Trabajo final de Estadística

Chain Alejandro

13 de junio de 2022

Índice

1. Introducción

El presente trabajo práctico se encuadra dentro del programa de la matería de Estadística de la maestría en Ciencia de Datos de la Facultad de Ingeniería de la Universidad Austral. El objetivo de este trabajo es la aplicación de los conceptos teóricos y prácticos desarrollados durante la cursada de la materia.

2. Ejercicio Nº1 - Estadística Descriptiva

Para el abordaje de este punto se utiliza como base de datos a la Encuesta Permanente de Hogares a nivel de individuos correspondiente al cuarto trimestre de 2021, esta encuesta es relevada por el instituto de estadísticas y Censo (INDEC), tiene una frecuencia trimestral y cubre 31 aglomerados urbanos y un área urbano-rural.

Para la descarga de esta base de datos se emplea la biblioteca de EPH (Kozlowski et al., 2020). La población que se tomará en cuenta serán todos los individuos encuestados que se encuentren con una condición de actividad de "Ocupados" y que hayan declarado haber recibido algun tipo de ingreso laboral. En total, esta población consta de 17,243 individuos de los cuales se tomará una muestra de 1.000 individuos de manera aleatoria, sin reposición y asignandole a cada individuo de la población la misma probabilidad de ser seleccionado (distribución uniforme de probabilidad).

Las variables que se tienen en cuenta para este análisis son tanto del tipo cuantitativo como cualitativo, a continuación se realiza una breve descripción de las mismas.

Variables Cualitativas:

- Sexo: Esta variable indica si el individuo es Hombre o Mujer.
- Calificación: Esta variable indica el nivel de calificación del individuo, esta variable puede ser "No calificados," "Operativos," "Técnicos" o "Profesionales."
- Nivel educativo: Esta variable indica el nivel educativo alcanzado por el individuo, este puede ser "Sin instrucción," "Primario Completo," "Secundario Completo" o "Superior Universitario Completo,"

Variables Cuantitativas:

- Edad: Esta variable numérica discreta indica la edad del individuo al momento del relevamiento.
- Horas de trabajo semanal: Esta variable numérica indica la cantidad de horas semanales que trabajo el individuo en su ocupación laboral principal.

- Ingreso total: Esta variable numérica discreta indica el ingreso total mensual en pesos que recibió la persona, este está compuesto por el ingreso laboral proveniente del desarrollo de su ocupación sumado a los ingresos no laborales percibidos.

Para realizar un análisis estadístico descriptivo desde un enfoque general de la muestra se pueden analizar tanto medidas de tendencia central, de variabilidad, de sesgo y de curtósis (Chao & Castaño, 1993) de las variables cuantitativas de los individuos muestreados.

2.1. Medidas de tendencia central

Una medida de tendencia central es aquel número que se toma como orientación para referirnos a un conjuto de datos. Dentro de las medidas de tendencia central, también conocidas como medidas de posición, se pueden encontrar:

- La Media aritmética (1), esta medida de posición es la que cuenta con mayor popularidad, esta medida calcula el centro físico del conjunto de datos y esta definida como la suma de los valores observados de una variable dividido por el total de observaciones.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1}$$

- La Mediana (2), esta medida de posición indica el valor que divide un conjunto de observaciones ordenadas respecto de la magnitud de los valores, de tal manera que el número de datos por encima de la mediana sea igual al número de datos por debajo de la misma.

$$X_m = \begin{cases} X_{(\frac{n+1}{2})} & \text{si n es un numero impar} \\ \\ \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2})+1}}{2} & \text{si n es un numero par} \end{cases} \tag{2}$$

- La Moda es una medida de posición que indica el valor que se da con mayor frecuencia en una sucesión de datos. En un conjunto de datos puede haber una moda, más de un moda (multimodal) o puede no haber ninguna moda.

Mediante el cálculo de estas medidas de posición para las variables cuantitativas de la muestra se puede obtener una descripción sencilla y simplificada de los datos. En el cuadro 1 se puede ver un resumen de estas medidas de posición para las variables de *Edad*, *Horas de Trabajo Semanal* e *Ingreso Total de los individuos*.

La media aritmética de la edad de los individuos de la muestra es igual a 41 años, por otro lado, el 50 % de los individuos de la muestra tiene 40 años o menos y la edad que es más frecuente entre

Cuadro 1: Medidas de posición

	Edad	Horas de trabajo semanal	Ingreso total
Media	41.143	35.318	$55,623
Mediana	40.000	40.000	$46,000
Moda	35.000	40.000	$60,000

estos individuos es de 35 años. Estos individuos trabajan en promedio 35 horas semanales, mientras que el 50 % de los individuos trabaja 40 horas semanales o menos, al mismo tiempo este es el valor más frecuente de horas semanales trabajadas por los individuos; para una jornada laboral de cinco días por semana esto totaliza una jornada laboral de ocho horas por día ("Jornada Full-time"). Por último, el ingreso total promedio de estos individuos es igual a $55,623 por mes, sin embargo, el 50 % de estas personas cobra $46,000 o menos y el salario más frecuente entre estas personas es igual a $60,000.

2.2. Medidas de variabilidad y asimetría

Anteriormente se presentaron multiples medidas de tendencia central para las variables de *Edad*, *Horas de Trabajo Semanal* e *Ingreso Total de los individuos. La utilidad de cada una de estas, el criterio de selección y sus limitaciones van a estar definidas por la variabilidad y la asimetría que presenten la distribución de los datos.

Tanto la media, la mediana y la moda pueden ser más o menos útiles para describir la tendencia central de los datos dependiendo de si esta distribución tiene una variación muy amplia o si la distribución es más o menos simétrica. Para esto es necesario en primer lugar dar una breve definición de las medidas de variabilidad y de simetría de un conjunto de datos. Una medida de variabilidad es una magnitud que indica el grado de dispersión de los datos, entre ellas las más comunes son la varianza, el desvio estandar y el coeficiente de variación. Una medida de simetría sirve para entender el grado de uniformidad a un lado y al otro del centro de la distribución, mediante esta medida se puede conocer si una una distribución es simétrica o está sesgada hacia la derecha (positivamente) o a la izquierda (negativamente) (Newbold et al., 2013).

A continuación se van a definir brevemente las medidas de variabilidad y de simetría que se utilizaron para el análisis descriptivo de la muestra de individuos.

- Varianza: esta medida de dispersión de los datos nace de la necesidad de poder calcular una desviación promedio de los datos al rededor de la media aritmética y de la limitación que impone la condición de que la suma de los desvios con respecto a la media de los datos es igual a cero. Para poder calcular esta desviación promedio se toma como alternativa realizar el promedio aritmético de las desviaciones con respecto a la media elevados al cuadrado (3).

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} \tag{3}$$

- Desvio Estandar: la transformación cuadrática que se realiza para el cálculo la varianza es necesaria, no obstante, esta pierde interpretabilidad debido a que brinda una magnitud de dispersión que se encuentra medida en unidades al cuadrado de los datos. El sentido del desvio estandar es poder transformar a la varianza para obtener una medida de dispersión que este representadaen la misma unidad de medida de los datos que la dieron origen, esto se logra calculando la raiz cuadrada de la varianza (4).

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}} \tag{4}$$

- Coeficiente de variación: esta medida, que representa el cociente entre el desvio estandar y la media aritmética (5), se emplea fundamentalmente para comparar la variabilidad realtiva de dos o más grupos de datos. El coeficiente de variación es muy útil porque no solo permite comparar la variabilidad de dos conjuntos de datos con distintas unidades de media (ej:ingresos y edad), sino que también es útil para comparar datos que tienen distinta media y para ver el grado de utilidad de la media como medida de tendencia central.

$$CV = \frac{S_X}{\bar{X}} \tag{5}$$

- Coeficiente de asimetría de Fisher: este coeficiente se define como el cociente entre el tercer momento en torno a la media y el cubo de la desviación estandar de los datos (7). El resultado de esta medida indica si la distribución de los datos tiene una asimetría positiva ($AS > 0$) o negativa ($AS < 0$), esto permite conocer si los valores se encuentran más lejanos a la media hacia la derecha o hacia la izquierda.

$$mc_3 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^3 * f_i}{n} \tag{6}$$

$$AS = \frac{mc_3}{S_x^3} \tag{7}$$

En el cuadro 2 se puede ver un resumen de estas medidas de dispersión y asimetría para las variables de *Edad*, *Horas de Trabajo Semanal* e *Ingreso Total de los individuos.*

A través del cálculo de estas medidas de dispersión y asimetría de los datos se puede concluir que los individuos de la muestra tienen una edad que varía, en promedio, 13 años por arriba y

Cuadro 2: Medidas de posición

	Edad	Horas de trabajo semanal	Ingreso total
Varianza	178.27	274.41	2,614,824,454.13
Desvío Estandar	13.35	16.57	51,135.35
Coeficiente Variación	0.32	0.47	0.92
Coef. Asimetría Fisher	0.39	0.05	6.46

13 años por debajo de los 41 años. Estos mismos individuos tienen una desviación promedio con respecto a la media de horas semanales trabajadas igual a 16 horas semanales mientras que su ingreso mensual varía en promedio al rededor de la media en $51,135. Las tres variables tienen una asimetría positiva, es decir, que la variabilidad de los datos es mayor hacia la derecha de la media. De estas tres variables, la que mayor variación relativa tiene es la de "Ingreso," en donde su desvio estandar representa el 92 % de la media, a su vez, esta es la que mayor asimetría posee en su distribución. La variable con menor dispersión relativa es la edad de los individuos con un coeficiente de variación igual al 32 % y la que mayor simetría posee son las horas semanales trabajadas por las personas de la muestra.

La asimetría positiva y la elevada dispersión de la variable *Ingreso total de los individuos* provoca que la utilización de la media aritmética como medida de tendencia central este lejos de ser una decisión óptima para lograr una buena representación del ingreso promedio de los individuos. En estos casos una medida como la mediana del ingreso total sería una mejor opción para resumir cual es el ingreso total promedio de las personas, en relación a que esta medida no esta siendo afectada por los valores extremos de la cola derecha de la distribución del ingreso. En conclusión, la media aritmética estaría estimando un ingreso total promedio mayor al ingreso que es recibido por el 50 % de las personas de la muestra.

2.3. Histograma de la distribución

Las medidas de tendencia central, de dispersión y de asimetría son útiles para resumir numéricamente las características de la distribución de las variables de interés. Sin embargo,los histogramas de las distribuciones brindan un método visual mediante el cual se pueden obtener conclusiones muy similares sobre las características de la distribución simplemente con una exploración gráfica.

En el gráfico 1 se encuentra el histograma de las variables de *Edad*, *Horas de Trabajo Semanal* e *Ingreso Total de los individuos* junto con sus respectivas medidas de tendencia central. Visualmente las tres variables presentan una distribución asimétrica positiva, siendo horas de trabajo semanal la que mayor grado de simetria presenta en su histograma mientras que el ingreso total es la variable que mayor grado de asimetria expone, a su vez es la que presenta una mayor cantidad de valores extremos hacia el lado derecho del histograma.

