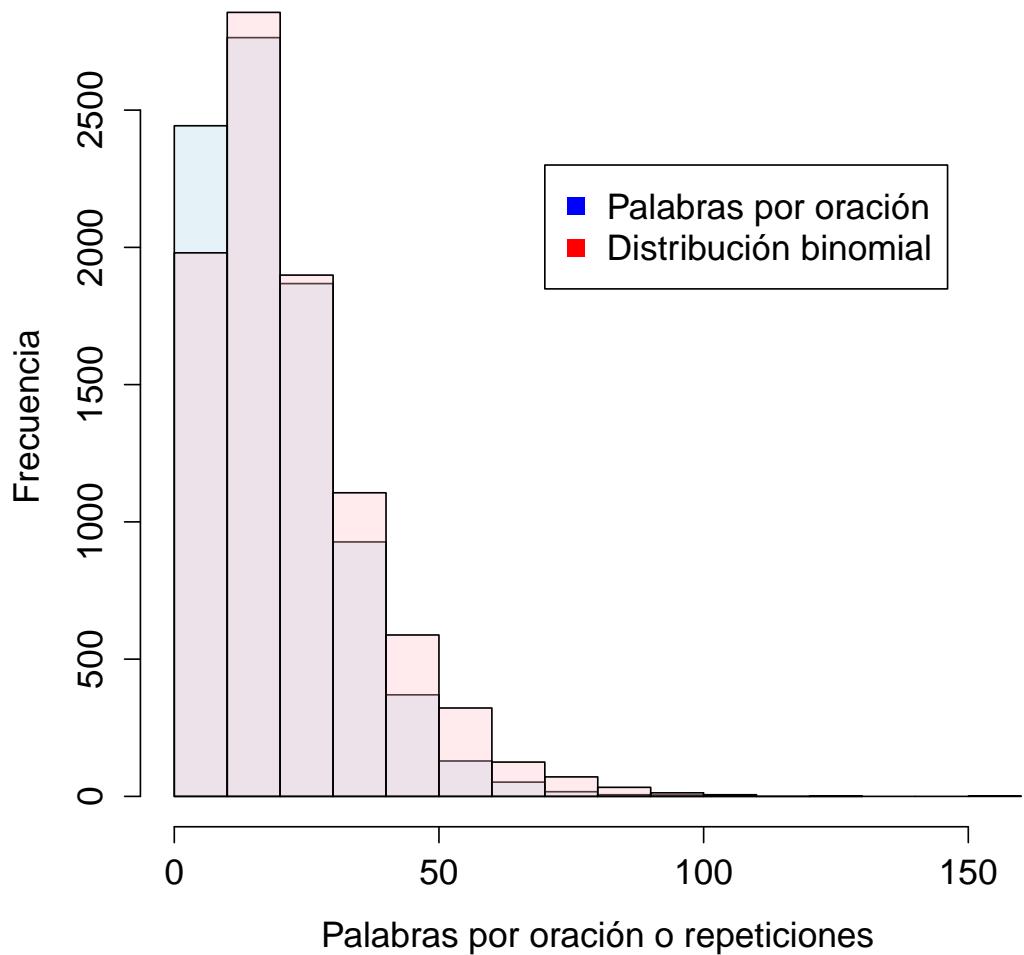


Figura 4.2: Histogramas de palabras por oración de *Dracula* (azul) y función de distribución binomial negativa (rojo).

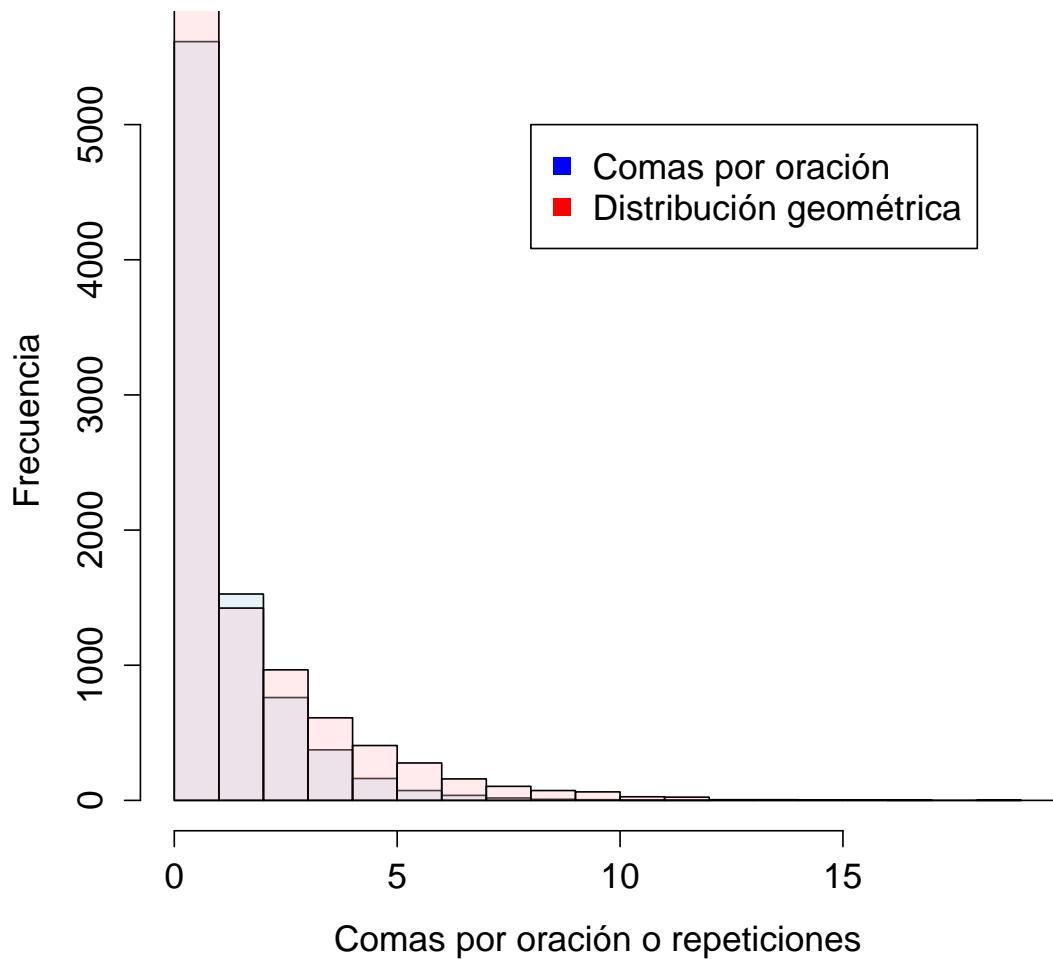


Figura 4.3: Histogramas de comas por oración de *Dracula* (azul) y función de distribución geométrica (rojo).

Distribución de Poisson

Alberto Benavides
29 de septiembre de 2020

1. INTRODUCCIÓN

La distribución de Poisson modela la probabilidad de que un evento con media λ se repita k dada la función

$$P(X = k) = \frac{\lambda^k \times e^{-\lambda}}{k!}. \quad (1.1)$$

En el presente reporte se explorarán maneras de generar esta función de distribución a partir de experimentos computacionales realizados en el lenguaje R [2] ejecutadas en un cuaderno de Jupyter [1].

2. VISUALIZACIÓN DE LA DISTRIBUCIÓN DE POISSON

Se puede utilizar la función `rpois(n, L)` para generar n valores aleatorios obtenidos de una distribución de Poisson con λ igual a L . Un ejemplo de 1000 números generados aleatoriamente de esta manera puede revisarse en la figura 2.1 (p. 2) en forma de histograma.

3. GENERACIÓN A PARTIR DE EXPERIMENTOS COMPUTACIONALES

Dada la definición de la distribución de Poisson como la probabilidad de que sucedan k eventos con media λ , se pueden realizar experimentos computacionales que generen histogramas de distribución similares a los de Poisson a partir de funciones que generen números pseudoaleatorios obtenidos de distribuciones normales, uniformes y exponenciales.

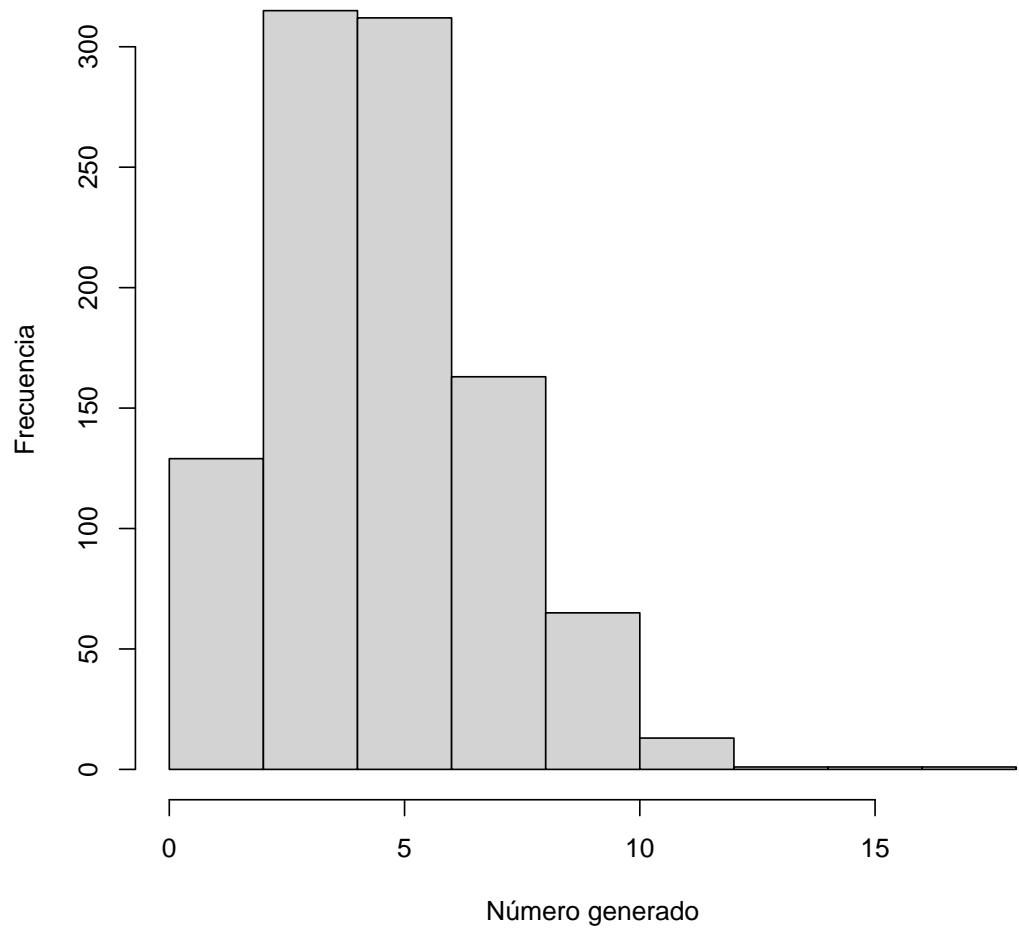


Figura 2.1: Histograma de 1000 valores aleatorios generados por una distribución de Poisson con $\lambda = 5$.

Una manera que se idea para lograrlo consiste en un ciclo que se repita $n = 10000$ veces en el que se sumen $m = 100$ números generados aleatoriamente a partir de una distribución normal con media $\mu = 5$ y desviación estándar $\sigma = 1$. Este procedimiento se refleja en el algoritmo mostrado en [1] y las n sumas generadas se despliegan en el histograma de la figura 3.1 (p. 4) junto al histograma obtenido por la función `rpois(n, m * mu)`. Es importante resaltar en este procedimiento que como ya se conoce de antemano μ y también la cantidad de veces m que se suman números pseudoaleatorios obtenidos de una distribución normal, se puede calcular la media de la distribución generada por la multiplicación de ambos valores, de modo que se puede utilizar en la distribución de Poisson $\lambda = m \times \mu$.

Algoritmo 1: Algoritmo para generar una distribución de Poisson a partir de sumas de números aleatorios generados a partir de una distribución normal.

```

 $n = 10000;$ 
 $m = 100;$ 
 $\mu = 5;$ 
 $\sigma = 1;$ 
resultados = [];
para  $i \in n$  hacer
    | Agregar a resultados  $\sum \mathcal{N}(\mu, \sigma)$ ;
fin
```

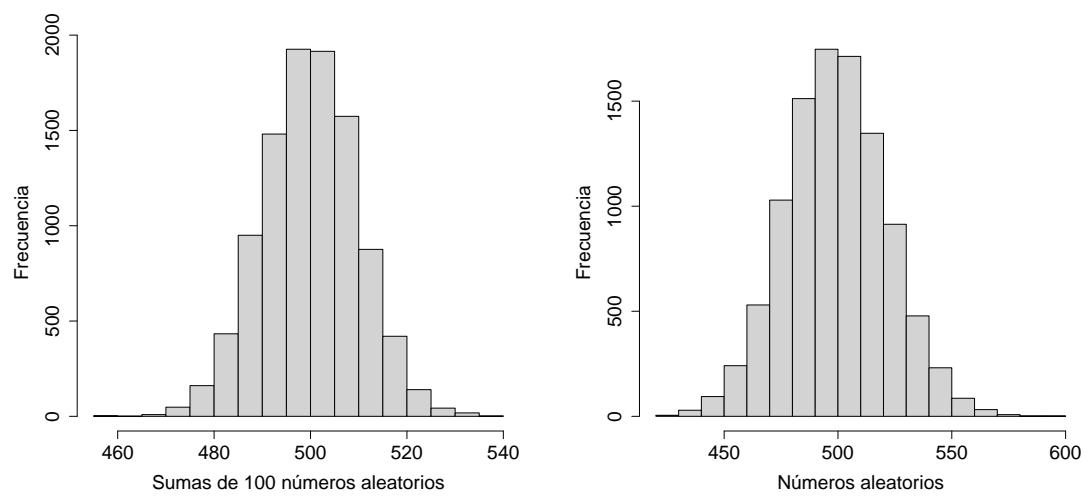
Este mismo procedimiento se puede utilizar con números generados a partir de una distribución uniforme $\mathcal{U}(0, 1)$. La diferencia en este caso es que la media de la suma de los valores debe dividirse entre dos debido a que una distribución uniforme tiene una $E(X) = 0.5$. Los histogramas de este generador y distribución de Poisson se hallan en las figuras 3.2a y 3.2b (p. 5). De manera análoga, se puede utilizar una distribución binomial $\mathcal{B}(m, p)$ en la que la multiplicación de $n \times m \times p$ da por resultado la media de la distribución generada y, por lo mismo, el parámetro λ que usa la distribución de Poisson. Los histogramas de estas funciones pueden verse en 3.2c y 3.2d (p. 5).

4. COMPARACIÓN ENTRE DISTRIBUCIONES

Los histogramas de distribuciones parecen semejantes, sin embargo una manera rápida de comprobar si las distribuciones son iguales consiste en usar diagramas de cajas y bigotes. En la figura 4.1 (p. 6) se puede constatar que la distribución uniforme es la que presenta una distribución distinta a las demás.

REFERENCIAS

- [1] Jupyter. Project Jupyter. <https://jupyter.org/>, 2020.



(a) Generador a partir de distribución normal. (b) Valores generados por distribución $\text{Poisson}(\lambda)$

Figura 3.1: Histograma de 10000 valores generados por la suma de 100 números aleatorios con una distribución $\mathcal{N}(5, 1)$ 3.1a y el obtenido por la función `rpois` de R 3.1b.

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

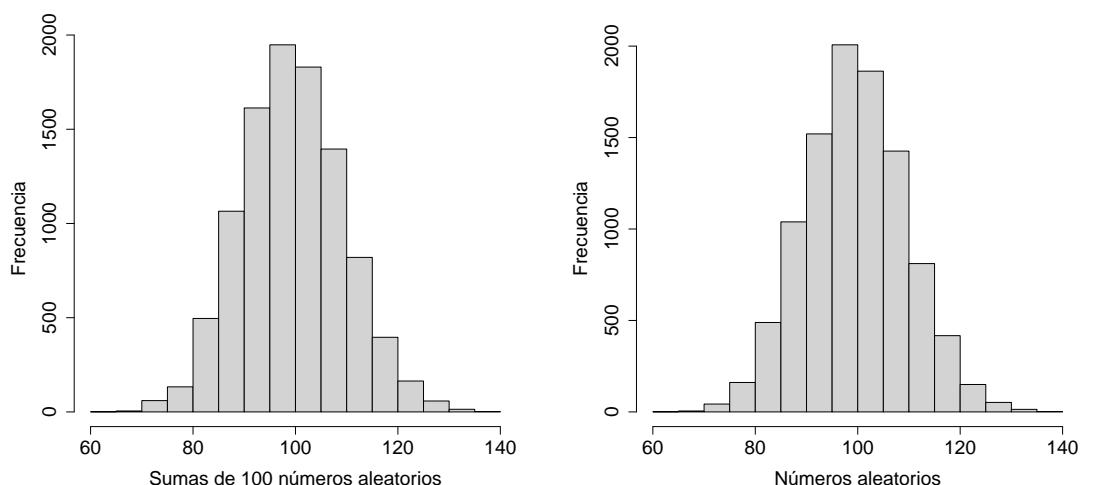
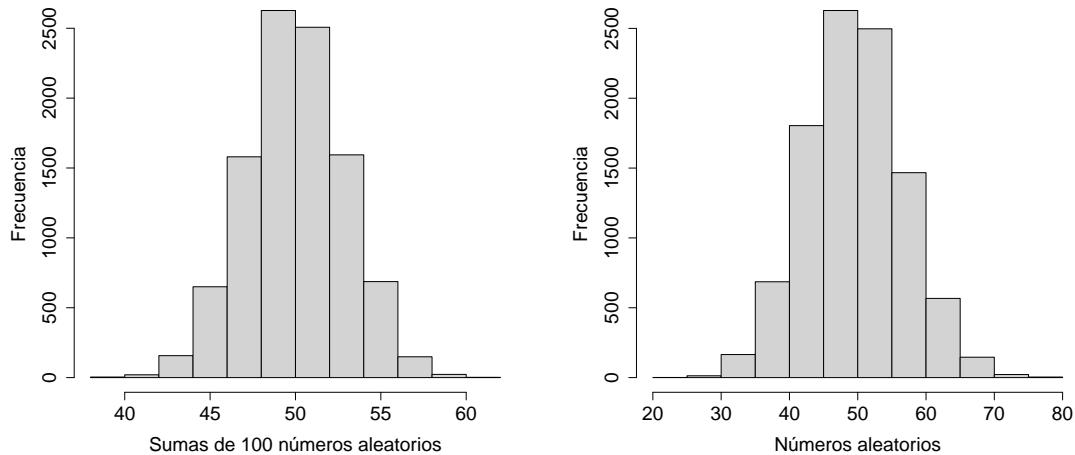


Figura 3.2: Histograma de 10000 valores generados por la suma de 100 números aleatorios con una distribución $\mathcal{U}(0, 1)$ 3.2a, una binomial $\mathcal{B}(100, 0.01)$ 3.2c y los obtenido por la función `rpois` de R para cada una respectivamente, 3.2b y 3.2d.

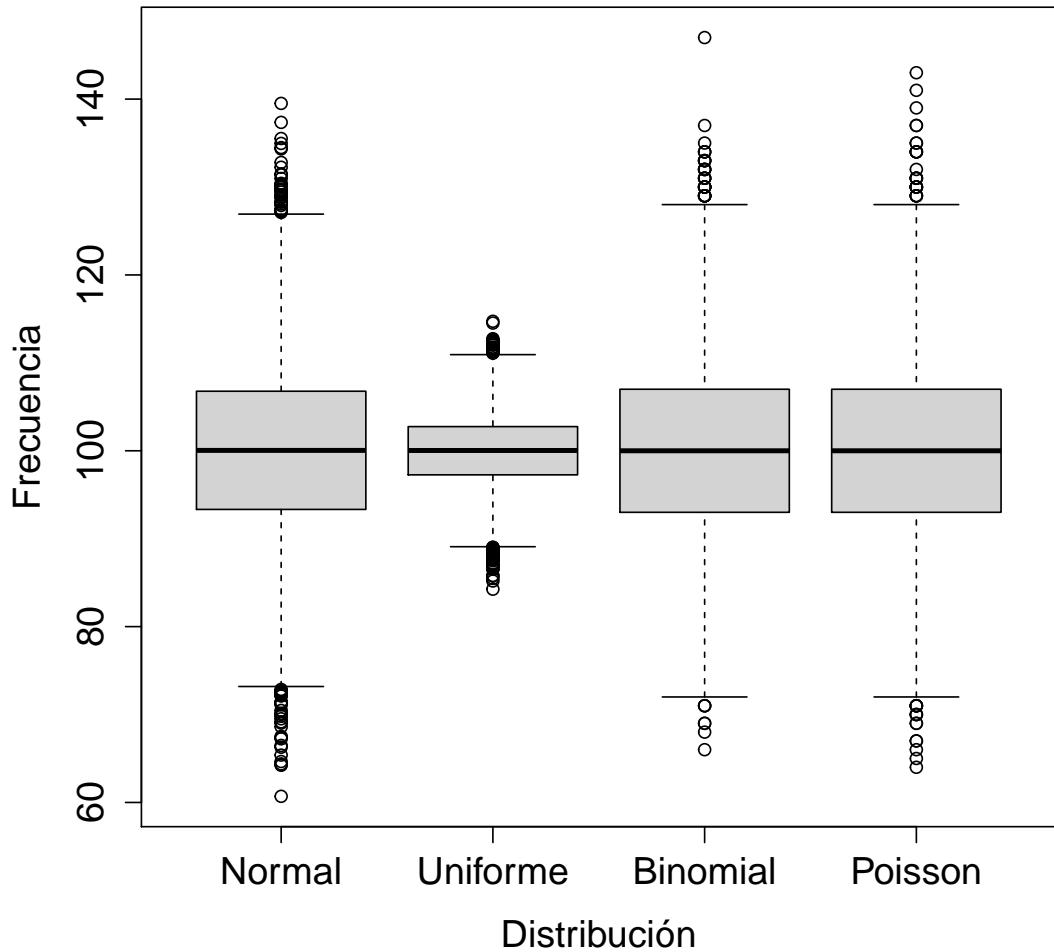


Figura 4.1: Diagramas de cajas y bigotes generados por la función detallada en `[1]` para las distribuciones normal, uniforme y binomial, mientras la distribución de Poisson utiliza la función `rpois` de R.

