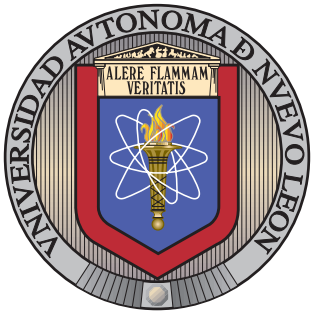
Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

M. C. DE LA ING. OR. SISTEMAS

Maestría



Portafolio de Evidencias

de

Joaquin Arturo Velarde Moreno

1649290

para el curso de Modelos Probabilistas Aplicados,

con la profesora Satu Elisa Schaeffer.

Semestre Agosto 2020 - Enero 2021.

https://github.com/joaquin3600/Modelos_Probabilistas_Aplicados

Universidad Autónoma de Nuevo León Facultad de Ingeniería Mecánica

TAREA 1:
MODELOS PROBABILISTAS APLICADOS. ANÁLISIS DE VISITANTES INTERNACIONALES QUE INGRESARON AL PAÍS EN EL 2020

Profesor:
DRA. SATU ELISA SCHAEFFER

Alumno:
Joaquín Arturo Velarde Moreno

Matería:
MODELOS PROBABILISTAS APLICADOS

Origen de los datos:

Los datos que a continuación se analizarán fueron publicados por el INEGI[1] en el apartado de Turismo y se obtuvieron mediante la aplicación de una serie de encuestas a viajeros que arribaron del extranjero a México durante el periodo de agosto del 2018 a junio del 2020. Dichos datos se publicaron en formatos XLSX.

Metodología:

Esta sección de datos se concentró, generó y editó en Microsoft Excel[2] proveyendo información preliminar de los meses enero, febrero, marzo, abril, mayo y junio.

Además. se creó un diagrama de cajas y bigotes que representa de manera visual el conjunto de visitantes por mes, obteniendo los siguientes datos: mínimo, máximo, media, primer cuartil y tercer cuartil. Los cuales fueron analizados utilizando el programa R versión 4.0.2[3].

En la figura 0.1 se puede observar la clasificación del tipo de turistas categorizándose en cuatro apartados:

1. turistas de internación
2. turistas fronterizos
3. excursionistas fronterizos
4. excursionistas en cruceros.

Turistas de internación:

Los turistas de internación son los que se adentran en el territorio nacional por vía aérea y terrestre, aunque la mayoría, como se puede observar, prefieren la internación aérea.

Turistas fronterizos:

Los turistas fronterizos son aquellos que viajan a las ciudades fronterizas de nuestro país desplazándose ya sea mediante automóvil o de manera peatonal.

Excursionistas fronterizos:

Los excursionistas fronterizos cruzan principalmente la frontera para hacer compras y retornan sin quedarse en el territorio nacional, ya sea en automóvil o de forma peatonal.

Excursionistas de crucero:

Los excursionistas de crucero son pasajeros que visitan un puerto mexicano sin quedarse la noche en México; es decir, llegan al puerto, conocer, hacen compras y regresan al crucero. Resalta en este apartado que, debido a la crisis sanitaria por la pandemia del COVID-19, en los últimos meses, el número de excursionistas de crucero ha sido de 0.

Gráfico:

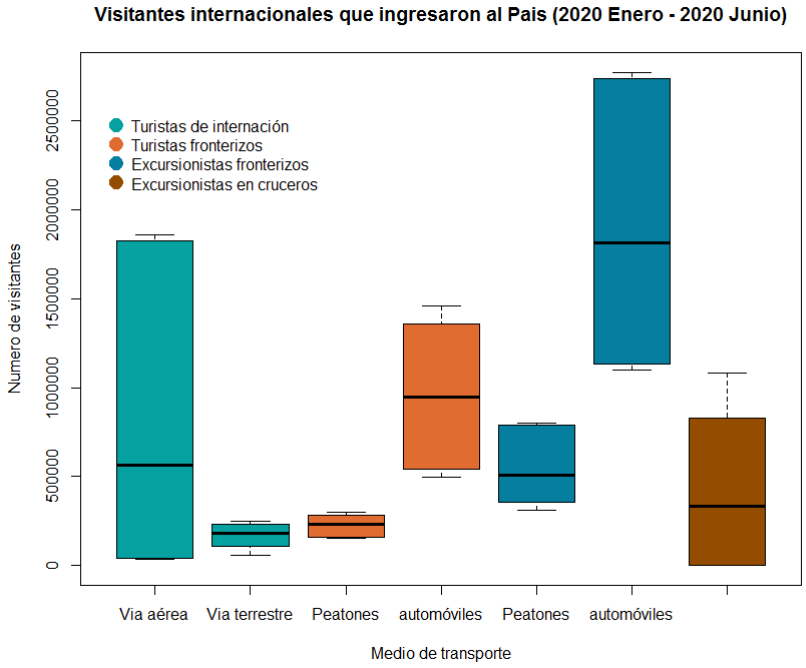


Figure 0.1: Visitantes internacionales por tipo

Referencias

[1] INEGI. *Visitantes internacionales que ingresaron al país por número, gasto total y gasto medio, según el mes*. `https://www.inegi.org.mx/contenidos/temas/economia/turismo/EVI_R_01.xlsx`. 2020.

[2] Microsoft Office. *Microsoft Excel*. `https://www.microsoft.com/es-mx/microsoft-365/excel`. 2011.

[3] The R Foundation. *The R Project for Statistical Computing*. `https://www.r-project.org/`. 2019.

La frecuencia de uso de palabras en el libro: Precepts in Practice; or, Stories Illustrating the Proverbs by A. L. O. E.

TAREA 2

Alumno:
Joaquín Arturo Velarde Moreno

1 Introducción:

En el presente trabajo se hizo el análisis de la frecuencia en el uso de una selección de palabras y letras contenidas en el libro de Precepts in Practice; or, Stories Illustrating the Proverbs del autor A. L. O. [1] Este libro se encuentra en la biblioteca digital de Gutenberg [2], el cual es el repositorio de más de 63,127 títulos en formato ebook.

Esta obra que revisamos está escrita en inglés y se conforma por un conjunto de relatos cuyas ilustraciones tienen el objetivo de complementar visualmente cada proverbio y, de acuerdo con este género literario, cada historia conlleva una enseñanza de la que podemos aprender en nuestra vida cotidiana, según las diferentes situaciones y asuntos de que se trate.

2 Metodología:

Para cumplir con nuestro objetivo, primeramente, procedimos a descargar el texto con el programa "R-4.0.2" [4]., dado que la biblioteca Gutenberg [2] permite tal descarga libremente. Como siguiente paso, se removió todo carácter especial no alfa numérico, es decir, los símbolos tipográficos (guiones, cursivas, negritas, etc.) para limpiar el texto. Debido a que la frecuencia de datos era dominada principalmente por artículos, pronombres personales, conjunciones y números, se hizo una hoja de datos en Microsoft Excel [3] con este tipo de palabras. Posteriormente, se le ordenó al programa limpiar el texto de dichos caracteres. Con el resto del léxico, se pudo calcular la frecuencia de dichas palabras en dos áreas:

1. Análisis de letras
2. Análisis de palabras

2.1 Análisis de grafías:

Para efectuar el análisis de grafías, se creó un diagrama de cajas y bigotes que representan la frecuencia de aparición de las letras en el texto Figure 2.1, obteniendo los siguientes datos: mínimo, máximo, media, primer cuartil y tercer cuartil utilizando, como se señaló antes, el programa R 4.0.2 [4].

Se decidió, para mayor claridad, dar un enfoque al conjunto que inicia desde el segundo quantil y cuarto quantil obtenidas con la función *summary()* para, de este modo, concentrar las frecuencias en un rango más cercano. Con esta información se elaboró una figura de barras en la que se muestra la frecuencia de las letras, descubriendo un predominio de una serie de consonantes como son: s, r, d, l, u, w, m, f, c, g, y, p ordenadas de mayor a menor frecuencia, tal y como se muestra en el gráfico Figure 2.2

2.2 Análisis de palabras:

Con la utilización del programa mencionado, también pudimos obtener información acerca de la frecuencia de las palabras y su cantidad contenidas del segundo quantil hasta el cuarto quantil del resumen dado por la function summary() para así concentrar el léxico más frecuente.

Figure 2.3

Referencias

[1] A. L. O. E. *Precepts in Practice; or, Stories Illustrating the Proverbs*. `https://www.gutenberg.org/files/58791/58791-0.txt`. 2019.

[2] Michael Hart. *Project Gutenberg*. `https://www.gutenberg.org/`. 1971.

[3] Microsoft Office. *Microsoft Excel*. `https://www.microsoft.com/es-mx/microsoft-365/excel`. 2011.

[4] The R Foundation. *The R Project for Statistical Computing*. `https://www.r-project.org/`. 2019.

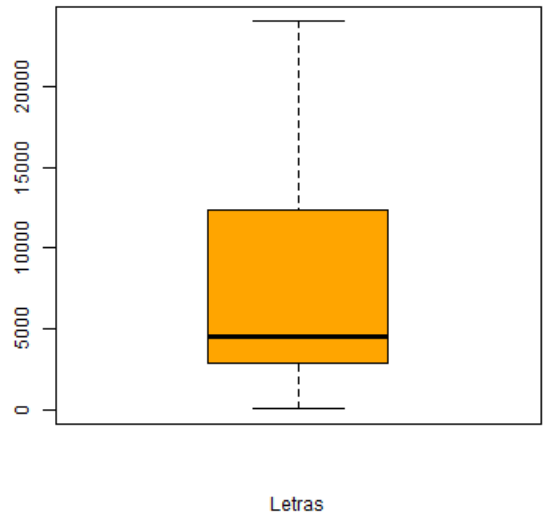


Figure 2.1: Frecuencia de letras representado en diagrama de cajas

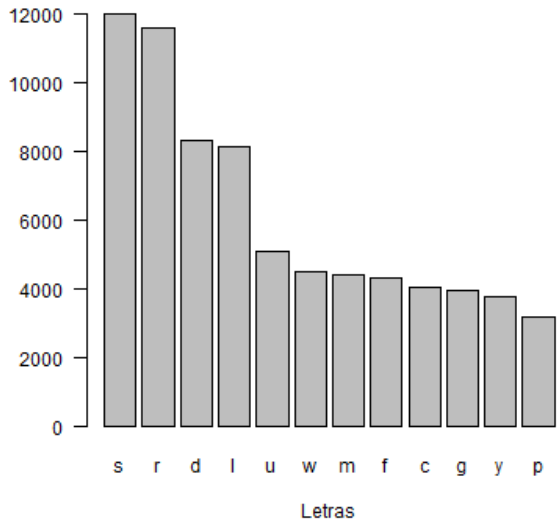


Figure 2.2: Frecuencia de letras representado en diagrama

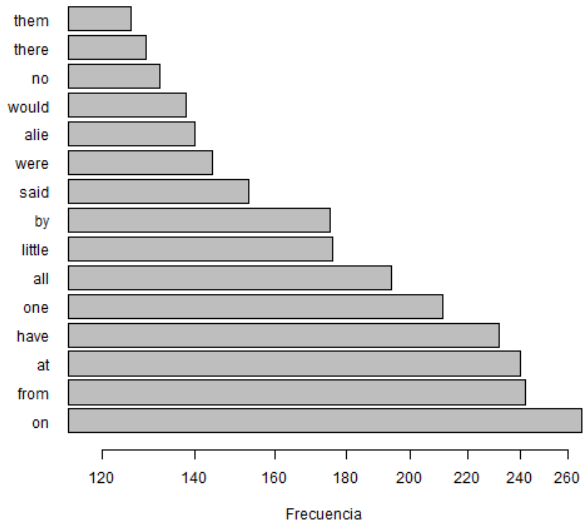


Figure 2.3: Frecuencia de Palabras representado en diagrama

distribucion de frecuencias en el numero de palabras del libro: Precepts in Practice; or, Stories Illustrating the Proverbs by A. L. O. E.

TAREA 3

Alumno:
Joaquín Arturo Velarde Moreno

1. Introducción

En el presente trabajo se hizo el análisis de la frecuencia en el uso de una selección de palabras contenidas en un texto literario con el objeto de calcular las probabilidades de que aparezcan determinadas palabras con cierto número de letras.

El libro que usamos es el de Precepts in Practice; or, Stories Illustrating the Proverbs del autor A. L. O. E. [1] Este libro se encuentra en la biblioteca digital de Gutenberg [2], el cual es el repositorio de más de 63,127 títulos en formato ebook.

Esta obra que revisamos está escrita en inglés y se conforma por un conjunto de relatos cuyas ilustraciones tienen el objetivo de complementar visualmente cada proverbio y, de acuerdo con este género literario, cada historia conlleva una enseñanza de la que podemos aprender en nuestra vida cotidiana, según las diferentes situaciones y asuntos de que se trate.

2. Metodología

Para cumplir con nuestro objetivo, primeramente, procedimos a descargar el texto con el programa R- 4.0.2 [4] , dado que la biblioteca Gutenberg [2] permite tal descarga libremente. Como siguiente paso, se removió todo carácter especial no alfa numérico, es decir, los símbolos tipográficos (guiones, cursivas, negritas, etc.) para limpiar el texto. Debido a que la frecuencia de datos era dominada principalmente por artículos, pronombres personales, conjunciones y números, se hizo una hoja de datos en Microsoft Excel [3] con este tipo de palabras. Posteriormente, se le ordenó al programa limpiar el texto de dicho contenido. Con el resto del léxico, se procedió a calcular la frecuencia de las palabras restantes con el programa R para ver las probabilidades de distribución cuantitativa. Finalmente, se llevó a cabo el análisis para realizar los histogramas utilizando las siguientes cuatro funciones de distribución de probabilidad.

1. Distribución geométrica
2. Distribución hypergeométrica
3. Distribución binomial negativa
4. Distribución regular

2.1. Distribución geométrica

Con este modelo pudimos observar la probabilidad de obtener palabras con menos de tres letras. Figure 2.1

2.2. Distribución hypergeométrica

Con esta distribución, que es primordial en el análisis de muestras pequeñas, se obtiene la probabilidad de que aparezcan cinco palabras que sean mayores de dos caracteres en una muestra de 100 palabras. Figure 2.2

2.3. Distribución binomial negativa

Con esta distribución se repite un experimento hasta juntar un número de casos requeridos de éxito. En nuestro trabajo, con este modelo se obtuvo la probabilidad de 15 palabras conformadas por más de 6 caracteres.Figure 2.3

2.4. Distribución regular

Es la distribución de probabilidad más importante por sus propiedades estadísticas. Supone que en experimentos repetidos, la mayor parte de los resultados coincidirán con un resultado promedio. Con este modelo se obtiene los resultados de un número determinado de experimentos. En nuestro análisis con la aplicación de dicho modelo se obtuvieron las probabilidades de sacar palabras con más de 7 caracteres. Figure 2.4

Referencias

[1] A. L. O. E. *Precepts in Practice; or, Stories Illustrating the Proverbs*. 2019.

[2] Michael Hart. *Project Gutenberg*. `https://www.gutenberg.org/`. 1971.

[3] Microsoft Office. *Microsoft Excel*. `https://www.microsoft.com/es-mx/microsoft-365/excel`. 2011.

[4] The R Foundation. *The R Project for Statistical Computing*. `https://www.r-project.org/`. 2019.