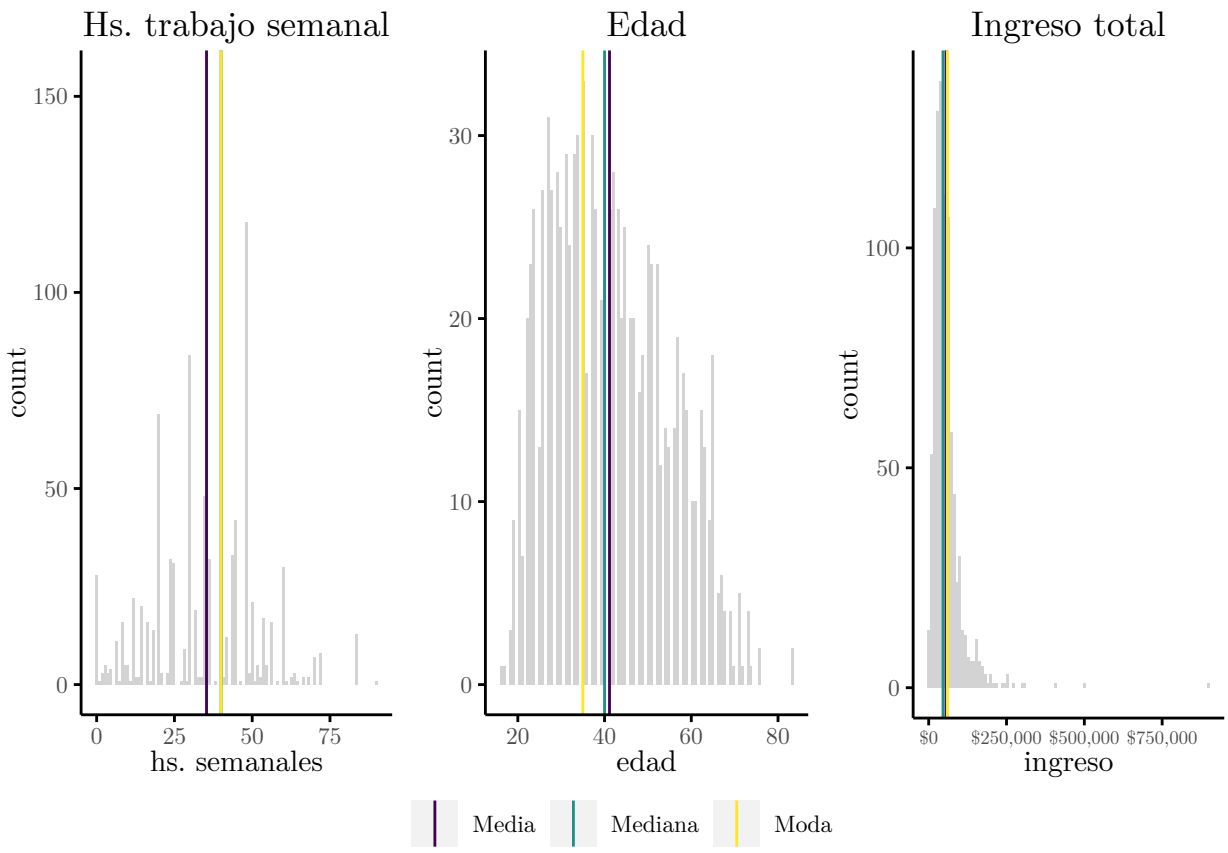


Figura 1: Histograma de variables

3. Ejercicio Nº2 - Probabilidad

3.1. Ejercicio 1

- Problema: Si lanzamos una moneda, ¿Cuál es la probabilidad esperada de obtener una cara?. Si lanzamos una moneda 10 veces, ¿Cuál es la cantidad esperada de caras?
- Solución: El lanzamiento único de una moneda "equilibrada" es un experimento aleatorio que tiene un espacio muestral igual a S={C,S}. Este experimento tiene dos eventos posibles, exhaustivos y mutuamente excluyentes: obtener una cara o una seca. Bajo estas condiciones, la probabilidad (definición clásica) de que salga una cara en el lanzamiento es igual al cociente de los casos favorables al evento "obtener una cara" sobre la cantidad de eventos posibles, en este caso solo hay dos eventos posibles y un evento favorable, por ello la probabilidad de obtener una cara es igual al 50 % (8).

$$P(cara) = \frac{\text{evento favorable=cara}}{\text{eventos posibles=cara+seca}} = \frac{1}{2} = 0.50 \tag{8}$$

En el caso de lanzar 10 veces la moneda, se puede pensar a cada lanzamiento como una distribución de Bernoulli con dos posibles sucesos (sale cara o sale seca) que son mutuamente excluyentes. Al repetir este experimento 10 veces en donde el resultado de cada repetición del experimento es independiente del anterior, se puede definir que la variable aleatoria (lanzamiento de la moneda) sigue una distribución Binomial de probabilidad (9) al ser una generalizacion de la distribución de Bernoulli con multiples repeticiones independientes (Newbold et al., 2013). Esta distribución Binomial tiene un n=10 que resulta de las 10 repeticiones del experimento y un P=0.5 que está dado por la probabilidad de éxito (obtener cara) en cada experimento.

$$X = 10 \sim B(n; P) \tag{9}$$

Para conocer la cantidad esperada de veces que sale cara es necesario el cálculo de la esperanza matemática de la distribución Binomial con 10 ensayos de Bernoulli y una probabilidad de encontrar el atributo en cada ensayo igual a 0.5. Por definición la esperanza de una distribución Binomial es igual a:

$$\mu = \text{E[x]= n*P}(\#eq : bino_2) \tag{10}$$

Reemplazando en la ecuación @(eq:bino_2) el valor de n=10 y de P=0.5 se obtiene que la cantidad de caras esperadas luego de lanzar una moneda 10 veces es igual a 5 caras.

- Problema: Lanzar una moneda 10 veces y contar el número de caras. Repetirlo 8 veces y almacenar el número de caras para cada una. Lanzar una moneda 10 veces, contar el número

de caras, almacenar el resultado y repetirlo 1000 veces. ¿Cómo difieren los resultados del experimento en (2) de los resultados en el experimento (3)? Justificar

- Solución:

```
set.seed(1118) #seteo de semilla

posibles<-c(0,1) #Resultados posibles del lanzamiento, en donde seca=0 y cara=1

proba<- c(0.5,0.5) # Probabilidad de obtencion de cada resultado

 # Simulación del lanzamiento de la moneda 10 veces a traves de un sampleo con la probabilidad
 # con una repetición de 8 veces
n_rep=1000
n=10
rep_mil<-list()
for (i in 1:n_rep) {
  data<-as.data.frame(t(matrix(sample(x = posibles,size=10, replace = TRUE,prob = proba))))
  names(data)<-as.factor(1:10)
  rep_mil[[i]]<-data
}

# Unnest de la base
rep_mil<- do.call(rbind,rep_mil)

# Cuenta de caras por cada repeticion

cantidad_caras_mil<-rowSums(rep_mil)
#Data frame con resultados

cantidad_caras_mil<-data.frame(cantidad_1000=summary(as.factor(cantidad_caras_mil)))
cantidad_caras_mil<- cantidad_caras_mil %>% mutate(porcentaje_1000=formattable::percent(cantid
```

```
## Joining, by = "Cantidad de caras C/ 10 lanzamientos"
## Joining, by = "Cantidad de caras C/ 10 lanzamientos"
```

El concepto de esperanza matemática que se aplico en la pregunta de cuantas caras se esperaría obtener si se tira 10 veces la moneda esta definido como el valor promedio que tomaría una variable aleatoria en un número muy grande de repeticiones de un experimento (Newbold et al., 2013). Cuando solamente hay ocho repeticiones del experimento de tirar 10 veces la moneda el valor más

frecuente que se obtuvo es el de 4 caras, sin embargo, cuando se repite el experimento 1,000 veces (un numero grande de repeticiones) el valor más frecuente de veces que salio cara es igual a 5, valor que se condice con la esperanza matemática del experimento aleatorio de lanzar 10 veces una moneda.

En el gráfico 2 se denota como la distribución de cantidad de veces que salio cara cuando el experimento se repitio 1,000 es muy similar a la distribución teórica que sigue una binomial con un P=0.5, mientras que la distribución cuando el experimento solo se repitio 8 veces es considerablemente distinta a la distribución teórica.

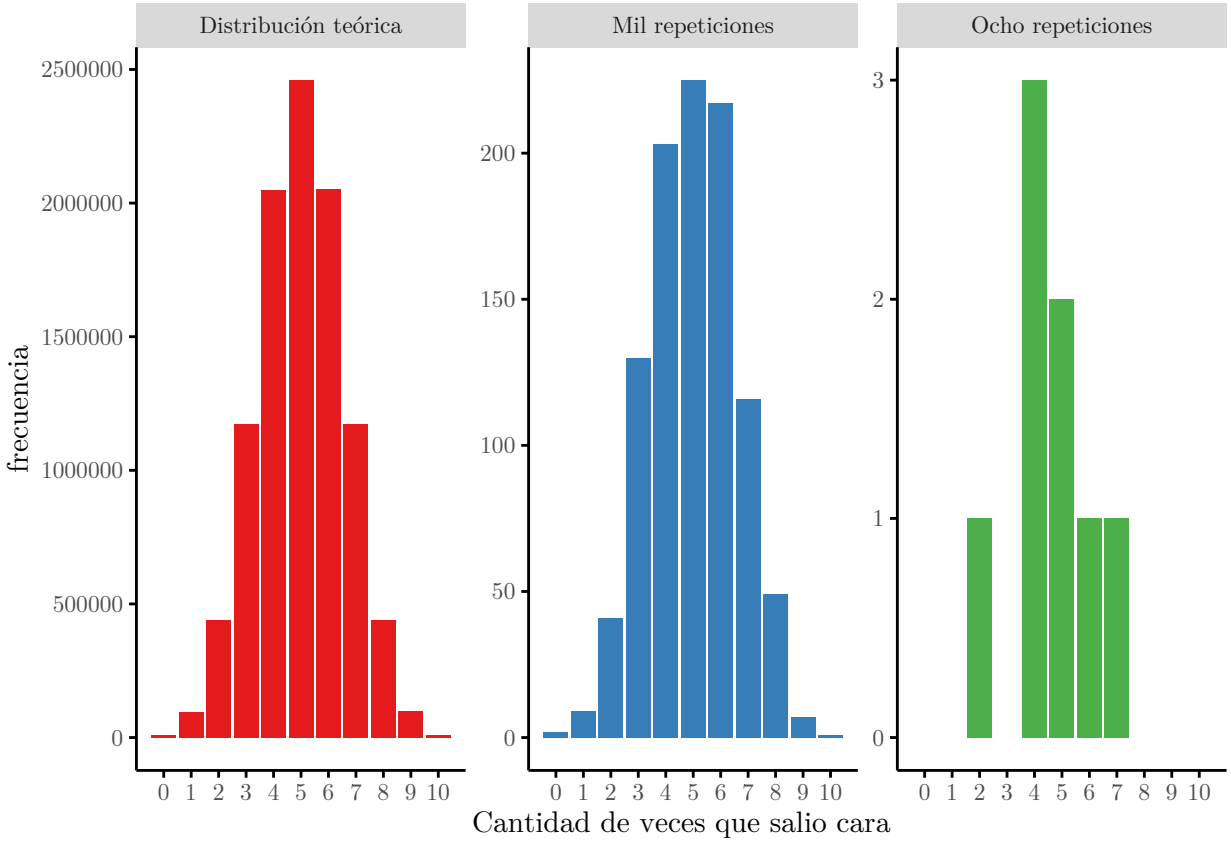


Figura 2: Distribución de veces que salio cara

- Problema: Una persona te propone jugar un juego con dados, el cual te solicita tirar 2 dados:
- Si sale un 7, te pagarán $ 3
- Si sale un 11, te pagarán $ 5.
- Si sale cualquier otra combinación, deberás pagar $ 0.70

1. ¿Cuál es la probabilidad de sacar un siete?
2. ¿Cuál es la probabilidad de sacar un once?
3. ¿Cuál es la probabilidad de sacar un siete o un once?
4. Simular tirar 2 dados mediante la función Roll1Dice(). Simular tirar 2 dados 100 veces y almacenar los resultados. Calcular los puntos anteriores (1, 2 y 3) a partir de los datos.
5. Suponga que jugó 10 veces y obtuvo una ganancia de $ 30. ¡Qué fácil parece ser el juego! ¡Debería seguir jugando! ¿Es correcta la suposición? Demostrar con una simulación.
6. Ahora dicha persona te ofrece disminuir el monto a pagar a $ 0.68. ¿Deberías aceptarlo?