Generadores pseudoaleatorios

Alberto Benavides
6 de octubre de 2020

1. INTRODUCCIÓN

Una máquina es determinista y, por sí misma, no puede generar números aleatorios. Existen estrategias para conseguirlo, como las utilizadas por RANDOM.ORG [1] que lo hace a partir de ruido ambiental. De todas formas, existen algoritmos que se utilizan para generar números pseudoaleatorios. Entre ellos figuran un generador definido por el *método de Box–Muller* o el llamado *generador lineal congruencial*.

2. MÉTODO BOX–MULLER

El método de Box–Muller parte de dos números pseudoaleatorios u_1 y u_2 obtenidos de una distribución uniforme con media 0 y desviación estándar 1. Estos valores u_1, u_2 son usados para generar dos números independientes

$$z_0 = \sqrt{-2 \log(u_1) \cos(2\pi u_2)}$$

$$z_1 = \sqrt{-2 \log(u_1) \sin(2\pi u_2)}$$

según el pseudocódigo descrito en la entrada dedicada a Box–Muller en Wikipedia [2], que luego serán multiplicados por una desviación estándar σ y se les sumará una media μ para generar números pseudoaleatorios provenientes de una distribución normal.

2.1. DIFERENCIA CUALITATIVA ENTRE z_0 Y z_1

Lo primero que se desea experimentar es si los valores obtenidos a partir de z_0 y z_1 son cualitativamente distintos. Para ello, se realiza un experimento en el que se obtienen 100000 pares de valores obtenidos por este algoritmo con $\mu = 7$ y $\sigma = 3$. Los diagramas de cajas y bigotes, mostrados en la figura 2.1 (p. 3), evidencian la similitud de los valores generados por ambos números z_0, z_1 .

2.2. CAMBIOS EN DISTRIBUCIONES DE u_1 Y u_2

Se probará ahora utilizar otras distribuciones de las que partan los valores de u_1 y u_2 . Se prueban las distribuciones normal, binomial y Poisson, normalizando posteriormente los valores obtenidos de cada distribución entre $[0, 1]$. Se comparan entre sí los diagramas de cajas y bigotes de estos resultados utilizando sólo los generados a partir de las z_0 s de cada distribución, lo cual puede revisarse en la figura 2.2 (p. 4).

2.3. VALORES DEPENDIENTES PARA u_1 Y u_2

Por último, se explora cómo se comportan los números generados cuando u_2 depende de u_1 . Se exploran las siguientes dependencias:

1. $u_2 = 1 - u_1$,
2. $u_2 = u_1/2$,
3. $u_2 = 0.1u_1$,
4. $u_2 = 0.9u_1$.

Los resultados para cada z_0, z_1 de estas dependencias se muestran en la figura 2.3 (5). Como puede constatarse, cuando los valores de u_1 y u_2 son dependientes, los números generados z_0, z_1 tienen distribuciones distintas que varían según la dependencia entre las variables.

3. CONCLUSIONES

Mientras los valores generadores u_1, u_2 sean independientes, los números pseudoaleatorios generados a partir de z_0, z_1 mantienen distribuciones similares, contrario a lo que sucede cuando u_1, u_2 no son independientes entre sí.

REFERENCIAS

- [1] RANDOM.ORG. RANDOM.ORG – True Random Number Service. <https://www.random.org>, 2020.
- [2] Wikipedia. Box–Muller transform. https://en.wikipedia.org/wiki/Box%E2%80%93Muller_transform, 2020.

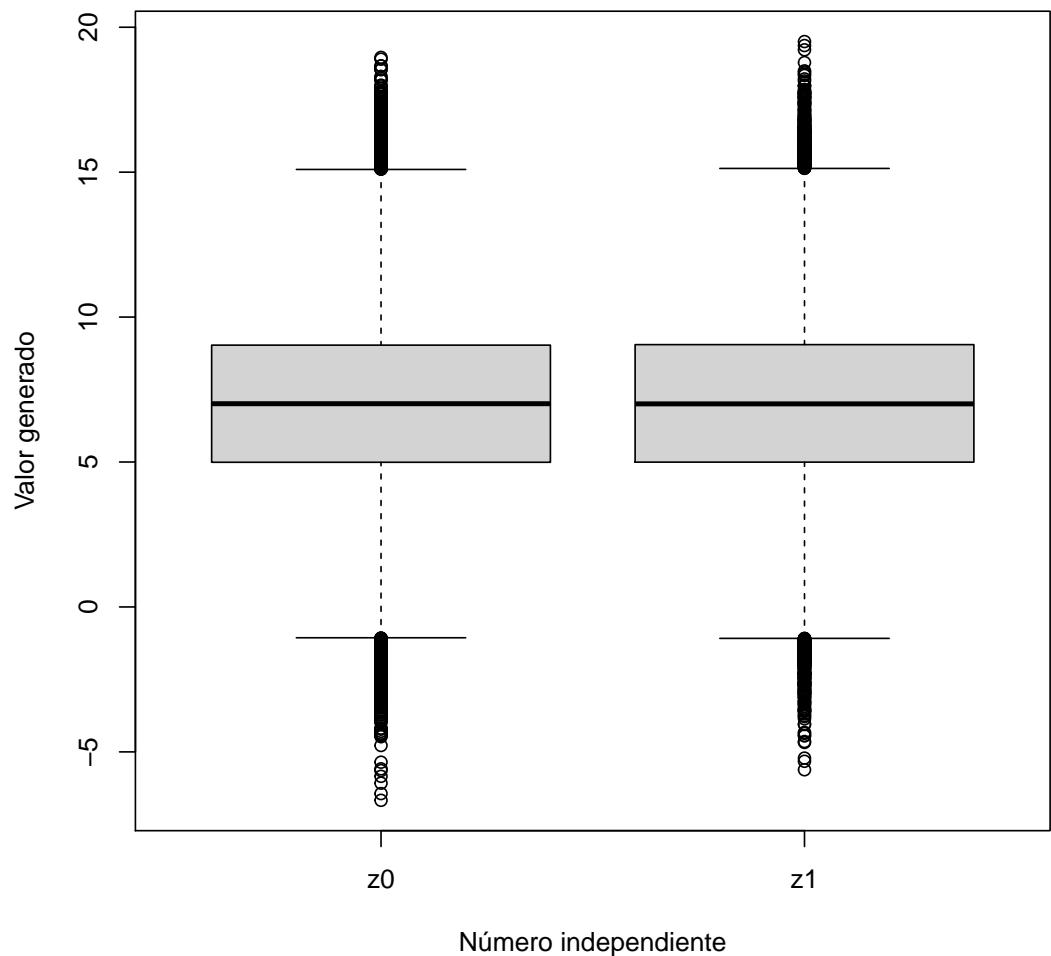


Figura 2.1: Diagramas de cajas y bigotes de 100000 valores generados de dos números independientes z_0, z_1 a partir de un algoritmo que parte del método de Box–Muller.

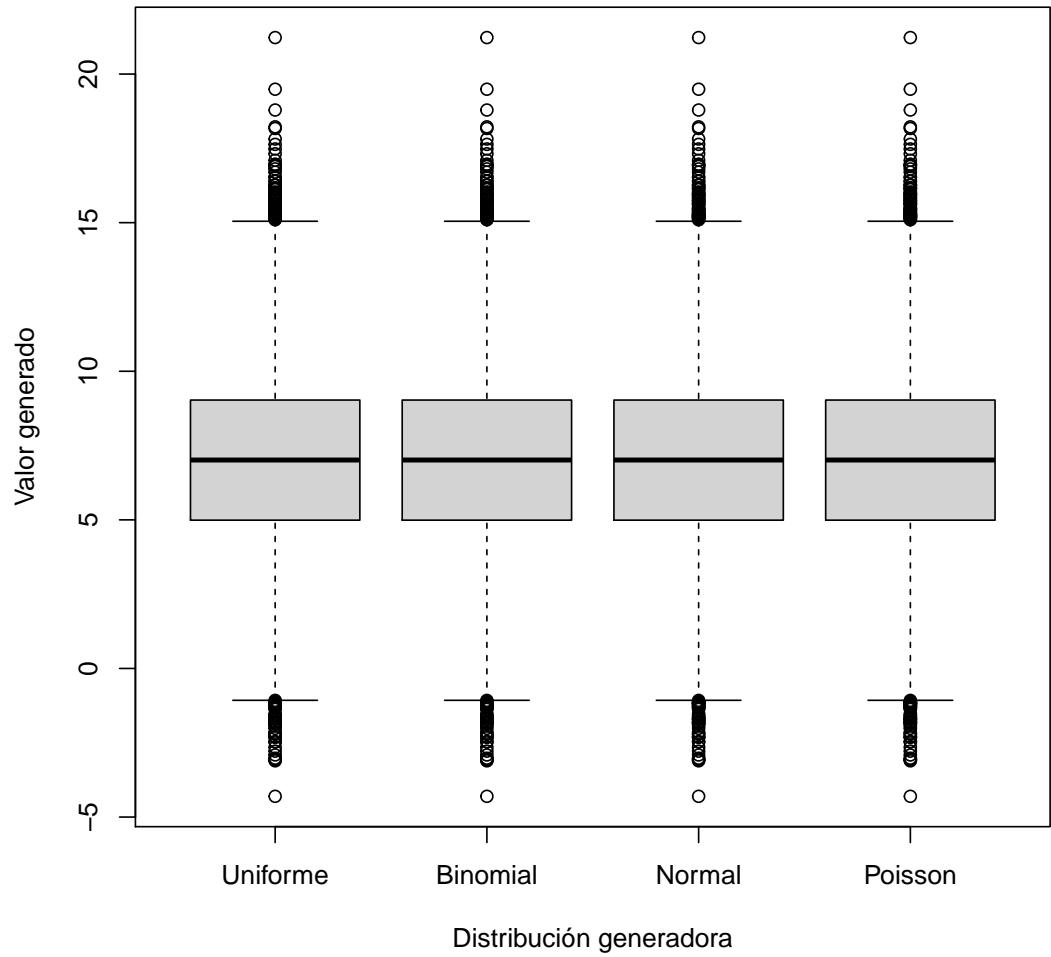


Figura 2.2: Diagramas de cajas y bigotes de 10000 valores generados de un números independientes z_0 a partir de un algoritmo que parte del método de Box–Muller con u_1, u_2 generados a partir de distintas distribuciones.

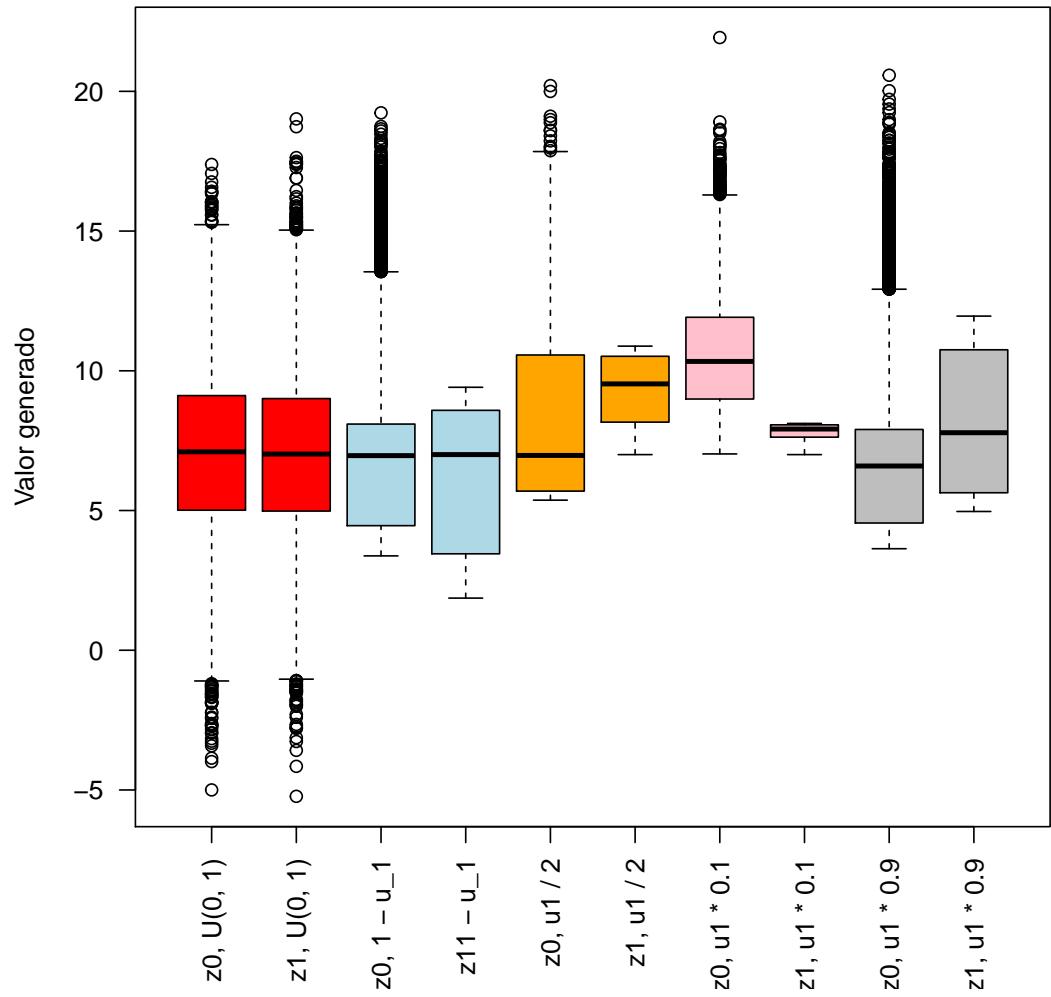


Figura 2.3: Diagramas de cajas y bigotes de 10000 valores generados de un números independientes z_0, z_1 a partir de un algoritmo que parte del método de Box-Muller con u_2 generado a partir de u_1 .

Pruebas estadísticas

Alberto Benavides
13 de octubre de 2020

1. PREGUNTAS

1.1. ¿RELACIÓN ENTRE CONTRASTE DE HIPÓTESIS Y PRUEBAS ESTADÍSTICAS?

Un contraste de hipótesis es un procedimiento en el que se prueba si una población se comporta como se espera a partir de una hipótesis. Existen dos tipos de hipótesis, la *hipótesis nula* H_0 que se quiere comprobar, y la *hipótesis alternativa* H_1 que se contrasta con H_0 . Las pruebas estadísticas ayudan a probar estas hipótesis mediante procedimientos que permiten aceptarlas o rechazarlas a partir de intervalos de confianza, comparaciones entre medias, regresiones, etc.

1.2. ¿QUÉ INDICARÍA RECHAZAR LA HIPÓTESIS NULA?

Dependiendo del intervalo de confianza, indicaría que no es verdadera o que no se tienen suficientes datos para aceptarla como verdadera.

1.3. ¿CÓMO SE INTERPRETA LA SALIDA DE UNA PRUEBA ESTADÍSTICA?

Generalmente se interpreta con un valor p que indica la probabilidad de que los resultados obtenidos por muestreo de la población estudiada sobre la o las variables de interés coincidan con los de H_0 , de modo que obtener valores menores a un α dado, generalmente de 0.05, se traduciría en que es casi imposible que las muestras de la población tengan valores similares a los de H_0 , por lo que se rechazaría ese supuesto.

1.4. ¿CÓMO SELECCIONAR EL α ?

El valor de α generalmente se establece en 0.05, sin embargo pueden usarse otros valores como 0.1 o 0.01 dependiendo del tipo de problema.

1.5. ¿CUÁLES SON LOS ERRORES FRECUENTES DE INTERPRETACIÓN DEL VALOR p ?

Son dos y se les llama *error tipo I* y *error tipo II*. El primero consiste en rechazar la hipótesis nula cuando ésta es verdadera y el segundo en aceptar la hipótesis nula cuando es falsa.

1.6. ¿QUÉ ES LA POTENCIA ESTADÍSTICA Y PARA QUÉ SIRVE?

La potencia estadística se utiliza para calcular la probabilidad de rechazar la hipótesis nula cuando la hipótesis alternativa es verdadera. Si se denomina β la probabilidad de ocurrencia de un error de tipo II, entonces la potencia estadística es $1 - \beta$.

1.7. EJEMPLOS DE PRUEBAS ESTADÍSTICAS PARAMÉTRICAS Y NO PARAMÉTRICAS.

- Paramétricas:
 - Prueba t;
 - análisis de varianza;
 - correlación lineal.
- No paramétricas
 - Wilcoxon,
 - Mann Whitney,
 - Kruskal Wallis.

1.8. ¿CUÁLES SON LOS SUPUESTOS PARA APLICAR TÉCNICAS PARAMÉTRICAS?

Que la población a la que se apliquen los datos debe tener una distribución normal y los errores independientes.

2. EJEMPLO

Se aplican distintas pruebas estadísticas a dos conjuntos de datos obtenidos del SIMA Sistema Integra del Monitoreo Ambiental [1] que contienen registros de calidad del aire para partículas de PM10 medidos en dos estaciones de monitoreo, la Norte y la Sur, ubicadas en Nuevo León, México en los municipios de Escobedo y Santiago, respectivamente. Una visualización de los datos en diagramas de cajas y bigotes junto a escalas de calidad del aire compartidas por el SIMA aparece en la figura 2.1 (p. 3).

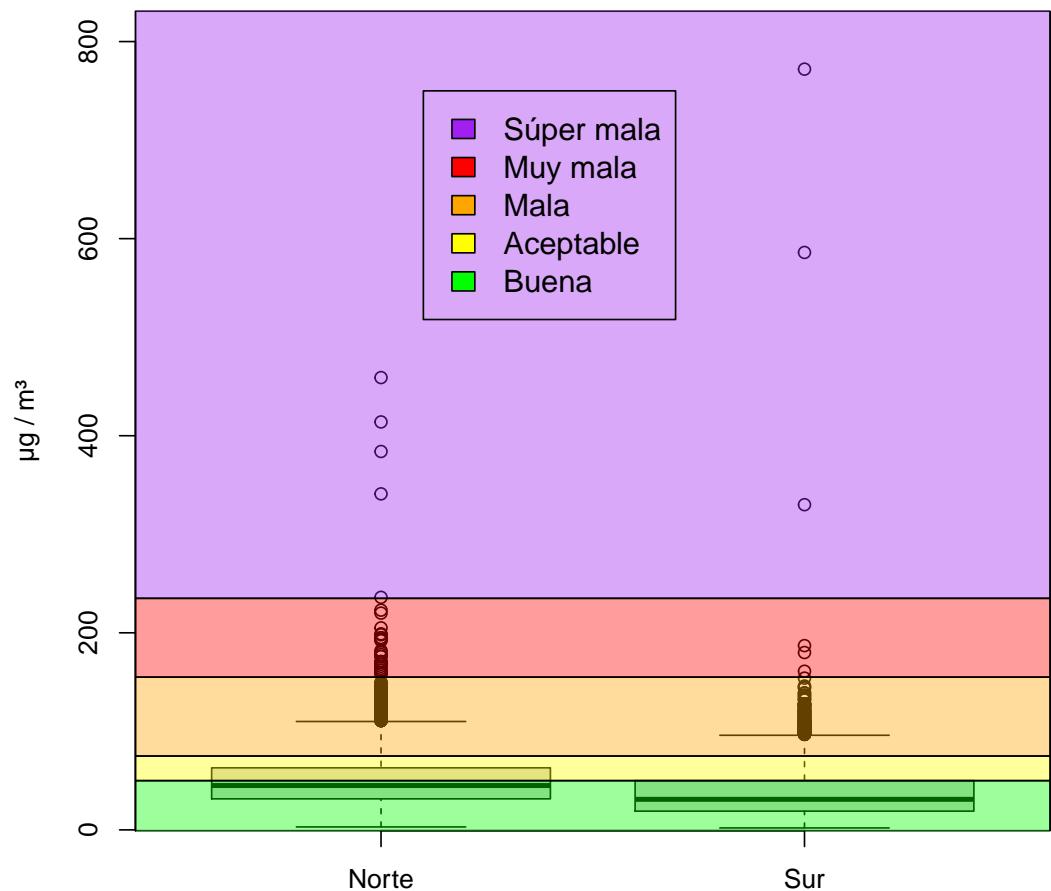


Figura 2.1: Diagramas de cajas y bigotes de las concentraciones de contaminantes con tamaño PM10 para las estaciones Norte y Sur durante el año 2018, medidas en $\mu\text{g}/\text{m}^3$.

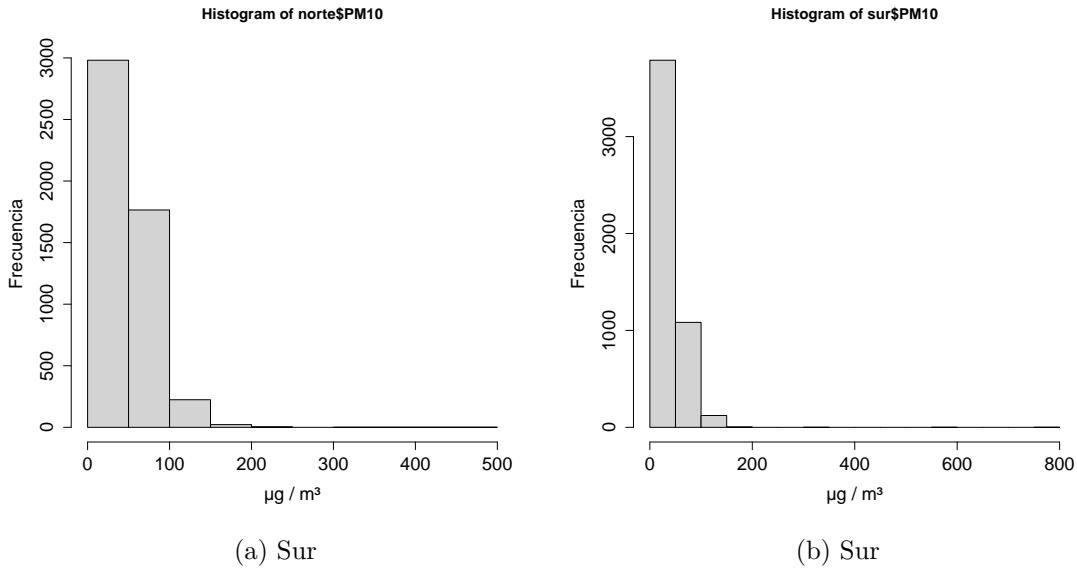


Figura 2.2: Histogramas para las concentraciones de contaminantes de tamaño PM10 para las estaciones Norte y Sur.