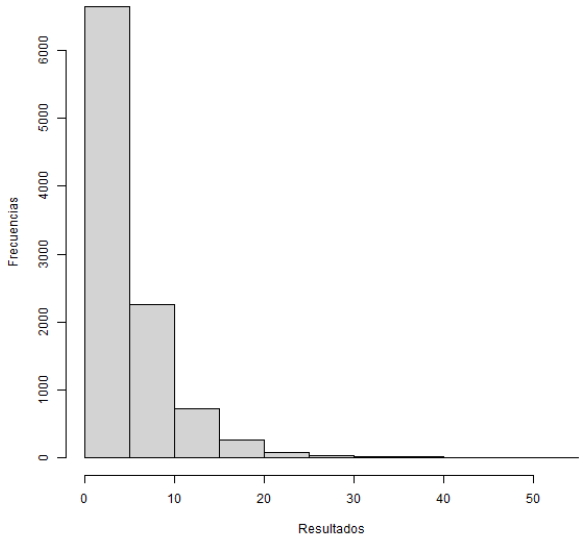


Figura 2.1: Probabilidad de que aparezca una palabra con menos de 3 letras.

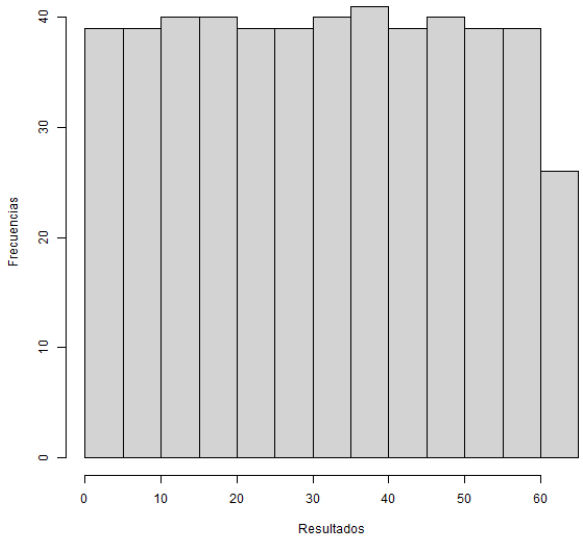


Figura 2.2: Probabilidad de que aparezcan 5 palabras con mas de 2 caracteres.

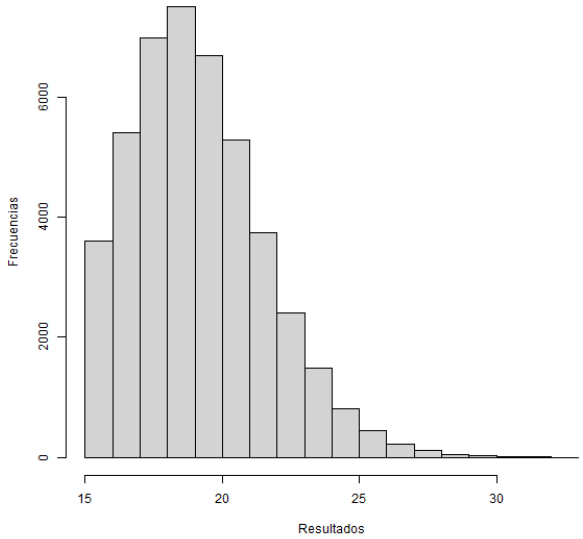


Figura 2.3: Casos de éxito en donde una palabra es mayor que 6 caracteres.

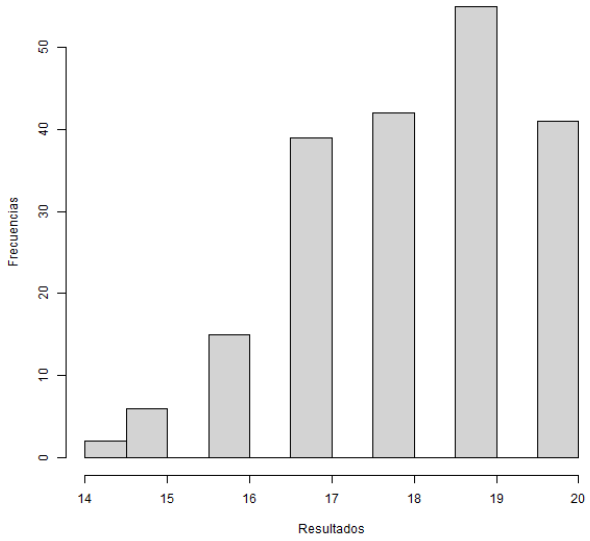


Figura 2.4: Probabilidad de sacar palabras con mas de 7 caracteres.

Distribución de probabilidad de normal

TAREA 5

Alumno:
Joaquín Arturo Velarde Moreno

1. Introducción

El objetivo del siguiente reporte es describir el comportamiento de una distribución normal, ver su representación matemática y la forma de su curva en histogramas, para lo cual usaremos el programa R 4.0.2 [2] y de este modo, haremos cálculos con conjuntos con la finalidad de mostrarlos gráficamente. Además, intentaremos simular una distribución normal a partir de valores uniformes utilizando la transformación Box-Muller. Para cumplir con esta finalidad, usaremos como apoyo el material de la Dra. Elisa Schaefer [1].

2. Definición

La distribución normal o Gaussiana, nombrada así en honor a Carl Friederich Gauss, es una distribución con forma de campana usada para aproximarse al valor de una variable aleatoria continua a una situación ideal. Esta distribución contiene dos parámetros que la definen:μ y σ donde

- μ es la media de la distribución y
- σ es la desviación estándar

Por ejemplo, si elaboráramos una encuesta de la edad de cada uno de los 150 estudiantes de una escuela y obtuviéramos que su promedio es de 18 años, tendríamos un histograma donde se muestra que hay una mayor cantidad de alumnos cerca del promedio o la media Figura 5.1, es decir, hay la tendencia de querer aproximarse a una curva de distribución normal Figura 5.2. Pero si esto lo aproximamos a una situación ideal, no tomando la edad de solo 150 individuos sino de 5000 estudiantes, entonces veremos que la forma de nuestra distribución se acerca a la curva de la distribución normal Figura 5.3, esta curva nos permite ver dos características de distribución, la asimétrica y la asintótica:

a) **Simétrica** , lo cual significa que la media es igual a los valores de la moda y la mediana, en otras palabras, los alumnos que estén más cerca del promedio, son más numerosos que los que están más alejados de este promedio, esto también implica que se cumple la siguiente ecuación:

$$f_X(\mu - x) = f_X(\mu + x)\, \forall x \in \mathbb{R}$$

b) **Asintótica** , significa que se acerca continuamente a la recta del eje x sin llegar nunca a encontrarla, es decir, puede extenderse hasta el infinito. La desviación estándar se representa con la letra griega σ), y nos indica lo lejos que están distribuidos los resultados de la media, lo cual puede producir dos situaciones: una situación homogénea o una heterogénea.

Si la situación es homogénea la desviación estándar σ es muy reducida y los sujetos están muy cercanos a la media μ y la curva será muy afilada Figura 5.3. En cambio, si la situación es heterogénea la desviación σ es muy alta y los sujetos están muy alejados a la media μ y la curva será menos afilada Figura 5.4.

3. Formula

La fórmula para la función de probabilidad normal está definida por la siguiente expresión:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} exp^{(\frac{-(x-\mu)^2}{2\sigma^2})}$$

donde:

- μ es la media de la distribución.
- σ es la desviación estándar.

Dado que esta ecuación es demasiado complicada, incluso haciendo uso de la calculadora, para este tipo de cálculos se suele hacer uso de la herramienta R [2], en la cual ya está predefinida con la siguiente función *dnorm()*, y si queremos generar un vector de una distribución normal, podemos utilizar el método *rnorm()*.

```
Promedio <- 18
Edad     <- rnorm(500000,Promedio,1)
```

Si esto lo graficamos mediante un histograma, tendremos la misma distribución normal que se espera Figura 5.3

4. Método de Box-Muller

También podemos generar una distribución normal por otros métodos tales como el método de Box-Muller, el cual es un método de generación de pares de números aleatorios independientes con distribución normal estándar, es decir con esperanza cero y varianza unitaria, expresada en las siguientes fórmulas.

$$Z_0 = R\,cos(\Theta) = \sqrt{-2\,log(U_1)}\,cos(2\pi\,U_2)$$

$$Z_1 = R\,sin(\Theta) = \sqrt{-2\,log(U_1)}\,sin(2\pi\,U_2)$$

donde:

- $U_1\,y\,U_2$ son variables aleatorias independientes con una distribución uniforme entre los valores 0 a 1.
- $Z_1\,y\,Z_2$ son variables aleatorias independientes con una distribución normal con una desviación estándar de 1.

Esto expresado en R puede ser definido como:

```
u = runif(2);
z0 = sqrt(-2 * log(u[1])) * cos(2 * pi * u[2]);
z1 = sqrt(-2 * log(u[1])) * sin(2 * pi * u[2]);
```

Si queremos tener nuestra propia media y desviación, podemos afectar la variable aleatoria:

```
mu      = 18;
sigma   = 1;
u       = runif(2);
z0      = sqrt(-2 * log(u[1])) * cos(2 * pi * u[2]);
z1      = sqrt(-2 * log(u[1])) * sin(2 * pi * u[2]);
z0      = sigma * z0 + mu);
z1      = sigma * z1 + mu);
```

Si utilizamos este método para generar nuestras variables aleatorias, se obtendrían resultados similares a la distribución normal Figura 5.5. De esta manera podemos verificar que efectivamente nuestras variables se comportan de acuerdo a una distribución normal teniendo la mayor frecuencia de valores en su media Figura 5.6.

Es necesario que ambas funciones se obtengan al calcular con dos variables aleatorias de distribución uniforme, en caso de cambiar el cálculo usando únicamente una sola variable uniforme, se rompería la distribución normal Figura 5.7 . Esto lo podemos verificar con el método *Shapiro.test()* el cual nos devuelve un valor *p-value*, si el valor obtenido es mayor a 0.05, entonces significa que nuestro vector muy probablemente viene de una distribución normal, sin embargo este método solo puede trabajar con un vector de 1 a 5000 valores por lo cual, es necesario utilizar el método *sample()* para tratar con vectores muy grandes.

```
test <- shapiro.test(runif(100))
print(test)
data:  runif(100)
    W = 0.95081, p-value = 0.0009379

    # prueba negativa 0.0009379 < 0.05

test <- shapiro.test(sample(Edades, 5000, replace = TRUE))
print(test)
    data:  sample(Edades, 5000, replace = TRUE)
    W = 0.99963, p-value = 0.5049

    # prueba positiva 0.5049 > 0.05
```

Si cambiáramos una variable uniforme del método de Box-Muller a por ejemplo esta fuera dependiente de la primera elevándola al cuadrado, también afectaría nuestro distribución Figura 5.8 .

```
u      <- runif(1);
u_2    <- u * u;
z0     <- sqrt(-2 * log(u)) * cos(2 * pi * u_2 );
z1     <- sqrt(-2 * log(u)) * sin(2 * pi * u_2 );
datos  <- c(z0, z1);
```

5. Método generador lineal congruencial

Nosotros podemos obtener nuestro propio vector de números distribuidos pseudoaleatoriamente por medio del método congruencial (GLC). Este algoritmo permite obtener una secuencia de números pseudoaleatorios calculados con una función lineal y es uno de los metodos mas antiguos para la generación de números pseudoaleatorios, este puede ser definido en las siguiente fórmulas.

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

donde:

- m es el modulo.
- X es la semilla.
- c es el incrementador.
- a es el multiplicador.

Esta función en R puede ser expresado como:

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

Si usáramos este método para simular 5000 números obtenidos pseudoaleatoriamente, podríamos ver que se comporta de la misma manera que una frecuencia de números aleatorios Figura 5.9 .

Referencias

[1] Satu Elisa Schaeffer. *Modelos probabilistas aplicados*. Sitio en, `https://elisa.dyndns-web.com/teaching/prob/pisis/prob.html`.

[2] The R Foundation. *The R Project for Statistical Computing*. `https://www.r-project.org/`. 2019.

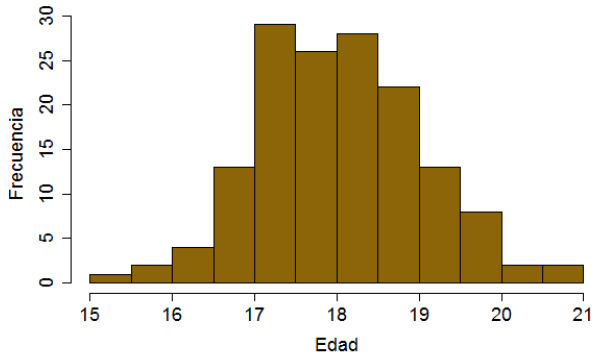


Figura 5.1: Distribución de frecuencia de la encuesta.

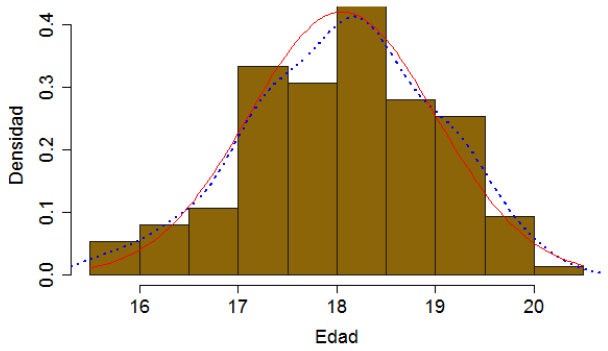


Figura 5.2: Densidad de resultados en la encuesta.

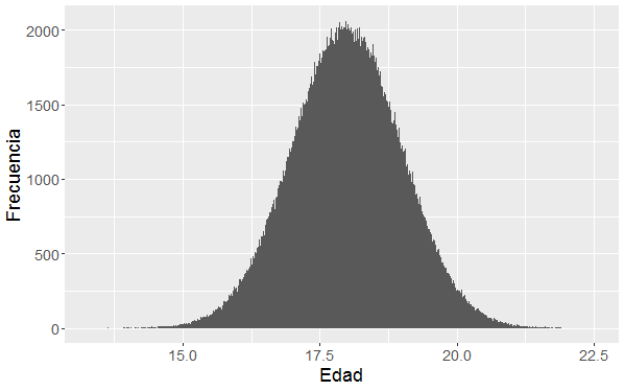


Figura 5.3: Distribución de frecuencia de la encuesta con 50000 estudiantes.

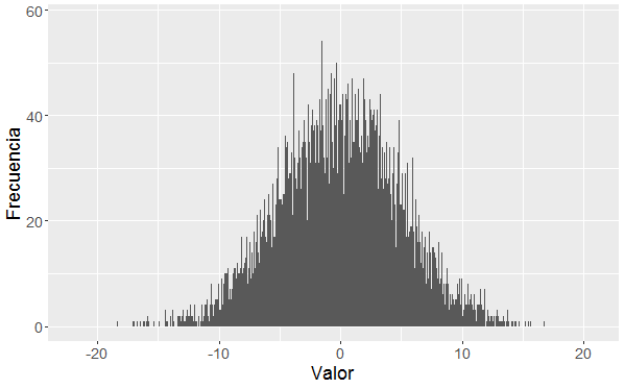
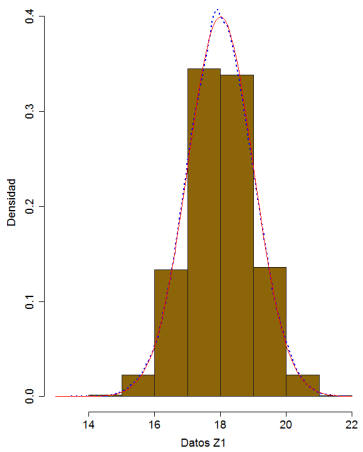
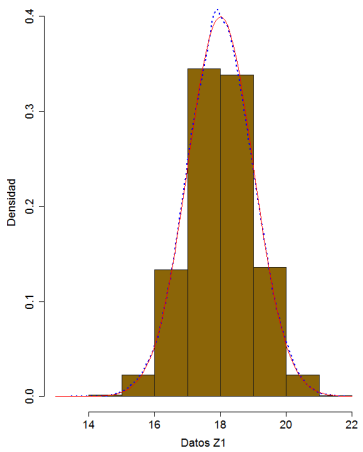


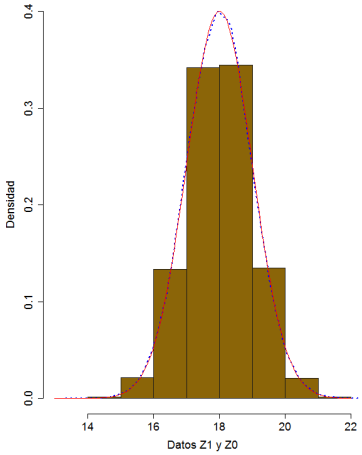
Figura 5.4: Distribución de frecuencia normal heterogénea.