- Solución:

```
#Permutaciones de los resultados posibles de tirar dos dados
resultados_posibles<-gtools::permutations(n = 6,r = 2,repeats.allowed = TRUE)
#Distribución de frecuencias de los resultados posibles de tirar dos dados
resultado_dados<-as.factor(rowSums(resultados_posibles))
epiDisplay::tab1( x0 =resultado_dados , graph = FALSE)
```

```
## resultado_dados :
##         Frequency Percent Cum. percent
## 2               1     2.8          2.8
## 3               2     5.6          8.3
## 4               3     8.3         16.7
## 5               4    11.1         27.8
## 6               5    13.9         41.7
## 7               6    16.7         58.3
## 8               5    13.9         72.2
## 9               4    11.1         83.3
## 10              3     8.3         91.7
## 11              2     5.6         97.2
## 12              1     2.8        100.0
##   Total        36   100.0        100.0
```

1. La probabilidad de sacar un 7 tirando dos dados es igual a 16.7 %.

2. La probabilidad de sacar un 11 es igual a 5.6 %.

3. La probabilidad de sacar un 7 o un 11 es igual a P(dice=7)+P(dice=11)=16.7+5.6= 22.3 %.

4. Simulación de tirar dos dados 100 veces:

```
## resultado_dados_100_veces :
##          Frequency Percent Cum. percent
## 2                3       3           3
## 3                9       9          12
## 4                9       9          21
## 5               16      16          37
## 6               10      10          47
## 7               16      16          63
## 8               12      12          75
## 9               11      11          86
## 10               8       8          94
## 11               5       5          99
## 12               1       1         100
##   Total        100     100         100
```

A partir de los datos simulados mediante 100 tiradas se puede calcular que la probabilidad de sacar un 7 es igual a 16 %, la de sacar un 11 es igual a 5 % y la de sacar un 7 o un 11 es igual a 21 %.

5. Simulación de ganancias tras tirar el dado unas 10 veces más:

```
ganancia_tras_jugar_10_veces<-sum(simulacion_ganancia_juego(n=20))
ganancia_tras_jugar_10_veces
```

```
## [1] 4.8
```

Tras jugar 20 veces más, el jugador gano $4.8, el jugador podría seguir jugando siempre y cuando no tenga que pagar ninguna prima para jugar y mientras no tenga un presupuesto que ante la acumulación de perdidas no pueda seguir jugando, esto debido a que la esperanza de la ganancia del juego es positiva e igual a $0.2371. De hecho, simulando que el jugador tira 10,000 veces el dado, este ganaría $2,690.4, lo cual es muy cercano al cálculo mediante la esperanza matemática de la ganancia tras jugar 1,000 veces (0.2371*10000) que es igual a $2371.

```
esperanza_juego<- (-0.777*0.70)+0.167*3+0.056*5
print(paste0('Ganancia simulada: ',sum(simulacion_ganancia_juego(10000)), ' Ganancia esperada:
```

```
## [1] "Ganancia simulada: 2690.4 Ganancia esperada: 2371"
```

6. Si la persona decide bajar la perdida ante el evento de sacar cualquier numero distinto a 7 o 11 sería aún mas conveniente seguir jugando, debido a que la esperanza de la ganancia es mayor.

4. Ejercicio Nº3 - Variables Aleatorias

- Problema:

1. Simular la suma de dos variables normales mediante la función rnorm(x,mu,sigma) en R con media igual a 0 y desvío igual a 1. Probar con el tamaño de muestra n = 10 y n = 100. ¿Qué observa si grafica ambos objetos?

- Solución:

En el gráfico 3 se encuentran los histogramas de la suma de las variables. La distribución con el n=100 posee un mayor grado de simetría y es más cercana a la forma de la distribución normal teórica (forma acampanada).

```
set.seed(1118)
norm1<- rnorm(10,mean = 0,sd=1)
norm2<- rnorm(10,mean = 0,sd=1)

norm3<- norm1+norm2

norm4<- rnorm(100,mean = 0,sd=1)
norm5<- rnorm(100,mean = 0,sd=1)

norm6<-norm4+norm5
```

- Problema:

2. Simular la suma de dos variables normales mediante la función rnorm(x,mu,sigma) en R con media igual a 0 y desvío igual a 1. Probar con distintos valores de tamaño de muestra. Podría probar con n = 100, n = 1000, n = 10000, n = 100000. Gráficar para estos distintos valores de muestra. Comprobar que la media (valor esperado) es igual a la suma de sus valores esperados y la varianza es igual a la suma de varianzas individuales.

- Solución: En el gráfico 4 se encuentran los histogramas de la suma de las variables. La distribuciónes ganan un mayor grado de simetría y son cada vez más cercanas a la forma de la distribución normal teórica (forma acampanada) conforme aumenta el n.

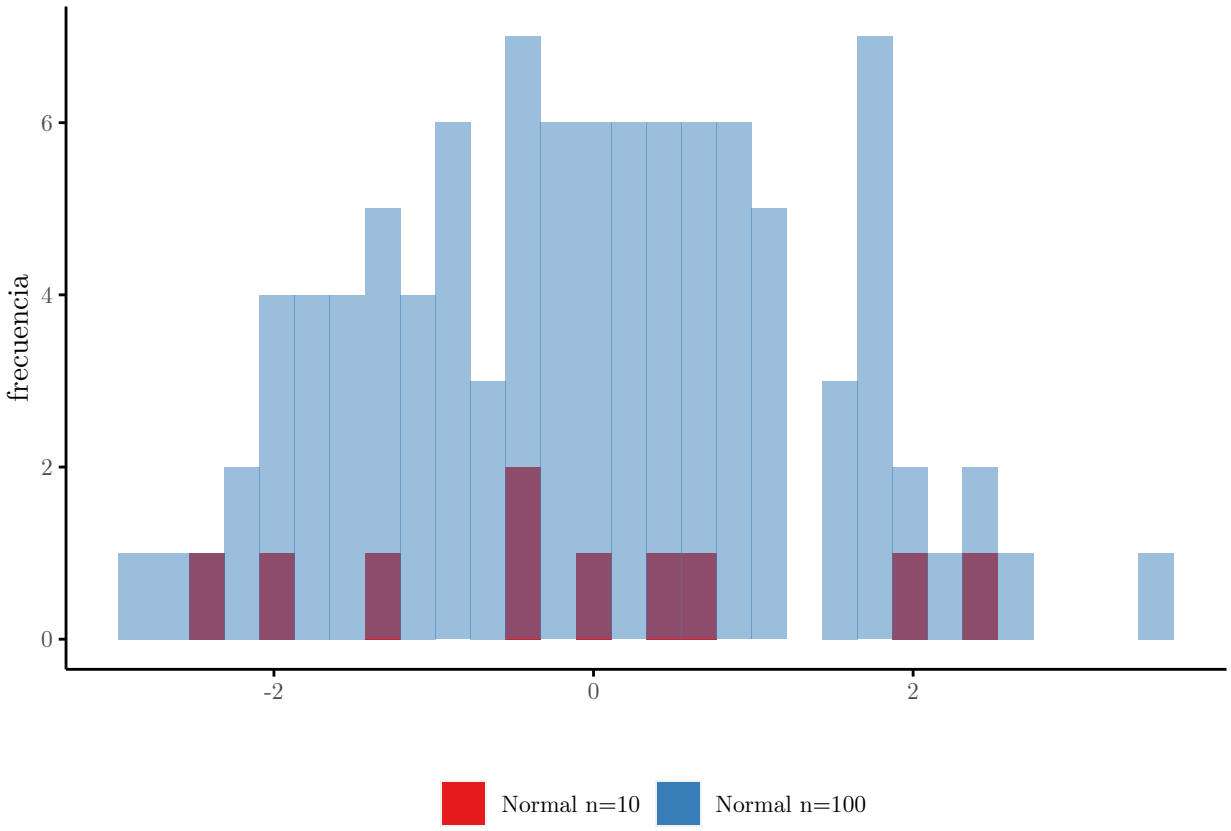


Figura 3: Histograma de suma de distribuciones normales

```r
set.seed(1118)
norm1<- rnorm(100,mean = 0,sd=1)
norm2<- rnorm(100,mean = 0,sd=1)

norm3<- norm1+norm2

norm4<- rnorm(1000,mean = 0,sd=1)
norm5<- rnorm(1000,mean = 0,sd=1)

norm6<-norm4+norm5

norm7<- rnorm(10000,mean = 0,sd=1)
norm8<- rnorm(10000,mean = 0,sd=1)

norm9<-norm7+norm8

norm10<- rnorm(100000,mean = 0,sd=1)
norm11<- rnorm(100000,mean = 0,sd=1)

norm12<-norm10+norm11
```

Las distribuciones, al ser suma de normales con media igual a 0 y varianza igual a 1 (1^2), deberían tener una media igual a la suma de sus medias y una varianza igual a la suma de sus varianzas.

La distribución que surge de la suma de dos normales, independientemente del tamaño del sampleo, deberían tener una media igual a 0 y una varianza igual a 2. A continuación se presentan los resultados obtenidos del cálculo de la media y la varianza calculadas de las distribuciones que surgen de la suma del sampleo para los distintos n.

- Media suma de distribución normal n=10: -0.02 y Varianza de suma de distribución normal n=10: 2.09.
- Media suma de distribución normal n=1,000: -0.04 y Varianza de suma de distribución normal n=1,000: 1.99.
- Media suma de distribución normal n=10,000: -0.02 y Varianza de suma de distribución normal n=10,000: 1.99.
- Media suma de distribución normal n=100,000: 0 y Varianza de suma de distribución normal n=100,000: 2.

Tanto la media como el desvio estandar tienden a los resultados esperados independientemente del n, sin embargo, cuando el n=100,000 la media y el desvio estandar se estabilizan en los valores esperados redondeando por los dos últimos decimales.

