Simulación de un dado justo lanzado 3 veces consecutivas. Estudio de variables aleatorias discretas.

Angel Granados

@aggranadoss

Abstract

El presente estudio se centra en la distribución de la media muestral de los resultados obtenidos al lanzar un dado no cargado tres veces. Se plantea determinar la media y la desviación estándar y encontrar la distribución de la muestra. Las simulaciones se harán en R-Software examinando e identificando la repetición de ciertos valores y su distribución tanto de forma teórica como observando en el histograma resultante de los valores simulados.

Palabras clave: distribución muestral, variables aleatorias discretas, simulación en R-Software, histogramas

1. Introducción

Consideraremos el lanzamiento de un dado justo (sin cargar) de tal forma que se lanzará 3 veces consecutivas en una muestra que consideraremos "grande" Toda la teoría se sustentará en el libro de Inferencia Estadística del Wackerly-Mendenhall-Scheaffer. Se responderán a las preguntas ¿Cuáles son la media $\mu_{\bar{Y}}$ y la desviación estándar $\sigma_{\bar{Y}}$, de $\bar{Y}$? ¿Cómo podemos determinar la distribución muestral de $\bar{Y}$? justificadas desde la teoría. El ejercicio se encuentra en el libro de Inferencia Estadística, solo que aca se justificará el detalle de las cuentas y la construcción del modelo hasta su solución. Se trabajará en una simulación para representar y comparar con el resultado teórico.

2. Soporte Teórico

Para el estudio del dado, la variable aleatoria que tomaremos será discreta ya que puede tomar sólo un número finito de valores distintos y la probabilidad de que $P(\bar{Y} = y_i)$ tome el valor de y_i se definirá como la suma de las probabilidades de todos los puntos muestrales a los que se le asigna el valor de y_i.

La distribución de probabilidad para una variable discreta la representaremos como un histograma y para cualquier distribución de probabilidad discreta, debe cumplirse que:

- $0 \leq p(y) \leq 1$ para todo y
- $\sum_y p(y) = 1$

Usaremos los siguientes teoremas, definiciones y ejemplos del texto de Inferencia Estadística del Wackerly:

Teorema 3.2
Sea Y una variable aleatoria discreta con función de probabilidad $p(y)$ y sea $g(Y)$ una función de valor real de Y. Entonces, el valor esperado de $g(Y)$ está dado por:

$$E[g(Y)] = g(y)p(y)$$

Varianza y desviación estándar
Sea Y una variable aleatoria con media $E(Y) = \mu$, la varianza de una variable aleatoria Y se define como el valor esperado de $(Y - \mu)^2$. Esto es,

$$V(Y) = E[(Y - \mu)^2]$$

Teorema 5.12
Sean $Y_1, Y_2, \cdots, Y_n$ y $X_1, X_2, \cdots, X_m$ variables aleatorias con $E(Y_i) = \mu_i$ y $E(X_j) = \alpha_j$ se define

$$U_1 = \sum_{i=1}^{n} a_i Y_i \qquad \text{y} \qquad U_2 = \sum_{j=1}^{m} b_j X_j$$

con constantes a_i y b_j. Entonces se cumple lo siguiente:

1. $E[U_1] = \sum_{i=1}^{n} a_i \mu_i$
2. $V(U_1) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2 \sum \sum_{1 \leq i \leq j \leq n} a_i a_j Cov(Y_i, Y_j).$
3. $Cov(U_1, U_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j Cov(Y_i, X_j)$

Ejemplo 5.27
Sean $Y_1, Y_2, \cdots, Y_n$ variables aleatorias independientes con $E(Y_i) = \mu$ y $V(Y_i) = \sigma^2$. (Estas variables pueden denotar los resultados de n intentos independientes de un experimento.) Se define

$$\bar{Y} = \frac{1}{n}\sum_{i=1} nY_i$$

se demuestra que $E(\bar{Y}) = \mu$ y $V(\bar{Y}) = \dfrac{\sigma^2}{n}$

3. Ejemplo Ilustrativo

Un dado sin estar cargado se lanza tres veces. Sean Y_1, Y_2 y Y_3 el número de puntos vistos en la cara superior para los tiros 1, 2 y 3, respectivamente. Suponga que estamos interesados en $\bar{Y} = (Y_1 + Y_2 + Y_3)/3$, el número promedio de puntos vistos en una muestra de tamaño 3. ¿Cuáles son la media $\mu_{\bar{y}}$ y la desviación estándar $\sigma_{\bar{y}}$, de $\bar{Y}$? ¿Cómo podemos determinar la distribución muestral de $\bar{Y}$?