(a) Distribución de frecuencia con una sola variable Z_1.



(b) Distribución de frecuencia con una sola variable Z_0.



(c) Distribución de frecuencia con ambas variables.

Figura 5.5: Densidad de las distribuciones generadas con el método de Box-Muller.

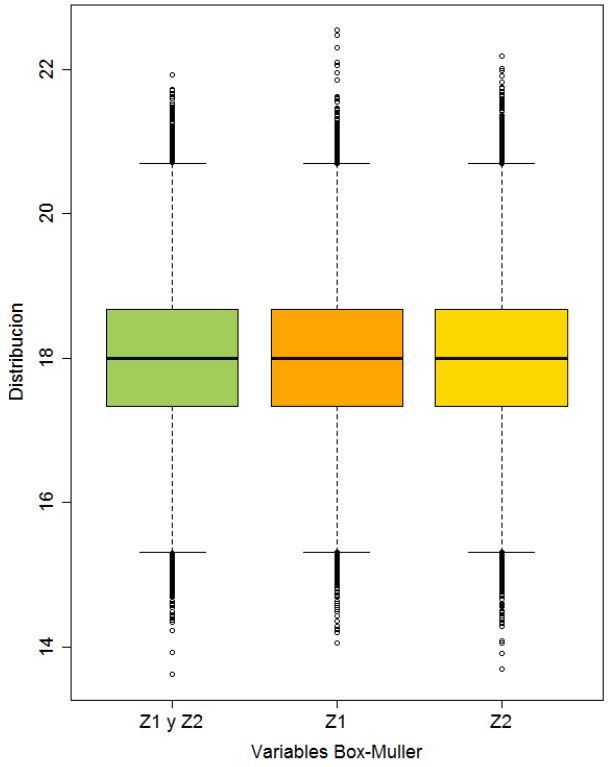
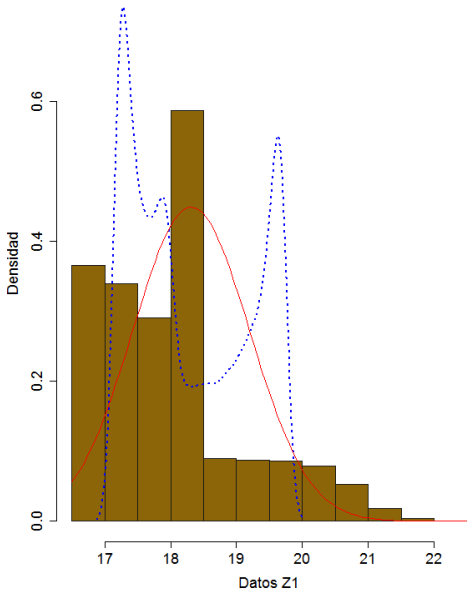
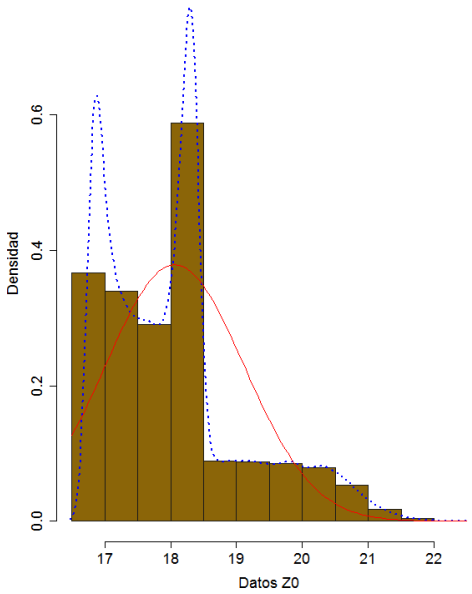


Figura 5.6: Distribución de normal de las variables del método Box-Muller por medio de boxplot.

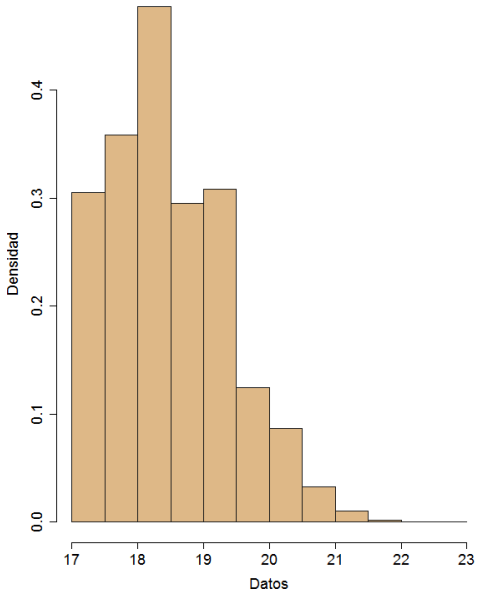


(a) Distribución de frecuencia de la variable Z_0con solo una variable uniforme.

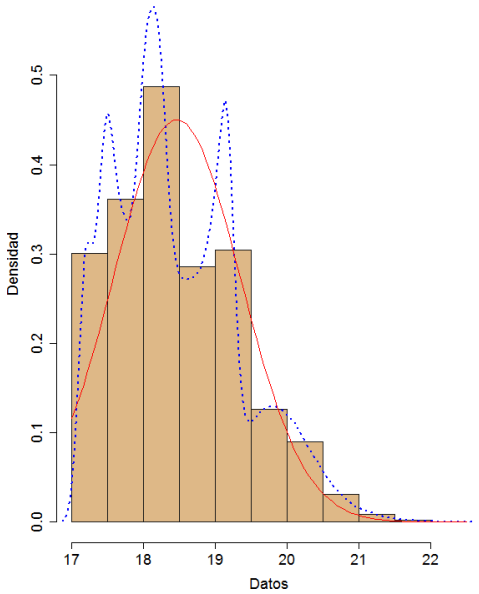


(b) Distribución de frecuencia de la variable Z_1con solo una variable uniforme.

Figura 5.7: Distribución de frecuencia de la variables de Box-Muller con solo una variable uniforme.

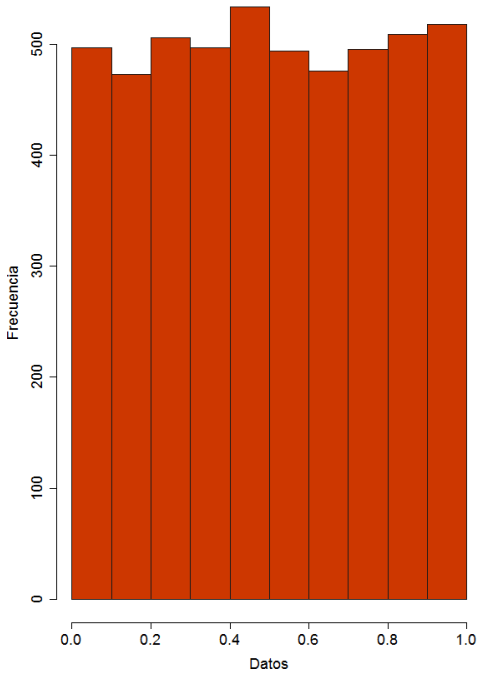


(a) Histograma que muestra la frecuencia de un vector obtenido por el método de Box-Muller.

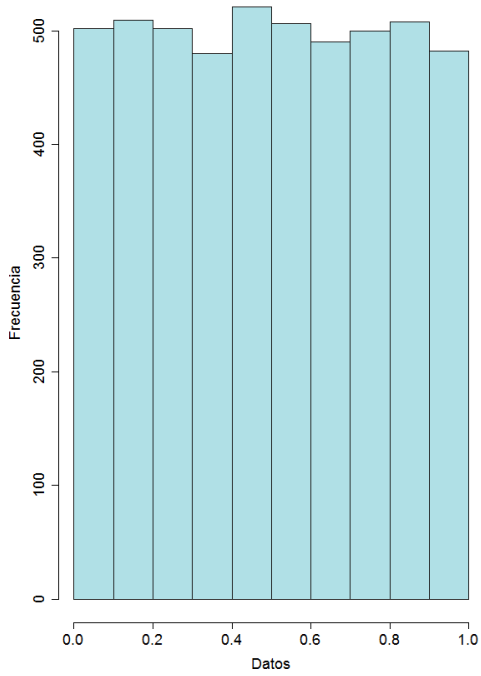


(b) Densidad de la distribución resultado de una variable dependiente en el método Box-Muller.

Figura 5.8: Distribución de frecuencia de los resultados de Box-Muller con una variable uniforme dependiente en una población de 50000 estudiantes.



(a) Histograma que muestra la frecuencia de un vector obtenido por el método generador lineal congruencial.



(b) Histograma que muestra la frecuencia de un vector obtenido por el método uniforme en R.

Figura 5.9: Frecuencia de distribución uniforme por el método generador lineal congruencial y el método en R con 5000 observaciones.

Pruebas estadísticas

TAREA 6

Alumno:
Joaquín Arturo Velarde Moreno

1. Introducción

Los objetivos de esta tarea son responder los cuestionamientos acerca del tema de pruebas estadísticas derivados de cuatro artículos diferentes dados por la Dra. Sara Verónica Rodríguez Sánchez y, además, aplicar una serie de pruebas estadísticas acerca del turismo internacional con el programa R [5].

2. Preguntas y respuestas

2.1. ¿Relación entre contraste de hipótesis y pruebas estadísticas?

Es el procedimiento para encontrar la veracidad de una hipótesis contrastándola con otra para lo cual es necesario realizar pruebas estadísticas y comprobar cuál de dichas pruebas es más factible.

2.2. ¿Qué indicaría rechazar la hipótesis nula?

Dicha hipótesis indicaría que nuestra prueba estadística nos dio un *valor-p* menor al nivel de significancia ($p < \alpha$), aceptando así la hipótesis alternativa, lo cual no significa necesariamente que la hipótesis alternativa sea correcta [2], pues podemos tener un error de tipo 1.

2.3. ¿Cómo se interpreta la salida de una prueba estadística?

La prueba estadística produce un número denominado *valor-p*, el cual tiene como limites *0* y *1*; el *valor-p* es la probabilidad de obtener los datos bajo la hipótesis nula, esta se debe comparar con el nivel de significancia, y en caso de tener $p < \alpha$ rechazamos la hipótesis nula [2].

2.4. ¿Cómo seleccionar el alfa?

El valor alfa, también denominado nivel de significación, debe ser primeramente un valor de *0* a *1* y, generalmente, dicho valor se fija en *0.05*, *0.01* o *0.001*. La elección de alfa debe depender de cuán peligroso sea rechazar la hipótesis nula. Por ejemplo, en un estudio que se proponga demostrar los beneficios de un tratamiento médico, deberá tener un alfa bajo. Por otro lado, si tomamos en cuenta la apreciación de un producto, podemos ser más moderados [2]. Además, es importante tomar en cuenta que tan factible es realizar el experimento múltiples veces para no ser muy exigentes con él.

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

Debido a que la prueba estadística suele usar muestras de una población aleatoriamente escogida, puede ocasionar dos situaciones: que una hipótesis nula que es verdadera sea rechazada, denominado error tipo 1 o que una hipótesis nula falsa sea aceptada denominado error tipo 2.

2.6. ¿Qué es la potencia estadística y para qué sirve?

Es la capacidad de un experimento para conducir al rechazo de la hipótesis nula [2]. La potencia de un experimento aumenta con alfa, con lo preciso que son las mediciones y con el numero de repeticiones del experimento, equivale a *1* menos el riesgo de ser errónea cuando se acepta H0 $(1-\beta)$, De este modo, mientras mayor sea la potencia, menor es el riesgo de equivocarse al aceptar H0.

2.7. Ejemplos de pruebas estadísticas paramétricas y no paramétricas.

1. Paramétricas:
   - *a*) "t"de student,
   - *b*) el coeficiente de correlación de Pearson,
   - *c*) la regresión lineal,
   - *d*) el análisis de varianza unidireccional (ANOVA Oneway),
   - *e*) análisis de varianza factorial (ANOVA),
   - *f*) análisis de covarianza (ANCOVA).
2. No paramétricas:
   - *a*) prueba chi-cuadrado,
   - *b*) coeficientes de correlación e independencia para tabulaciones cruzadas,
   - *c*) coeficientes de correlación por rangos ordenados Spearman y Kendall,
   - *d*) prueba de suma de rango Wilcoxon.

2.8. Resume la guía para encontrar la prueba estadística que buscas

Al escoger una prueba estadística, debemos preguntarnos primeramente si nuestro objetivo es asociar o comparar, pues, aunque ambas establecen relaciones, la comparación evalúa dichas relaciones entre uno o varios grupos. También, hay analizar qué tipo de muestras tenemos, si es que son independientes o relacionadas. Es necesario diferenciar entre ambas muestras, ya que las relacionadas pueden ser del tipo antes-después, como por ejemplo el estudio de pacientes donde se comparan antes y después de la aplicación de un tratamiento. Otra pregunta que debemos hacernos es qué tipo de datos tenemos, ¿son variables cualitativas o cuantitativas? Al elegir entre una y otra las variables debemos corroborar si nuestros datos cumplen o no con los supuestos (normalidad, homogeneidad, independencia), de este modo, podremos elegir entre pruebas paramétricas, no paramétricas y robustas [3]. Como paso final, hay que analizar si podemos contar con los resultados dados por nuestra prueba estadística *p-value* graficándolos para decidir .

3. Pruebas estadísticas

A continuación veremos las mas comunes pruebas estadísticas [4] y su aplicación en R [5], junto con datos obtenidos de INEGI [1].

3.1. One sample t-Test.

Esta es una prueba estadística de tipo paramétricas y es usada para probar si la media de una muestra de una distribución normal puede ser un valor especifico [4]. Como primer ejemplo utilizaremos el numero de visitantes que ingresan al país por mes por la vía aérea (Figura 3.1), según los datos obtenidos del INEGI [1].

Lo expresaremos en R como:

```
t.test(VisitantesPorAvion, mu = 1540000)

data:  VisitantesPorAvion
t = 0.076095, df = 11, p-value = 0.9407
alternative hypothesis: true mean is not equal to 1540000
```

En esta prueba estadística, quisimos demostrar que la media de personas que entran al país por avión al mes es de *1,540,000* siendo esta nuestra hipotesis nula, sin embargo la prueba produjo un *p-value* mayor al nivel de significancia de *0.05*, por lo que no rechazamos nuestra hipotesis nula.

3.2. Wilcoxon Signed Rank Test.

Esta prueba estadística es un método no paramétrico que evalúa la media de una muestra sin asumir que esta distribuida normalmente, este puede ser una alternativa al t-Test, especialmente cuando no se tiene información de la distribución en que sigue [4]. usaremos como muestra el numero de visitantes que ingresan al país por mes por la vía terrestre el cual asumimos no tiene una distribución normal (Figura 3.3), según los datos obtenidos del INEGI [1]. Lo expresaremos en R como:

```
wilcox.test(VisitantesTerrestres, mu=640049, conf.int = TRUE)

data:  VisitantesTerrestres
V = 40, p-value = 0.9697
alternative hypothesis: true location is not equal to 640049
```

En esta prueba estadística quisimos demostrar que la media de personas que entran al país por la vía terrestre al mes es de *640,049* siendo esta nuestra hipotesis nula, sin embargo la prueba produjo un *p-value* mayor al nivel de significancia de *0.05*, por lo que no rechazamos nuestra hipotesis nula.