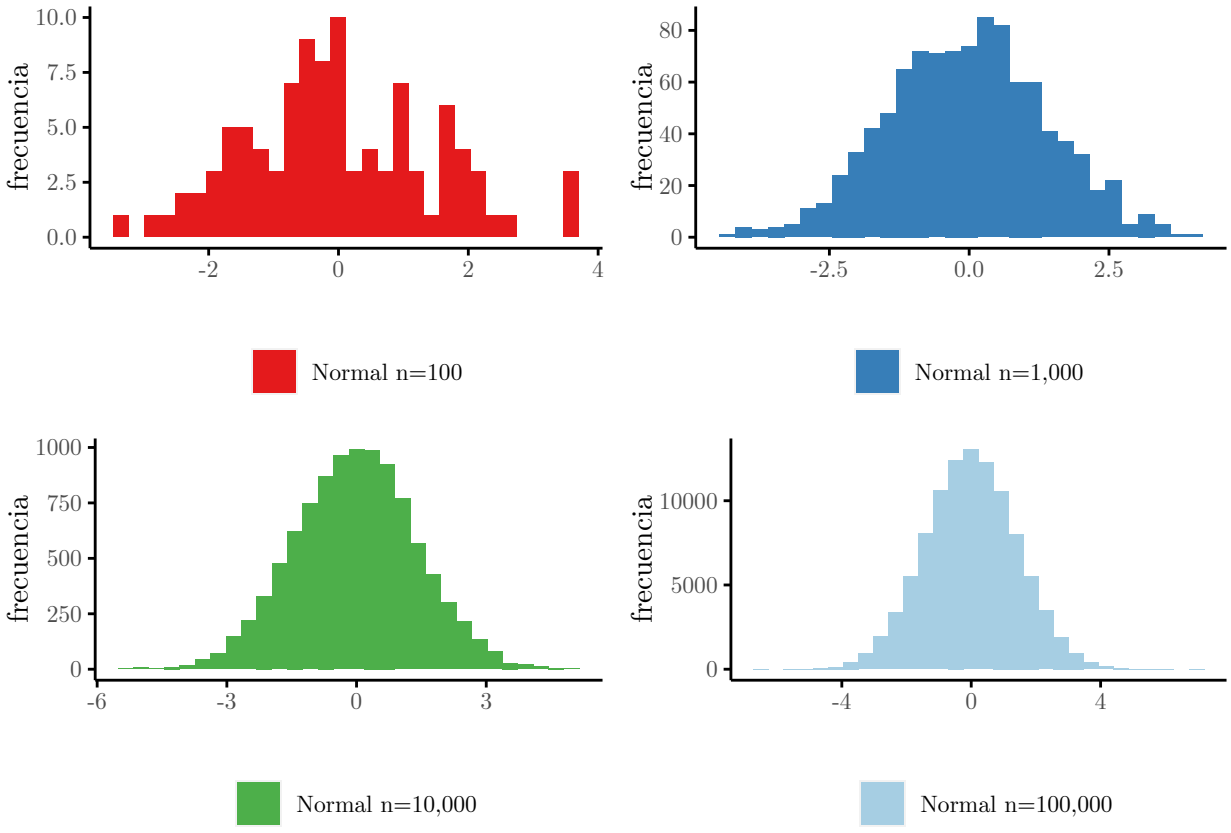


Figura 4: Histograma de suma de distribuciones normales

- Problema:

3. Simular la suma de diez variables normales mediante la función rnorm en R con media igual a 0 y desvío igual a 1. Probar con distintos valores de tamaño de muestra. Podría probar con n = 100, n = 1000, n = 10000, n = 100000. Gráficar para estos distintos valores de muestra. Comprobar que la media (valor esperado) es igual a la suma de sus valores esperados y la varianza es igual a la suma de varianzas individuales.

- Solución:

Mientras mayor es el número de distribuciones normales que se suman, mayor es la variabilidad de la distribución (gráfico 5). Tanto la media como el desvio estandar tienden a los resultados esperados independientemente del n (media=0 y varianza=10):

- Media suma de distribución normal n=10: 0.54 y la Varianza de suma de distribución normal n=10: 11.23.
- Media suma de distribución normal n=1,000: 0.14 y la Varianza de suma de distribución normal n=1,000: 10.47.
- Media suma de distribución normal n=10,000: -0.03 y la Varianza de suma de distribución normal n=10,000: 9.98.
- Media suma de distribución normal n=100,000: 0.01 y la Varianza de suma de distribución normal n=100,000: 10.02.

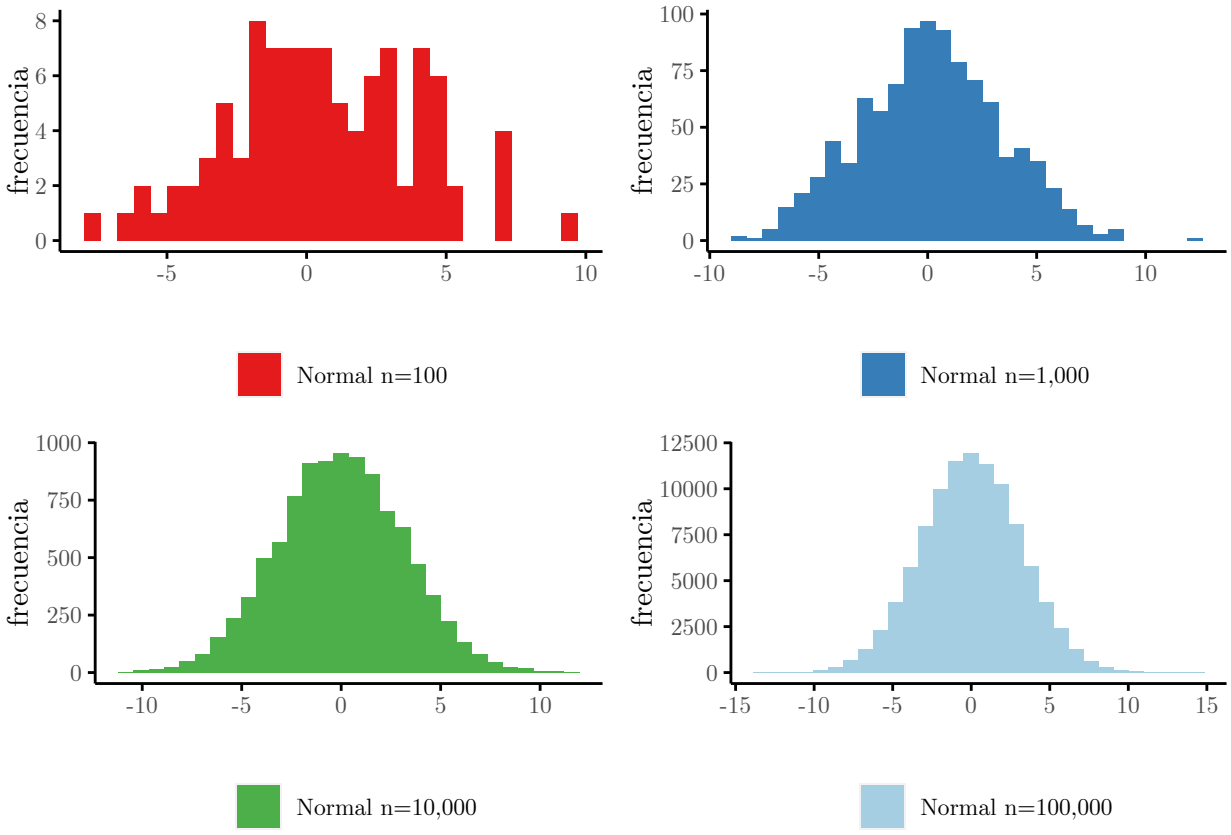


Figura 5: Histograma de suma de diez distribuciones normales

5. Ejercicio Nº4: Teorema Central del Límite

-Problema:

En clase hemos visto que la media de variables Normales es una Normal. Ahora bien, ¿Ocurrirá lo mismo si las variables que se promedian no son normales?. Se plantea el siguiente ejercicio para que intenten resolver, de forma tal que descubran al Teorema Central del Límite. Repetir el proceso visto en clase mediante R cuando la variable aleatoria original se distribuye de la forma siguiente:

1. Poisson de parámetro lambda = 1.3
2. Exponencial de parámetro mu = 1.5
3. Uniforme en el intervalo [5,10]
4. Weibull de parámetros shape = 1.2 y scale = 0.5 Realizar un Gráfico de Histograma y plotear la densidad de una variable normal para cada caso.

- Solución

1. Demostración del TCL para una distribución de Poisson de parametro lambda=1.3:

En el gráfico 6 se puede ver como la media muestral de una muestra que se extrae de una variable que se distribuye Bernoulli, mientras más grande es el tamaño de la muestra, más tiende a distribuirse como una distribución normal. Como se puede ver, para esta distribución en particular, con un n=1,000 la distribución de las medias es casi identica a la distribución normal con misma media y varianza.

2. Demostración del TCL para una distribución de Exponencial de parametro μ=1.5:

En el gráfico 7 se puede ver como la media muestral de una muestra que se extrae de una variable que se distribuye Exponencial, mientras más grande es el tamaño de la muestra, más tiende a distribuirse como una distribución normal. Como se puede ver, para esta distribución en particular, con un n=1,000 la distribución de las medias es casi identica a la distribución normal con misma media y varianza.

3. Demostración del TCL para una distribución de Uniforme en el intervalo [5,10]:

En el gráfico 8 se puede ver como la media muestral de una muestra que se extrae de una variable que se distribuye Uniforme, mientras más grande es el tamaño de la muestra, más tiende a distribuirse como una distribución normal. Como se puede ver, para esta distribución en particular, con un n=1,000 la distribución de las medias es casi identica a la distribución normal con misma media y varianza.