Solución del ejercicio:

Como el dado no está cargado, la probabilidad de que caiga un número arbitrario del conjunto $y_i \in \{1, \cdots, 6\}$ es $p(y_i) = \dfrac{1}{6}$ con $i \in \{1, \cdots, 6\}$ definido de la forma $y_i = i$. Respondiendo a la pregunta ¿Cuál es la media $\mu_{\bar{y}}$? se tiene que el cálculo de la media está dado:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (3.1)$$

por (3.1) se tiene:

- $\mu = \dfrac{1}{6}\displaystyle\sum_{i=1}^{6} y_i$

  $= \dfrac{1}{6}[1 + \cdots + 6]$

  $\mu = 3.5.$

Pero, para determinar $\mu_{\bar{Y}}$ se tiene que:

- $\mu_{\bar{Y}} = \dfrac{1}{3}\displaystyle\sum_{i=1}^{3} y_i$ con $y_i \in 1, 2, 3$

  $= \dfrac{1}{3}[1 + 2 + 3]$

  Lo cuál

$$\mu_{\bar{Y}} = 2.$$

Ahora, para determinar desviación estándar, por definición de Varianza y desviación estándar (Inferencia estadística) dada en el soporte teórico:

Varianza y desviación estándar
Si Y es una variable aleatoria con media $E(Y) = \mu$, la varianza de una variable aleatoria Y se define como el valor esperado de $(Y - \mu)^2$. Esto es,

$$V(Y) = E[(Y - \mu)^2] \qquad (3.2).$$

La *desviación estándar* de Y es la raíz cuadrada positiva de $V(Y)$.

Lo cuál, necesitamos probar que $E(Y)$ debe ser igual a μ para tener la hipótesis de la definición de Varianza y desviación estándar.

Por definición Valor esperado (Inferencia estadística).

Valor esperado
Sea Y una variable aleatoria discreta (Ya que el dado solo toma 6 valores enteros positivos) con la función de probabilidad $p(y_i)$. Entonces el valor esperado de Y, $E(Y)$, se define como

$$E(Y) = \sum_i y_i p(y_i) \qquad (3.3).$$

Como $p(y_i) = \frac{1}{6}$ para cualquier $i \in \{1, \cdots, 6\}$ ya que el dado es un dado justo (no cargado) y por la ecuación (3) de la definición de Valor esperado entonces

- $$\begin{aligned} E(Y) &= \sum_{i=1}^{6} y_i p(y_i) \\ &= \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) \\ &= \frac{1}{6}(1 + \cdots + 6) \\ &= \mu \text{ Lo cuál obtenemos la media (o el promedio)} \end{aligned}$$

Ahora, respondiendo la pregunta particular para los tres los lanzamientos del dado con $Y_1 = 1$, $Y_2 = 2$ y $Y_3 = 3$, ¿cuál es la desviación estándar $\sigma_{\bar{Y}}$?, y además por el teorema 5.12 parte 2 (Inferencia estadística) dada en el soporte teórico se tiene que:

- $V(\bar{Y}) = \sum_{i=1}^{n} a_i V(Y_i) + 2\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(Y_i, Y_j)$ con $i \neq j$.

  Como es un dado justo (ya que todos sus lanzamientos son independientes) entonces su $Cov(Y_i, Y_j) = 0$ lo cuál:

  $V(\bar{Y}) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \cancel{Cov(Y_i, Y_j)}.$

  $V(\bar{Y}) = \sum_{i=1}^{n} a_i^2 V(Y_i)$

  $V(\bar{Y}) = \sum_{i=1}^{n=3} (\frac{1}{n})^2 \sigma_i^2$ Dado que n tiene 3 lanzamientos.

  $V(\bar{Y}) = \frac{1}{9}\sum_{i=1}^{n=3} \sigma^2$

$V(\bar{Y}) = \frac{1}{\cancel{3}}(\cancel{3}\sigma^2)$

Lo cuál, la varianza de $\bar{Y}$ está denotada como:

$$V(\bar{Y}) = \frac{\sigma^2}{3} \qquad (3.4)$$

.