3.3. Two Sample t-Test and Wilcoxon Rank Sum Test.

Tanto t-Test como Wilcoxon rank pueden ser usados para comparar la media de 2 muestras, la diferencia como ya dijimos es que t-Test asume que la muestra sigue una distribución normal mientras que Wilcoxon rank no [4]. Usaremos nuestras 2 muestras anteriores que son el numero de visitantes que ingresan al país por mes a través de la vía terrestre y aérea (Figura 3.3 y Figura 3.1), según los datos obtenidos del INEGI [1]. Lo expresaremos en R como:

```
wilcox.test(VisitantesTerrestres, VisitantesPorAvion, paired = TRUE)

data:  VisitantesTerrestres and VisitantesPorAvion
V = 0, p-value = 0.0004883
alternative hypothesis: true location shift is not equal to 0
```

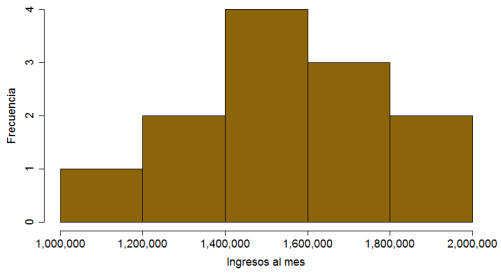
En esta prueba estadística quisimos demostrar nuestra hipótesis nula el cual es que la media de personas que entran al país por la vía terrestre al mes es la misma que el promedio de personas que entran por la vía aérea, sin embargo la prueba produjo un *p-value* menor al nivel de significancia de *0.05*, por lo que rechazamos nuestra hipótesis nula.

3.4. Shapiro Test.

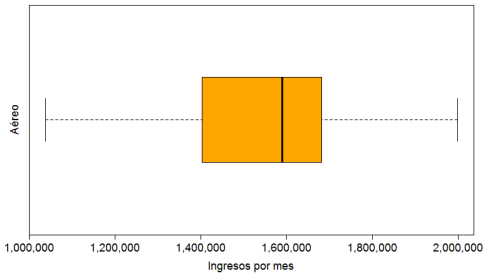
Esta prueba estadística evalúa si una muestra sigue una distribución normal [4]. Usaremos nuestra muestra de numero de visitantes que ingresan al país por mes a través de la vía marítima (Figura 3.2), según los datos obtenidos del INEGI [1]. Lo expresaremos en R como:

```
shapiro.test(VisitantesMaritimos)
data:  VisitantesMaritimos
W = 0.82518, p-value = 0.01838
```

En esta prueba estadística quisimos demostrar que nuestra población de personas que entran al país por la vía marítima al mes sigue una distribución normal, sin embargo la prueba produjo un *p-value* menor al nivel de significancia de *0.05*, por lo que rechazamos nuestra hipótesis nula.

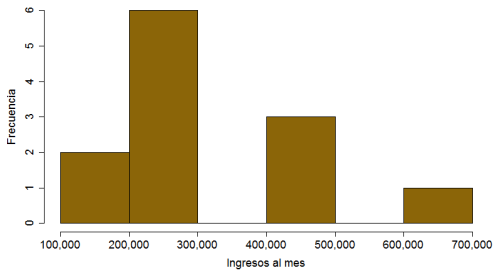


(a) Histograma que muestra la frecuencia de turistas ingresados al mes por la vía aérea.

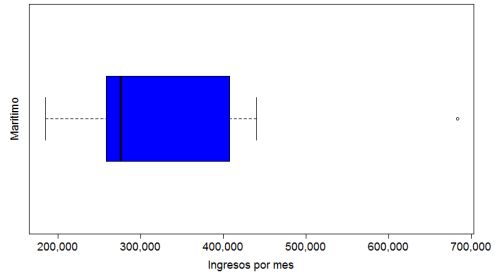


(b) Gráfico con información de la media de ingresos turísticos por la vía aérea.

Figura 3.1: Gráficos que muestran el ingreso de turistas al país cada mes por medio de la vía aérea.

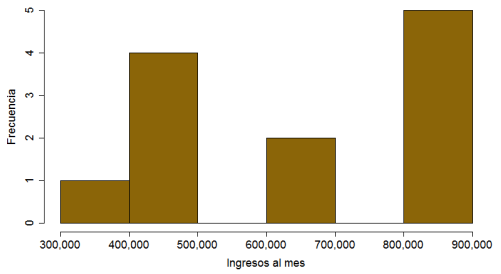


(a) Histograma que muestra la frecuencia de turistas ingresados al mes por la vía marítima.

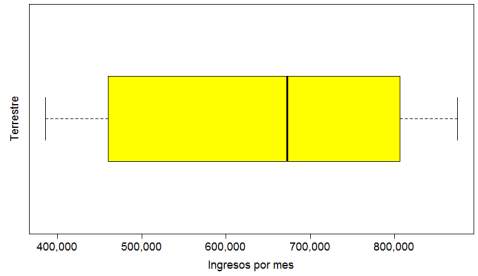


(b) Gráfico con información de la media de ingresos turísticos por la vía marítima.

Figura 3.2: Gráficos que muestran el ingreso de turistas al país cada mes por medio de la vía marítima.



(a) Histograma que muestra la frecuencia de turistas ingresados al mes por la vía terrestre.



(b) Gráfico con información de la media de ingresos turísticos por la vía terrestre.

Figura 3.3: Gráficos que muestran el ingreso de turistas al país cada mes por medio de la vía terrestre.

Referencias

[1] INEGI. *Numero de visitantes internacionales que ingresaron al país por mes*. https://www.inegi.org.mx/contenidos/temas/economia/turismo/EVI_R_01.xlsx. 2020.

[2] *Pruebas estadísticas*. Sitio en,https://help.xlstat.com/s/article/que-es-una-prueba-estadistica?language=es.

[3] Rosana Ferrero. *Guía definitiva para encontrar la prueba estadística que buscas*. Sitio en, https://www.maximaformacion.es/blog-dat/guia-para-encontrar-tu-prueba-estadistica/.

[4] Selva Prabhakaran. *Statistical Tests*. Sitio en,http://r-statistics.co/Statistical-Tests-in-R.

[5] The R Foundation. *The R Project for Statistical Computing*. https://www.r-project.org/. 2019.

Ajuste de curvas

TAREA 7

Alumno:
Joaquín Arturo Velarde Moreno

1. Introducción

El objetivo del siguiente reporte es describir cómo se obtiene la correlación lineal a través de varios ejemplos de funciones matemáticas; ver su representación gráfica y la forma de su curva. Para cumplir con este objetivo, usaremos el programa R 4.0.2 [2] y de este modo, haremos cálculos con una serie de datos. Usaremos como apoyo el material de la Dra. Elisa Schaefer [1].

2. Definición correlación lineal

La correlación es una medida de la presencia de una relación lineal de un conjunto de datos que provienen de dos variables medidas al mismo tiempo sobre una serie de individuos, también se le conoce como datos bivariados. Estos datos se pueden representar por medio de una gráfica, por ejemplo, en la (Figura 2.1) se muestra gráficamente los datos de una encuesta que se hizo a alumnas de una escuela para ver la relación entre su peso y estatura.

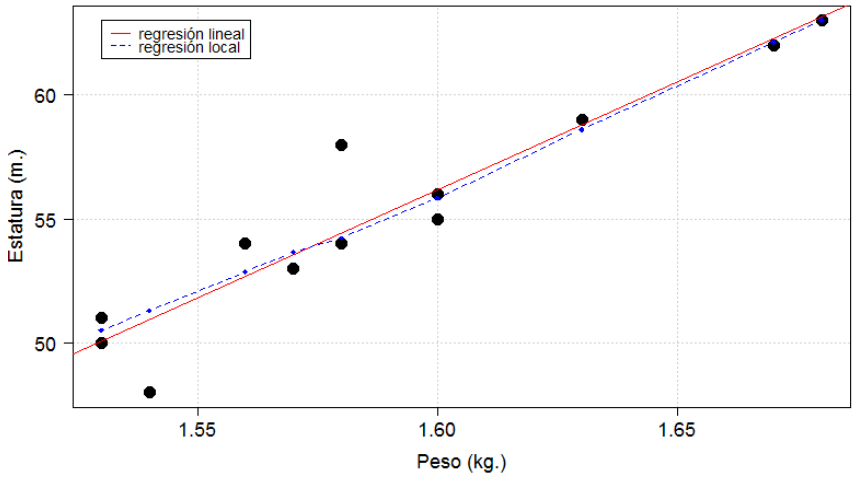


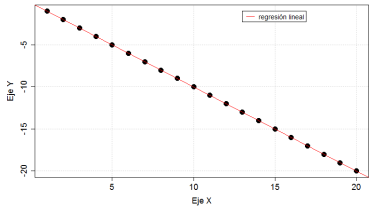
Figura 2.1: Relación entre peso(kg.) y altura(m.) de alumnas.

En el presente reporte usaremos la correlación de Pearson, la cual denotaremos por r; el rango de la correlación r va de *1* a *-1*, lo cual puede generar tres casos (Figura 4.1):

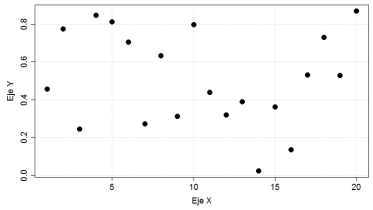
- r igual a 1: positiva,

- r igual a -1: negativa,
- r igual a 0: sin correlación.

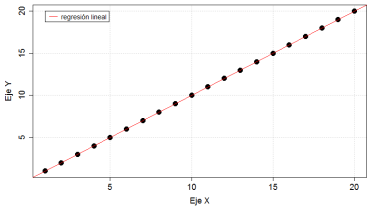
La mayoría de las veces no obtendremos mediciones exactas para cada caso, sino serán aproximadas y de estas, podremos decidir si existe una relación entre los datos o si son independientes entre sí.



(a) Correlación negativa Y disminuye cuando X crece.



(b) No hay relación alguna en los datos.



(c) Correlación positiva Y aumenta cuando X crece.

Figura 2.2: Correlaciones en sus valores negativo, positivo y 0.

3. Fórmula

La correlación de Pearson puede ser obtenida por la siguiente expresión:

$$\frac{\Sigma\left(x \times y\right) - \frac{1}{n}\left(\Sigma x \times \Sigma y\right)}{\sqrt{\left(\Sigma x^2 - \frac{1}{2}\left(\Sigma x\right)^2\right) \times \left(\Sigma y^2 - \frac{1}{2}\left(\Sigma y\right)^2\right)}}.$$

donde:

- x y y son nuestros datos bivariados,
- n es el número de datos.

Esta ecuación puede ser expresada en R de la siguiente manera:

```
n                 <- length(VectorX)
SumatoriaX        <- sum(VectorX)
SumatoriaY        <- sum(VectorY)

numerador         <- sum(VectorX * VectorY) - (SumatoriaX * SumatoriaY) / n
denominadorX      <- sum(VectorX **2) - (SumatoriaX**2) / n
denominadorY      <- sum(VectorY **2) - (SumatoriaY**2) / n
denominador       <- sqrt(denominadorX * denominadorY)
correlacion       <- numerador / denominador
```

Existe una manera más sencilla de hacer este cálculo obteniendo los promedios de cada vector, a los cuales denotaremos como X y Y, pudiéndose expresar de la siguiente manera:

$$\frac{\Sigma\left(x \times y\right)}{\sqrt{\Sigma x^2 \times \Sigma y^2}}.$$

donde:

- $y = Y - \mu Y$,
- $x = X - \mu X$.

Esta ecuación puede ser expresada en R de la siguiente forma:

```
x         <- VectorX - mean(VectorX)
y         <- VectorY - mean(VectorY)

correlacion      <- sum(x * y) / sqrt(sum(x**2) * sum(y**2))
```

4. Transformadas

No en todos los casos los conjuntos de datos aparecen de manera lineal; algunos se presentan muy dispersos y otros forman curvas, como es el caso del ejemplo de la función $f(x) = x^2$ (Figura 4.3). En estos casos lo mejor es trabajar con transformadas, que son manipulaciones que se hacen a un vector de los datos para poder acomodarlos en una relación lineal. Una opción sencilla es la escalera de transformaciones de Tukey:

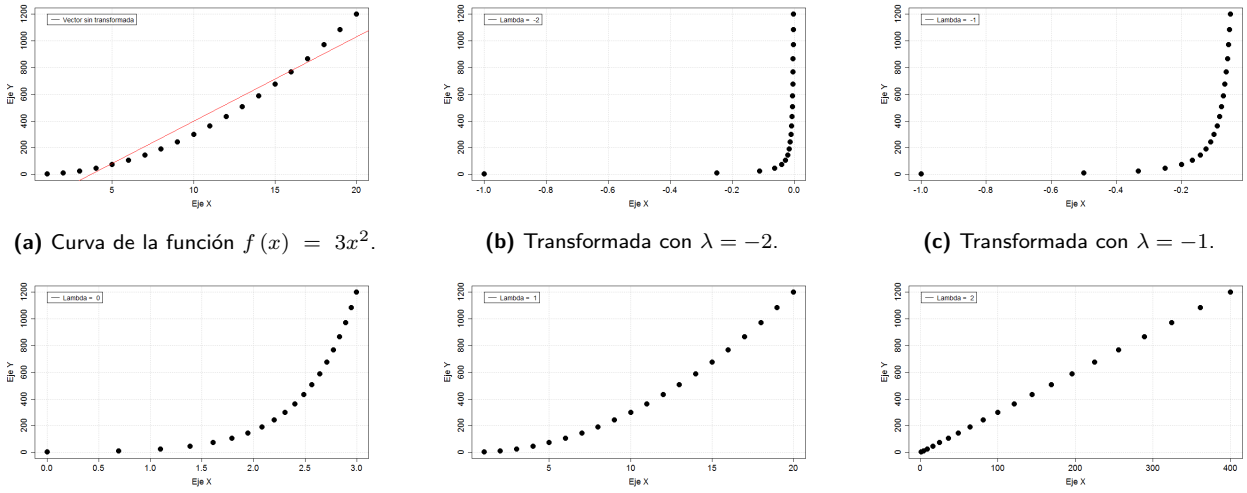
Cuadro 1: Escalera de Tukey con un rango de -2 a 2.

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

Podemos calcular y comparar cada transformada en R del siguiente modo:

```
for(i in seq(-2,2,1))
{
    if(i > 0)
    { z <- x^i}
    else if(i < 0)
    { z <- -1*x^i}
    else
    { z <- log(x)}
    cor(y,z)
 }
```

Dada la función $f(x) = 3x^2$ podemos buscar una transformada por la escalera de Tukey que nos encuentre el λ que mejore una relación lineal en nuestros conjuntos.



(a) Curva de la función $f(x) = 3x^2$. **(b)** Transformada con $\lambda = -2$. **(c)** Transformada con $\lambda = -1$.

(d) Transformada con $\lambda = 0$. **(e)** Transformada con $\lambda = 1$. **(f)** Transformada con $\lambda = 2$.

Figura 4.1: Transformadas de la función $f(x) = 3x^2$.