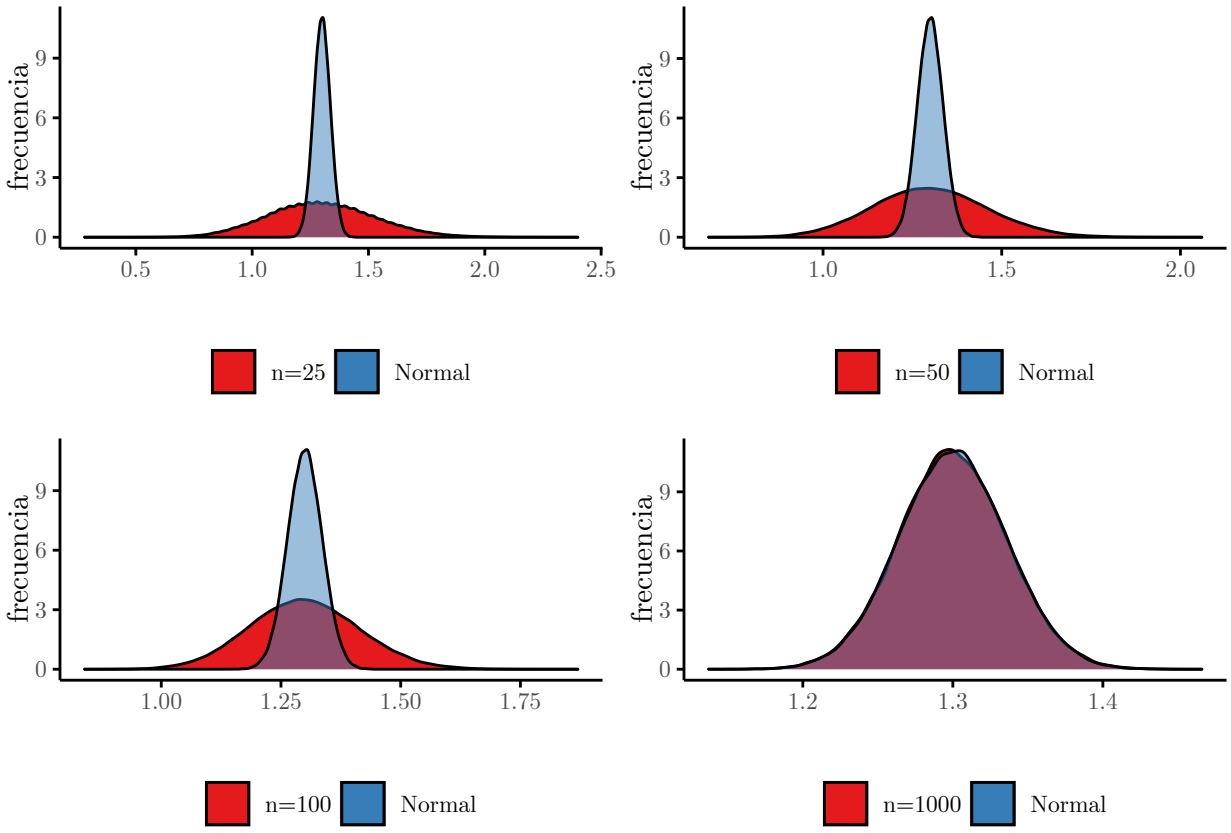


Figura 6: Aproximación al TCL con Distribución Bernoulli

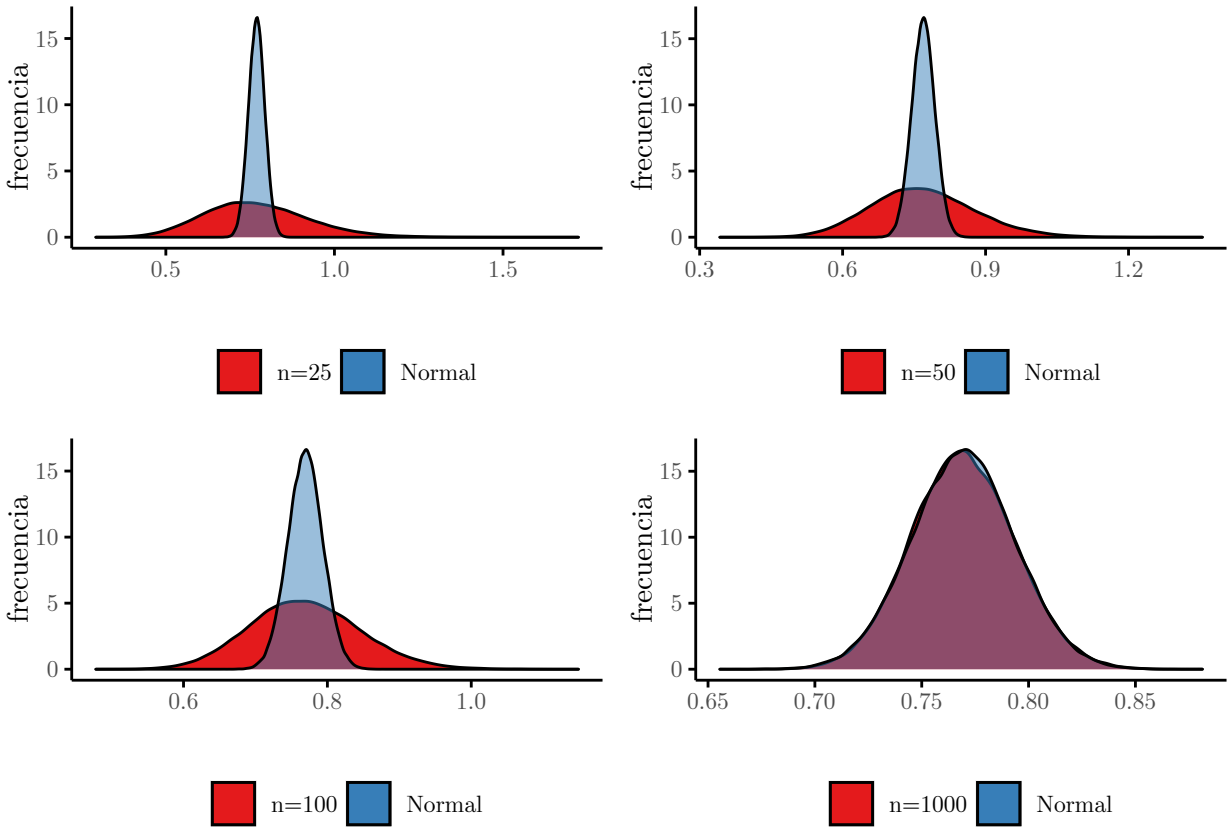


Figura 7: Aproximación al TCL con Distribución Exponencial

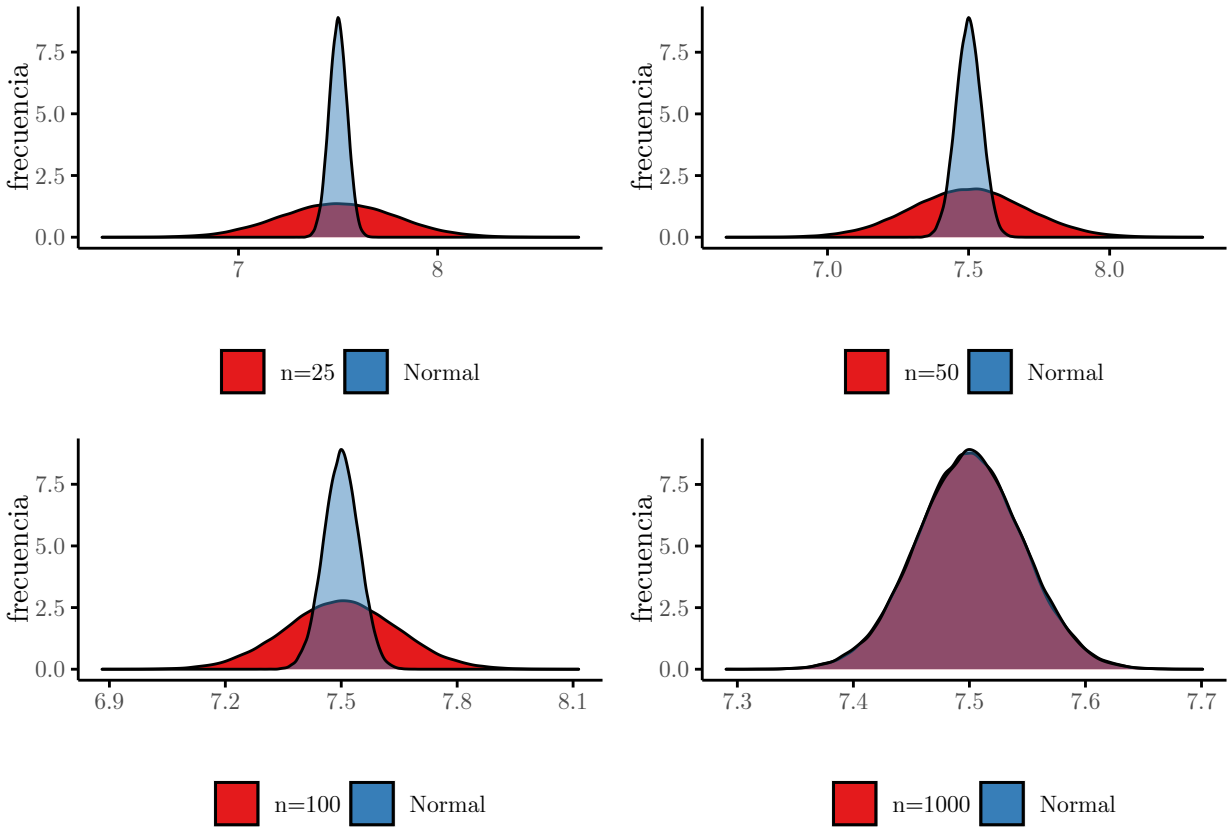


Figura 8: Aproximación al TCL con Distribución Exponencial

3. Demostración del TCL para una distribución Weibull de parámetros shape=1.2 y scale=0.5:

En el gráfico 9 se puede ver como la media muestral de una muestra que se extrae de una variable que se distribuye Weibull, mientras más grande es el tamaño de la muestra, más tiende a distribuirse como una distribución normal. Como se puede ver, para esta distribución en particular, con un n=1,000 la distribución de las medias es casi identica a la distribución normal con misma media y varianza.

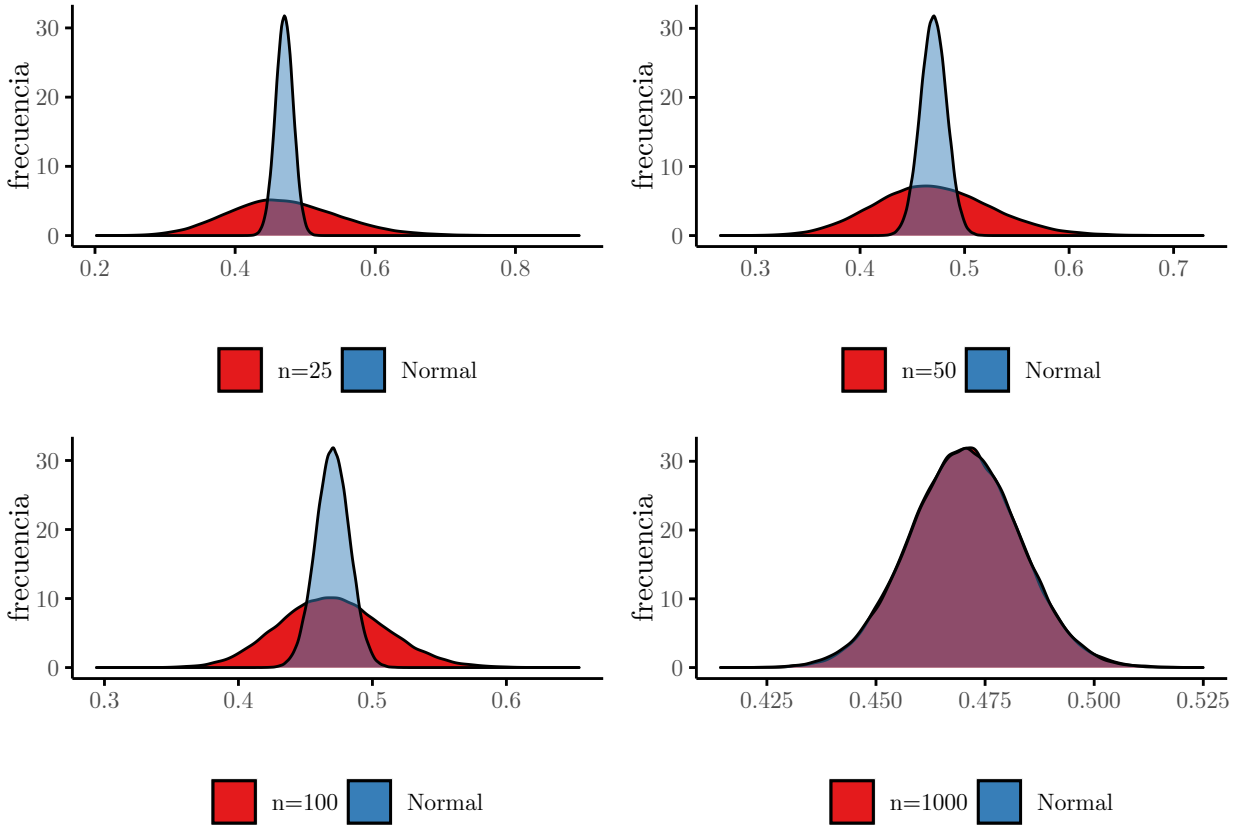


Figura 9: Aproximación al TCL con Distribución Weibull

6. Ejercicio Nº5: Regresión lineal simple

- Problema:

Realizar los siguientes ejercicios. Interprete todas sus respuestas. a) Considerar solamente las observaciones que van desde la 2 hasta la 35 y definir el data frame "datos2a35." Verificar su tamaño, variables y estructura.