A estos conjuntos de datos se les aplica una prueba de Shapiro para determinar si tienen distribución normal, mas es sólo con fines demostrativos, pues los histogramas ya evidencian una distribución geométrica para ambos conjuntos de datos, como se muestra en la figura 2.2 (p. 4). Los valores p para ambos conjuntos de datos dan 2.2×10^{-6} por lo que se puede rechazar la hipótesis nula de que sigan distribuciones normales.

Ahora se utiliza una prueba para determinar si de alguno de los conjuntos de datos se puede extraer una media dada. Esto se hace mediante una prueba t cuando los datos están normalmente distribuidos, pero como estos conjuntos no lo están, se utiliza la prueba de Wilcoxon. Como valor de media se elige 50 por ser el valor máximo del rango de valores de buena calidad del aire. La prueba tiene valores $p < 0.01$ para ambos conjuntos, por lo que se puede rechazar la hipótesis nula en ambos casos, además de que estima medias para ambos conjuntos de $47 \mu\text{g}/\text{m}^3$ y $34 \mu\text{g}/\text{m}^3$, por lo que se podría decir que en 2018 la calidad del aire en cuanto a este tipo de contaminante fue bastante buena. Sin embargo, cuando se comparan las medias para ambos conjuntos de datos mediante Wilcoxon, con el valor $p = 2.2 \times 10^{-16}$, se puede concluir que las medias son distintas, lo cual también se había constatado visiblemente en los diagramas de cajas y bigotes de la figura 2.1 (p. 3).

Ahora bien, mediante una prueba F puede determinarse si ambos conjuntos tienen la misma varianza, sin embargo se puede concluir que no es así dado que el valor $p = 0.07$, por lo que se rechaza la hipótesis nula de esta prueba que supone ambas distribuciones tienen la misma varianza. Por último, resulta interesante determinar si las estaciones tienen alguna relación entre sí. Para ello se puede realizar una prueba de correlación entre ambos conjuntos que tiene como hipótesis nula que la correlación entre los conjuntos es igual a 0. El valor p obtenido para esta prueba fue 1.99×10^{-5} , de modo que se puede

afirmar que existe una correlación estadísticamente significativa entre los conjuntos de datos. Se muestra un diagrama de dispersión del logaritmo de los dos conjuntos de datos con un valor de correlación de 0.07 en la figura 2.3 (p. 6).

REFERENCIAS

- [1] Sistema Integra del Monitoreo Ambiental. aire.nl.gob.mx. <http://aire.nl.gob.mx/>, 2019.

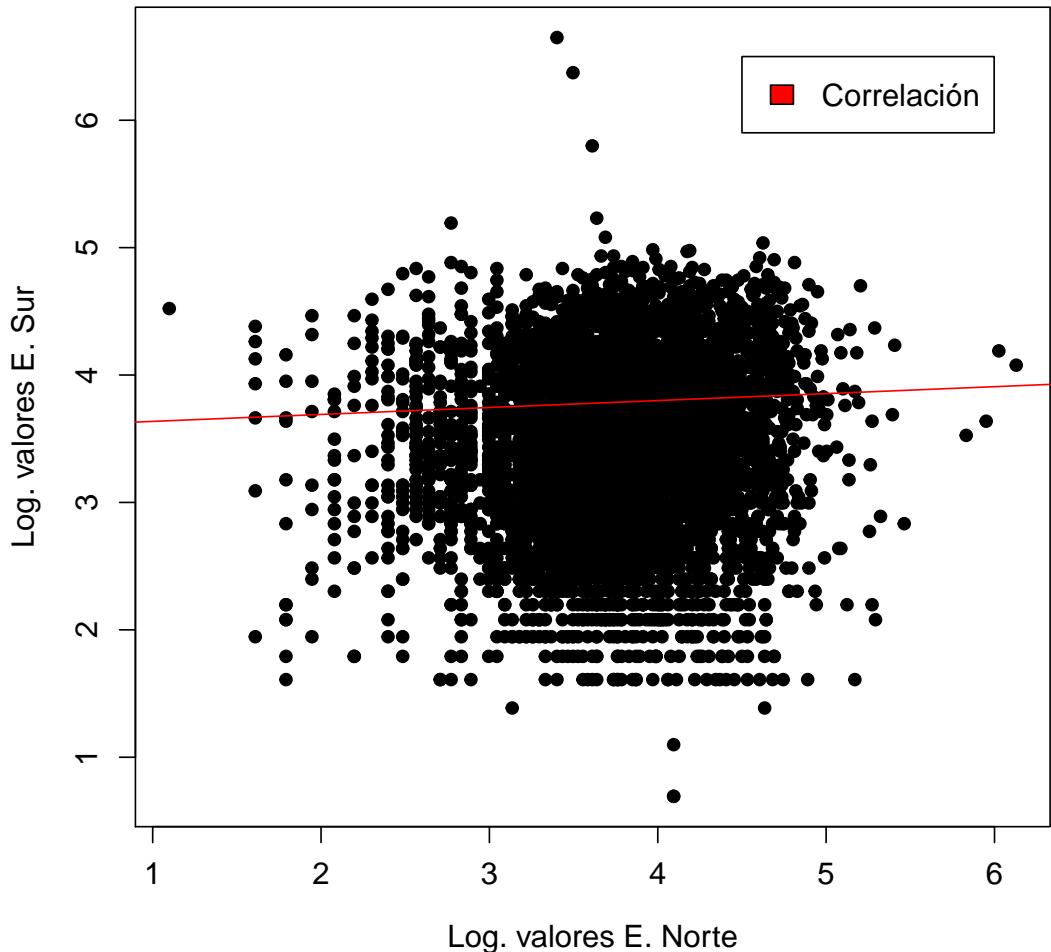


Figura 2.3: Diagramas de cajas y bigotes de las concentraciones de contaminantes con tamaño PM10 para las estaciones Norte y Sur durante el año 2018, medidas en $\mu\text{g}/\text{m}^3$.

Transformaciones de datos

Alberto Benavides
20 de octubre de 2020

1. INTRODUCCIÓN

En esta práctica se ha desarrollado una metodología computacional que permite comparar las correlaciones de ciertas funciones después de transformar sus variables de entrada y salida mediante las transformadas de Tukey y de Box–Cox.

2. TRANSFORMADAS

Las *transformadas de Tukey* [1] parten de la idea de transformar variables independientes X o dependientes $Y = f(X)$ en potencias de dichas variables X^λ, Y^λ a partir de los distintos valores que pueda tomar λ , tal que

$$z_\lambda = \begin{cases} x^\lambda, & \text{si } \lambda > 0, \\ \log(x), & \text{si } \lambda = 0, \\ -(x^\lambda), & \text{si } \lambda < 0. \end{cases} \quad (2.1)$$

Por otro lado, las *transformadas de Box–Cox* [2], por su cuenta, se realizan a partir de la ecuación

$$z_\lambda = \frac{x^\lambda - 1}{\lambda}. \quad (2.2)$$

3. CORRELACIONES

Ambas transformadas suelen usarse para mejorar la correlación de las funciones resultantes con respecto a la función original. Para ello, es posible transformar sólo X, Y

o ambas al mismo tiempo y luego calcular las correlaciones entre dichas variables. Así, una manera de encontrar la mejor transformada para una determinada función sería definir algunos valores de λ , transformarla mediante ambas transformadas y graficar las correlaciones de las funciones con transformaciones en X , Y o ambas variables.

4. DISEÑO DE EXPERIMENTOS

Se realiza un diseño de experimentos para comparar las correlaciones de las funciones transformadas a partir de las transformaciones de Tukey y Box–Cox. Como ejemplo, se utilizan las funciones

- $1/x$,
- x^2 ,
- x^3 ,
- $\log(x)$,
- e^x ,
- $\sin(x \cdot \frac{180}{\pi})$,
- $\cos(x \cdot \frac{180}{\pi})$,
- $\tan(x \cdot \frac{180}{\pi})$.

La variable $\lambda = [-3.0, -2.5, -2.0, \dots, 2.0, 2.5, 3.0]$, mientras que las transformaciones se aplican sobre sólo X , sólo Y y ambas simultáneamente. Se generan mil valores de X a partir de una distribución uniforme $\mathcal{U}(-100, 100)$ y se calculan Y para cada función. Luego, se grafican las funciones de las transformadas de Tukey y Box–Cox para cada una de estas variantes. Por último, se grafican las correlaciones con los distintos λ tanto para las transformaciones de Tukey como las de Box–Cox.

5. RESULTADOS

Se muestran algunos ejemplos de esta práctica. Primero $1/x$. La función se muestra en la figura 5.1 (p. 3). El despliegue de correlaciones para las transformaciones de X , Y y ambas se muestra en la figura 5.2 (p. 4). En dicha figura se puede apreciar que los valores de las correlaciones para ambas transformadas con los mismos valores λ son iguales. Además, una animación de las diferentes transformaciones a lo largo de los cambios en λ puede consultarse en <https://tinyurl.com/yxjepzdf>, <https://tinyurl.com/y468ba9o> y <https://tinyurl.com/y2e6a6w2> para las transformaciones de X , Y y ambas, respectivamente. El resto de las funciones, comparaciones entre correlaciones, animaciones y el código se hallan en <https://github.com/jbenavidesv87/probabilidad/tree/master/tema7>.

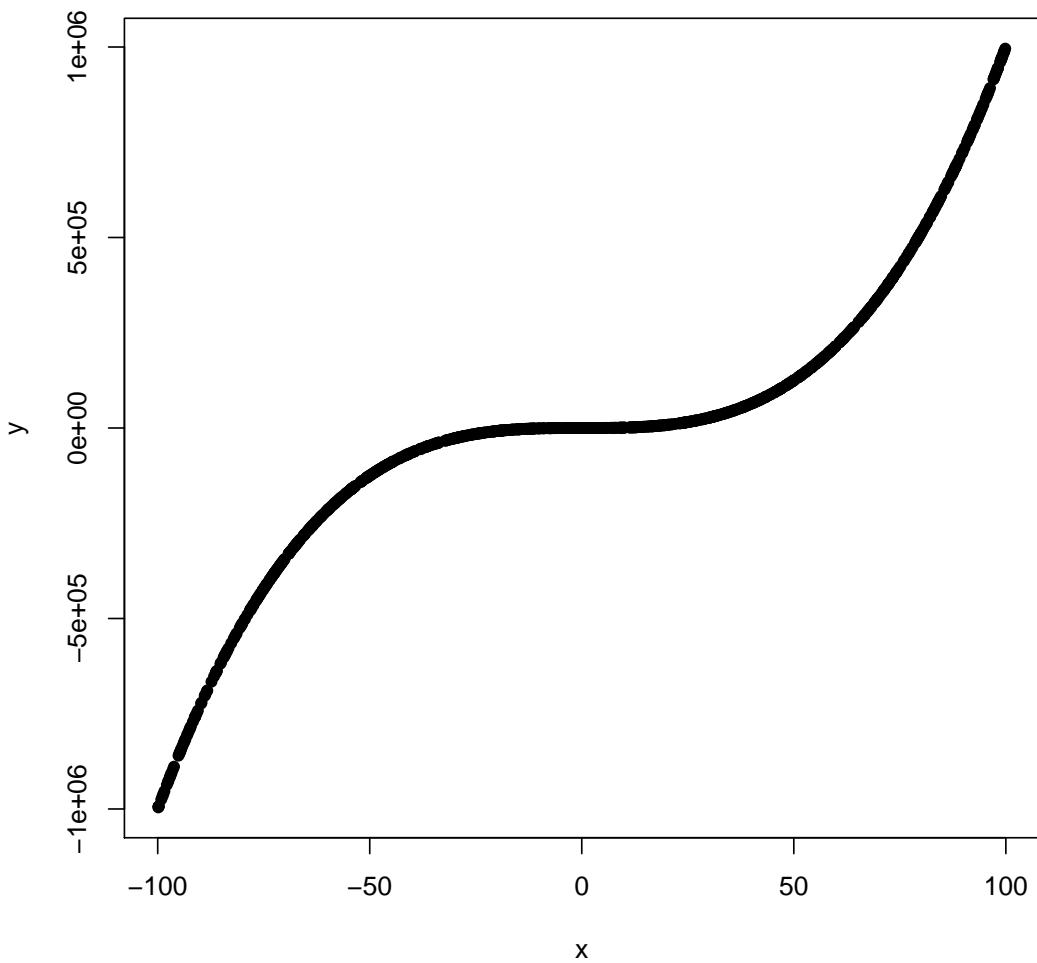


Figura 5.1: Función $1/x$ a partir de mil valores de X desde una distribución uniforme con valores $[-100, 100]$.

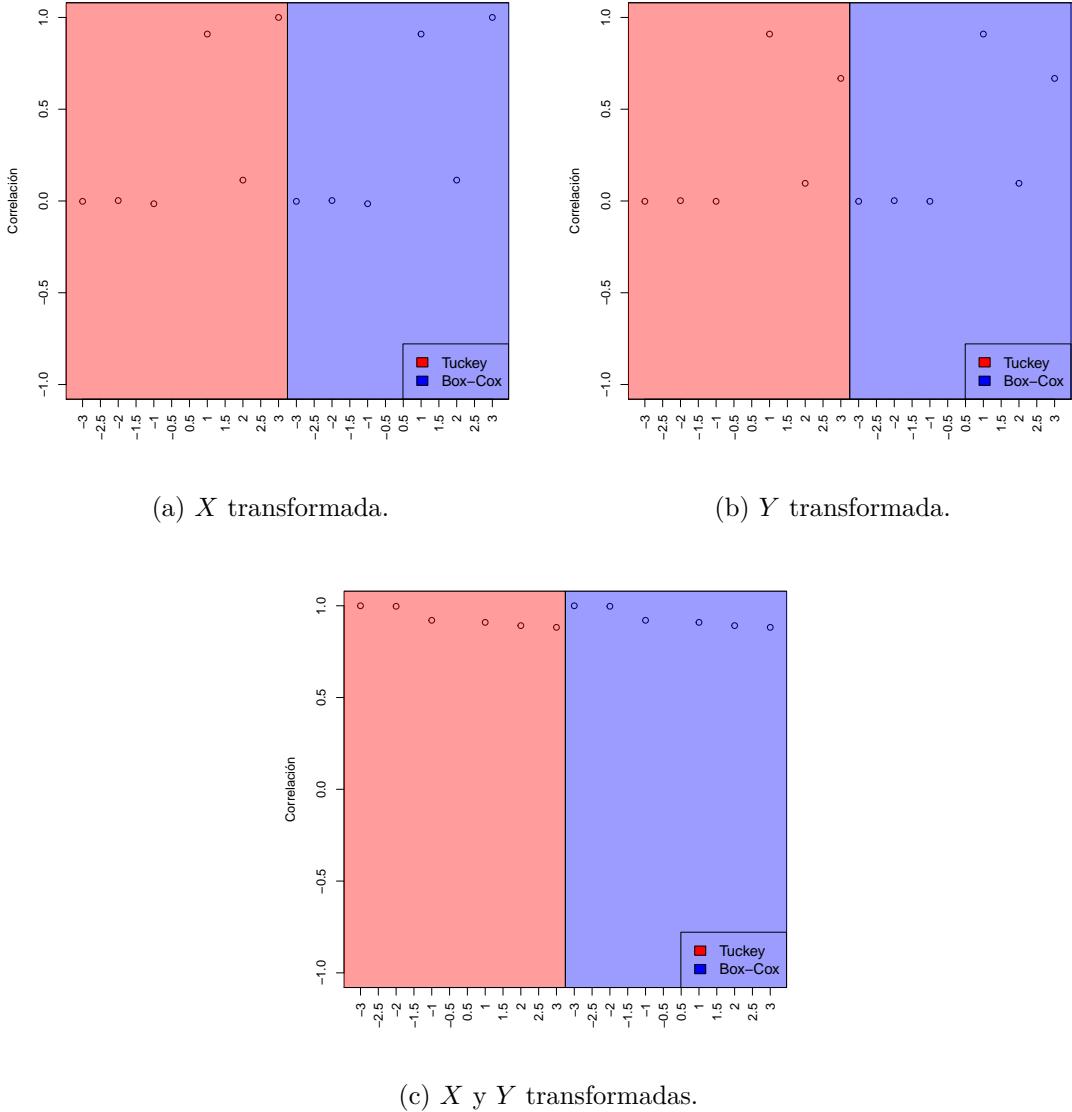


Figura 5.2: Gráficas de las correlaciones para la función $1/x$ con transformaciones para X , Y y ambas. En el eje horizontal se grafican los valores de λ , la mitad roja corresponde a transformación de Tukey, mientras que la azul a la de Box–Cox.

REFERENCIAS

- [1] Elisa Schaeffer. Modelos Probabilísticos Aplicados – Curso en Línea. <https://elisa.dyndns-web.com/teaching/prob/pisis/prob.html>, 2020.
- [2] David Scott. Box–Cox Transformations. <http://onlinestatbook.com/2/transformation/box-cox.html>, 2020.

Teorema de Bayes para datos de Covid-19

Alberto Benavides
26 de octubre de 2020

1. TEOREMA DE BAYES

Según Grinstead y Snell [4], en problemas en los que hay m hipótesis $H_i, i = [1, m]$ que pueden ser confirmadas por alguna evidencia E , se puede utilizar el teorema de Bayes para conocer la probabilidad de una determinada hipótesis dada la evidencia $P(H_i | E)$ según la fórmula

$$P(H_i | E) = \frac{P(H_i)P(E | H_i)}{\sum_{k=1}^m P(H_k)P(E | H_k)}. \quad (1.1)$$

2. PRUEBAS DE DETECCIÓN DE COVID-19

Estos conceptos de la probabilidad se pueden aplicar a ramas de la epidemiología donde se usan pruebas para detectar enfermedades. En 2020 se vive una pandemia dada por el contagio del virus SARS-CoV-2, responsable de la enfermedad Covid-19 [7]. Existen diferentes tipos de pruebas para detectar el contagio presente o pasado de esta enfermedad [1], las cuales tienen algunas características [6] que ayudan a determinar la posibilidad de que sus resultados sean correctos. En este respecto, existen cuatro posibilidades (véase la tabla 2.1 en la p. 2):

- Verdadero positivo (VP): El paciente está infectado p y fue diagnosticado positivo p' .
- Falso positivo (FP): El paciente no está infectado n y fue diagnosticado positivo p' .
- Falso negativo (FN): El paciente está infectado p y fue diagnosticado negativo n' .

Tabla 2.1: Representación visual de las combinaciones entre diagnóstico e infección de pacientes de Covid–19.

		Infección	
		p	n
Diagnóstico	p'	Verdadero positivo	Falso positivo
	n'	Falso negativo	Verdadero negativo

- Verdadero negativo (VN): El paciente no está infectado n y fue diagnosticado negativo n' .

Además, a partir de estas posibilidades se pueden calcular

- Exactitud: $\frac{VP+VN}{P+N}$,
- Precisión: $\frac{VP}{VP+FP}$,
- Sensibilidad: $\frac{VP}{VP+FN}$,
- Especificidad: $\frac{FN}{VN+FP}$.

3. DATOS UTILIZADOS

Con todo esto, se puede aplicar este teorema a valores de contagio obtenidos de datos reales. Se extraen de la Secretaría de Salud de México [3] datos de contagios de Covid–19 en México. De estos datos se extraen solamente los casos que se consideran positivos y negativos, o sea $p' = 864,696$ positivos y $n' = 1,072,760$ negativos; con un total de 1,937,456 pruebas.

4. EXACTITUD DE LAS PRUEBAS

Entre las distintas pruebas que existen, el porcentaje de exactitud puede variar desde un 20 % a un 80 % según las características y tipos de pruebas aplicadas. Los porcentajes más bajos corresponden a pruebas consideradas rápidas que se realizan a partir de muestras de sangre. Los resultados altos para las pruebas se asocian a pruebas virales (de mucosas y tejidos del sistema respiratorio). Para pacientes asintomáticos, las pruebas virales tienen una exactitud de 30 % a 50 %, mientras que en los pacientes con síntomas, están en el rango de 60 % a 80 % [2, 5]. Generalmente, el porcentaje de exactitud para pruebas bayesianas que se suele elegir es de 70 %, por lo que se tomará dicho valor en este análisis también.

5. ESTIMACIONES A PARTIR DE LOS DATOS

De los valores que se tienen a partir de los datos de contagios obtenidos de la Secretaría de Salud de México y del 70 % estimado de exactitud de las pruebas, se pueden obtener la

Tabla 5.1: Diagnósticos e infecciones a partir de los datos de la Secretaría de Salud de México y el estimado de 70 % de exactitud en las pruebas realizadas.

	p	n
p'	605,287	259,409
n'	321,828	750,932

cantidad de verdaderos positivos $VP = 0.7p'$, falsos positivos $FP = 0.3n'$, falsos negativos $FN = 0.3p'$, y verdaderos negativos $VN = 0.7n'$. Esto queda representado en la tabla 5.1. Con estos resultados, se puede estimar el número de contagiados en México es $p = VP + FN = 927,115$, en tanto el número de personas no contagiadas que se hicieron las pruebas sería $n = p = VN + FP = 1,010,341$. Estos estimados incrementan un 7.2 % los casos positivos y disminuyen un 5.9 % los casos negativos, en relación a los reportados.

6. TEOREMA DE BAYES APLICADO A PRUEBAS DE COVID-19

En el caso de las pruebas de Covid-19, se pueden definir algunas variables. Por ejemplo, H_1 : “estoy contagiado de Covid-19”; E : “la prueba salió positiva”. De este modo, se podría desear averiguar si un paciente tiene Covid-19 dado que recibió una prueba con resultado positivo, esto es $P(H_1 | E)$. Por otro lado, H_2 : “no estoy contagiado de Covid-19”. De aquí, se tiene $P(H_1) = \frac{927,115}{1,937,456} = 0.4785$ como la proporción de pacientes con Covid-19 entre todos los que se hicieron la prueba; y $P(H_2) = 1 - P(H_1) = 0.5215$ en tanto la proporción de pacientes no contagiados de Covid-19 de entre los que se hicieron la prueba.