Una alternativa optimizada es la transformada Box-Scott, en esta expresion X se transforma a $\frac{(x^\lambda - 1)}{\lambda}$. Podemos calcular y comparar cada transformada en R de la siguiente manera:

```
for(i in seq(-2,2,0.05))
{
    z <- (abs(x)^i-1)/i
    w <- c(w,cor(y,z))
    v <- c(v,i)
}
plot(v, w)
```

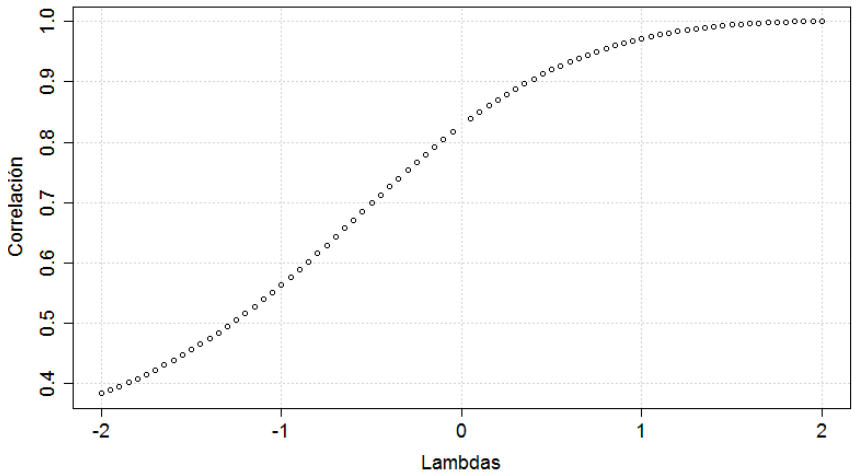


Figura 4.2: Lambdas y su correlación usando la transformada de Box-Scott en la función $f(x) = 3x^2$.

Referencias

[1] Satu Elisa Schaeffer. *Modelos probabilistas aplicados*. Sitio en, `https://elisa.dyndns-web.com/teaching/prob/pisis/prob.html`.

[2] The R Foundation. *The R Project for Statistical Computing*. `https://www.r-project.org/`. 2019.

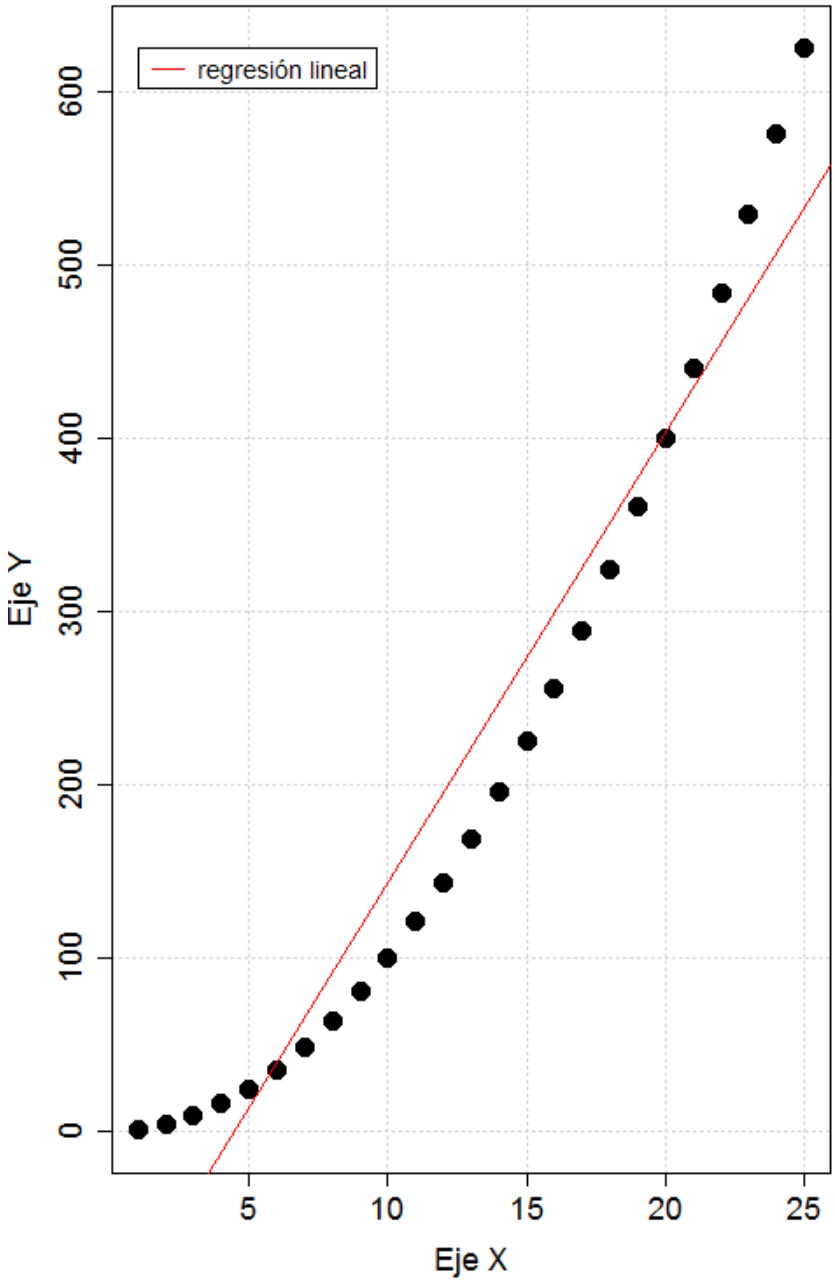


Figura 4.3: Relación entre el conjunto X y el Y al aplicar $f(x) = x^2$.

Tarea 9
Joaquín Arturo Velarde Moreno

1 Exercise 1, page 247

A card is drawn at random from a deck consisting of cards numbered 2 through 10. A player wins 1 dollar if the number on the card is odd and losses 1 dollar if the number is even. ¿What is the expected value of his winnings?
$E = \sum x * P(X = x)$.
$E = (-1)P(x = -1) + 1(1)P(x = 1)$.
$P(X = 1) = \frac{4}{9}$.
$P(X = -1) = \frac{4}{9}$.
$E = (-1)(\frac{4}{9}) + (1)(\frac{4}{9})$.
$E = -\frac{1}{9}$.

2 Exercise 6, page 247

A die is rolled twice. Let X denote the sum of the two numbers that turn up, and Y the difference of the numbers (specifically, the number on the first roll minus the number on the second). Show that $E(XY) = E(X)E(Y)$. ¿Are X and Y independent?
A = result of the first die.
B = result of the second die.
$X = A + B$.
$Y = A - B$.
$E(A) = E(B)$.
$E(A^2) = E(B^2)$.
We will now show $E(XY) = E(X)E(Y)$.
$E(XY) = E(X)E(Y)$.
$E(XY) = E((A + B)(A - B))$.
$E(XY) = (A^2 - B^2 + BA - AB)$.
$E(XY) = (A^2 - B^2)$.
$E(XY) = E(A^2) - E(B^2)$.
$E(XY) = 0$.
$E(X)E(Y) = E(A + B)E(A - B)$.
$E(X)E(Y) = (E(A) + E(B))(E(A) - E(B))$.
$E(X)E(Y) = (E(A))^2 - (E(B))^2$.
$E(X)E(Y) = 0$.
then we have $E(XY) = E(X)E(Y)$.
We will show now that X and Y are independent. By definition we have $P(X|Z) = P(x)$ if they are independent. We set $X = 2$ if we calculate the probability $P(X = 2) = \frac{1}{36}$ And now we calculate $P(X = 2|Y = 5) = 0$. Since $P(X = 2) \neq P(X = 2|Y = 5)$,then, X and Y are not independent.

3 Exercise 18, page 249

Exactly one of six similar keys opens a certain door. If you try the keys, one after another, ¿what is the expected number of keys that you will have to try before success?
$E(X) = \sum_{x=1}^{6} x * P(X = x)$.
$E(X) = \sum_{x=1}^{6} x * \frac{1}{6}$.
$E(X) = (1 + 2 + 3 + 4 + 5 + 6) * \frac{1}{6}$.

$E(X) = 3.5$.

4 Exercise 1, page 263

A number is chosen at random from the set $S = \{-1, 0, 1\}$. Let X be the number chosen. Find the expected value, variance, and standard deviation of X.
$E(X) = \sum_{x=1}^{3} x * P(X = x)$.
$E(X) = \sum_{x=1}^{3} x * \frac{1}{3}$.
$E(X) = (-1, 0, 1) * \frac{1}{3}$.
$E(X) = 0$.
Now for the variance: $\sigma^2 = V(X) = E((x - E(x))^2)$.
$\sigma^2 = E((x - 0)^2)$.
$\sigma^2 = E(x^2)$.
$\sigma^2 = \sum x^2 * P(X^2 = x^2)$.
$\sigma^2 = (0 * \frac{1}{3}) + (1 * \frac{2}{3})$.
$\sigma^2 = \frac{2}{3}$.
Now for standar deviation: $\sigma = \sqrt{\frac{2}{3}}$.

5 Exercise 9, page 264

A die is loaded so that the probability of a face coming up is proportional to the number on that face. The die is rolled with outcome X. Find $V(X)$ and $D(X)$.
The probability of each face is $P(X = x) = \frac{x}{21}$.
With this we can get our expected value.
By definition we have
$E(X) = \sum x * P(X = x)$.
$E(X) = \sum x * \frac{x}{21}$.
$E(X) = \frac{1}{21} + \frac{4}{21} + \frac{9}{21} + \frac{16}{21} + \frac{25}{21} + \frac{36}{21}$.
$E(X) = \frac{91}{21}$.
Using the definition we have $V(cX) = E((cX)^2) - E((cX))^2$,
then, we need to get $E(X^2)$.
$E(X^2) = \sum X^2 * P(X^2 = X^2)$.
However since it will be the same probability. $E(X^2) = \sum X^2 * P(X = X)$.
$E(X^2) = \sum X^2 * \frac{x}{21}$.
$E(X^2) = \sum X^3 * \frac{x}{21}$.
$E(X^2) = \frac{1}{21} + \frac{8}{21} + \frac{27}{21} + \frac{64}{21} + \frac{125}{21} + \frac{216}{21} = 21$.
With this we can get our variance. $V(X) = E(x^2) - (E(x))^2$.
$V(X) = 21 - (\frac{91}{21})^2$.
$V(X) = \frac{981}{441}$.
$V(X) = 2.22$.
Finally the standar deviation.
$D(X) = \sqrt{2.2}$

6 Exercise 12, page 264

Let X be a random variable with $\mu = E(X)$ and $\sigma^2 = V(X)$. Define $X^* = (X - \mu)/\sigma$. The random variable X^* is called the *standardized random variable associated* with X. Show that this standardized random variable has expected value 0 and variance 1.
First we will get our expected value.

$E(X) = E(\frac{x-\mu}{\sigma})$.
$E(X) = E(\frac{1}{\sigma}(x-\mu))$.
$E(X) = \frac{1}{\sigma}E(x-\mu)$.
$E(X) = \frac{1}{\sigma}(E(x) - E(\mu))$.
$E(X) = \frac{1}{\sigma}E(\mu-\mu)$.
$E(X) = 0$.
Next the variance.
$V(X) = V(\frac{x-\mu}{\sigma})$.
$V(X) = V(\frac{1}{\sigma}(x-\mu))$.
$V(X) = (\frac{1}{\sigma})^2 V(x-\mu)$.
$V(X) = (\frac{1}{\sigma^2})V(x)$.
$V(X) = (\frac{1}{\sigma})(\sigma^2)$.
$V(X) = 1$.

7 Exercise 3, page 278

The lifetime, measure in hours, of the ACME super light bulb is a random variable T with density function $f_T(t) = \lambda^2 t \exp^{-\lambda t}$, where $\lambda = 0.05$. ¿What is the expected litetime of this light bulb? ¿What is its variance?
By definition we have $E(T) = \int_0^\infty (t)(f_T(t)dt)$ to get our expected value of a continuous variable.
$E(T) = \int_0^\infty (t)(\lambda^2 t \exp^{-\lambda t})dt)$.
$E(T) = \int_0^\infty t^2\lambda^2 \exp^{-\lambda t}\, dt$.
$E(T) = 40$.
By definition we have $V(cX) = E((cX)^2) - E((cX))^2$,
then, we need to get $E(T^2)$.
$E(T^2) = \int_0^\infty (t^2)(\lambda^2 t \exp^{-\lambda t})dt$.
$E(T^2) = \int_0^\infty t^3\lambda^2 \exp^{-\lambda t}\, dt$.
$E(T^2) = 2400$.
With this we can get the variance.
$V(X) = E((X)^2) - E((X))^2$.
$V(X) = 2400 - 1600^{\cdot}$ $V(X) = 800$.

8 Exercise 15, page 249

A box contains two gold balls and three silver balls. You are allowed to choose successively balls from the box at random. You win 1 dollar each time you draw a gold ball and lose 1 dollar each time you draw a silver ball. After a draw, the ball is 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.
We need the probability of each case.
$P(W = 1) = \frac{1}{2}$.
$P(W = 0) = \frac{1}{5}$.
$P(W = -1) = \frac{3}{10}$.
By definition we have:
$E(W) = \sum(g)P(G = g)$.
$E(W) = (-1)\frac{3}{10} + (0)\frac{1}{5} + (1)(\frac{1}{2})$.
$E(W) = -\frac{3}{10} + (1)(\frac{2}{10})$.
$E(W) = \frac{1}{5}$.
with this our strategy gives a expected value greater that 0, which means is favorable.

Practica Ejercicios en R

TAREA 10

Alumno:
Joaquín Arturo Velarde Moreno

1. Introducción

En este reporte hago uso del programa R 4.0.2 [2] para poder resolver 3 problemas escogidos del libro de introducción a la probabilidad [1] y representare gráficamente los resultados para ver como se comportan las variables.

2. Ejercicio 1, página 247

Se saca una carta al azar de una baraja que consta de cartas numeradas del 2 al 10. Un jugador gana 1 dólar si el número de la carta es impar y pierde 1 dólar si el número es par. ¿Cuál es el valor esperado de sus ganancias?

$E = \sum x * P(X = x)$.
$E = (-1)P(x = -1) + 1(1)P(x = 1)$.
$P(X = 1) = \frac{4}{9}$.
$P(X = -1) = \frac{4}{9}$.
$E = (-1)(\frac{4}{9}) + (1)(\frac{4}{9})$.
$E = -\frac{1}{9}$.

Analíticamente, obtenemos que nuestro valor esperado es de $-\frac{1}{9}$ o $-0{,}11111111$. Ahora pasaremos a aplicarlo en nuestro programa R,de acuerdo con nuestro problema tenemos una baraja con cartas numeradas del 2 al 10.

```
Baraja <- c(2,3,4,5,6,7,8,9,10)
```

de acuerdo con la instrucción debemos sacar una carta al azar de la baraja, para esto haremos uso de la función *sample()*.

```
carta  <- sample(Baraja, 1)
```

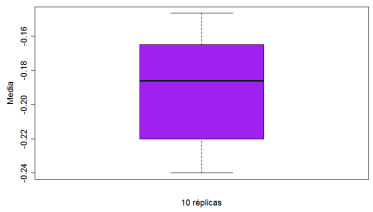
Lo siguiente es validar si nuestro número es par o impar, para lo cual usaremos la operación módulo y lo guardaremos en un arreglo.

```
if((carta %% 2) == 0)
{
  Ganancias = - 1;
}else {
  Ganancias =   1;
}
    ArregloGanancias = c(ArregloGanancias,Ganancias)
```

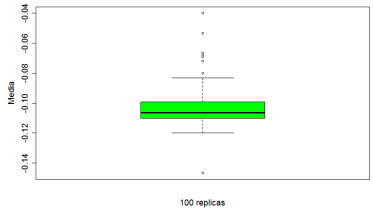
Por último, replicaremos nuestro proceso cuántas veces queramos para comprobar si se obtiene nuestro valor esperado.

```
for(j in 1:Replicas)
{
  Ganancias <- 0;
  for(i in 1:50)
  {
    carta  <- sample(Baraja, 1)
    if((carta %% 2) == 0)
    {
      Ganancias = - 1;
    }else {
      Ganancias =   1;
    }
    ArregloGanancias = c(ArregloGanancias,Ganancias)
  }
  Media = c(Media,mean(ArregloGanancias))
}
```

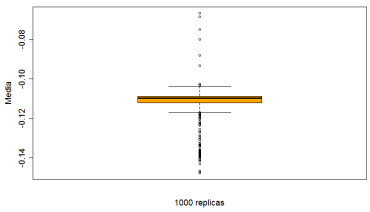
Lo anterior nos dará como resultado que, a medida que el número de réplicas aumenta el valor de nuestra media, se aproxima más a $-\frac{1}{9}$ (Figura 2.1).