Todas los puntos siguientes resolverlo con el dataset "datos2a35," b) Definir el objeto "Sexo" (género de los estudiantes). Conviértalo en factor y diga cuáles son sus respectivos niveles. c) Construir una tabla de frecuencias para la variable Sexo y el diagrama de barras correspondiente. d) Determinar la proporción de mujeres. e) Mediante el método de la región crítica: Al nivel del 5 %, determine si el porcentaje poblacional de mujeres es menor o igual que el 30 %. Escribir un resumen del enunciado del problema, verificar los supuestos, concluya, diga cuál es la fórmula, el valor de prueba, el valor crítico, la región crítica e interprete. f) Mediante el método del P-valor: Determine si el porcentaje poblacional de mujeres es menor o igual que el 30 %. Halle el P-valor, interprete y compare su decisión con el inciso (e). g) Realizar la misma prueba del inciso (h) con la función prop.test y compare los resultados obtenidos. h) Construir un intervalo del 95 % de confianza para la proporción poblacional de mujeres y compare los resultados obtenidos en los incisos anteriores.

```
## Downloading data from: https://github.com/hllinas/DatosPublicos/blob/main/Estudiantes.Rdata
```

```
## SHA-1 hash of the downloaded data file is:
## 6bf9d5a19779293538bd61d55d0662bdaf8100a1
```

```
## [1] "Estudiantes"
```

- Solución:

a) Definición de dataset "datos2a35" teniendo en cuenta solamente los datos de la observación 2 a la 35:

```
datos2a35<- datos %>% filter(Observacion>=2&Observacion<=35)

paste0('Cantidad de observaciones del DF: ', nrow(datos2a35))
```

```
## [1] "Cantidad de observaciones del DF: 34"
```

```r
paste0('Cantidad de variables del DF: ', ncol(datos2a35))
```

```
## [1] "Cantidad de variables del DF: 46"
```

b) Definición de la variable Sexo como factor con niveles 'Femenino' y 'Masculino':

```r
#Definición de niveles:
levels<- c('Femenino','Masculino')
datos2a35$Sexo<- factor(datos2a35$Sexo,levels = levels)

paste0('Clase de variable Sexo: ', class(datos2a35$Sexo))
```

```
## [1] "Clase de variable Sexo: factor"
```

```r
#Cantidad de alumnos por clase
summary(datos2a35$Sexo)
```

```
##  Femenino Masculino
##        20        14
```

c) Tabla de frecuencia de la variable "Sexo" en la tabla 3 y diagrama de barras en el gráfico 10:

```r
# Tabla de frecuencias:
sexo<-datos2a35 %>% group_by(Sexo) %>% summarise(cantidad=n())
```

Cuadro 3: Tabla de frecuencias por Sexo

Sexo	cantidad
Femenino	20
Masculino	14

c) La proporción de mujeres se puede ver en la tabla 4 y es igual a 58.82 %.

g) La hipotesis nula de que la proporción de las mujeres es menor a 0.30 plantea una prueba en donde se puede realizar un test de proporción para la dsitribución de la base de datos, a continuación se realiza el test de proporciones con una hipótesis alternativa de que el verdadero valor de p (poblacional) es mayor a 0.30.

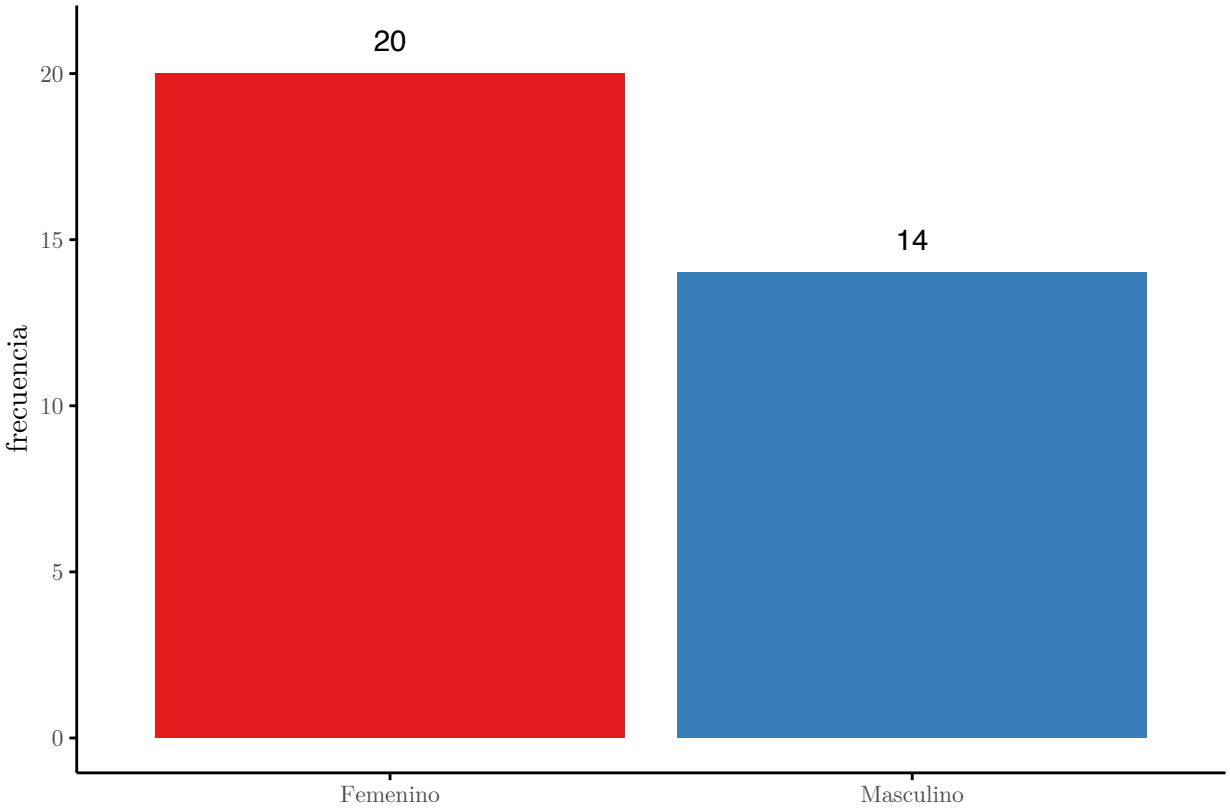


Figura 10: Tabla de frecuencia de la variable Sexo

Cuadro 4: Tabla de frecuencias y proporción por Sexo

Sexo	cantidad	proporcion
Femenino	20	58.82 %
Masculino	14	41.18 %

```
##
##  1-sample proportions test with continuity correction
##
## data:  x out of n, null probability p_test
## X-squared = 12.113, df = 1, p-value = 0.0002503
## alternative hypothesis: true p is greater than 0.3
## 95 percent confidence interval:
##  0.4337288 1.0000000
## sample estimates:
##         p
## 0.5882353

##
##  2-sample test for given proportions with continuity correction
##
## data:  c(20, 14) out of c(34, 34), null probabilities c(0.3, 0.7)
## X-squared = 24.227, df = 2, p-value = 5.485e-06
## alternative hypothesis: two.sided
## null values:
## prop 1 prop 2
##    0.3    0.7
## sample estimates:
##    prop 1    prop 2
## 0.5882353 0.4117647
```

Se puede concluir mediante el test que se rechaza la hipótesis nula de que la proporción de mujeres es menor o igual a 0.30, de hecho, con un nivel de confianza del 95 % se puede estimar que el verdadero valor poblacional de la proporción de mujeres se encuentra en el intervalo de p=[0.43,1].

7. Ejercicio Nº6: Regresión lineal simple

- Problema:

Una analista de deportes quiere saber si existe una relación entre la cantidad de bateos que realiza un equipo de béisbol y el número de runs que consigue. En caso de existir y de establecer un modelo, podría predecir el resultado del partido.

Se solicita responder lo siguiente, 1) Realizar una visualización gráfica que permita determinar la relación entre ambas variables. 2) Poner a prueba la conjetura de significancia estadística del coeficiente de correlación lineal poblacional entre ambas variables con un nivel de significación del 5 %. 3) Construir un modelo de regresión lineal simple. Identifique la variable respuesta y el regresor del modelo. Interpretar los resultados de los parámetros estimados. 4) Realizar una estimación por intervalos de confianza para la predicción del modelo obtenido con una confianza del 95 %. Interpretar los resultados. 5) Incorporar al gráfico del punto 1) el modelo estimado. 6) Verificar los supuestos del modelo de regresión lineal.

- Solución:

1) En el gráfico 11 se encuentra el gráfico de dispersión entre la cantidad de bateos (variable dependiente) y la cantidad de runs (variable independiente). A nivel gráfico, la variable cantidad de bateos pareciera estar correlacionada positivamente con la variable cantidad de runs. Para aportar aun más a este análisis se agrega una función que, a través de una regresión local polinomial, ajusta una función que relaciona el numero de bateos con la cantidad de runs, y esto aporta aun más evidencia de que la correlación es positiva debido a que la función relaciona positivamente la cantidad de bateos con la cantidad de runs.