Así, se pueden obtener algunos otros datos de interés como la proporción de pruebas positivas para los pacientes que sí están infectados $P(E | H_1) = 0.7$ y para los que no lo están $P(E | H_2) = 0.3$.

Con estos valores, se puede calcular $P(H_1 | E)$ a partir de la ecuación 1.1 como sigue:

$$P(H_1 | E) = \frac{P(H_1)P(E | H_1)}{P(H_1)P(E | H_1) + P(H_2)P(E | H_2)} = 0.6816,$$

es decir que se tiene una probabilidad de estar contagiado de un 68.16 % si es que una prueba sale positiva, a partir de los supuestos aquí descritos.

7. NOTAS ADICIONALES

El teorema de Bayes es una manera de conocer probabilidades de eventos dependientes entre sí que puede aplicarse para dar estimaciones más certeras a partir de resultados observados. En este caso, es posible repetir los cálculos con valores que reflejen la prueba usada para la detección del Covid-19 y sus porcentajes de exactitud asociados, además del número de pruebas realizadas, los casos positivos y negativos, tanto falsos como verdaderos.

REFERENCIAS

- [1] FOOD, U. S. y DRUG ADMINISTRATION (2020), «Conceptos básicos de las pruebas para el coronavirus», <https://www.fda.gov/consumers/articulos-en-espanol/conceptos-basicos-de-las-pruebas-para-el-coronavirus>.
- [2] GARCÍA, S. (2020), «¿Qué tan certero puede ser el resultado de una prueba para diagnosticar Covid-19?», <https://verificado.com.mx/que-tan-certero-puede-ser-elresultado-de-una-prueba-para-diagnosticar-la-covid-19/>.
- [3] GOBIERNO DE MÉXICO (2020), «Covid-19 México», <https://datos.covid-19.conacyt.mx/#DownZCSV>.
- [4] GRINSTEAD, C. M. y J. L. SNELL (1997), *Introduction to probability*, American Mathematical Society.
- [5] LISBOA BASTOS, M., G. TAVAZIVA, S. K. ABIDI, J. R. CAMPBELL, L.-P. HARAOUI, J. C. JOHNSTON, Z. LAN, S. LAW, E. MACLEAN, A. TRAJMAN, D. MENZIES, A. BENEDETTI y F. AHMAD KHAN (2020), «Diagnostic accuracy of serological tests for Covid-19: systematic review and meta-analysis», *BMJ*, **370**.
- [6] RANJAN, A. (2020), «Covid-19, Bayes' theorem and taking probabilistic decisions», <https://towardsdatascience.com/covid-19-bayes-theorem-and-taking-data-driven-decisions-part-1-b61e2c2b3bea>.
- [7] SCHNIPPER, J. L. y P. E. SAX (2020), «Covid-19 test accuracy supplement: The math of Bayes' theorem», <https://www.statnews.com/2020/08/20/covid-19-test-accuracy-supplement-the-math-of-bayes-theorem/>.

Ejercicios de valor esperado y varianza

Alberto Benavides
2 de noviembre de 2020

P. 247, 1

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

Hay cuatro números impares 3, 5, 7, 9 y cinco números pares 2, 4, 6, 8, 10, así que:

$$E(X) = -1 \times 5(1/9) + 1 \times 4(1/9) = -1/9.$$

P. 247, 6

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?

El espacio muestral de X está dado por

$$\begin{bmatrix} 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 6 & 7 & 8 & 9 & 10 \\ 6 & 7 & 8 & 9 & 10 & 11 \\ 7 & 8 & 9 & 10 & 11 & 12 \end{bmatrix},$$

de donde

$$\begin{aligned}
 E(X) &= \\
 &2\left(\frac{1}{36}\right) + 3\left(\frac{1}{18}\right) + 4\left(\frac{1}{12}\right) + 5\left(\frac{1}{9}\right) \\
 &+ 6\left(\frac{5}{36}\right) + 7\left(\frac{1}{6}\right) + 8\left(\frac{5}{36}\right) + 9\left(\frac{1}{9}\right) + 10\left(\frac{1}{12}\right) + 11\left(\frac{1}{18}\right) + 12\left(\frac{1}{36}\right) \\
 &= 7;
 \end{aligned}$$

el de Y por

$$\begin{bmatrix} 0 & -1 & -2 & -3 & -4 & -5 \\ 1 & 0 & -1 & -2 & -3 & -4 \\ 2 & 1 & 0 & -1 & -2 & -3 \\ 3 & 2 & 1 & 0 & -1 & -2 \\ 4 & 3 & 2 & 1 & 0 & -1 \\ 5 & 4 & 3 & 2 & 1 & 0 \end{bmatrix},$$

así que

$$\begin{aligned}
 E(Y) &= (-5)\left(\frac{1}{36}\right) + (-4)\left(\frac{1}{18}\right) + (-3)\left(\frac{1}{12}\right) + (-2)\left(\frac{1}{9}\right) + (-1)\left(\frac{5}{36}\right) \\
 &+ 0\left(\frac{1}{6}\right) + 1\left(\frac{5}{36}\right) + 2\left(\frac{1}{9}\right) + 3\left(\frac{1}{12}\right) + 4\left(\frac{1}{18}\right) + 5\left(\frac{1}{36}\right) \\
 &= 0;
 \end{aligned}$$

y el de XY por

$$\begin{bmatrix} 0 & -3 & -8 & -15 & -24 & -35 \\ 3 & 0 & -5 & -12 & -21 & -32 \\ 8 & 5 & 0 & -7 & -16 & -27 \\ 15 & 12 & 7 & 0 & -9 & -20 \\ 24 & 21 & 16 & 9 & 0 & -11 \\ 35 & 32 & 27 & 20 & 11 & 0 \end{bmatrix}$$

que, como sigue un patrón similar al espacio muestral de Y , puede verse que la suma de las variables multiplicadas por su probabilidad se van a contrarrestar, por lo que

$$E(XY) = 0,$$

de donde resulta que $E(XY) = E(X)E(Y)$ y por el teorema 6.4 se puede concluir que son independientes.

P. 249, 15

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.

Suponiendo que O son bolas de oro y P de plata, el orden en que se pueden sacar las bolas, señalado en rojo hasta donde se acaba o detiene el juego y seguido por la ganancia total es

- $\textcolor{red}{O} \text{ O P P P} = 1,$
- $\textcolor{red}{O} \text{ P O P P} = 1,$
- $\textcolor{red}{O} \text{ P P O P} = 1,$
- $\textcolor{red}{O} \text{ P P P O} = 1,$
- $\textcolor{red}{P} \text{ O O P P} = 1,$
- $\textcolor{red}{P} \text{ O P O P} = 0,$
- $\textcolor{red}{P} \text{ O P P O} = -1,$
- $\textcolor{red}{P} \text{ P O O P} = 0,$
- $\textcolor{red}{P} \text{ P O P O} = -1,$
- $\textcolor{red}{P} \text{ P P O O} = -1.$

A partir de esto, el $E(X) = 1(\frac{5}{10}) - 1(\frac{3}{10}) = \frac{1}{5}.$

P. 249, 18

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?

Para esto, primero se calculan las distribuciones para cada caso:

$$\begin{aligned} P(X = 0) &= \frac{1}{6}, \\ P(X = 1) &= \frac{5}{6} \cdot \frac{1}{5} = \frac{1}{6}, \\ P(X = 2) &= \frac{5}{6} \cdot \frac{4}{5} \cdot \frac{1}{4} = \frac{1}{6}, \\ P(X = 3) &= \frac{5}{6} \cdot \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{6}, \\ P(X = 4) &= P(X = 5) = \frac{5}{6} \cdot \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{6}. \end{aligned}$$

Con ello, se puede calcular

$$\begin{aligned} E(X) &= 0P(X = 0) + 1P(X = 1) + 2P(X = 2) + 3P(X = 3) + 4P(X = 4) + 5P(X = 5) \\ &= \frac{1}{6}(15) \\ &= \frac{5}{2}. \end{aligned}$$

P. 249, 19

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.

Los subconjuntos que se pueden elegir son de:

- 0 respuestas: Nunca gana puntos; $E(X = 0) = 3(0) = 0$.
- 1 respuesta: Gana 3 puntos $1/4$ de las veces y pierde 1 punto $3/4$ de las veces; $E(X = 1) = 3\frac{1}{4} - 1\frac{3}{4} = 0$.
- 2 respuestas: La mitad de las veces gana 2 puntos y la otra mitad pierde 2 puntos; $E(X = 2) = 2\frac{1}{2} - 2\frac{1}{2} = 0$ porque
 - elige una respuesta correcta (3 puntos) y otra incorrecta (-1 punto), con total de 2 puntos;
 - elige dos respuestas incorrectas, con total de -2 puntos.
- 3 respuestas: Tres cuartos de las veces elige entre las respuestas una correcta y un cuarto de las veces la deja fuera; $E(X = 3) = 1\frac{3}{4} - 3\frac{1}{4} = 0$ porque
 - elige una respuesta correcta (3 puntos) y dos incorrectas (-2 puntos), con total de 1 punto;
 - elige todas las respuestas incorrectas, con total de -3 puntos
- 4 respuestas: Siempre elige la correcta y las tres incorrectas; $E(X = 4) = 3(1) - 3(1) = 0$.

P. 263, 1

A number is chosen at random from the set $S = \{-1, 0, 1\}$. Let X be the number chosen. Find the expected value, variance, and standard deviation of X .

- $E(X) = \frac{1}{3}(-1 + 0 + 1) = 0$.
- $V(X) = E(X^2) - E(X)^2 = \frac{1}{3}((-1)^2 + 0^2 + 1^2) - 0^2 = \frac{2}{3}$.
- $D(X) = \sqrt{V(X)} = \sqrt{\frac{2}{3}}$.

P. 264, 9

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 $D(X)$.

Supongamos que t es el total de la suma de proporciones de los dados, de modo que, por ejemplo, la probabilidad de que salga uno es $\frac{1}{t}$, dos sería $\frac{2}{t}$ hasta $\frac{6}{t}$. De esta forma, $t = 1 + 2 + 3 + 4 + 5 + 6 = 21$. Con esto se puede calcular

- $E(X) = 1\frac{1}{21} + 2\frac{2}{21} + 3\frac{3}{21} + 4\frac{4}{21} + 5\frac{5}{21} + 6\frac{6}{21} = \frac{13}{3}.$
- $V(X) = E(X^2) - E(X)^2 = 1^2\frac{1}{21} + 2^2\frac{2}{21} + 3^2\frac{3}{21} + 4^2\frac{4}{21} + 5^2\frac{5}{21} + 6^2\frac{6}{21} - \frac{13}{3}^2 = 21 - \frac{13}{3}^2 = 20/9.$
- $D(X) = \sqrt{V(X)} = \sqrt{\frac{20}{9}} = \sqrt{5}\frac{2}{3}.$

P. 278, 3

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

En este caso, en lugar de calcular desde $-\infty$, se calcula desde 0 puesto que no existen tiempos negativos para duraciones, así que

$$\begin{aligned} E(T) &= \int_0^\infty t \cdot \lambda^2 t e^{-\lambda t} dt \\ &= \lambda^2 \left(-\frac{e^{-\lambda t} (\lambda^2 t^2 + 2\lambda t + 2)}{\lambda^3} \right) \Big|_0^\infty \\ &= \frac{2}{\lambda} = \frac{2}{0.05} = 40. \end{aligned}$$

Por último,

$$\begin{aligned} E(T^2) &= \int_0^\infty t^3 \cdot \lambda^2 t e^{-\lambda t} dt \\ &= \lambda^2 \left(-\frac{e^{-\lambda t} (\lambda^3 t^3 + 3\lambda^2 t^2 + 6\lambda t + 6)}{\lambda^4} \right) \Big|_0^\infty \\ &= \frac{6}{\lambda^2} \\ &= 2400, \end{aligned}$$

y la varianza $V(T) = E(T^2) - E(T)^2 = 2400 - 40^2 = 800$.

Ejercicios de valor esperado y varianza en R

Alberto Benavides
9 de noviembre de 2020

P. 247, 1

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

En una variable **ganancias** se suman los dólares que se pueden ganar según las reglas descritas del problema que se traducen en la instrucción de R: `-1 + 2 * sample(2:10, 1) %% 2`. Al realizar este experimento $r = [1, 2, \dots, 10\ 000]$ repeticiones, se tienen promedios cercanos a $E(X) = -1/9$ como se ve en la figura 1 (p. 2).

P. 247, 6

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?

Se generan 100 000 tiradas de dos dados d_1 y d_2 , las que se suman en X , se restan $Y = d_1 - d_2$ y se obtiene XY . Una muestra de estos resultados se halla en la tabla 1 (p. 2).

Además, se muestran los histogramas de los valores de X , Y así como XY en la figura 2 (p. 3), donde se puede constatar que los valores esperados son $E(X) = 7$, $E(Y) = 0$, $E(XY) = 0$ y $E(X) \cdot E(Y) = 0 = E(XY)$.

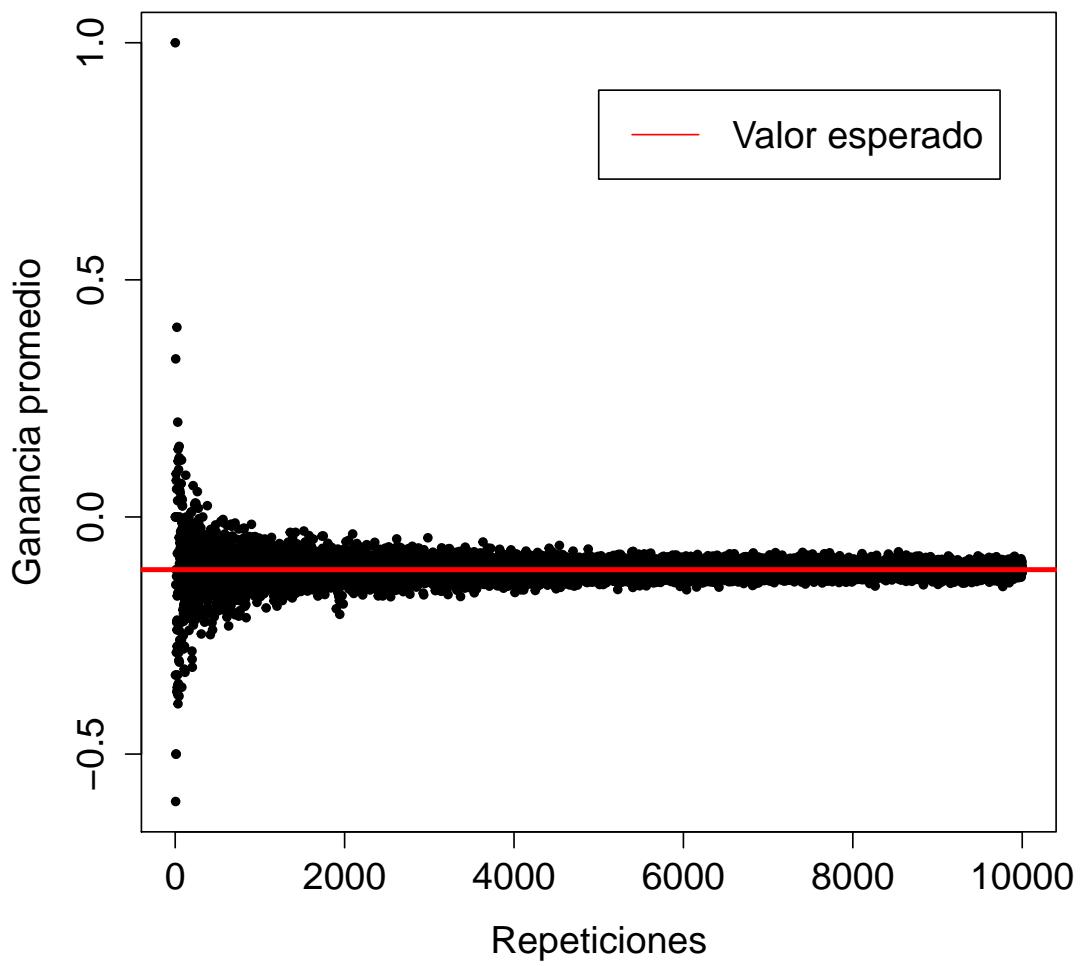


Figura 1: P. 274, 1.

Tabla 1: Muestra de resultados de p. 247, 6.

d_1	d_2	X	Y	XY
4	4	8	0	0
3	6	9	-3	-27
2	1	3	1	3
4	1	5	3	15
5	1	6	4	24

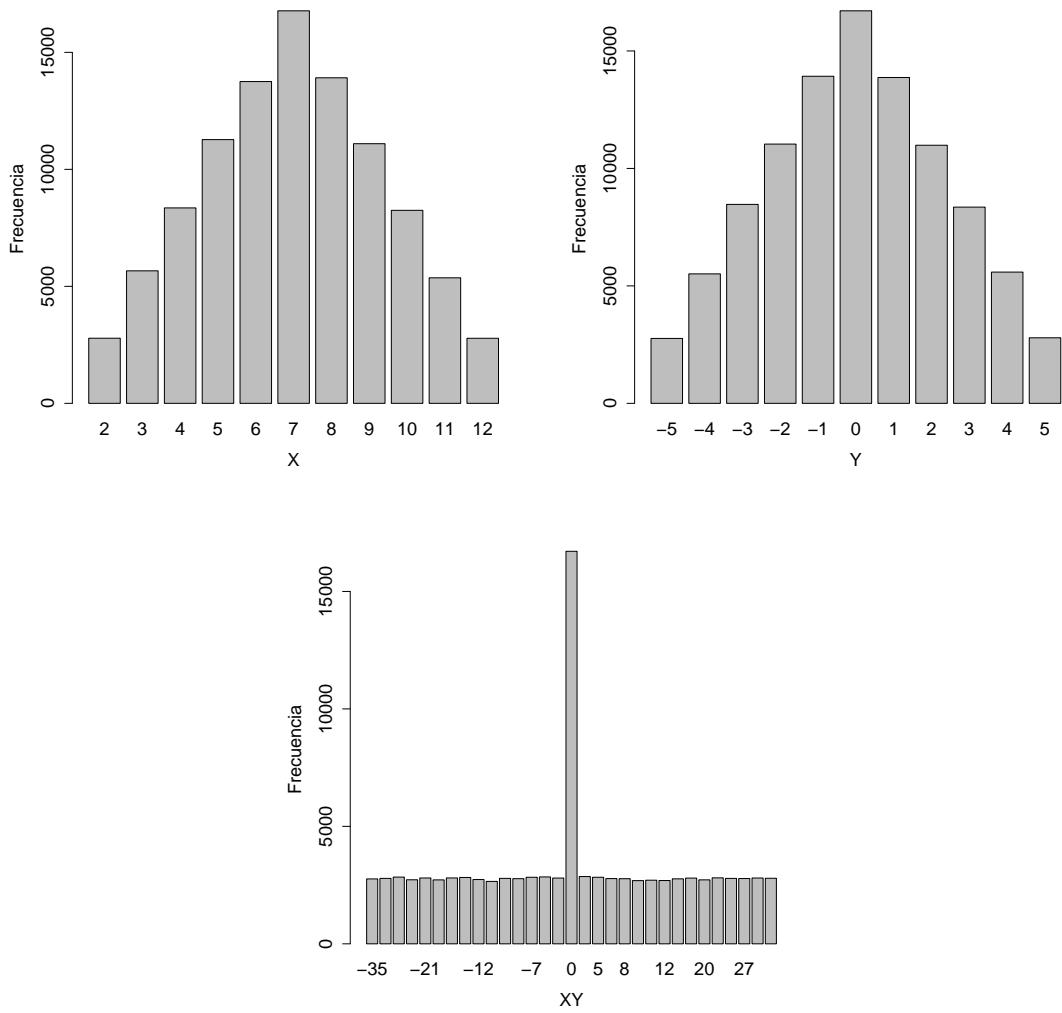


Figura 2: Histogramas de p. 247, 6.

P. 249, 15

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.

En este caso, se realiza un diseño de experimentos en el que se obtienen cincuenta promedios de 10 000 repeticiones del experimento mostrado en el código 1 entre las líneas 6 a la 24. En la línea 1 se crea la variable `caja` que contiene un vector $[-1, -1, -1, 1, 1]$ análogo a los dólares que se obtienen por extraer una bola plateada, -1 , y dorada 1 . Esta `caja` se desordena en la línea 7 y se simula la extracción de una en una de las bolas en la línea 10. La variable `ganancias` declarada en la línea 3 se utiliza para almacenar las ganancias (calculadas en la línea 11) del experimento descrito, de modo que cuando se va ganando por un dólar (línea 12) o cuando han salido las dos bolas de oro (contadas en la línea 15 y revisadas en la 16), se termina el juego y se almacenan las ganancias en la línea 22.

Código 1: P. 249, 15

```
1 caja = c(rep(-1, 3), rep(1, 2))
2 orden = c()
3 ganancias = c()
4 promedios = c()
5 for (k in 1:50){
6   for (i in 1:10000){
7     muestra = sample(caja)
8     ganancia = 0
9     oro = 0
10    for (j in 1:length(caja)){
11      ganancia = ganancia + muestra[j]
12      if (ganancia == 1){
13        break
14      } else if (muestra[j] == 1){
15        oro = oro + 1
16        if (oro == 2){
17          break
18        }
19      }
20    }
21    orden = c(orden, toString(caja))
22    ganancias = c(ganancias, ganancia)
23  }
24  promedios = c(promedios, mean(ganancias))
25 }
```

La variable `orden` almacena el acomodo en que se desordenan al azar las cajas de bolas. Cinco muestras de estas combinaciones y sus ganancias dadas las reglas pueden verse en la tabla 2 (p. 5). Adicionalmente, se muestra un diagrama de cajas y bigotes con los resultados en la figura 3 (p. 5).

Tabla 2: Muestra de resultados de p. 249, 15.