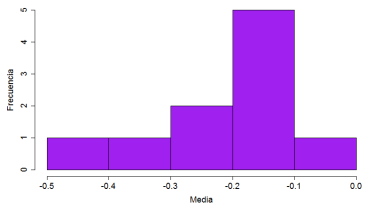
(a) Media del evento replicada 10 veces.



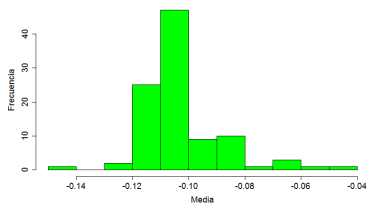
(b) Media del evento replicada 100 veces.



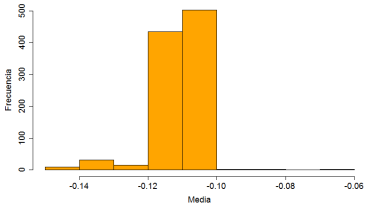
(c) Media del evento replicada 1000 veces.



(d) Media del evento replicada 10 veces.



(e) Media del evento replicada 100 veces.



(f) Media del evento replicada 1000 veces.

Figura 2.1: Distribuciones de medias en los experimentos.

3. Ejercicio 18, página 249

Exactamente una de las seis llaves similares abre una puerta determinada. Si prueba las teclas, una tras otra, ¿cuál es la cantidad esperada de teclas que tendrá que probar antes de tener éxito?

$E(X) = \sum_{x=1}^{6} x * P(X = x)$.
$E(X) = \sum_{x=1}^{6} x * \frac{1}{6}$.
$E(X) = (1 + 2 + 3 + 4 + 5 + 6) * \frac{1}{6}$.
$E(X) = 3{,}5$.

Analíticamente, obtenemos que nuestro valor esperado es de 3,5. Para este ejercicio tomaremos un vector de 6 elementos para representar nuestras 6 llaves y estableceremos una llave correcta.

```
Llaves          <- c(1:6)
LlaveCorrecta   <- sample(Llaves, 1)
```

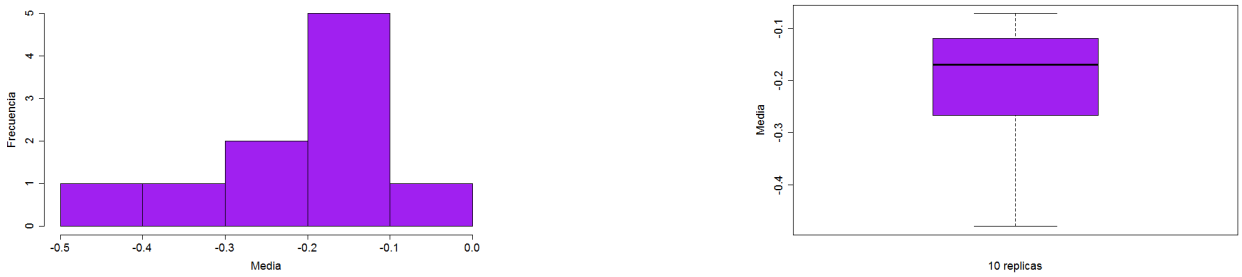
Después, recorreremos nuestro vector en busca de la llave correcta, contando la veces que nos tomó encontrarla.

```
    Llaves          <- c(1:6)
    LlaveCorrecta   <- sample(Llaves, 1)
for (LLave in LLaves)
{
  if(LLave == LLaveCorrecta)
  {
    break;
  }else{
    contador = contador + 1;
  }
}
```

Por último, calcularemos la media de las veces que nos tomó encontrarlas.

```
for(i in 1:100)
{
  LLaveCorrecta   <- sample(LLaves, 1)
  contador        <- 0;
  for (LLave in LLaves)
  {
    if(LLave == LLaveCorrecta)
    {
      break;
    }else{
      contador = contador + 1;
    }
  }
  contadores = c(contadores,contador)
}
Medias   <- c(Medias,mean(contadores))
```

Para nuestra sorpresa, el valor de la media obtenido más frecuentemente es el de 2,5 por lo que son necesarias más pruebas.



(a) Frecuencia de medias.

(b) Distribuciones de media.

Figura 3.1: Distribución de la media para el número de llaves necesarias en 10 eventos.

4. Ejercicio 1, página 263

Se elige un número al azar del conjunto $S = \{-1, 0, 1\}$. Sea X el número elegido. Encuentre el valor esperado, la varianza y la desviación estándar de X.

$E(X) = \sum_{x=1}^{3} x * P(X = x)$.
$E(X) = \sum_{x=1}^{3} x * \frac{1}{3}$.
$E(X) = (-1, 0, 1) * \frac{1}{3}$.
$E(X) = 0$.
Ahora para la varianza: $\sigma^2 = V(X) = E((x - E(x))^2)$.
$\sigma^2 = E((x - 0)^2)$.
$\sigma^2 = E(x^2)$.
$\sigma^2 = \sum x^2 * P(X^2 = x^2)$.
$\sigma^2 = (0 * \frac{1}{3}) + (1 * \frac{2}{3})$.
$\sigma^2 = \frac{2}{3}$.
Finalmente, la desviación estándar: $\sigma = \sqrt{\frac{2}{3}}$.

Dado que, analíticamente conseguimos nuestro valor de la media como 0, la varianza como $0,6$ y la desviación estándar como $0,81$, aplicaremos el ejercicio en R, para lo cual, primero necesitamos un arreglo que vaya de -1 a 1 y obtener un número al azar.

```
Numeros    <- c(-1,0,1)
Numero     <- sample(Numeros, 1)
```

Realizaremos este proceso 100 veces y guardaremos nuestros resultados en un vector.

```
Numeros          <- c(-1,0,1)

  DistribucionNumeros <- c()
  for(i in 1:100)
  {
    Numero     <- sample(Numeros, 1)
    DistribucionNumeros <- c(DistribucionNumeros,Numero)
  }
```

Por último, obtendremos la media, la varianza y la desviación estándar de cada uno y repetiremos este proceso 1000 veces para ver el resultado.

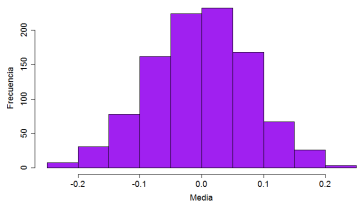
```
Numeros          <- c(-1,0,1)
for(j in 1:1000)
{
  DistribucionNumeros <- c()
  for(i in 1:100)
  {
    Numero     <- sample(Numeros, 1)
    DistribucionNumeros <- c(DistribucionNumeros,Numero)
  }
  ArrayMedias     = c(ArrayMedias,mean(DistribucionNumeros))
  ArrayVarianza   = c(ArrayVarianza,var(DistribucionNumeros))
  ArrayDestandar  = c(ArrayDestandar,sd(DistribucionNumeros))
}
```

Con lo anterior, se muestra que nuestros valores se acercan a los obtenidos analíticamente Figura 4.1.

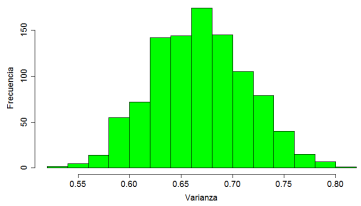
Referencias

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

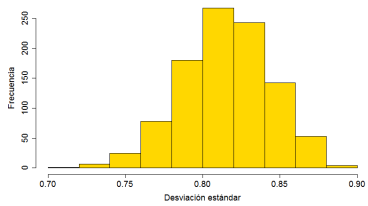
[2] The R Foundation. *The R Project for Statistical Computing*. `https://www.r-project.org/`. 2019.



(a) Frecuencia de medias obteniendo en promedio 0.



(b) Frecuencia de la varianza obteniendo en promedio $0{,}6$.



(c) Frecuencia de la desviación estándar obteniendo en promedio $0{,}81$.

Figura 4.1: Frecuencias de la media, varianza y desviación estándar.

Convolución

TAREA 11

Alumno:
Joaquín Arturo Velarde Moreno

1. Introducción

En este reporte, con el uso del programa R 4.0.2 [2], me propongo demostrar y probar algunas propiedades de la covarianza propuestos en el material del curso [1].

2. Convolución

Supongamos que tenemos dos variables independientes discretas, X y Y, también tenemos la suma de estas dos variables como $Z = X + Y$, para encontrar el valor de Z, podemos usar la convolución, la cual es la operación matemática que muestra la probabilidad de que la suma de dos variables independientes sea un número específico.

$$P(Z = j) = \sum_{i=-\infty}^{\infty} P(X = i)P(Y = j - i).$$

Tomemos por ejemplo un caso real en donde nos interesa saber la suma de las 2 caras de un dado al ser lanzadas, nuestras variables X y Y serian el resultado de cada cara por lo que $Z = X + Y$. Debido a que solo existen seis posibles resultados, al caer los dados la ecuación sería:

$$P(Z = j) = \sum_{i=1}^{6} P(X = i)P(Y = j - i).$$

3. Covarianza

La covarianza es un dato básico que existe para determinar una dependencia de dos variables aleatorias. A diferencia de los coeficientes de correlación, este no está estandarizado, por lo que puede tomar valores de ∞ hasta $-\infty$ y se representa como $Cov(X, Y)$, esto es igual a $E[(X - E[X])(Y - E[Y])]$, por lo tanto:

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])].$$

3.1. Primera propiedad

Probaremos la primera propiedad de la covarianza de manera numérica empleando la herramienta R [2]. Si X, Y son variables aleatorias y a, b, c, d son constantes, tenemos que:

$$Cov(aX + b, cY + d) = acCov(X, Y).$$

Primero estableceremos nuestras variables aleatorias por medio de una distribución uniforme para X y asignando a nuestra Y una operación de X.

```
X <- runif(100)
Y <- X*2/3
```

después declaramos nuestras constantes.

```
X <- runif(100)
Y <- X*2/3
a <- 1
b <- 2
c <- 3
d <- 4
```

Ahora, representaremos el primer miembro de nuestra ecuación $Cov(aX + b, cY + d)$.

```
X <- runif(100)
Y <- X*2/3
a <- 1
b <- 2
c <- 3
d <- 4
PrimerMiembro = cov((a * X) + b, (c * Y) + d)
```

Por último, obtenemos el segundo miembro de nuestra ecuación $acCov(X, Y)$.

```
X <- runif(100)
Y <- X*2/3
a <- 1
b <- 2
c <- 3
d <- 4
PrimerMiembro  = cov((a * X) + b, (c * Y) + d)
SegundoMiembro = a * c * cov(X,Y)
#> print(PrimerMiembro)
#[1] 0.1518915
#> print(SegundoMiembro)
#[1] 0.1518915
```

con esto podemos demostrar que ambos miembros de la ecuación son lo mismo. para la prueba analítica tenemos que recordar que la $Cov(X, Y)$ es $E(XY) - E(X)E(Y)$. Si las variables son afectadas por nuestras constantes, entonces tenemos que.

$$\begin{aligned} Cov(aX + b, cY + d) &= E[(aX + b)(cY + d)] - E(aX + b)E(cY + d) \\ &= E(acXY + adX + bcY + bd) - (acE(X)E(Y) + cbE(Y) + adE(X) + bd) \\ &= acE(XY) + adE(Y) + bcE(Y) + bd - (acE(X)E(Y) + cbE(Y) + adE(Y) + bd) \\ &= acE(XY) - acE(x)E(Y) \\ &= ac(E(XY) - E(X)E(Y)) \\ &= acCov(X, Y) \end{aligned}$$

Por lo cual resulta que nuestra propiedad analítica es correcta.

3.2. Segunda propiedad

Probaremos ahora una segundo propiedad de la varianza de manera numérica empleando la herramienta R [2]. Si X, Y son variables aleatorias, entonces la varianza de la suma de las variables es igual a la varianza de X y Y más 2 veces la covarianza de X, Y :

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y).$$

Primero estableceremos nuestras variables aleatorias por medio de una distribución uniforme para X y asignando a nuestra Y una operación de X.

```
X <- runif(100)
Y <- X*2/3
```

Ahora representaremos los miembros de nuestra ecuación $Var(X) + Var(Y) + 2Cov(X, Y)$.

```
X <- runif(100)
Y <- X*2/3
PrimerMiembro   = var(X + Y)
SegundoMiembro  = var(X) + var(Y) + (2 * cov(X,Y))
#> print(PrimerMiembro)
#[1] 0.2186215
#> print(SegundoMiembro)
#[1] 0.2186215
```

con esto podemos demostrar que ambos miembros de la ecuación son lo mismo. Para la prueba analítica debemos que recordar que la $V(X)$ es $E((X)^2) - (E(X))^2$. Si tenemos ahora la suma de X y Y entonces:

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

por lo cual resulta que nuestra propiedad analíticamente es correcta.

Referencias

[1] Satu Elisa Schaeffer. *Modelos probabilistas aplicados*. Sitio en, `https://elisa.dyndns-web.com/teaching/prob/pisis/prob.html`.

[2] The R Foundation. *The R Project for Statistical Computing*. `https://www.r-project.org/`. 2019.

Práctica. Ejercicios

TAREA 12

Alumno:
Joaquín Arturo Velarde Moreno

1. Introducción

El objetivo de esta tarea es resolver una serie de problemas seleccionados del libro Introducción a la probabilidad[1] con el uso del Wolfram Aplha[2].

2. Ejercicio 1, página 393

Sea $Z_1, Z_2, ..., Z_N$ describa un proceso de ramificación en el que cada padre tiene j ramas con probabilidad p_j. Encuentre la probabilidad d de que el proceso finalmente se extinga
(a) $p_0 = \frac{1}{2}, p_1 = \frac{1}{4}, p_2 = \frac{1}{4}$.

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

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

(d) $p_j = \frac{1}{2}, p_1 = \frac{1}{4}, p_2 = \frac{1}{4}$.

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

(f) $p_j = \exp^{-2} 2^j$, for $j = 0, 1, 2, ...$(estime d numéricamente)..

Para el punto (a) usamos un teorema el cual nos puede dar dos casos, sea N el número de hijos y d la probabilidad que el proceso muera, tenemos: $E[N] \leq 1$, entonces $d = 1$ o
$E[N] > 1$, entonces $d = \text{mín}(d = h(d))$.

Por lo que, primeramente, podemos obtener la esperanza de N.

$E(N) = \sum_{i=0}^{2} N_i P(N_i)$.

$E(N) = 0 + \frac{1}{4} + 2 * \frac{1}{4} = \frac{3}{4}$.

De acuerdo con nuestro teorema, si $E[N] \leq 1$ entonces $d = 1$.

Para el punto (b) podemos usar el mismo teorema, podemos obtener primero la esperanza de N.
$E(N) = \sum_{i=0}^{2} N_i P(N_i)$.

$E(N) = 0 + \frac{1}{3} + 2 * \frac{1}{3} = \frac{3}{3} = 1$.

De acuerdo con nuestro teorema, si $E[N] \leq 1$ entonces $d = 1$.

Para el punto (c) intentaremos el mismo teorema, obteniendo la esperanza de N.
$E(N) = \sum_{i=0}^{2} N_i P(N_i)$.

$E(N) = 0 + 0 + 2 * \frac{2}{3} = \frac{4}{3}$.