2) Para probar la significancia estadistica se utiliza un test de correlación entre muestras pareadas a través del coeficiente de Pearson, este testo tiene como hipotesis nula de que el coeficiente de correlación entre las variables es igual a cero y que la hipotesis alternativa es que el coeficiente de correlación entre las variables no es igual a 0 (existe correlación).

```
# Se realiza el test de correlación
cor.test(datos$numero_bateos,datos$runs,method = 'pearson',conf.level = 0.95)
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos$numero_bateos and datos$runs
## t = 4.0801, df = 28, p-value = 0.0003388
```

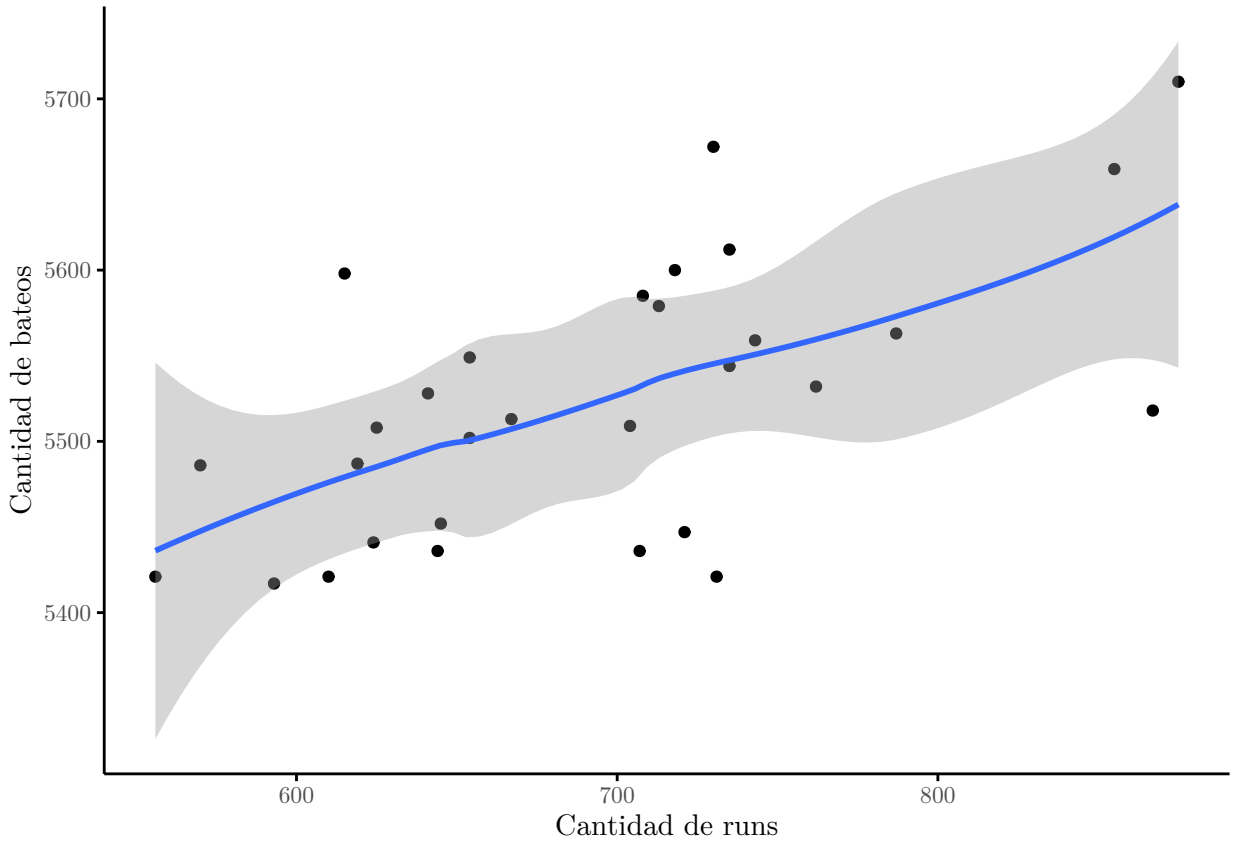


Figura 11: Análisis gráfico de correlación entre cantidad de bateos y runs

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3209675 0.7958231
## sample estimates:
##      cor
## 0.610627
```

El test de correlación indica que hay evidencias suficientes para rechazar la hipóteis nula al 95 % de nivel de confianza, es decir, se rechaza la nula de que el coeficiente de correlación lineal de Pearson es igual a 0. Otra información que brinda el test es una estimación puntual del coeficiente de correlación y una estimación por intervalos, el test indica que el coeficiente de correlación estimado entre estas variables es igual a 0.61 (correlación positiva), con un nivel de confianza del 95 % la correlación poblacional se encuentra entre 0.32 y 0.79 (rango de valores positivos al 95 % de nivel de confianza).

3) El analista quiere conocer si existe una relación entre la cantidad de bateos y la cantidad de runs en un partido de beisbol, puesto de esta forma lo que intenta es poder estimar la cantidad de bateos mediante una dependencia a la variable de runs que tiene en un partido cada equipo. Para esto se realiza una regresión lineal con la variable cantidad de bateos como regresando (variable dependiente) y la variable cantidad de runs como regresor (variable independiente).

Para el logro de este objetivo se plantea el modelo de regresión lineal entre la cantidad de bateos y la cantidad de runs:

```
##
## Call:
## lm(formula = datos$numero_bateos ~ datos$runs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.616  -42.407    3.573   43.036  126.975
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5113.3510   101.2081   50.52  < 2e-16 ***
## datos$runs     0.5913     0.1449    4.08 0.000339 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.37 on 28 degrees of freedom
```

```
## Multiple R-squared:  0.3729, Adjusted R-squared:  0.3505
## F-statistic: 16.65 on 1 and 28 DF,  p-value: 0.0003388
```

El modelo indica que existe una relación positiva estadísticamente significativa entre la cantidad de bateos y la cantidad de runs, por cada run que haga un equipo aumenta la cantidad de bateos en 0.59, esto puede no llegar a tener sentido práctico debido a que el numero de bateos es una variable discreta pero este coeficiente servira para calcular las predicciones de la cantidad de bateos con un número especifico de runs. Por otro lado, La variable cantidad de runs explica la variabilidad de la variable cantidad de bateos en un 35 % (R cuadrado ajustado), esta métrica podría mejorarse, pero para eso sería necesario obtener más datos que esten relacionados con la cantidad de bateos y que no tengan una alta colinealidad con la variable runs (para evitar la multicolinealidad).

4) La predicción del número de bateos por intervalos con un 95 % de nivel de confianza se realiza a continuación:

```
#Predicción de numero de bateos mediante el modelo con un intervalo de confianza al 95%
datos_fit<-data.frame(predict(object=modelo, newdata=datos, interval="confidence", level=0.95))
```

En el gráfico 12 se encuentra las predicciones de los bateos dependiendo de la cantidad de runs con su respectivo limite superior e inferior al 95 % de nivel de confianza.

5) En el gráfico **??** se encuentra la visualización del punto 1 en conjunto con la recta definida por la regresión lineal entre el número de bateos y de runs descripto anteriormente.

6) A continuación se testean algunos requisitos:

En primer lugar, se testea la linealidad, la distribución normal de los residuos y la media de los residuos igual a 0.

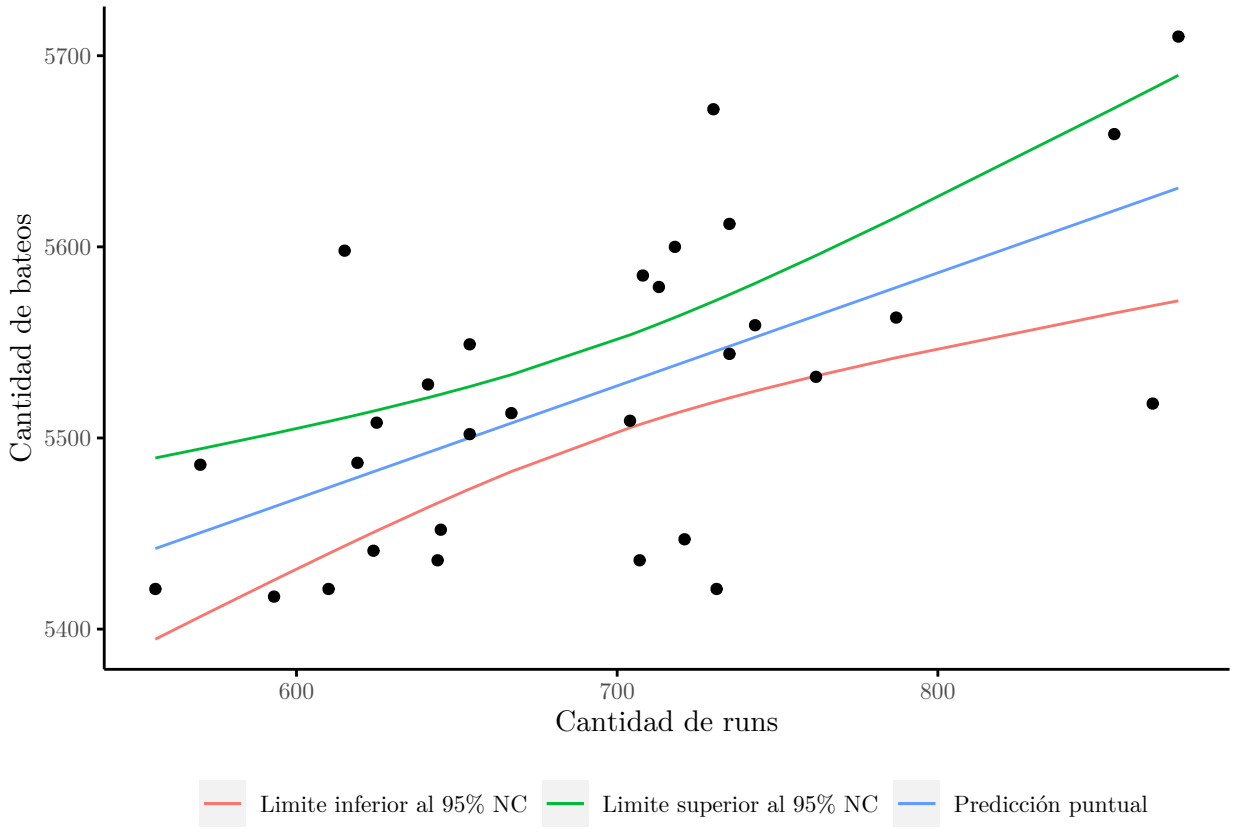


Figura 12: Predicciones de bateos por intervalo de confianza

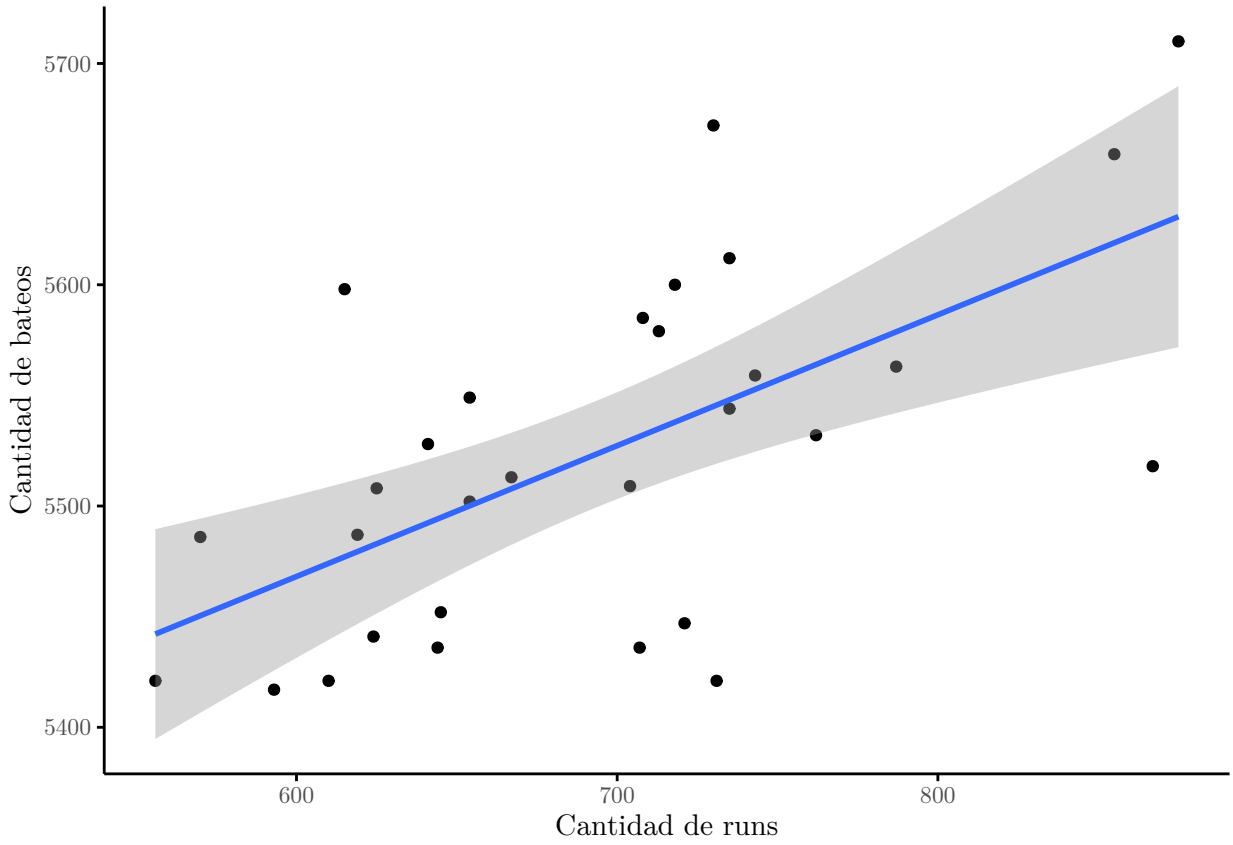
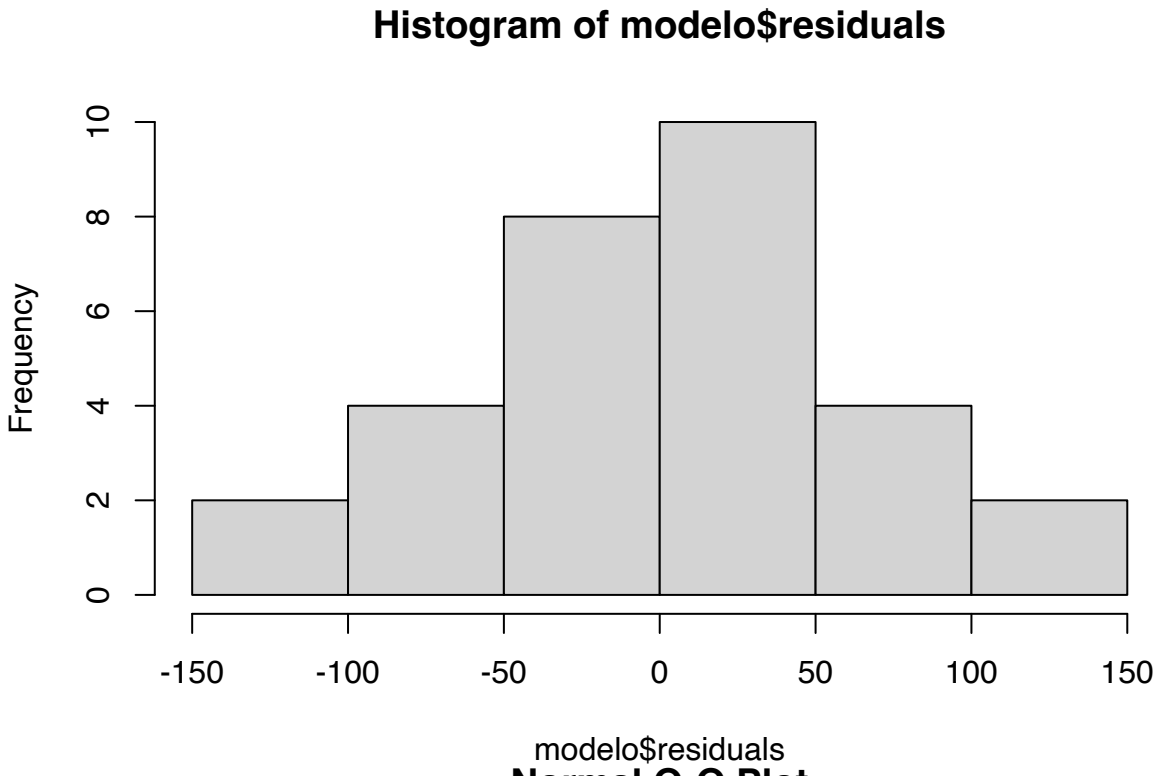
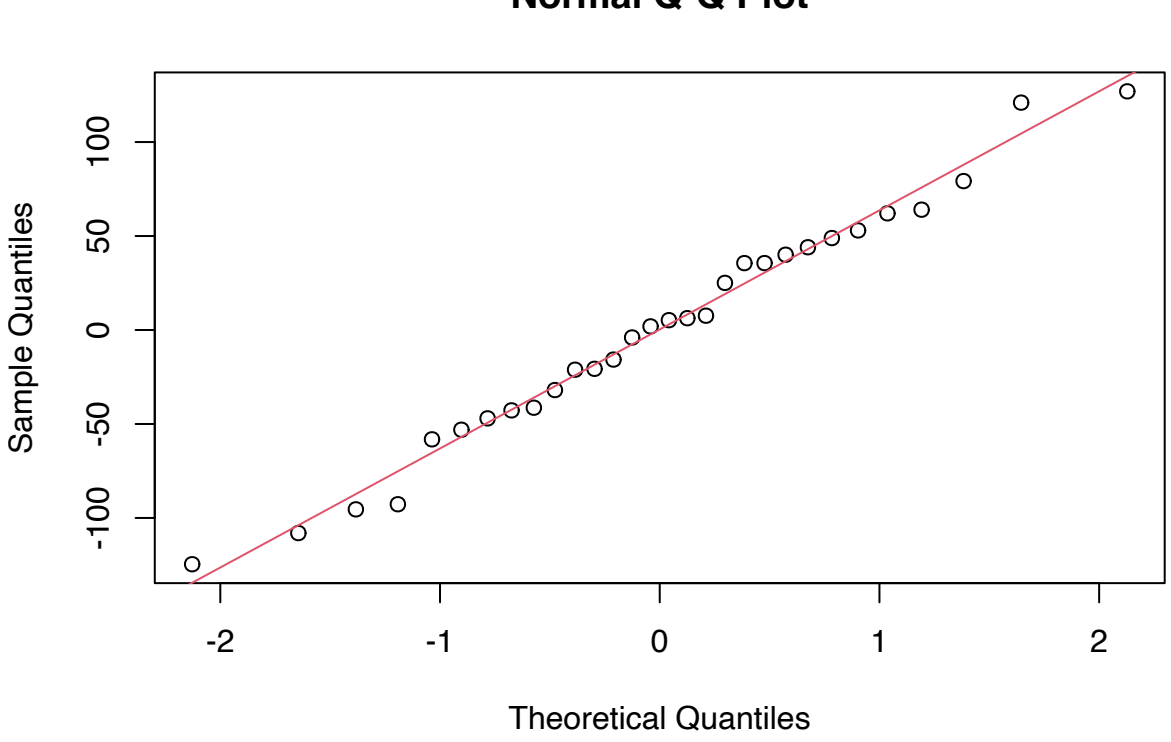
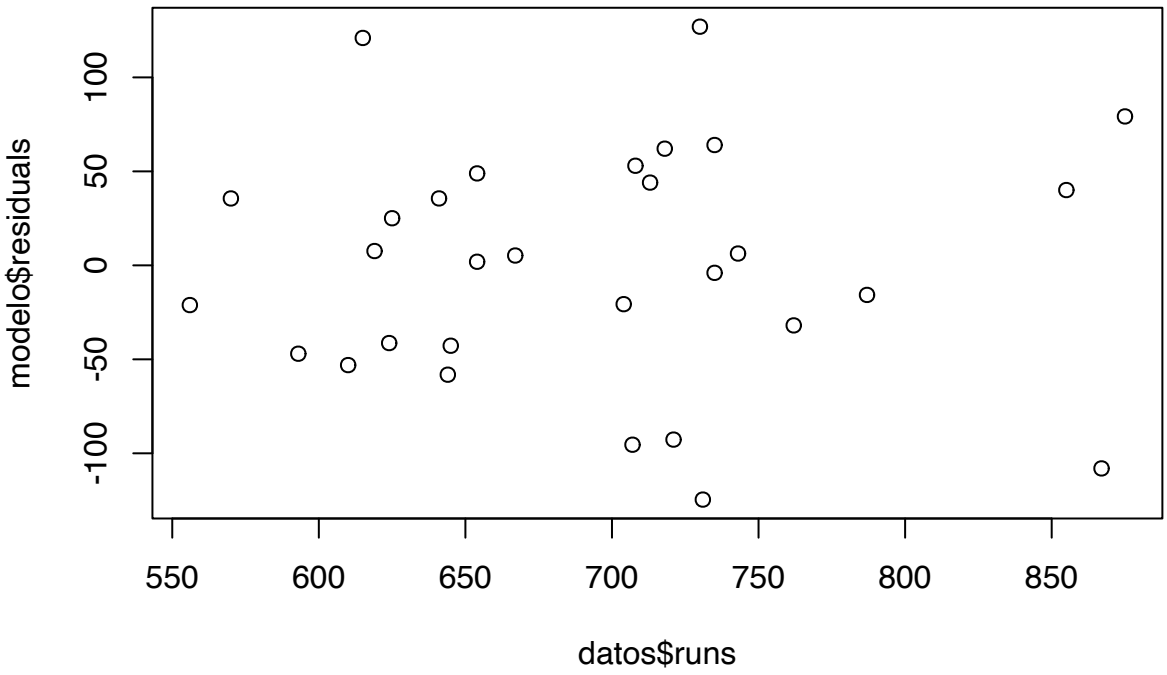


Figura 13: Gráfico de correlación entre cantidad de bateos, runs y recta de regresión


Histogram of modelo$residuals
Frequency
0
2
4
6
8
10
-150
-100
-50
0
50
100
150
modelo$residuals


Normal Q-Q Plot
Sample Quantiles
-100
-50
0
50
100
-2
-1
0
1
2
Theoretical Quantiles



Los residuos se distribuyen normalmente en torno a 0, sin embargo, la variabilidad de los residuos no parece ser constante e independiente a los valores de x (homocedasticidad).

Bibliografía

Chao, L. L., & Castaño, J. M. (1993). *Estadísticas para las ciencias administrativas* (3a. ed). McGraw-Hill.

Kozlowski, D., Tiscornia, P., Weksler, G., Rosati, G., & Shokida, N. (2020). *Eph: Argentina's permanent household survey data and manipulation utilities*. https://doi.org/10.5281/zenodo.3462677

Newbold, P., Carlson, W. L., & Thorne, B. M. (2013). *Estadística para administración y economía* (8a Edición). Pearson Educación S.A.