Orden	Ganancias (USD)
-1, -1, -1, 1, 1	1
-1, -1, -1, 1, 1	-1
-1, -1, -1, 1, 1	1
-1, -1, -1, 1, 1	1
-1, -1, -1, 1, 1	1

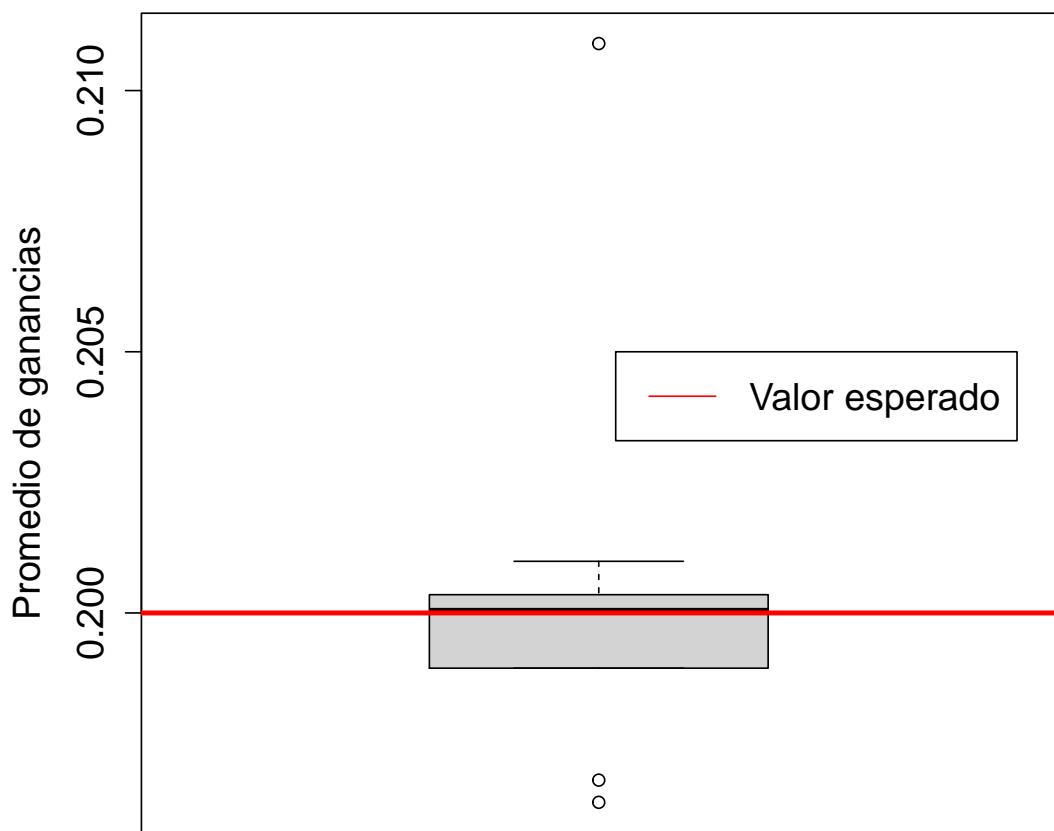


Figura 3: Diagrama de cajas y bigotes de 50 promedios de 10 000 repeticiones para el problema P. 279, 15.

P. 249, 18

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?

Se generan 100 000 números aleatorios entre 0 y 5 que corresponden a los intentos que se harían antes de abrir la puerta mediante `sample(0:5, 100000, replace = TRUE)`. La media de esos intentos es aproximadamente 2.5, lo que coincide con su valor esperado calculado analíticamente.

P. 249, 19

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.

Esto se puede lograr a partir del procedimiento mostrado en el código 2, mediante el cálculo del promedio (línea 1) de 100 000 repeticiones (línea 3) de la suma (línea 4) de los valores de respuestas (línea 6) de un subconjunto elegido al azar de esas respuestas (línea 7). La función `unlist` permite convertir la lista resultante de la función `lapply` en un vector al que se le puede aplicar la función `mean`.

Código 2: P. 249, 19

```
1 mean(  
2   unlist(  
3     lapply(1:100000,  
4       function(x) {  
5         sum(  
6           sample(c(-1, -1, -1, 3),  
7             sample(1:4)  
8           )  
9         )  
10      })  
11    )  
12  )  
13 )
```

P. 263, 1

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 .

Es posible usar las funciones `mean`, `var` y `sd` de R en una muestra de un millón de valores del arreglo $S = \{-1, 0, 1\}$, dado por la instrucción `sample(c(-1, 0, 1), 1000000, replace=TRUE)`, para calcular aproximadamente la esperanza $E^*(X)$, varianza

$V^*(X)$ y desviación estándar $D^*(X)$, en ese orden. Estos valores son $E^*(X) = -0.0007 \approx E(X) = 0$; $V^*(X) = 0.6669 \approx V(X) = 2/3$; y $D^*(X) = 0.8164 \approx D(X) = \sqrt{\frac{2}{3}}$.

P. 264, 9

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 $D(X)$.

La función `sample` de R permite asignar probabilidades a los elementos de un vector sobre el que se desean extraer muestras con el uso del atributo `prob`, de modo que `sample(1:6, prob = (1/21 * 1:6), 100000, replace = T)` devuelve 100 000 valores entre $[1, 2, 3, 4, 5, 6]$ con probabilidades $[\frac{1}{21}, \frac{2}{21}, \frac{3}{21}, \frac{4}{21}, \frac{5}{21}, \frac{6}{21}]$ respectivamente, a los cuales se les puede calcular el promedio, varianza y desviación estándar con `mean`, `var` y `sd`. Los resultados, en ese orden, son 4.3388, 2.2129, 1.4876, muy cercanos a los calculados analíticamente.

P. 278, 3

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

La probabilidad de una variable aleatoria continua x en un intervalo $[a, b]$ se calcula mediante

$$P(a \leq x \leq b) = \int_a^b f_x(x) dx,$$

mientras que su valor esperado es

$$E(X) = \int x f_X(x) dx$$

y la varianza

$$E(X^2) = \int (x - \mu_X)^2 f_X(x) dx.$$

En este problema,

$$f_T(t) = 0.05^2 t e^{-0.05t}. \quad (1)$$

Esto se puede calcular en R con el uso de la función `integrate`, lo cual se realiza en el código 3, donde se definen `f` como $f_T(t)$, `g` como $E(T)$ y `h` - g^2 como $E(T^2)$. Posteriormente se usa `integrate` para calcular las integrales de las variables `g` y `h`, de donde se obtienen el valor esperado y la varianza en las líneas 5 y 6 respectivamente.

Código 3: P. 278, 3; solución analítica

```

1 f <- function(t) 0.05^2 * t * exp(-0.05 * t)
2 g <- function(t) t * f(t)
3 h <- function(t) t^2 * f(t)
4
5 E <- integrate(g, lower = 0, upper = Inf)$value
6 V <- integrate(h, lower = 0, upper = Inf)$value - E^2

```

Como la ecuación 1 es una función de distribución, es posible calcular las probabilidades de cualquier valor t que reciba. Así, se generan un millón de valores aleatorios con distribución $\mathcal{U}(0, 150)$ usando la instrucción `a = runif(100000, 0, 150)`, de las que se calculan sus probabilidades almacenadas en `p = f(a)`. Un histograma de la función de distribución de esta variable se encuentra en la figura 3 (p. 7). Posteriormente, se obtienen un millón de valores aleatorios del vector `a` con las probabilidades almacenadas en el vector `p` mediante la variable `s` que recibe el vector resultante de `sample(a, 100000, prob=p, replace=TRUE)`, con lo que se puede calcular tanto el valor esperado como el promedio de esos valores, obtenidos por la función `mean`, y la varianza con `var`, como se observa en el código 4.

Código 4: P. 278, 3; aproximación experimental

```

1 a = runif(100000, 0, 1000)
2 p = f(a)
3 s = sample(a, 100000, prob=p, replace=TRUE)
4 mean(s)
5 var(s)

```

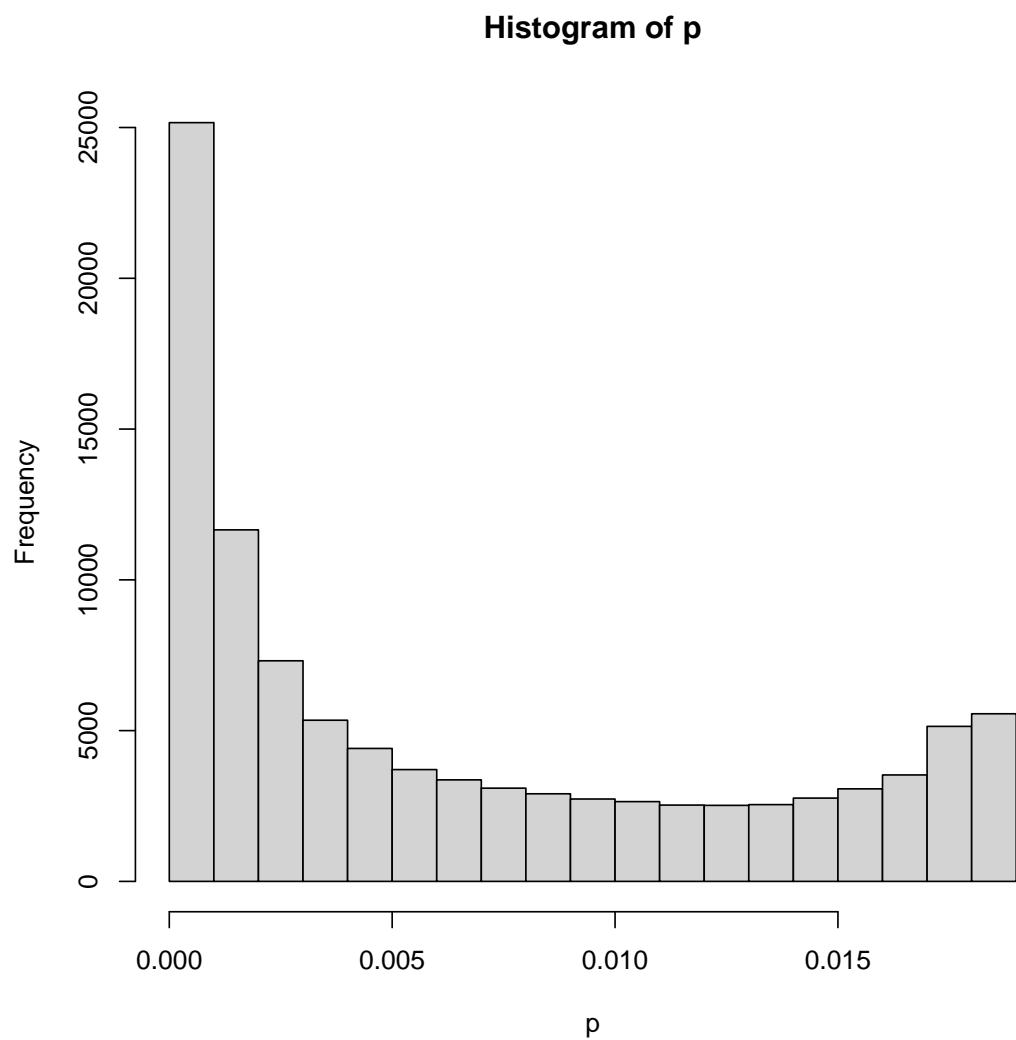


Figura 4: Histograma de la función de distribución de probabilidades dada por la ecuación [\[1\]](#) para un millón de variables aleatorias con distribución $\mathcal{U}(0, 150)$.

Convolución y Chi cuadrada

Alberto Benavides

17 de noviembre de 2020

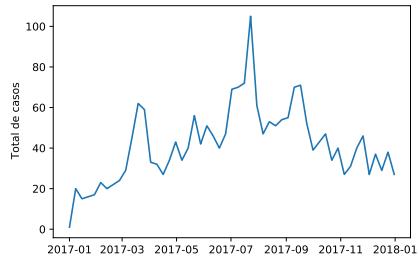
1. SOBRE LOS DATOS

Los datos para esta tarea provienen del tema de tesis que actualmente trabajo: relación entre contaminantes y enfermedades en el *área metropolitana de Monterrey* (AMM). Los datos de los contaminantes fueron obtenidos del SIMA [3], mientras que los datos de las enfermedades provienen de la página de la Secretaría de Salud de México [2]. Los datos de contaminantes fueron tomados cada segundo por las distintas trece estaciones de monitoreo ambiental ubicadas en el AAM e incluyen concentraciones de algunos contaminantes, de los que se destacan los que tienen tamaños de partículas de 10 PM medidas en $\mu\text{g} / \text{m}^3$, en tanto que los datos de las enfermedades contienen información sobre la edad, género, talla, enfermedad y, dato de principal interés para esta tarea, la fecha en la que se dio la consulta de cada caso.

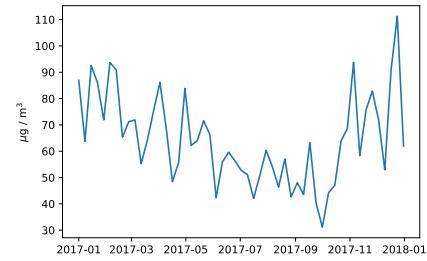
2. PREPROCESAMIENTO DE LOS DATOS

A ambos conjuntos de datos se les eliminaron las inconsistencias como valores fuera de rangos permitidos o la ausencia de valores, posteriormente se obtuvieron sólo concentraciones de contaminantes y consultas del año 2017 por ser el año en que se hallan datos más completos. Después, se agruparon estos datos semanalmente, obteniendo el promedio semanal de las concentraciones de partículas de 10 PM y la suma de todos los casos de consultas registradas, ambos datos mostrados en la figura [1] (p. [2]).

Una práctica común para comparar series de tiempo es normalizarlas y diferenciarlas mediante la resta de cada valor, menos su valor en una unidad de tiempo anterior, lo que hace que la serie de tiempo resultante tenga valores en torno a cero entre $[-1, 1]$, como puede verse en la figura [2] (p. [2]).



(a) Casos.



(b) Partículas de 10 PM.

Figura 1: Series de tiempo de casos registrados de enfermedades y concentraciones de partículas de 10 PM en el AMM durante 2017.

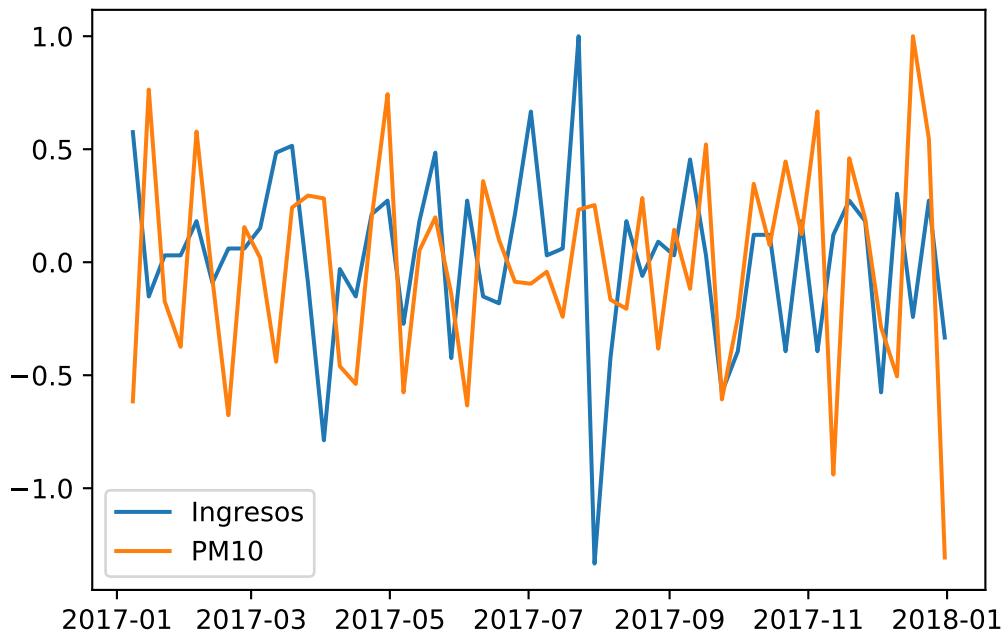


Figura 2: Series de tiempo diferenciadas de casos registrados de enfermedades y concentraciones de partículas de 10 PM en el AMM durante 2017.

3. CHI CUADRADA

La prueba de Chi cuadrada tiene dos usos principales: Determinar la independencia entre conjuntos de datos y conocer si valores observados son similares a valores esperados (a esto último se le llama prueba de ajuste). En este caso, se utiliza la prueba de Chi cuadrada para determinar si los valores diferenciados y normalizados (como lo sugieren Hyndman y Athanasopoulos [1]) de los contaminantes son independientes de los de las consultas de las enfermedades registradas y, también, si los casos registrados observados son similares a los esperados, tomados estos últimos como las concentraciones en promedio de las partículas de 10 PM.

Para el caso de la independencia, la hipótesis nula es que ambas variables son independientes. El p -valor obtenido es 1 por lo que se acepta la hipótesis nula. Mientras que para la prueba de ajuste, la hipótesis nula es que los valores observados y esperados no tienen diferencia, pero ésta hipótesis se rechaza dado que el p -valor = 6.07×10^{-175} .

4. CONVOLUCIÓN

Una convolución muestra la distribución de probabilidad Z de la suma de dos variables aleatorias X y Y tal que $P(Z = j) = \sum_{i=-\infty}^{\infty} P(X = i) \times P(Y = j - i)$. Para el caso discreto, esto se puede calcular como $f_c(i) = \sum_j f_1(j) \times f_2(i - j)$, en tanto para el continuo se tiene $(f * g)(z) = \int_{-\infty}^{\infty} f(z - x) \times g(x) dx$. Este concepto de convolución también se puede utilizar para conocer las interacciones entre señales o, en este caso, series de tiempo de modo que se obtiene el área común a ambas series de tiempo, es decir, el grado de relación de una (fija) respecto a la otra (que se desplaza en el tiempo). La convolución de las series de tiempo diferenciadas puede verse en la figura 3 (p. 4).

5. PROPIEDADES DE LA VARIANZA

Existen dos propiedades de la varianza, a saber $\text{Cov}[aX + b, cY + d] = ac\text{Cov}[X, Y]$ y $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$ que pueden demostrarse numéricamente al generar dos conjuntos de variables aleatorias $X \sim \mathcal{U}(0, 1)$ y $Y \sim \mathcal{B}(100, 0.05)$ de mil valores cada uno. Los histogramas de estos conjuntos aparecen en la figura 4 (p. 5). Para comprobar esto, se toman enteros aleatorios entre $[-10, 10]$ para a, b, c, d y mediante las funciones `var` y `cov` de R, que se utilizan para calcular la varianza y covarianza en ese orden, se comprueba que ambas propiedades se cumplen, como se muestra en el código 1.

Código 1: Demostración numérica de propiedades de varianza y covarianza

```
1 X = runif(1000)
2 Y = rbinom(1000, 100, 0.05)
3
4 a = sample(-10:10, 1)
5 b = sample(-10:10, 1)
6 c = sample(-10:10, 1)
7 d = sample(-10:10, 1)
8
```

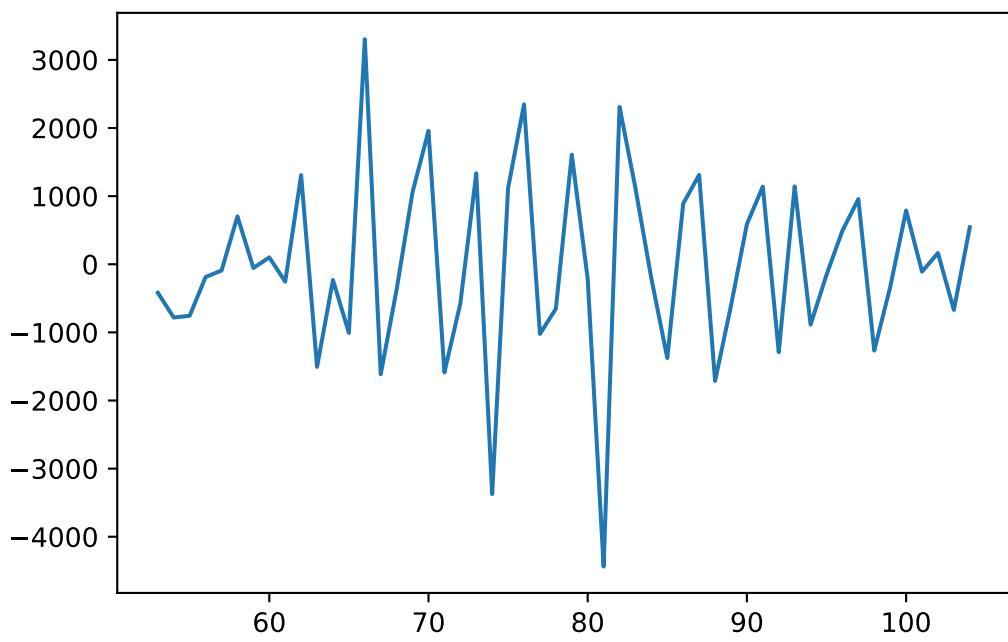


Figura 3: Convolución de las series de tiempo diferenciadas de casos registrados de enfermedades y concentraciones de partículas de 10 PM en el AMM durante 2017.