De acuerdo con nuestro teorema si $E[N] > 1$ entonces d sera la más pequeña de la ecuación d = h(d).
Esto también se puede representar como:
$d = P_0 + P_1 d + P_2 d^2$.
$d = \frac{1}{3} + (0)d + \frac{2}{3}d^2$.
Si lo igualamos a 0 tenemos,
$\frac{2}{3}d^2 - d + \frac{1}{3} = 0$.
Este ecuación podemos resolverla con la fórmula general la cual es:

$$d = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Para obtener nuestro valor tenemos que obtener la solución más pequeña de la fórmula general.
$d = \frac{-1+\sqrt{-1^2-4(\frac{2}{3})(\frac{1}{3})}}{2(\frac{2}{3})}$.

$d = \frac{-1+\sqrt{-1-\frac{8}{9}}}{\frac{4}{3}}$.
$d = \frac{0,37}{\frac{4}{3}}$.
$d = 0,277$.

$d = \frac{-1-\sqrt{-1^2-4(\frac{2}{3})(\frac{1}{3})}}{2(\frac{2}{3})}$.

$d = -0,66$.

Con esto obtenemos el d que salió menor, $-0,66$, sin embargo resultó negativo, por lo cual es necesario seguir haciendo pruebas.

Para el punto (d) obtendremos la esperanza de N.

$E(N) = \sum_{N=0}^{+\infty} N(\frac{1}{2^{N+1}})$.

Debido a su naturaleza, es un poco más complicado obtener la siguiente esperanza, por lo cual utilizamos el siguiente código en wolframalpha para calcular la probabilidad.

*sum(n * (1/2**(n+1))) from 0 to infinity.*

Esto nos da que $d = 1$.

Para el punto (e) obtendremos la esperanza de N.

$E(N) = \sum_{N=0}^{+\infty} N((\frac{1}{3})(\frac{2}{3})^N)$.

Debido a su naturaleza es un poco más complicado obtener la siguiente esperanza, por lo que utilizamos el siguiente código en wolframalpha para calcular la probabilidad.

*sum(n * (1/3)(2/3)**n) from 0 to infinity.*

Esto nos da que $d = 2$, Por lo que será necesario revisar aún más este problema. Para el punto (f) ob-

tendremos la esperanza de N.

$E(N) = \sum_{N=0}^{+\infty} N(\exp^{-2}(2^N))$.

Debido a su naturaleza es un poco más complicado obtener la siguiente esperanza, por lo que utilizamos el siguiente código en wolframalpha para poder obtener la probabilidad.

*sum(n * ((exp**-2)(2**n)) from 0 to infinity.*

Esto nos da que $d = 1$.

3. Ejercicio 3, página 393

En el problema de las letras encadenadas (vea el ejemplo 10.14) encuentre el beneficio esperado si
(a) $p_0 = \frac{1}{2}, p_1 = 0, p_2 = \frac{1}{2}$.

(b) $p_0 = \frac{1}{6}, p_1 = \frac{1}{2}, p_2 = \frac{1}{3}$.

Demuestre que si $p_0 > \frac{1}{2}$, no puedes esperar obtener ganancias.

Para el punto (a) tenemos que sacar la esperanza $E(Z_1) = \sum_{n=0}^{2} P_n N$.

Esto es $E(Z_1) = 2(\frac{1}{2})$ por lo tanto nuestra esperanza es $E(Z_1) = 1$.
Esto lo sustituimos en la ganancia esperada $50m + 50m^{12} - 100$, por lo que tenemos $50 + 50 * 1^{12} - 100 = 0$

Para el punto (b) también obtenemos la esperanza $E(Z_1) = \sum_{n=0}^{2} P_n N$.

Esto es $E(Z_1) = \frac{1}{2} + 2(\frac{1}{3})$ por lo tanto nuestra esperanza es $E(Z_1) = 1{,}16$.
Esto lo sustituimos en la ganancia esperada $50m + 50m^{12} - 100$, por lo que tenemos $50 * (1{,}16) + 50(1{,}16)^{12} - 100 = 254$.
Demostraremos ahora que si $P_0 > \frac{1}{2}$ entonces no tendremos ganancias, si tenemos que $P_0 > \frac{1}{2} entonces P_1 + P_2 < \frac{1}{2}$ lo cual tendremos que nuestro valor esperado es $E(Z_1) = P_1 + P_2$ y esto a su vez seria que $E(Z_1) < 1$, por lo que al obtener nuestras ganancias tendríamos $50m + 50m^{12} - 100 < 0$ puesto que $m = E(Z1)$.

4. Ejercicio 6, página 403

X una variable aleatoria continua cuya función característica Sea X una variable aleatoria continua cuya función característica $k_X(t)$ es
$k_X(t) = \exp^{-|t|}, -\infty < t < +\infty$.

Muestre de manera directa que la densidad fx de X es
$fx(x) = \frac{1}{\pi(1+x^2)}$.
$fx(x) = \frac{1}{2x} \int_{-\infty}^{+\infty} (e^{itx}) e^{-|t|} dt$.

5. Ejercicio 1, página 402

Sea X una variable aleatoria continua con valores en [0,2] y densidad fx.Encuentra la función generatriz de momentos $g(t)$ para X si
(a) $fx(x) = \frac{1}{2}$.

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

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

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

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

Sugerencia: utilice la definición integral, como en los ejemplos 10.15 y 10.16

Para el punto (a) obtenemos $g(t)$ integrando de la siguiente manera $\int_0^2 exp^{tx}(\frac{1}{2})dx$.
Utilizamos el siguiente código en wolframalpha para poder obtener la función generadora.

*integrate (e**(tx)*(1/2) dx) from 0 to 2.*

Obtenemos $g(t) = \frac{\exp^{2t}-1}{2t}$
Para el punto (b) obtenemos $g(t)$ integrando de la siguiente manera $\int_0^2 exp^{tx}(\frac{1}{2})xdx$.
Utilizamos el siguiente código en wolframalpha para poder obtener la función generadora.

*integrate (e**(tx)*(1/2)*x dx) from 0 to 2.*

Obtenemos $g(t) = \frac{e^{2t}(2t-1)+1}{2t^2}$.
Para el punto (c) obtenemos $g(t)$ integrando de la siguiente manera $\int_0^2 exp^{tx}(1 - \frac{1}{2}x)dx$.
Utilizamos el siguiente código en wolframalpha para poder obtener la función generadora.

*integrate (e**(tx)*(1-(1/2)*x) dx) from 0 to 2.*

Obtenemos $g(t) = \frac{-2t+\exp^{2t}-1}{2t^2}$.

Para el punto (d) obtenemos $g(t)$ integrando de la siguiente manera $\int_0^2 exp^{tx}(|1 - x|)dx$.
Utilizamos el siguiente código en wolframalpha para poder obtener la función generadora.

*integrate (e**(tx)*(—1-x—) dx) from 0 to 2.*

Obtenemos $g(t) = \frac{(\exp^2-1)(\exp^t(t-1)+t+1)}{t^2}$.
Para el punto (e) obtenemos $g(t)$ integrando de la siguiente manera $\int_0^2 exp^{tx}(\frac{3}{8}x^2)dx$.
Utilizamos el siguiente código en wolframalpha para poder obtener la función generadora.

*integrate (e**(tx)*(3/8)*x**2 dx) from 0 to 2.*

Obtenemos $g(t) = \frac{3(e^{2t}(4t^2-4t+2)-2)}{8t^3}$.

6. Ejercicio 10, página 404

Sea X_1,X_2, ...,X_n un proceso de ensayos independientes con densidad
Sea $S_n = X_1 + X_2 + ... + X_n$.
Sea $A_n = \frac{S_n}{n}$.
Sea $S_n^* = (S_n - n\mu)/\sqrt{n\sigma^2}$.
$f(x) = \frac{1}{2}\exp^{-|x|}, -\infty < t < +\infty$.

(a) Encuentre la media y la varianza de $f(x)$.

(b) Encuentre la función generadora de momentos para X_1, S_n, A_n, y S_n^*.

(c) qué se puede decir sobre la función generadora de momentos de S_n^* con $n \to \infty$.

(c) qué se puede decir sobre la función generadora de momentos de A_n con $n \to \infty$.

Para el punto (a) tenemos que encontrar la varianza $V = E(X^2) - E(X)^2$, por lo que primero obtendremos la esperanza.
Para la esperanza de X tenemos $\int_{-\infty}^{+\infty} x(\frac{1}{2}\exp^{-|x|})dx$.

Utilizamos el siguiente código en wolframaplha para obtener la esperanza.

integrate x(1/2 * exp**-x)dx from 0 to positive infinity.*

Esto nos da 0.25, por lo que $E(X)^2 = 0{,}0625$.

Para la esperanza de X^2 tenemos $\int_{-\infty}^{+\infty} x^2(\frac{1}{2}\exp^{-|x|})dx$.

Utilizamos el siguiente código en wolframaplha para obtener la esperanza.

*integrate (x**2)*(1/2 * exp**-x)dx from 0 to positive infinity.*

Esto nos da 0.2215, por lo que $E(X^2) = 0{,}2215$.

Por ultimo obtenemos nuestra varianza $V = 0{,}2215 - 0{,}0625 = 0{,}159$.

Para el punto (b) empezamos con encontrar nuestra función generadora X_1 por lo que $g_{X_1}(t) = \int_{-\infty}^{+\infty} \exp^{xt}(\frac{1}{2}\exp^{-|x|})dx$.

Referencias

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

[2] *WolframAlpha computational intelligence*. `https://www.wolframalpha.com/`. Accessed: 2020-11-24.

Ley de números grandes

TAREA 13

Alumno:
Joaquín Arturo Velarde Moreno

1. Introducción

El objetivo de esta tarea es explicar conceptos de probabilidad introducidos en el material del curso de modelos probabilistas aplicados[2] y ademas con el uso de R poder comprobar numéricamente sus propiedades[3], este documento se encuentra alojado en el repositorio[1] como recurso libre.

2. La desigualdad de Markov

Es un teorema de la teoría probabilidad, bautizado así por el matemático ruso Andrei Markov, que afirma que la probabilidad que una variable aleatoria X sea como mínimo igual a cierto valor a, es a lo más la esperanza de esta variable entre el valor a, lo cual se puede describir con la fórmula:

$$P(|X| \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Donde:

- X es una variable aleatoria.
- a es un número positivo diferente de 0 ($a \in R_{+}^{*}$).
- $\mathbb{E}(X)$ es la esperanza de X.

Podemos comprobar esta propiedad numéricamente, supongamos que tenemos una distribución de poisson con n datos, un promedio de 3 y asignamos un número cualquiera para a.

```
a <- 2
n <- 1000
E <- 3
X <- rpois(n,E)
```

Ahora, por medio de un *for* obtendremos la probabilidad de que nuestros resultados sean mayor o igual a nuestra a.

```
a <- 2
n <- 1000
E <- 3
X <- rpois(n,E)
contador = 0
for(i in 1:n)
{
  if(abs(X[i]) >= a)
  {
    contador = contador + 1
  }
}
```

Ahora para probar obtendremos la razón de nuestra esperanza entre nuestro número a.

```
> P <- contador/n
> print(P)
#[1] 0.809
> Cociente <- E/a
> print(Cociente)
#[2] 1.5
```

3. La desigualdad de Chebyshev

Es un teorema de probabilidad formulado por el matemático ruso Pafnuty Chebyshev y maestro de Andrei Markov. Esta desigualdad es la herramienta básica para demostrar resultados como la ley de los grandes números. Establece que la probabilidad de que la distancia de una variable aleatoria X a su promedio sea como mínimo un cierto número a, es a lo más la varianza entre el cuadrado del número a, el cual se puede describir como:

$$P(|X-\mu|\geq a) \leq \frac{\sigma^2}{a^2}.$$

Donde:

- σ^2 es la varianza.
- μ es el promedio esperado.
- a es un número positivo diferente de 0 ($a \in R_+^*$).

En resumen, establece que la probabilidad de que la distancia de la media sea mayor a cierto número, es a lo más la varianza entre el cuadrado de cierto número. Así que, si la varianza es pequeña, la probabilidad de estar lejos en la media también es pequeña. Este teorema puede ser probado por el teorema de Markov, Si analizamos su teorema podremos ver que podemos sustituir elementos de la desigualdad con los de Chebyshev.

$$P(|X|\geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Donde:

- $X = (X-\mu)^2$.
- $a = \epsilon^2$.

Si sustituimos las variables nos da algo esencialmente igual al teorema de Chebyshev.

$$P(|(X-\mu)^2|\geq \epsilon^2) \leq \frac{\mathbb{E}(X)}{a}.$$

Nosotros sabemos que $\sigma^2 = V[X] = \mathbb{E}[(X-\mu)^2]$ y $P(|(X-\mu)^2|\geq \epsilon^2)$ es esencialmente lo mismo que $P(|(X-\mu)|\geq \epsilon)$, por lo tanto nos queda:

$$P(|X-\mu|\geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

4. Ley de números grandes

Es un teorema fundamental de la probabilidad que indica que si repetimos muchas veces un mismo experimento(que tiende al infinito), este se acercará a nuestro promedio teórico. Es decir, la probabilidad de que la diferencia entre el promedio analítico y el experimental sea mayor o igual a un número positivo cuando n se aproxima al infinito es de 0.

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

Donde:

- n es el números de datos.
- P es la probabilidad.
- ϵ es un número positivo diferente de 0 ($\epsilon \in R_+^*$).
- μ es el promedio analítico.
- $\frac{S_n}{n}$ es el promedio experimental.

También se establece que la posibilidad de que la diferencia del promedio analítico y experimental sea menos que un número positivo cuando n se aproxima al infinito es de 1.

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

Donde:

- n es el números de datos.
- P es la probabilidad.
- ϵ es un número positivo diferente de 0 ($\epsilon \in R_+^*$).
- μ es el promedio analítico.
- $\frac{S_n}{n}$ es el promedio experimental.

Este comportamiento se puede comprobar en R[3]; imaginemos una distribución normal con n datos con un promedio de *mu* y una desviación estándar de 2.

```
n  <- 1000
mu <- 7
de <- 2
x  <- rnorm(n,mu,de)
```

Si por medio de la función *sample* obtenemos el promedio de esos datos, tendremos un valor menor o mayor que nuestro promedio analítico.

```
Sample   <- Sample(x,50)
Promedio <- sum(Sample)/length(Sample)
print(Promedio)
>   print(Promedio)
#[1] 6.574619
```

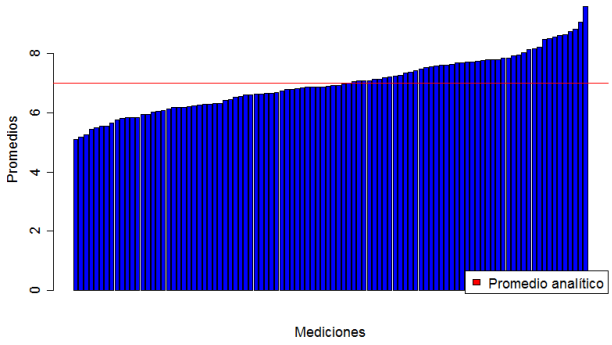
Si aumentáramos el número de elementos que tomamos en nuestra muestra, eventualmente nuestro promedio experimental y analítico serán cada vez mas cercanos (Figura 4.1).

Referencias

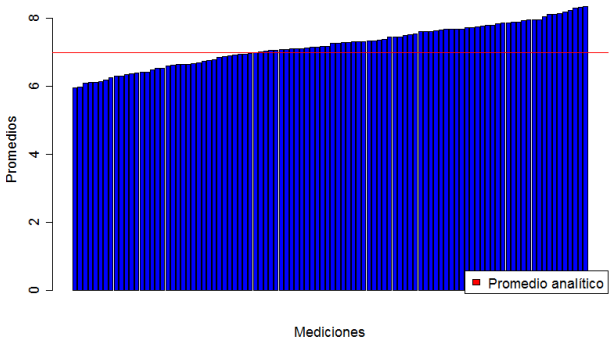
[1] Joaquin Arturo Velarde Moreno. *Repositorio con material de la clase de probabilidad.* Recursos libre, disponible en `https://github.com/joaquin3600/Modelos_Probabilistas_Aplicados`. 2020.

[2] Satu Elisa Schaeffer. *Modelos probabilistas aplicados.* Sitio en, `https://elisa.dyndns-web.com/teaching/prob/pisis/prob.html`.

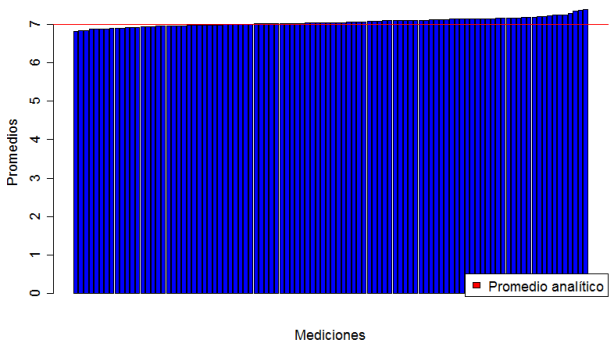
[3] The R Foundation. *The R Project for Statistical Computing.* `https://www.r-project.org/`. 2019.



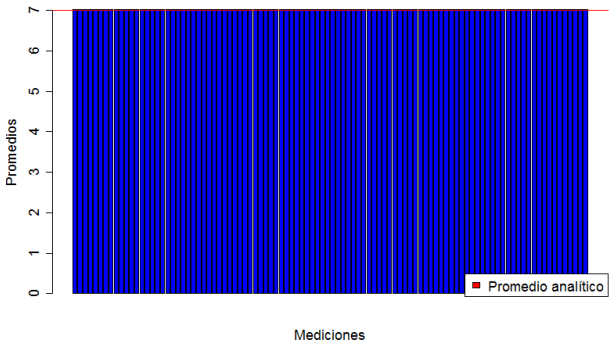
(a) Distribución del promedio de una muestra de 5 elementos aleatorios.



(b) Distribución del promedio de una muestra de 10 elementos aleatorios.



(c) Distribución del promedio de una muestra de 200 elementos aleatorios.



(d) Distribución del promedio de una muestra de 1000 elementos aleatorios.

Figura 4.1: Distribuciones de promedio en las muestras obtenidas al ir aumentando el número de muestras.

TEOREMA DEL LÍMITE CENTRAL

TAREA 14

Alumno:
Joaquín Arturo Velarde Moreno

1. Introducción

El objetivo de esta tarea es explicar como se puede aplicar el teorema de limite central, para analizar una población de la que no se conoce ni su media ni su varianza, las cuales queremos aproximar, puesto que nos interesa hacer inferencia sobre esta población en general, siendo estos dos parámetros los mas importantes para realizar inferencia. Estos conceptos de probabilidad son introducidos en el material del curso de modelos probabilistas aplicados[2] y ademas con el uso de R poder comprobar numéricamente sus propiedades[3], este documento se encuentra alojado en el repositorio[1] como recurso libre.

2. Teorema del límite central

Una muestra de tamaño n de una población por definición tiene una media:

$$p = \frac{X_1 + X_2 + ...X_n}{n}.$$

Donde:

- n es el numero de elementos de la muestra,
- X_n cada elemento de la muestra,
- p es la media muestral.

Y una varianza muestral:

$$\sigma_p^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - p)^2.$$

Donde:

- n es el numero de elementos de la muestra,
- X_i cada elemento de la muestra,
- σ_p^2 es varianza.

Ademas la desviación estándar de la muestra esta relacionada con la desviación típica de la distribución de probabilidad real de la población.

$$\sigma_p = \frac{\sigma}{\sqrt{n}}.$$

Donde:

- σ es la desviación estándar de la población,

- n tamaño de la muestra,
- σ_p es la desviación estándar de la muestra.

El teorema central del limite dice que, bajo condiciones más bien generales, las medias obtenidas de muestras aleatorias de una población tienden a tener una distribución que se aproxima a la normal.
Teorema: Tenemos por $f(x)$ una densidad con μ y varianza finita σ^2 y sea p la media de una muestra aleatoria de tamaño n de $f(x)$ definimos la variable aleatoria Y por la distribución de Y se aproxima a la normal de media cero y varianza uno cuando n crece indefinidamente.

$$Y_p = \frac{p_n - \mu}{\sigma}\sqrt{n}.$$

Podemos observar este comportamiento en R [3], supongamos que tenemos una distribución exponencial con n datos(Figura 2.1).

```
n       <- 500
vector <- rexp(n,2)
```

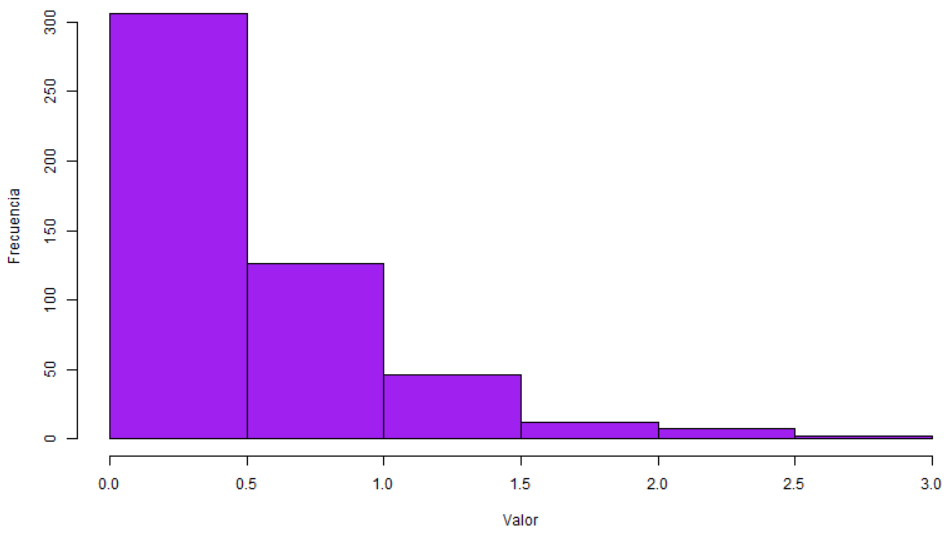
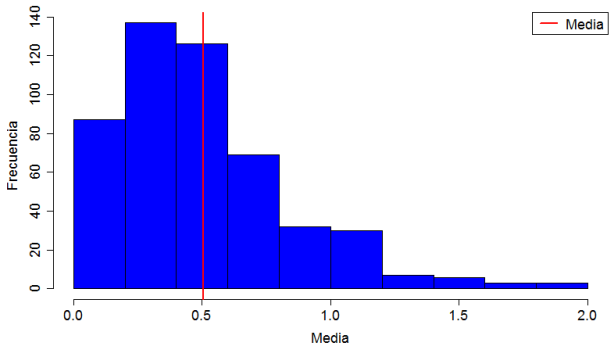


Figura 2.1: Distribución exponencial de n muestras.

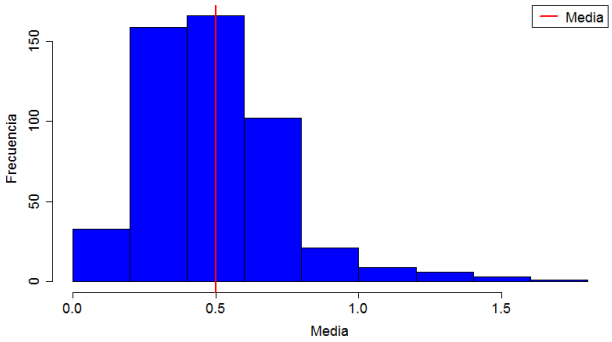
Ahora obtendremos una muestra aleatoria de elementos de esta y sacaremos su media muestral un numero n de veces.

```
n           <- 500
vector      <- rexp(n,2)
nmuestral   <- 50
for(i in 1:n)
{
  sample    <- sample(vector, nmuestral)
  medias    <- c(medias, sum(sample)/nmuestral)
}
```

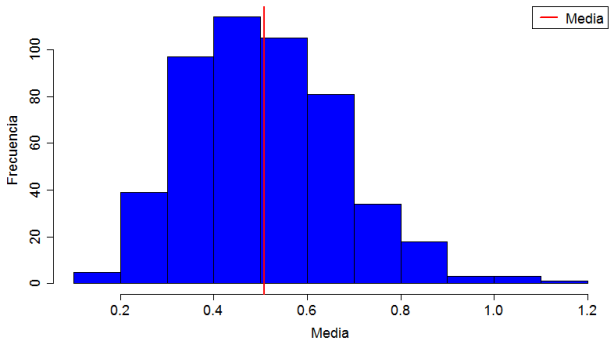
Si observamos la distribución de las medias obtenidas su distribución se aproximara cada vez mas a una normal (Figura 2.2).



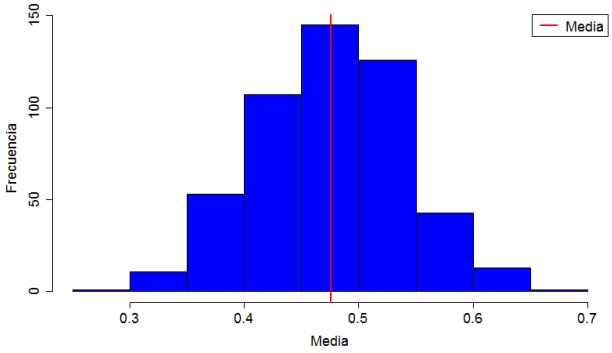
(a) Distribución del promedio de una muestra de 2 elementos aleatorios.



(b) Distribución del promedio de una muestra de 5 elementos aleatorios.



(c) Distribución del promedio de una muestra de 10 elementos aleatorios.



(d) Distribución del promedio de una muestra de 50 elementos aleatorios.

Figura 2.2: Distribuciones de promedio en las muestras obtenidas al ir aumentando el número de muestras.

3. Aplicación Pizzería

Un local de pizzas que opera en la Ciudad de México tarda una media de 25 minutos en llevar una pizza, con una desviación típica de 7 minutos. Supongamos que durante el día de hoy han repartido 150 pizzas.

- ¿Cuál es la probabilidad de que la media de los tiempos de entrega de hoy esté entre 20 y 25 minutos?.

Consideremos la variable X como tiempo de entrega, y sabemos que su media es 25 minutos y su desviación típica, 5. Pero en general esta variable no sigue una distribución normal. Durante el día de hoy se han entregado n = 150 paquetes. Es decir, tenemos una muestra $X_1, X_2, ..., X_n$ de nuestra variable. Por el teorema del límite central sabemos que la media muestral se comporta como una normal de esperanza 25 y desviación típica: $\frac{7}{\sqrt{150}} = 0{,}571$.

Con estos datos podemos usar la aproximación a la normal con nuestra variable p para calcular la probabilidad buscada de la siguiente manera:

$$P(20 \leq p \leq 25) = P(\frac{20-25}{0{,}571} \leq \frac{p-25}{0{,}571} \leq \frac{25-25}{0{,}571}).$$

Que es igual a la siguiente probabilidad:

$$P(\frac{20-25}{0{,}572} \leq Y \leq P(\frac{25-25}{0{,}572}) = P(-8{,}74 \leq Y \leq 0) = P(Y \leq 0) - P(Y \leq -8{,}74) = 0{,}5 - 0.$$

Referencias

[1] Joaquin Arturo Velarde Moreno. *Repositorio con material de la clase de probabilidad*. Recursos libre, disponible en `https://github.com/joaquin3600/Modelos_Probabilistas_Aplicados`. 2020.

[2] Satu Elisa Schaeffer. *Modelos probabilistas aplicados*. Sitio en, `https://elisa.dyndns-web.com/teaching/prob/pisis/prob.html`.

[3] The R Foundation. *The R Project for Statistical Computing*. `https://www.r-project.org/`. 2019.

Propuestas para proyecto final

TAREA 15

Alumno:
Joaquín Arturo Velarde Moreno

1. Manejo de tickets y tiempos de respuesta

Una empresa consultora quiere expandir su base de clientes, se tiene una base de datos con clientes actuales donde se puede extraer la media de del tiempo de respuesta a los tickets solicitados por fechas, utilizando el teorema central del límite trataremos de calcular qué personal se requerirá para atender el triple de los clientes que ahora tiene. Además de este objetivo, se propone saber cómo se distribuyen la horas de creación un ticket, su media, varianza y usar pruebas estadísticas para comprobar su tipo de distribución.

2. Cálculo de exceso de muertes por Covid-19 en México en 2020

En este año se ha presentado la pandemia del Covid-19 en México, que ha causado la muerte de 114,000 personas hasta el momento. Algunos medios de comunicación y el público en general, afirman que el número de muertos es mayor, por eso existe la necesidad de tener una manera de estimar el número de muertos que se dan en México anualmente. Usando los datos de mortalidad de los últimos 25 años en México, se pretende ajustar dos o más distribuciones de probabilidad para estimar el número de muertes esperado en 2020, y poder calcular exceso de muertes por Covid-19 no registradas.

3. Depreciación del peso respecto al dólar

Las empresas necesitan tomar préstamos en el extranjero en dólares, por ello necesitan tener una idea del nivel de depreciación del peso a largo plazo, que les permita valorar el riesgo de tomar estos préstamos, para ello se va a hacer una regresión lineal con el objetivo de calcular el valor del dólar a 15 años, basándonos en los datos de los últimos 40 años dados por el banco de México.

Retroalimentación de propuestas para proyecto final

TAREA 16

Alumno:
Joaquín Arturo Velarde Moreno

1. Erick, Galletas de mantequilla

Un negocio peque no dedicado a la venta de galletas de mantequilla, desea saber los niveles óptimos para hornear sus galletas, ya que ha decidido innovar y probar nuevos ingredientes y así poder ofertar otro tipo de producto, algunos factores que se suponen afecta la calidad de sus galletas son: el tiempo de horneado, el tipo de horno a utilizar, el tipo de harina (trigo o almendra), la temperatura del horno, el grosor de las galletas. Por lo que es necesario determinar mediante un diseño de experimentos, qué factores son los que afectan más al proceso de horneado de un tipo de galleta y el nivel requerido para cada uno de ellos.

Considero también pensar en el tiempo de horneado en relación a la temperatura horno, puesto que a mayor temperatura se tendrá menor tiempo de horneado. No encuentro otro factor que pueda incluirse en el estudio que pueda incluirse en el estudio e influya en la mejoría de las galletas.

2. Fabiola, mercado de valores

Las fluctuaciones de precios en los mercados de valores, no corresponden a modelos deterministas. El uso de caminatas aleatorias ha sido usado para tratar de entender y modelar estos fenómenos [2]. Este proyecto se centrara en usar modelos de caminatas aleatorias para tratar de modelar las variaciones reales de precios en un mercado de acciones.

Seria conveniente definir que valores de las acciones podrían ser mas afines, parece obvio que hay algunos que pueden tener un impacto que no es tan aleatorio, por ejemplo el precio del petroleo, si no se mueven los autos por la pandemia entonces el petroleo disminuye. Por lo que hay que ver factores aleatorios que cambia los precios.

3. Johana, Encuesta nacional (ENIM)

Mediante información de la Encuesta Nacional de Niños, Niñas y Mujeres (ENIM) 2015 en México, se construye la variable binaria que tomará el valor de 1 si el niño (entre 0 y 5 años) está desnutrido y cero, en caso contrario. La desnutrición es medida a través del indicador talla para la edad. Se utilizarán variables explicativas como la edad de la madre, número de hermanos, región de nacimiento y lactancia por parte de la madre, ingresos de la familia y, se utilizará un modelo de regresión Logit para tratar de explicar la probabilidad de si un niño es desnutrido o no.

Seria bueno que se comparara con una binomial múltiple, para comparar los resultados con la regresión Logit. específicamente con los ingresos, la lactancia y el número de hermanos.