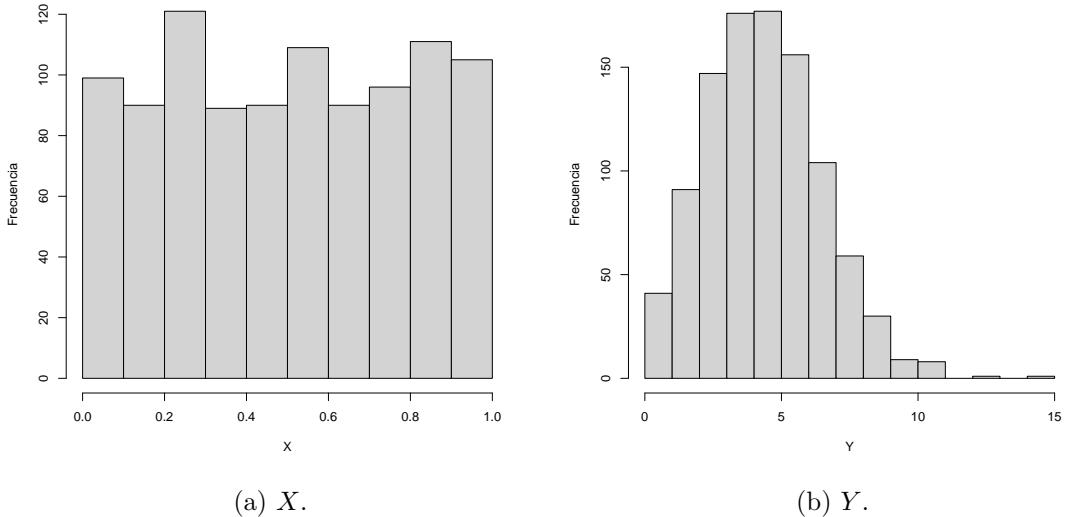


Figura 4: Histogramas de mil valores aleatorios para $X \sim \mathcal{U}(0, 1)$ y $Y \sim \mathcal{B}(100, 0.05)$.

```

9 cov(a * X + b, c * Y + d) == a * c * cov(X, Y)
10 # TRUE
11
12 var(X + Y) == var(X) + var(Y) + 2 * cov(X, Y)
13 # TRUE
14

```

REFERENCIAS

- [1] HYNDMAN, R. J. y G. ATHANASOPOULOS (2018), «Forecasting: Principles and Practice», [oTexts.com/fpp2](http://otexts.com/fpp2), [Accedido 13/nov/2020].
- [2] SECRETARÍA DE SALUD (2020), «Egresos Hospitalarios. Datos abiertos», URL http://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_egresoshosp_gobmx.html.
- [3] SISTEMA INTEGRAL DE MONITOREO AMBIENTAL [SIMA] NUEVO LEÓN (2020), «aire.nl.gob.mx | Home», <http://aire.nl.gob.mx/>, [Accedido 14/may/2020].

Ejercicios Procesos de ramificación

Alberto Benavides
24 de noviembre de 2020

Exercise 1 (P. 392, e. 1). *Let Z_1, Z_2, \dots, Z_n describe a branching process in which each parent has j offspring with probability p_j . Find the probability d that the process eventually dies out if*

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

Para este caso, el número esperado de hijos es $m = h'(1) = p_1 + 2p_2 = 1/4 + 2(1/4) = 3/4 \leq 1$, por lo que por el teorema 10.2, $d = 1$ así que el proceso o herencia o apellido se acabará.

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

Igual que el inciso anterior, $m = 1/3 + 2(1/3) = 1 \leq 1$, así que $d = 1$.

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

Aquí, $m = 0 + 2(2/3) = 4/3 > 1$, pero como $p_0 < p_2$ se puede obtener $d = p_0/p_2 = \frac{1/3}{2/3} = 1/2$.

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

En este inciso,

$$\begin{aligned}
h(z) &= p_0 + p_1 z + p_2 z^2 + p_3 z^3 + \dots \\
&= 1/2^{0+1} + 1/2^{1+1} z + 1/2^{2+1} z^2 + 1/2^{3+1} z^3 + \dots \\
&= 1/2^1 + 1/2^2 z + 1/2^3 z^2 + 1/2^4 z^3 + \dots \\
&= \frac{1}{2}(1/2^1 + 1/2^2 z + 1/2^3 z^2 + 1/2^4 z^3 + \dots)/\frac{1}{2} \\
&= \frac{1}{2}(1 + 1/2^1 z + 1/2^2 z^2 + 1/2^3 z^3 + \dots) \\
&= \frac{1}{2} \left(\frac{1}{1 - \frac{1}{2}z} \right) \\
&= \frac{1}{2 - z}.
\end{aligned}$$

Si esto es verdad, entonces

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

por lo que, como $m = h'(1) = \frac{1}{(2-1)^2} = 1 \leq 1$, $d = 1$.

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

De manera análoga al inciso precedente,

$$\begin{aligned}
h(z) &= p_0 + p_1 z + p_2 z^2 + p_3 z^3 + \dots \\
&= \frac{1}{3} \left(\frac{2}{3} \right)^0 + \frac{1}{3} \left(\frac{2}{3} \right)^1 z^1 + \frac{1}{3} \left(\frac{2}{3} \right)^2 z^2 + \frac{1}{3} \left(\frac{2}{3} \right)^3 z^3 \dots \\
&= \frac{1}{3} \left[1 + \left(\frac{2}{3} \right)^1 z^1 + \left(\frac{2}{3} \right)^2 z^2 + \left(\frac{2}{3} \right)^3 z^3 \dots \right] \\
&= \frac{1}{3} \left(\frac{1}{1 - \frac{2}{3}z} \right) \\
&= \frac{1}{3 - 2z}
\end{aligned}$$

de donde

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

por lo que $m = h'(1) = \frac{2}{(3-2)^2} = \frac{2}{(1)^2} = 2$ y $d < 1$ cuando $z \neq 1$. Para calcular esta d se obtienen las raíces a partir de igualar $z = h(z)$, así que

$$z = \frac{1}{3 - 2z}$$

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

de donde $z_1 = 1$ (ya conocida) y $z_2 = 1/2 = d$.

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

Finalmente, $d = 0.2032$. Esto se obtiene mediante el código [1]

Código 1: Aproximación

```

1 p = function(j){
2   return( exp(-2) * 2 ** j / factorial(j) )
3 }
4 d = p(0)
5 for (m in 1:1000){
6   sum = 0
7   for (j in 0:100){
8     sum = sum + p(j) * (d ** j)
9   }
10  d = sum
11 }
12 d
13 # 0.2031878699799799

```

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

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

Como $m = p_1 + 2p_2 = 0 + 2(1/2) = 1$, entonces se espera una ganancia de $50(1 + 1^{12}) - 100 = 0$.

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

Aquí $m = p_1 + 2p_2 = 1/2 + 2(1/3) = 7/6$, entonces se espera una ganancia de $50(7/6 + (7/6)^{12}) - 100 \approx 276.26$.

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

Exercise 3 (P. 401, e. 1). *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*

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

$$\begin{aligned} g_X(t) &= \int_0^2 e^{tx} \cdot \frac{1}{2} dx \\ &= \int_0^2 \frac{e^{tx}}{2} dx; u = tx \rightarrow \frac{du}{dx} = t \rightarrow dx = \frac{du}{t} \\ &= \frac{1}{2t} \int_0^2 e^u du \\ &= \frac{1}{2t} [e^u]_0^2 = \frac{1}{2t} (e^{2t} - e^0) \\ &= \frac{e^{2t} - 1}{2t}. \end{aligned}$$

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

$$\begin{aligned} g_X(t) &= \int_0^2 e^{tx} \cdot \frac{1}{2} x dx \\ &= \frac{1}{2} \int_0^2 x e^{tx} dx \\ &= \frac{1}{2} \left[\frac{x e^{tx}}{t} - \int_0^2 \left((1) \frac{e^{tx}}{t} \right) dx \right] u = tx \rightarrow \frac{du}{dx} = t \rightarrow dx = \frac{du}{t} \\ &= \frac{1}{2} \left[\frac{x e^{tx}}{t} - \int_0^2 \frac{e^u}{t^2} du \right] \\ &= \frac{1}{2} \left[\frac{x e^{tx}}{t} - \frac{1}{t^2} \int_0^2 e^{tx} dx \right] \\ &= \frac{1}{2} \left[\frac{x e^{tx}}{t} - \frac{e^{tx}}{t^2} \Big|_0^2 \right] \\ &= \frac{1}{2} \left[\frac{t x e^{tx} - e^{tx}}{t^2} \Big|_0^2 \right] \\ &= \frac{1}{2} \left[\frac{e^{tx}(tx - 1)}{t^2} \Big|_0^2 \right] \\ &= \frac{1}{2} \left[\frac{e^{2t}(2t - 1)}{t^2} - \frac{(1)(-1)}{t^2} \right] \\ &= \frac{e^{2t}(2t - 1) + 1}{2t^2}. \end{aligned}$$

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

$$\begin{aligned}
g_X(t) &= \int_0^2 e^{tx} \cdot \left(1 - \frac{1}{2}x\right) dx \\
&= -\frac{1}{2} \int_0^2 (x-2)e^{tx} dx \\
&= -\frac{1}{2} \left[(x-2) \frac{e^{tx}}{t} - \int_0^2 (1) \frac{e^{tx}}{t} dx \right]; u = tx \rightarrow \frac{du}{dx} = t \rightarrow dx = \frac{du}{t} \\
&= -\frac{1}{2} \left[\frac{(x-2)e^{tx}}{t} - \frac{1}{t^2} \int_0^2 e^u du \right] \\
&= -\frac{1}{2} \left[\frac{(x-2)e^{tx}}{t} - \left. \frac{e^{tx}}{t^2} \right|_0^2 \right] \\
&= -\frac{1}{2} \left[\frac{(x-2)te^{tx} - e^{tx}}{t^2} \right]_0^2 \\
&= -\frac{1}{2} \left[\frac{e^{tx}[t(x-2) - 1]}{t^2} \right]_0^2 \\
&= -\frac{1}{2} \left[\left(\frac{e^{2t}[t(2-2) - 1]}{t^2} \right) - \left(\frac{e^{0t}[t(0-2) - 1]}{t^2} \right) \right] \\
&= -\frac{1}{2} \left[\left(\frac{-e^{2t}}{t^2} \right) - \left(\frac{-2t-1}{t^2} \right) \right] \\
&= -\frac{1}{2} \left[\frac{-e^{2t} + 2t + 1}{t^2} \right] \\
&= \frac{e^{2t} - 2t - 1}{2t^2}.
\end{aligned}$$

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

$$\begin{aligned}
g_X(t) &= \int_0^2 e^{tx} |1 - x| dx \\
&= \int_0^1 (1 - x) e^{tx} dx + \int_1^2 (-1 + x) e^{tx} dx \\
&= \frac{e^t - 1}{t} - \frac{e^t}{t} + \frac{e^t}{t^2} - \frac{1}{t^2} - \frac{e^{2t} - e^t}{t} + \frac{2e^{2t}}{t} - \frac{e^{2t}}{t^2} - \frac{e^t}{t} + \frac{e^t}{t^2} \\
&= \frac{t(e^t - 1) - t(e^t) + e^t - 1 - t(e^{2t} - e^t) + t(2e^{2t}) - e^{2t} - t(e^t) + e^t}{t^2} \\
&= \frac{te^t - t - te^t + e^t - 1 - te^{2t} + te^t + 2te^{2t} - e^{2t} - te^t + e^t}{t^2} \\
&= \frac{-t + 2e^t - 1 + te^{2t} - e^{2t}}{t^2} \\
&= \frac{2e^t - t - 1 + e^{2t}(t - 1)}{t^2}.
\end{aligned}$$

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

$$\begin{aligned}
g_X(t) &= \int_0^2 e^{tx} \left(\frac{3}{8}x^2\right) dx \\
&= \frac{3}{8} \cdot \int_0^2 e^{tx} x^2 dx \\
&= \frac{3}{8} \left[\frac{e^{tx} x^2}{t} - \int \frac{2e^{tx} x}{t} dx \right]_0^2 \\
&= \frac{3}{8} \left[\frac{e^{tx} x^2}{t} - \frac{2}{t} \left(\frac{e^{tx} x}{t} - \frac{e^{tx}}{t^2} \right) \right]_0^2 \\
&= \frac{3}{8} \left(\frac{4e^{2t}}{t} - \frac{2}{t} \left(\frac{2e^{2t}}{t} - \frac{e^{2t}}{t^2} \right) - \frac{2}{t^3} \right) \\
&= \frac{3}{8} \left(\frac{4e^{2t}}{t} - \frac{4e^{2t}}{t^2} + \frac{2e^{2t}}{t^3} - \frac{2}{t^3} \right) \\
&= \frac{3}{8} \left(\frac{4t^2 e^{2t} - 4te^{2t} + 2e^{2t} - 2}{t^3} \right) \\
&= \frac{3}{8} \left(\frac{2e^{2t}(2t^2 - 2t + 1) - 2}{t^3} \right) \\
&= \frac{3}{4} \left(\frac{e^{2t}(2t^2 - 2t + 1) - 1}{t^3} \right).
\end{aligned}$$

Exercise 4 (P. 402, e. 6). 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

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

$$\begin{aligned}
f_X(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ix\tau} e^{-|\tau|} d\tau \\
&= \frac{1}{2\pi} \left[\int_{-\infty}^0 e^{-ix\tau - (-\tau)} d\tau + \int_0^{\infty} e^{-ix\tau - \tau} d\tau \right] \\
&= \frac{1}{2\pi} \left[\int_{-\infty}^0 e^{\tau - ix\tau} d\tau + \int_0^{\infty} e^{-ix\tau - \tau} d\tau \right] \\
&= \frac{1}{2\pi} \left[\frac{1}{1 - ix} \int_{-\infty}^0 e^u du - \frac{1}{1 + ix} \int_0^{\infty} e^v dv \right] \\
&= \frac{1}{2\pi} \left[\frac{1}{1 - ix} - \frac{1}{1 + ix} \right] \\
&= \frac{1}{2\pi} \left[\frac{1 + ix}{i^2 x^2 - 1^2} - \frac{1 - ix}{1^2 - i^2 x^2} \right] \\
&= \frac{1}{2\pi} \left[\frac{ix + 1 - ix + 1}{1^2 - i^2 x^2} \right] \\
&= \frac{1}{2\pi} \left[\frac{2}{1 + x^2} \right] \\
&= \frac{1}{\pi(1 + x^2)}.
\end{aligned}$$

Exercise 5 (P. 403, e. 10). Let X_1, X_2, \dots, X_n be an independent trials process with density

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

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

Ley de los grandes números

Alberto Benavides
1 de diciembre de 2020

1. INTRODUCCIÓN

La ley de los grandes números afirma que mientras un experimento aleatorio se repita n veces, el promedio de los resultados $\frac{X_1, X_2, \dots, X_n}{n}$ de ese experimento se aproximará a su valor esperado $E(X)$ conforme n se aproxime a $+\infty$.

Los ejemplos más usados para mostrar la ley de los grandes números consisten en lanzar una moneda al aire o la tirada de un dado de seis caras.

Para el lanzamiento de una moneda, el $E(X) = 0.5$. Los lenguajes computacionales permiten experimentar computacionalmente mediante la generación de valores pseudoaleatorios¹. En R² se puede realizar un experimento que simule mil lanzamientos de monedas para, de ellos, calcular la media, con el procedimiento `mean(sample(0:1, 1000, replace=TRUE))`. Con una semilla fija `set.seed(33)`, se obtiene por resultado $0.507 \approx E(X)$. En este caso, el `0:1` simula los valores que puede tomar la variable aleatoria X : 0 para cara y 1 para cruz.

En el caso de mil tiradas de un dado de seis caras, se puede simular con `mean(sample(1:6, 1000, replace=TRUE))`, donde el valor obtenido con la misma semilla es $3.554 \approx E(X) = \frac{1+2+3+4+5+6}{6} = 3.5$.

Una representación gráfica de la parte en la que n se approxima a $+\infty$ consiste en graficar las medias conforme n crece. Para el ejemplo de la tirada del dado de seis cara en la que se registran las tiradas en un rango $i \in [1, 10\,000]$, se logra mediante el código [\[1\]](#).

Código 1: Medias experimentales de las tiradas de un dado de seis caras

```
1  medias <- c()
```

¹https://es.wikipedia.org/wiki/Generador_de_n%C3%BAmeros_pseudoaleatorios

²<https://www.r-project.org/>

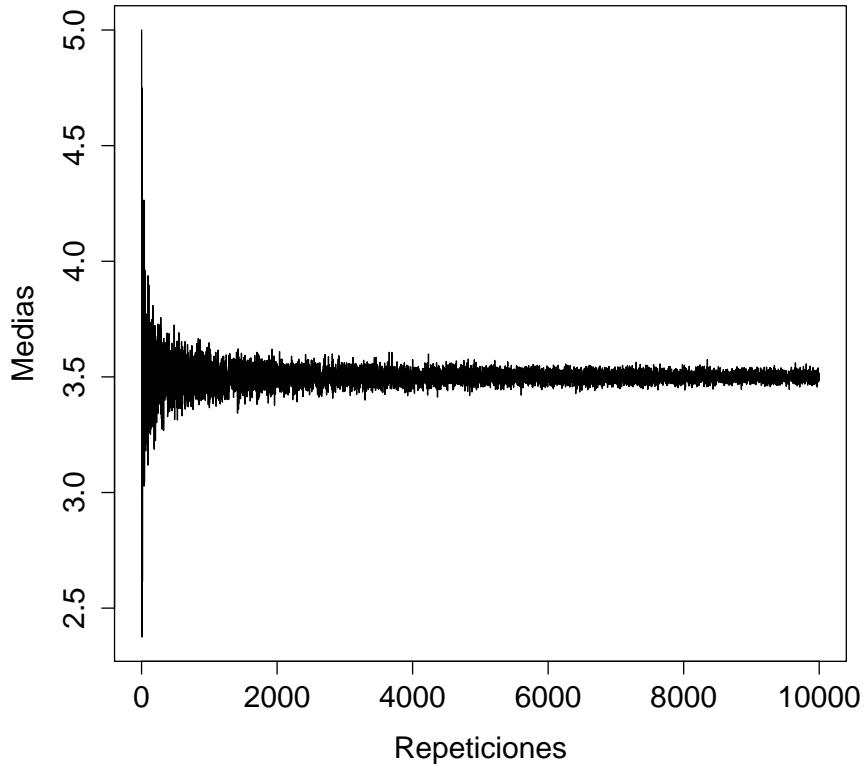


Figura 1: Medias experimentales de distintas tiradas de dados.

```

2   for (i in 1:10000){
3     medias <- c(medias, mean(sample(1:6, i, replace=TRUE)))
4   }

```

Las medias de cada i número de experimentos se almacenan en una variable cuyos valores, mostrados en la figura 1 (p. 2), permite ver cómo converge al $E(X)$ el resultado de las medias conforme el número de repeticiones del experimento se hace más grande.

2. PROBLEMA

Actualmente, investigo la relación que hay entre los contaminantes del aire y los reportes de enfermedades por parte de los centros de salud, estudio delimitado al Área Metropolitana de Monterrey durante el año 2017. Los datos de contaminantes para ese año se obtuvieron del Sistema Integral de Monitoreo Ambiental (SIMA) [2]. Un ejemplo de los datos obtenidos puede verse en la tabla 1 (p. 3), donde se constata que existen entre los datos la fecha en que fueron registrados, una de las trece estaciones que hizo el registro, los valores de ciertos contaminantes. Además, hay una columna llamada Válida,

Tabla 1: Muestra de mediciones capturadas por los sensores del área metropolitana de Monterrey.

Fecha	Estación	CO	NO	...	Válida
19-Aug-16 16	Centro	0.43	1.90	...	1
23-Mar-97 0	Suroeste	1.23	1.25	...	1
21-Oct-11 0	Norte	2.16	40.00	...	1

Elisa Schaeffer³ agregó para diferenciar las mediciones inválidas conforme a lineamientos que el SIMA también proporcionó. Si la medición es válida, se asigna un valor de 1, mientras que si es inválida en alguno de los valores reportados, se le asigna un valor de 0.

Uno de los problemas que se ha tenido con estos datos es la cantidad de errores en medición que presentan y una de las preguntas que se plantean por esta situación es qué estrategias se pueden implementar para mejorar la precisión de los sensores. En esta tarea se presenta una respuesta que utiliza la ley de los grandes números para encontrar un número de mediciones que se deberían hacer a partir de la media de valores inválidos registrados para saber si un cambio o reparación en los sensores mejora su certeza.

3. SOLUCIÓN

Los errores de medición detectados por Elisa Schaeffer se agruparon por año y estación de monitoreo, obteniendo por resultado registros como los que se muestran en la tabla 2 (p. 3).

Tabla 2: Cantidad de errores agrupados

Año	Estación	Errores	Datos por grupo
1999	Noroeste	40	8760
2003	Suroeste	0	8760
2000	Centro	35	8784

La estación con más errores de medición, con un total de 41.47 %, es la Suroeste, cuya ubicación geográfica se muestra en la figura 2 (p. 4).

Análogo al ejemplo 4.3 descrito por Sedor [4], se puede calcular la cantidad de mediciones que se deben hacer para saber, con un 95 % de confianza, que el porcentaje de errores de medición se mantiene. En el caso de la estación Suroeste, se sabe que la probabilidad de obtener una medición errónea es $\mu = 0.4147 = p$, de donde $\sigma^2 = p(1 - p) = 0.2427$. Ahora, se puede definir un error de medición $\epsilon = 0.02$, lo que quiere decir que se probará si una medida es errónea con probabilidad entre $[0.4147 - 0.02, 0.4147 + 0.02]$. Por la ley de los grandes números, se tiene que

$$P[|\bar{X} - \mu| > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2},$$

³<https://elisa.dyndns-web.com/>

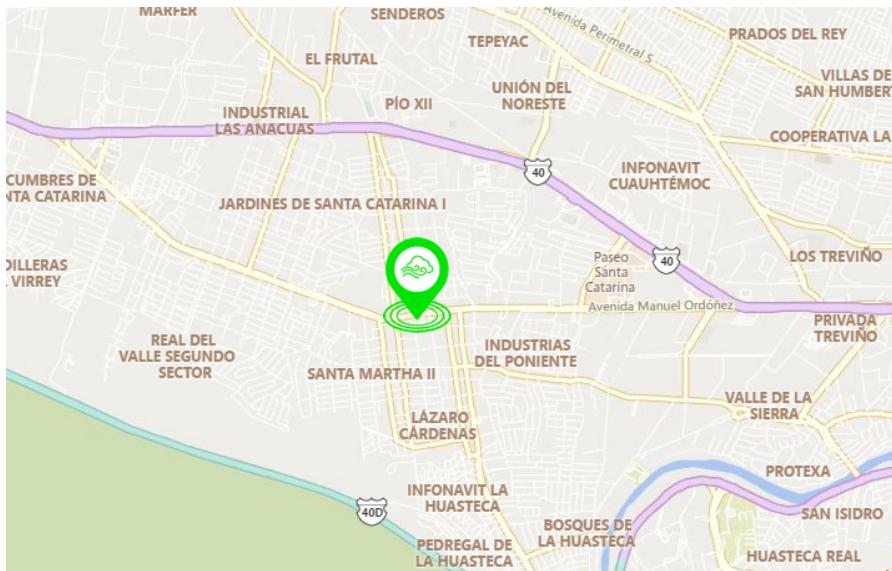


Figura 2: Mapa que muestra la ubicación de la estación de monitoreo Suroeste con un marcador de posición verde. Imagen obtenida de la página del SIMA [2].

al sustituir

$$P[|\bar{X} - 0.4147| > 0.02] \leq \frac{0.2427^2}{n(0.02)^2}.$$

Como se desea obtener un intervalo de confianza de 95 %, entonces se debe cumplir

$$\frac{0.2427^2}{n(0.02)^2} = 0.05,$$

por lo que

$$n = \frac{0.2427^2}{(0.05)(0.02)^2} = 2945.1645 \approx 2945.$$

Así, se podría recomendar realizar $n = 2945$ mediciones y comparar el porcentaje de errores encontrados, para saber si alguna modificación en los sensores ha modificado la certeza de sus mediciones.

REFERENCIAS

- [1] SEDOR, K. (2015), «The Law of Large Numbers and its Applications», en L. University (editor), *Department of mathematical sciences (honours seminar)*.
- [2] SISTEMA INTEGRAL DE MONITOREO AMBIENTAL [SIMA] NUEVO LEÓN (2020), «aire.nl.gob.mx | Home», <http://aire.nl.gob.mx/>, [Accedido 14/may/2020].

Teorema del límite central

Alberto Benavides
8 de diciembre de 2020

1. INTRODUCCIÓN

El **teorema del límite central** expresa que conforme la media \bar{X}_n de n variables aleatorias independientes e idénticamente distribuidas X_i de una distribución D cualquiera, se aproxima a una distribución normal con media μ y varianza σ^2 relacionados con la distribución D conforme n tiende a $+\infty$. Esto se puede mostrar experimentalmente partiendo, por ejemplo, de una distribución exponencial generada a partir de la función `rexp` del lenguaje de programación R. El histograma de mil valores obtenidos al azar de dicha distribución aparece en la figura 1 (p. 2). De dicha distribución se pueden obtener muestras de distintos tamaños, en este caso $n = [2, 10, 30, 50]$, con lo que mostrar cómo estas distribuciones se aproximan a una distribución normal. Para este fin, se plasman los histogramas para cada n en la figura 2 (p. 3).

2. APLICACIÓN

Actualmente en México existe una normativa que se utiliza para comunicar la calidad del aire, publicada por la Secretaría de Gobernación de México [1]. De entre los contaminantes que regula, se encuentra el material particulado con tamaño menor o igual a $10 \mu\text{g}$, contaminante denominado como PM10. Este tipo de partículas están asociadas con riesgos pulmonares, cardiovasculares e incluso con muertes prematuras. El Sistema Integral de Monitoreo Ambiental [SIMA] Nuevo León [2] registra cada segundo, mediante trece estaciones de monitoreo meteorológico, la concentración de PM10 medida en $\mu\text{g}/\text{m}^3$. La norma establece que a partir de $50 \mu\text{g}/\text{m}^3$ de concentración de PM10 promediadas por cada doce horas es que aparecen síntomas y afecciones que son clasificados como *malas*.

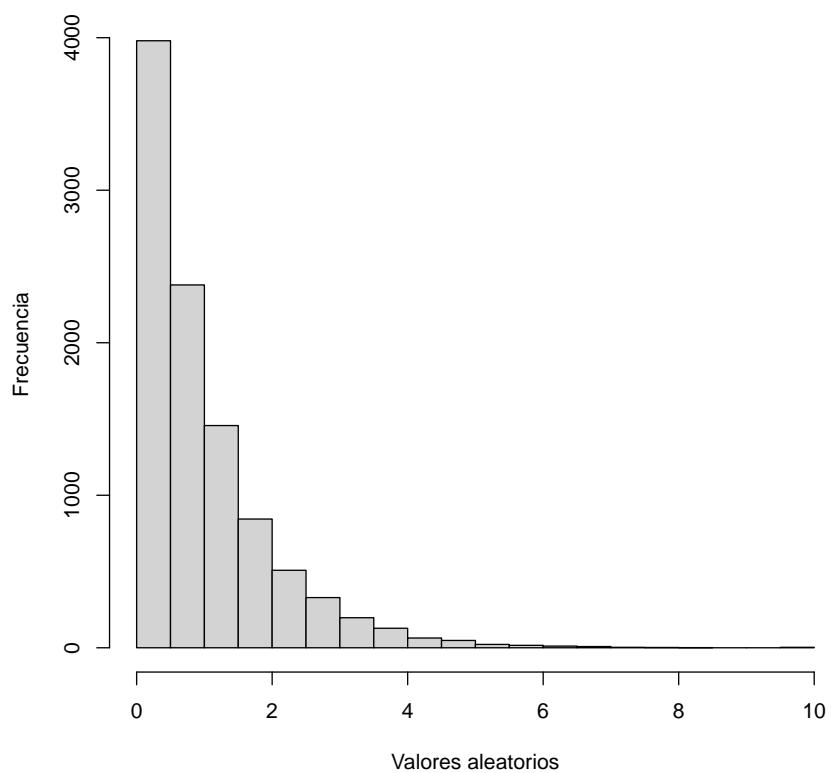
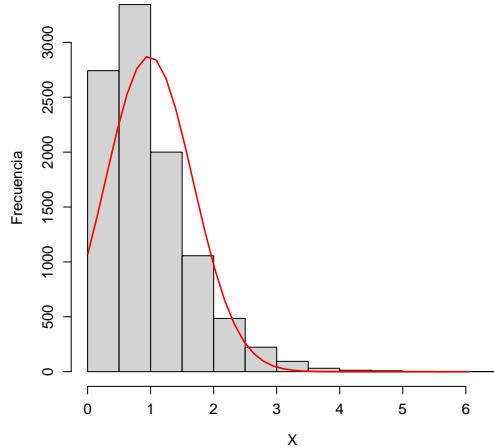
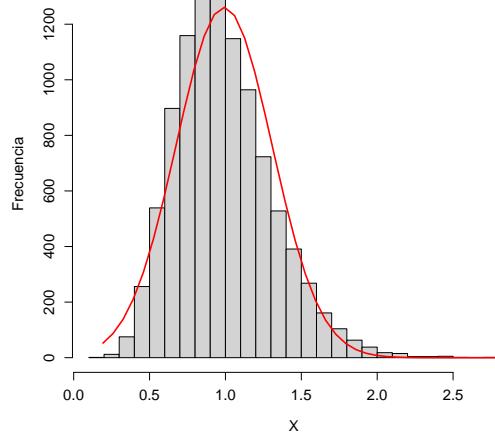


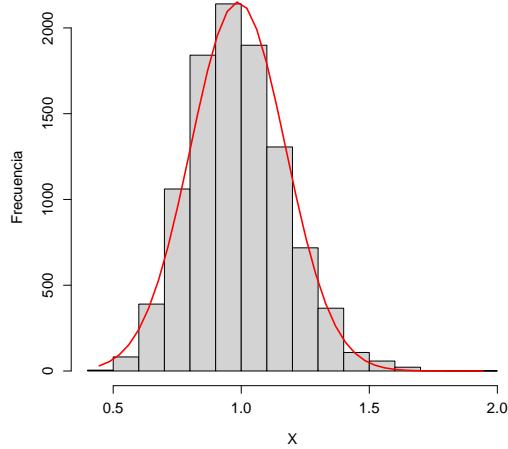
Figura 1: Histograma de mil valores obtenidos de una distribución exponencial con $\lambda = 1$.



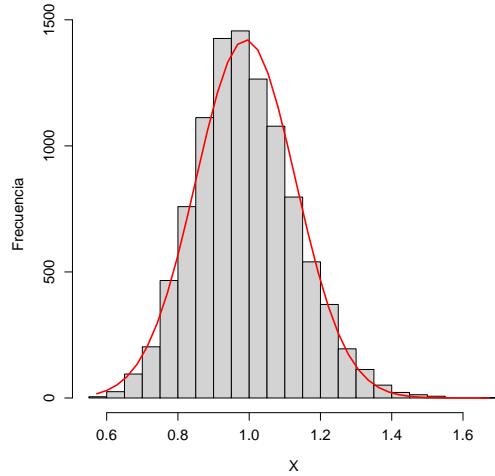
(a) $n = 2$.



(b) $n = 10$.



(c) $n = 30$.



(d) $n = 50$.

Figura 2: Distribuciones de muestras con $n = [2, 10, 30, 50]$ variables aleatorias, con densidad teórica (línea roja) superpuesta.

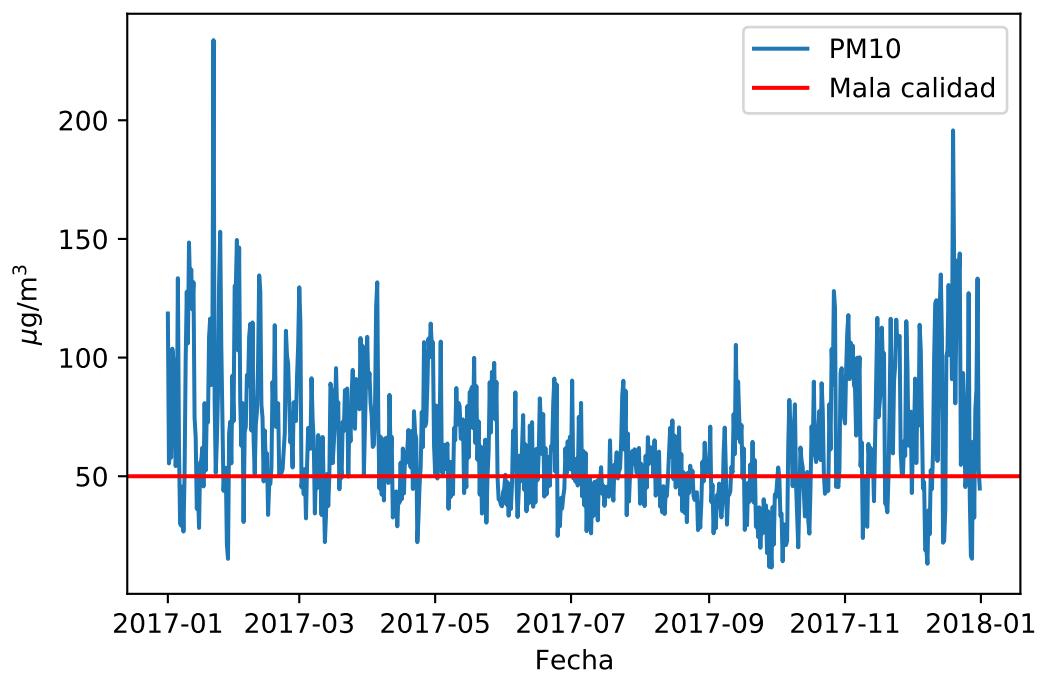
En la figura 3 (p. 5) aparecen la serie de tiempo¹ y el histograma de estas concentraciones, acompañadas por una recta roja que indica el nivel a partir del cual se consideran *mala* la cantidad de contaminantes.

Una justificación de un estudio sobre estos contaminantes podría obtenerse mediante el teorema del límite central. Así, se puede preguntar cuál es la probabilidad de que de una muestra suficientemente grande de las concentraciones de esos valores esté por encima de lo considerado como *malo*, o sea $50\mu\text{g}/\text{m}^3$. A partir del teorema del límite central, se puede considerar que dicha muestra tendrá una distribución normal con $\mu = 64.03$ y $\sigma^2 = 27.82$ calculadas a partir de los registros. Finalmente, mediante `pnorm(50, 64.03, 27.82, lower.tail = F)` se calcula $P\left(\frac{\bar{X}-64.03}{27.82} \geq 50\right) = 0.69$, es decir que teóricamente se puede esperar con un 69 % de probabilidades, encontrar concentraciones *malas* al tomar muestras de manera aleatoria entre los registros de PM10 durante 2017 en el Área Metropolitana de Monterrey, lo que motiva su análisis y exploración de la relación entre estos contaminantes y las afecciones que pudieran ocasionar.

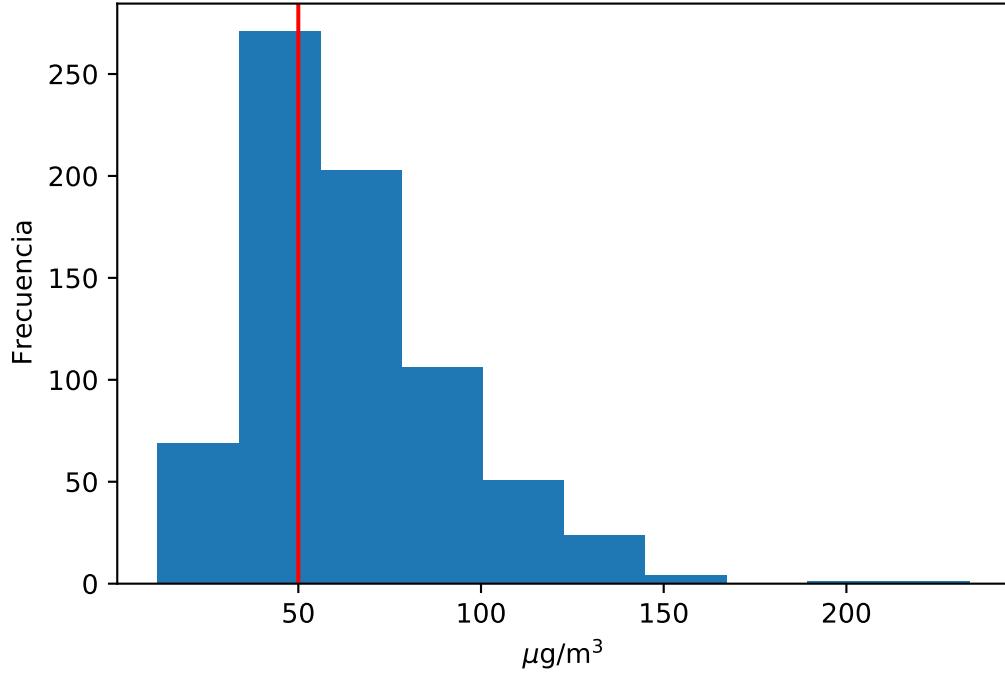
REFERENCIAS

- [1] SECRETARÍA DE GOBERNACIÓN DE MÉXICO (), *Norma Oficial Mexicana NOM-172-SEMARNAT-2019, Lineamientos para la obtención y comunicación del Índice de Calidad del Aire y Riesgos a la Salud.*, SEMARNAT, URL https://www.dof.gob.mx/nota_detalle.php?codigo=5579387&fecha=20/11/2019.
- [2] SISTEMA INTEGRAL DE MONITOREO AMBIENTAL [SIMA] NUEVO LEÓN (2020), «aire.nl.gob.mx | Home», <http://aire.nl.gob.mx/>, [Accedido 14/may/2020].

¹https://es.wikipedia.org/wiki/Serie_temporal



(a) Serie de tiempo.



(b) Histograma.

Figura 3: Serie de tiempo e histograma de las concentraciones en promedio por cada doce horas de PM10 durante 2017 en el área metropolitana de Monterrey.

Propuesta de proyecto final

Alberto Benavides

Posgrado en Ingeniería de Sistemas
Facultad de Ingeniería Mecánica y Eléctrica
Universidad Autónoma de Nuevo León

14 de diciembre de 2020

1. Relación entre dos series de tiempo

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.

2. Pronosticabilidad de una serie de tiempo a partir de otra

También inspirado por el tema de tesis que actualmente trabajo, otra propuesta es pronosticar una serie de tiempo a partir de otra, particularmente pronosticar la serie de tiempo d e consultas de una enfermedad respiratoria a partir de series de tiempo de algún contaminante del aire, mediante la prueba de causalidad de Wiener-Granger que implica aproximaciones autorregresivas.

3. Selección de modelos para series de tiempo por AIC o BIC

Seleccionar un modelo adecuado para pronóstico de series de tiempo es otro de los temas que captaron mi atención porque generalmente parece una cuestión de prueba y error, sin embargo existen metodologías que utilizan el criterio de información de Akaike o Bayesiano para determinar qué modelo explica mejor una serie de tiempo. Las ventajas de estos dos criterios es que funcionan como selección de características en las series de tiempo, por lo que también se pueden usar para determinar autocorrelaciones más significativas para usar en pronósticos o relaciones entre series de tiempo.

Retroalimentación a propuestas de compañeros

Alberto Benavides

Posgrado en Ingeniería de Sistemas
Facultad de Ingeniería Mecánica y Eléctrica
Universidad Autónoma de Nuevo León

15 de diciembre de 2020

1. Johana

Propuesta: *Realizar un diseño de experimentos para evaluar entre un grupo de personas (por medio de una encuesta) la probabilidad estimada que tienen de ingresar a la Universidad considerando factores como la edad, genero, raza, ingresos mensuales, experiencia laboral, años de educación y si están trabajando.*

Retro: En este caso, me parece que habría que definir sobre qué universidad se desea hacer el estudio, pues una primera dificultad que encuentro es que a veces no se puede acceder a los datos que se propone estudiar ya sea porque no se capturan consistentemente o porque no se comparten por motivos de protección de privacidad. En cuanto a las variables que se desean medir, me parece conveniente revisar la relevancia de las presentes y también si existen otros factores que pueden ser considerados, especialmente porque existe mucha bibliografía y aproximaciones en torno a este tema. Respuesta: La encuesta se haría a cualquier persona, ya que se quiere estimar la probabilidad de que se ingrese a la Universidad y escogí las variables porque son los factores que he visto que más influyen en la decisión de entrar o no a la Universidad.

2. Óscar

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

Retro: Este tema resulta de especial interés para muchas empresas, pero generalmente el pronóstico a través de estos modelos depende en gran parte de sus autocorrelaciones parciales, por lo que habría que especificar la empresa y el tipo de datos con que se cuenta. Centralmente es importante determinar la frecuencia de captura de datos y la frecuencia con que se desea pronosticar dichos datos. De todas formas, convendría contemplar la posibilidad de toparse con series de tiempo no estacionarias y también proponer estrategias donde se pasen por alguna metodología para volverlas estacionarias o contemplar otros modelos con los que hacer el pronóstico, como el de Holter-Winters.

3. Palafox

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

Retro: Esta idea de aplicar caminatas aleatorias como metodologías para pronóstico de series de tiempo que se consideran irregulares suena bastante interesante en tanto aproximación estocástica para un proceso que también se entiende como estocástico. Me quedan dudas sobre si en esta investigación se abordarán algunos presupuestos de las series de tiempo tales como estacionariedad o relación con ruido blanco, además de otros modelos que usualmente se utilizan para pronóstico de series de tiempo relacionadas con el mercado de valores.

Seleccionar modelos de pronóstico para series de tiempo de contaminantes de PM10 por criterio de Akaike y bayesiano

Alberto Benavides

Nuevo León, México

Abstract

Selecting time series by selection criteria models

Keywords: series de tiempo, AM, MA, ARMA, modelos de selección de criterios, Akaike, Bayes, AIC, BIC, PM10, Monterrey

1. Introduction

La capacidad de pronosticar acertadamente es una de las habilidades más valoradas en muchos de los ámbitos humanos que incluso aparece elogiado desde narraciones bíblicas y otras provenientes del periodo griego clásico [1]. La certeza de estos pronósticos ha sido relevante en la prevención de desastres naturales [2], el tratamiento preventivo de determinadas enfermedades (principalmente cáncer [3, 4]) o la elección de estrategias ventajosas en operaciones bursátiles [5, 6].

En los modelos de pronóstico es común utilizar estrategias intuitivas para elegir los parámetros con los que se harán las predicciones, sin embargo estas aproximaciones pueden considerarse poco formales como lo fueron las artes adivinatorias o proféticas representadas popularmente por Nostradamus [7].

En este artículo se pretenden utilizar criterios robustos para realizar la selección de parámetros en el pronóstico de contaminantes en el Área Metropolitana de Nuevo León, México (AMM). Para ello, se abordan en la sección 2 los fundamentos relacionados a las series de tiempo, sus características, modelos de pronóstico y los criterios con los que hará la selección de sus parámetros. Posteriormente, en la sección 3 se describen los datos, su origen, manipulación y preprocesamiento. Luego, se muestra en la sección 4 la meto-

dología a la que se sometieron y los resultados obtenidos para, en la sección 5, mostrar las conclusiones obtenidas.

2. Marco teórico

Las series de tiempo son conjuntos de observaciones tomadas a lo largo del tiempo sobre algún evento, formalmente definidas [8] a partir del concepto de una familia de variables aleatorias $Z(\omega, t)$, con un espacio muestral ω y un conjunto de índices temporales t , en las que para una determinada t , $Z(\omega, t)$ es una variable aleatoria, y para una ω dada, $Z(\omega, t)$ es una **serie de tiempo** que, por comodidad, se denomina Z_t .

De estas series de tiempo se puede calcular su media

$$\mu_t = E(Z_t), \quad (1)$$

y varianza

$$\sigma_t^2 = E(Z_t - \mu_t)^2; \quad (2)$$

a partir de estas, la covarianza entre dos tiempos t_1, t_2

$$\gamma(t_1, t_2) = E(Z_{t_1} - \mu_{t_1})(Z_{t_2} - \mu_{t_2}), \quad (3)$$

y la correlación también entre dos tiempos t_1, t_2

$$\rho(t_1, t_2) = \frac{\gamma(t_1, t_2)}{\sigma_{t_1}\sigma_{t_2}}. \quad (4)$$

Esto permite definir la **función de autocorrelación** (ACF) ρ_k como la correlación que tiene una serie consigo misma en los tiempos $t_1 = 0, t_2 = k$, es decir

$$\rho_k = \frac{\gamma_k}{\gamma_0} \quad (5)$$

Ahora se introduce el concepto de **regresión lineal** entendida como la relación lineal entre la variable dependiente X_i y la variable independiente Y_i para $i = [1, \dots, n]$, tal que

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \quad (6)$$

Aquí, $\beta_0 + \beta_1 X_i$ son la función lineal que mejor se ajusta a X_i con base en el menor de los errores cuadrados, mientras que ϵ_i es la distancia entre

dicha recta y el valor de Y_i para determinada i . Cuando se tienen p variables dependientes $X_{ij}, j = [1, \dots, p]$, la regresión lineal se escribe

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \epsilon_i. \quad (7)$$

Con esto, la **función de autocorrelación parcial** (PACF) para un retraso de k unidades, queda definida a partir de la ecuación 7 como

$$Z_t = \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \dots + \beta_k Z_{t-k} + \epsilon_t, \quad (8)$$

siendo el coeficiente β_k el que define la interacción del retraso k en la serie de tiempo Z_t .

Ahora bien, un **modelo autorregresivo** (AR) es uno de los modelos utilizados para el pronóstico de series de tiempo. Este modelo se basa en la idea de que una serie de tiempo Z_t puede ser pronosticada Y_t a partir de la información proporcionada por las regresiones de momentos pasados de dicha serie de tiempo. A saber:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \epsilon_t \quad (9)$$

es el modelo autorregresivo de orden p , abreviado AR(p). En este tipo de modelos, es una práctica común utilizar los valores significativos de la PACF de una serie de tiempo como los valores p para obtener pronósticos.

Otro modelo utilizado para pronóstico es el **modelo de media móvil** en que en lugar de utilizar valores pasados de Z_t para realizar el pronóstico, como en los AR, se usan los errores de pronóstico ϵ_t de las predicciones con diferentes retrasos y coeficientes θ , por lo que

$$Y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}. \quad (10)$$

Estos dos modelos se combinan para formar el **modelo autorregresivo de media móvil** (ARMA) que tiene la forma

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}, \quad (11)$$

denominado ARMA(p, q).

También, una serie de tiempo se puede entender a partir de su descomposición en la *tendencia* T_t , los *residuales* R_t y la **estacionalidad** S_t [9]. La tendencia es equivalente a β_0 y los residuales a los ϵ_t , mientras que la

estacionalidad explica los patrones o ciclos que se hallan en las serie. Estas componentes pueden encontrarse sumadas

$$Z_t = T_t + R_t + S_t \quad (12)$$

o bien, multiplicadas:

$$Z_t = T_t \times R_t \times S_t. \quad (13)$$

Los modelos de pronóstico basados en el AR y MA son eficientes cuando se parte de una serie de tiempo **estacionaria**, esto es, aquella que tiene una media μ_t y varianza σ_t^2 constantes y no es estacional. Para determinar si una serie de tiempo es estacionaria, se cuenta con las pruebas de hipótesis de Dickey-Fuller aumentada (ADF) y la de Kwiatkowski-Phillips-Schmidt-Shinn (KPSS).

La hipótesis nula de la prueba ADF es que la serie es no estacionaria con un valor $p = 0.05$. Cuando no se puede rechazar la hipótesis nula, es posible hacer una diferenciación Z'_t de la serie de tiempo Z_t mediante

$$Z'_t = Z_t - Z_{t-d} \quad (14)$$

donde d es el retraso dado. Al realizar este proceso, el modelo ARMA se llama ARIMA, donde la I viene de *integrated* en inglés, que puede traducirse como diferenciada en este contexto. Se dice que se aplica un modelo ARIMA(p, d, q) a partir de un modelo AR(p), MA(q) con una serie de tiempo diferenciada d unidades.

A partir de la ACF, PACF de una serie de tiempo se pueden elegir las variables p, f de un modelo ARMA o ARIMA a partir de los valores estadísticamente significativos de dichas series. Las combinaciones de estos valores pueden ser variadas e incluir más o menos parámetros en los modelos o más o menos exactitud respecto a la serie que se desea pronosticar, por lo que se utilizan algunos criterios para determinar cuáles son las mejores combinaciones de parámetros para este tipo de pronósticos. Los dos modelos más usados son el Akaike (AIC) y el bayesiano (BIC).

El criterio de información de Akaike se describe como

$$\text{AIC} = -2 \log L + 2(p + q + d + 1), \quad (15)$$

donde L es la similitud (definida en la ecuación 7.7.2 de [I]) entre la series de tiempo Z_t y el modelo Y_t . Por último, el criterio de información bayesiano

[10] en series de tiempo depende del AIC, pero también toma en cuenta el número de muestras n en el modelo, así que se puede escribir

$$\text{BIC} = -2 \log(L) + \ln(n) \cdot (p + q + d + 1). \quad (16)$$

Para estos criterios, es preferible un valor pequeño respecto a otro mayor porque esto implica el uso de menos parámetros ($p + q + d + 1$) y una mayor semejanza $2 \log L$ entre serie de tiempo y función pronosticada.

3. Datos

La serie de tiempo que se desea pronosticar proviene de los registros de calidad de aire obtenidos del Sistema Integral de Monitoreo Ambiental de Nuevo León (Méjico) [11] que cuenta con trece estaciones de monitoreo ambiental (cuya ubicación se muestra en la figura 1, p. 6) que registran fecha y hora, estación meteorológica, presión atmosférica, precipitaciones, humedad, radiación solar, temperatura, velocidad, dirección del viento y los contaminantes CO, NO, NO₂, O₃, SO₂, PM10, PM2.5. Algunas de estas estaciones registran datos desde 1993, y sólo coinciden su operación a partir de 2017, como se muestra en la figura 2 (p. 7).

Se llaman contaminantes PM10 a los que tienen que ver con partículas suspendidas con tamaño menor o igual a $10\mu\text{m}$. Las altas concentraciones de estos contaminantes están relacionadas con enfermedades respiratorias [12] y muertes prematuras en población de riesgo [13].

Este estudio resulta interesante porque porque la Secretaría de Gobernación de México publicó la norma Norma NOM-172-SEMARNAT-2019 [14] en la que determina las concentraciones en las que el contaminante PM10 se consideran malas, mismas que se hallan en el cuadro 1 (p. 6). Al extraer los datos de PM10 durante el 2017 en el AMM, se comprueba que al menos un 65 % de los registros tienen una calidad considerada mala por la Secretaría de Gobernación de México, lo cual puede constatarse en la figura 3 (p. 7).

4. Metodología y resultados

La serie de tiempo se descompuso en las componentes de tendencia, estacionalidad y residuales, como se muestra en la figura 4 (p. 8). En estas imágenes, se puede observar que la tendencia β_0 no se mantiene constante, por lo que se realiza la prueba de Dickey-Fuller aumentada y se obtiene un valor $p = 0.51$, por lo que no se puede rechazar la hipótesis de que la serie

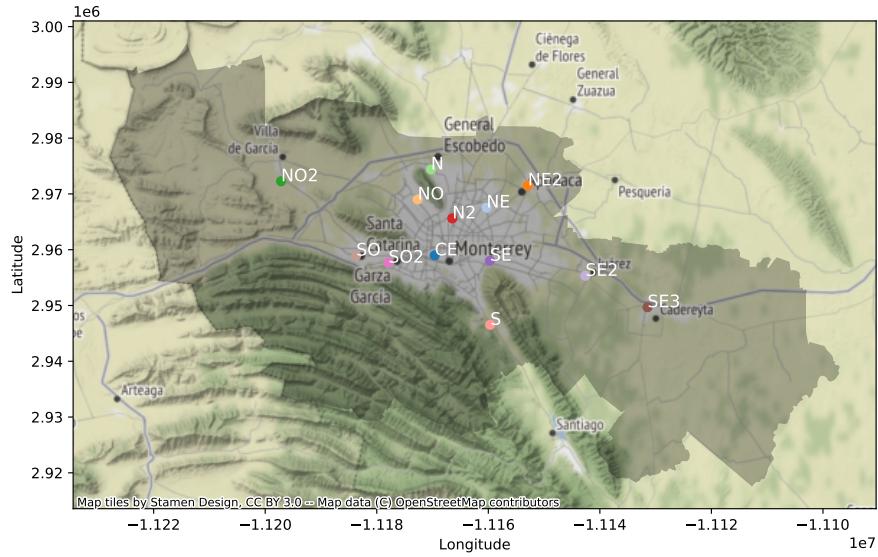


Figura 1: Trece estaciones de monitoreo del Área Metropolitana de Monterrey (Nuevo León, México)

Tabla 1: Índice de aire y salud para PM10.

Calidad del aire	Nivel de riesgo	12 horas ($\mu\text{g} / \text{m}^3$)
Buena	Bajo	[0, 50)
Aceptable	Moderado	[50, 75)
Mala	Alto	[75, 155]
Muy mala	Muy alto	[155, 235]
Extremadamente mala	Extremadamente alto	> 235

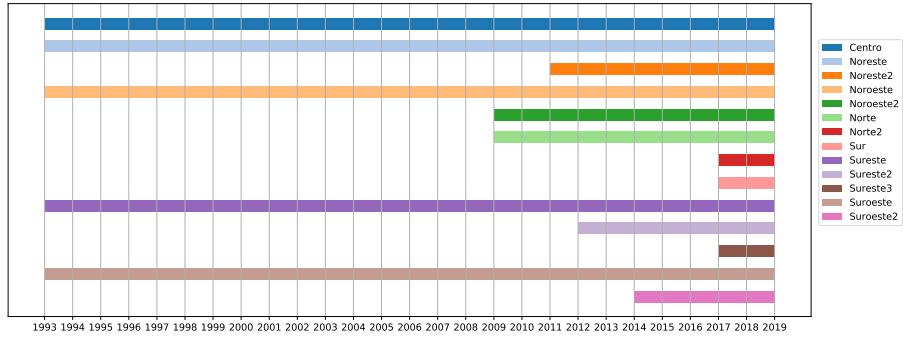


Figura 2: Barras de tiempo en que han estado activas las estaciones de monitoreo del AMM.

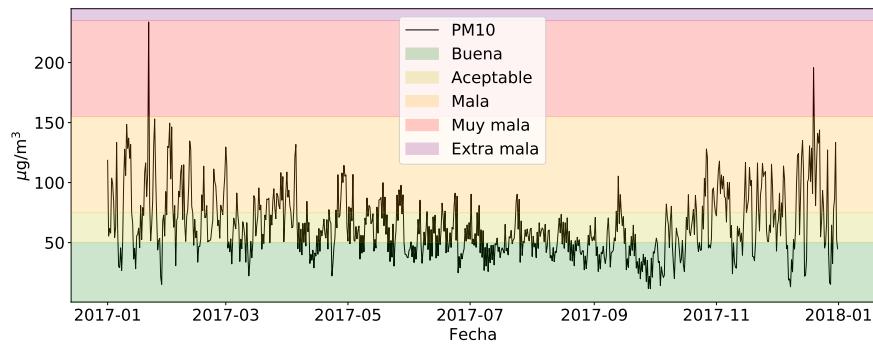


Figura 3: Serie de tiempo de PM10, durante el año 2017 para el AMM.

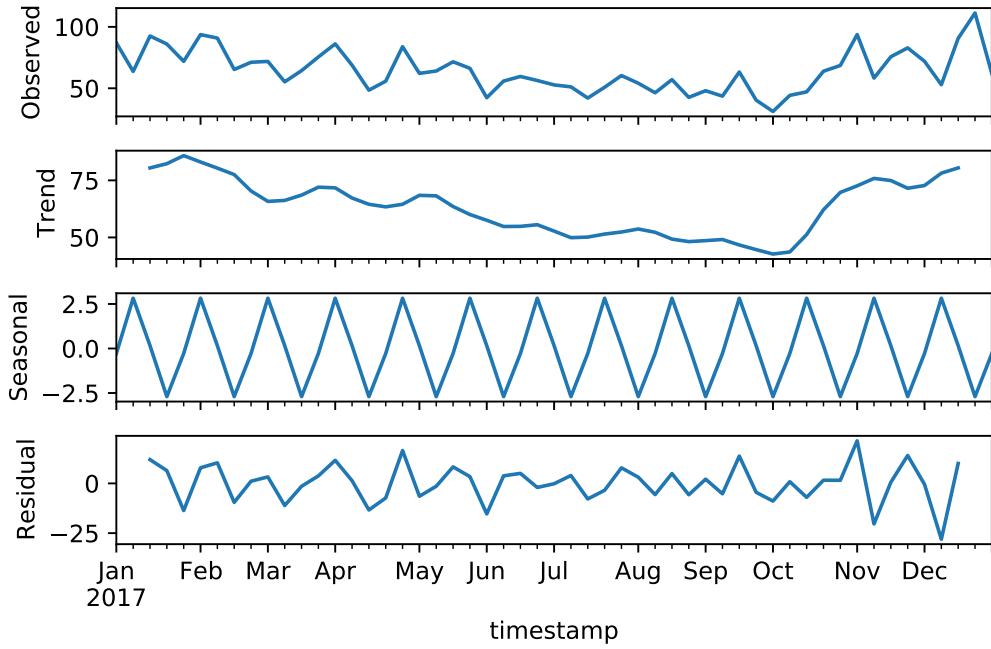


Figura 4: Serie de tiempo de PM10 de 2017 descompuesta en tendencia, estacionalidad y residuales.

no es no estacionaria, de modo que se debe hacer una diferenciación de la misma aplicando la ecuación [14]. La serie tuvo que ser diferenciada dos veces, por lo que $d = 2$. La serie de tiempo original y la diferenciada con $d = 2$ están plasmadas en la figura [5].

También se obtuvieron sus ACF y PACF, incluidas en las figuras [6] y [6] (p. [10]). En este caso, no hay manera de seleccionar por ninguna de las funciones de correlación un buen conjunto de parámetros, por lo que se procederá a generar modelos ARIMA($p, 2, q$) con $p = [1, \dots, 10]$, $q = [1, \dots, 10]$, de los que se calcula el AIC y BIC, además de la suma de ambos y luego se muestran los diez menores valores de AIC + BIC y sus configuraciones de p y q en la tabla [2] (p. [10]). En estos datos, se ve que la mejor combinación de valores es el modelo ARIMA(2, 2, 5).

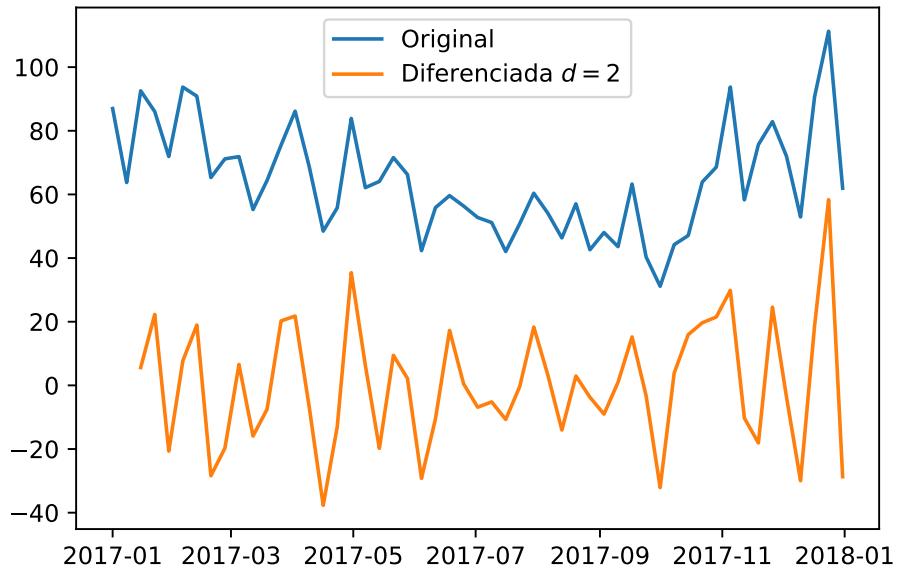


Figura 5: Serie de tiempo de PM10 de 2017 (azul) y la diferenciada en $d = 2$ (naranja).

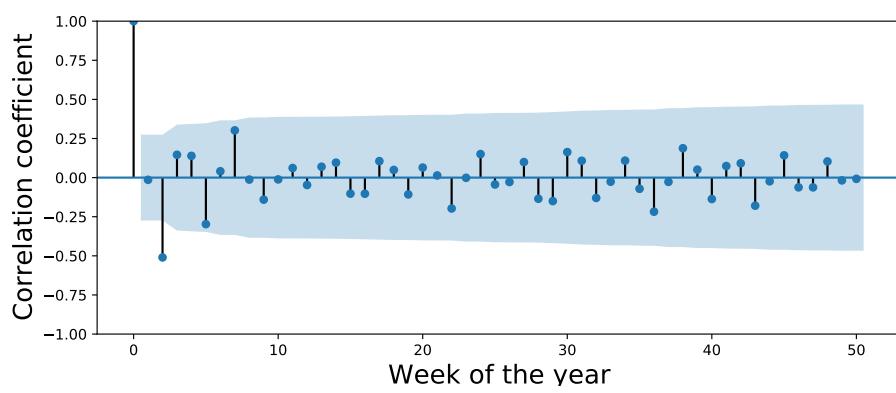


Figura 6: Coeficientes de autocorrelación para PM10 durante 2017 en el AMM.

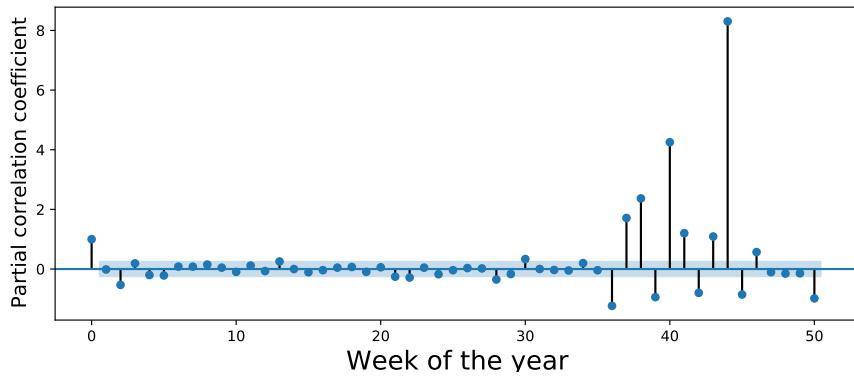


Figura 7: Coeficientes de autocorrelación parciales para PM10 durante 2017 en el AMM.

Tabla 2: Diez mejores combinaciones de p y q para ARIMA con base en la suma de AIC + BIC.

p	q	AIC	BIC	AIC + BIC
2	5	422.78	439.80	862.58
5	2	424.32	441.35	865.66
4	3	424.36	441.38	865.74
2	3	427.56	440.80	868.36
1	4	427.60	440.84	868.45
6	6	421.36	447.85	869.21
5	6	423.18	447.77	870.95
2	4	428.02	443.16	871.18
6	1	430.17	447.20	877.37
5	1	431.18	446.32	877.50

5. Conclusiones

Las dificultades que los modelos de pronóstico, como el ARIMA, presentan es que el tanteo de los parámetros del modelo de pronóstico pueden no conseguirse intuitiva ni directamente a partir de las ACF, PACF y diferenciación necesaria por prueba de estacionaridad. Los criterios de selección de modelos de Akaike y bayesiano resultan una herramienta estadística eficiente para solventar estas limitaciones. En general, sin embargo, es recomendable seguir la metodología descrita para tener una mejor información de la serie de tiempo que se desee pronosticar, pues desde la descomposición se tiene información relevante en la planeación de los modelos y estrategias a seguir en el pronóstico.

Referencias

- [1] R. Hyndman, Forecasting : principles and practice, OTexts, Melbourne, 2018.
- [2] H.-W. Cheng, Arima models for forecasting poisson process observations: Application to the volcanoes worldwide (2007). [doi:10.25669/1RU9-PIWG](https://doi.org/10.25669/1RU9-PIWG).
- [3] A. Earnest, S. M. Evans, F. Sampurno, J. Millar, Forecasting annual incidence and mortality rate for prostate cancer in australia until 2022 using autoregressive integrated moving average (ARIMA) models, BMJ Open 9 (8) (2019) e031331. [doi:10.1136/bmjopen-2019-031331](https://doi.org/10.1136/bmjopen-2019-031331).
- [4] N. Kumar, P. Kumari, P. Ranjan, A. Vaish, ARIMA model based breast cancer detection and classification through image processing, in: 2014 Students Conference on Engineering and Systems, IEEE, 2014. [doi:10.1109/sces.2014.6880070](https://doi.org/10.1109/sces.2014.6880070).
- [5] B. A. R. SK, Exchange rate forecasting using ARIMA, neural network and fuzzy neuron, Journal of Stock & Forex Trading 04 (03) (2015). [doi:10.4172/2168-9458.1000155](https://doi.org/10.4172/2168-9458.1000155).
- [6] Y. Xiao, J. Xiao, J. Liu, S. Wang, A multiscale modeling approach incorporating ARIMA and anns for financial market volatility forecasting, Journal of Systems Science and Complexity 27 (1) (2014) 225–236. [doi:10.1007/s11424-014-3305-4](https://doi.org/10.1007/s11424-014-3305-4).

- [7] R. H. Popkin, Predicting, prophecyng, divining and foretelling from nostradamus to hume, *History of European Ideas* 5 (2) (1984) 117–135. [doi:10.1016/0191-6599\(84\)90063-9](https://doi.org/10.1016/0191-6599(84)90063-9).
- [8] W. Wei, Time series analysis univariate and multivariate methods, Pearson, Boston, 2019.
- [9] P. Brockwell, Introduction to time series and forecasting, Springer, New York, 2002.
- [10] Introduction to bayesian thinking, in: Bayesian Computation with R, Springer New York, pp. 19–37. [doi:10.1007/978-0-387-71385-4_2](https://doi.org/10.1007/978-0-387-71385-4_2).
- [11] SISTEMA INTEGRAL DE MONITOREO AMBIENTAL [SIMA] NUEVO LEÓN, aire.nl.gob.mx — Home, <http://aire.nl.gob.mx/>, [Accessed 14/may/2020] (2020).
- [12] C. A. Pope, D. W. Dockery, J. D. Spengler, M. E. Raizenne, Respiratory health and PM10pollution: A daily time series analysis, *American Review of Respiratory Disease* 144 (3_pt_1) (1991) 668–674. [doi:10.1164/ajrccm/144.3_pt_1.668](https://doi.org/10.1164/ajrccm/144.3_pt_1.668).
- [13] K. Ito, G. Thurston, Daily pm10/mortality associations: an investigations of at-risk subpopulations., *Journal of exposure analysis and environmental epidemiology* 61 (1996) 79–95.
- [14] Secretaría de Gobernación de México, Norma Oficial Mexicana NOM-172-SEMARNAT-2019, Lineamientos para la obtención y comunicación del Índice de Calidad del Aire y Riesgos a la Salud., SEMARNAT. URL https://www.dof.gob.mx/nota_detalle.php?codigo=5579387&fecha=20/11/2019