Por definición de Varianza y desviación estándar, $\sigma^2 = V(Y_i)$ con $Y_i = y_i \in \{1, \cdots, 6\}$ para todo $i \in \{1, \cdots, 6\}$ lo cuál

$V(Y_i) = E[(Y_i - \mu)^2]$ por la ecuación (3.2)

$\sigma^2 = E[(Y_i - \mu)^2]$ Transitividad de la igualdad en las dos ecuaciones ($\sigma^2 = V(Y_i)$ y $V(Y_i) = E[(Y_i - \mu)^2]$).

$\sigma^2 = \sum_{i}^{n}(y_i - \mu)^2 p(y_i)$ Por el teorema 3.2 enunciado en soporte teórico .

$= \sum_{i=1}^{6}(y_i - \mu)^2 p(y_i)$ con $\mu = 3.5$ dado en el cálculo de la media en (3.1).

$= (1-\mu)^2(\frac{1}{6}) + (2-\mu)^2(\frac{1}{6}) + (3-\mu)^2(\frac{1}{6}) + \cdots + (6-\mu)^2(\frac{1}{6})$

$= V(Y_i) = 2.91667$ Varianza de Y_i

Ahora, por (3.4) se tiene que

$V(\bar{Y}) = \frac{\sigma^2}{3} = \frac{V(Y_i)}{3} = \frac{2.91667}{3} \approx .972222.$

Lo cuál, hemos usado implicitamente el ejemplo 5.27 para llegar al resultado:

$$\sigma_{\bar{Y}} = \sqrt{V(\bar{Y})} \approx .986013$$

Ahora, responderemos la preguna **¿Cómo podemos determinar la distribución muestral $\bar{Y}$?**

Para determinar la distribución de la variable aleatoria $\bar{Y}$, que es el promedio de tres tiradas de un dado, es útil comprender las posibles combinaciones de resultados para Y_1, Y_2 y Y_3, y cómo se relacionan con los posibles valores de $\bar{Y}$.

Como el dado está equilibrado, hay 6.6.6 posibles lanzamientos, es decir, hay 6 posibles puestos para el primer lanzamiento, y de forma similar, para el segundo y tercer lanzamiento, por lo tanto, existen $6^3 = 216$ lanzamientos, tal que el resultado de su suma se puede repetir. Ahora observaremos las posibles combinaciones del dado lanzado 3 veces, sea W la suma de tres lanzamientos consecutivos de la forma:

$$W = Y_1 + Y_2 + Y_3 \qquad (3.5)$$

Ya que por hipótesis se obtiene:

$$\bar{Y} = \frac{Y_1 + Y_2 + Y_3}{3} \qquad (3.6)$$

Entonces, teniendo las combinaciones y usando la ecuación (3.6) observamos que la mínima combinación es que $W = 3$ y la máxima $W = 18$:

- $P(\bar{Y} = \frac{3}{3} = 1) = P(W = 3) = Y_1 + Y_2 + Y_3 = 3$ lo cuál la posibilidad de que $W = 3$ es que $p(1,1,1) = \frac{1}{216}$
- $P(\bar{Y} = \frac{4}{3}) = P(W = 4) = Y_1 + Y_2 + Y_3 = 4$ lo cuál $p(1,1,2) + p(1,2,1) + p(2,1,1) = \frac{3}{216}$
- $P(\bar{Y} = \frac{5}{3}) = P(W = 5) = Y_1 + Y_2 + Y_3 = 5$ lo cuál $p(1,1,3) + p(1,3,1) + p(3,1,1) + p(2,2,1) + p(2,1,2) + p(1,2,2) = \frac{6}{216}$
- $P(\bar{Y} = \frac{6}{3} = 2) = P(W = 6) = Y_1 + Y_2 + Y_3 = 6$ lo cuál $p(1,1,4) + p(1,4,1) + p(4,1,1) + p(1,2,3) + p(2,3,1) + p(3,1,2) + p(3,2,1) + p(2,1,3) + p(1,3,2) + p(2,2,2) = \frac{10}{216}$

$\vdots$

- $P(\bar{Y} = \frac{18}{3} = 6) = P(W = 18) = Y_1 + Y_2 + Y_3 = 6$ lo cuál $p(6,6,6) = \frac{1}{216}$

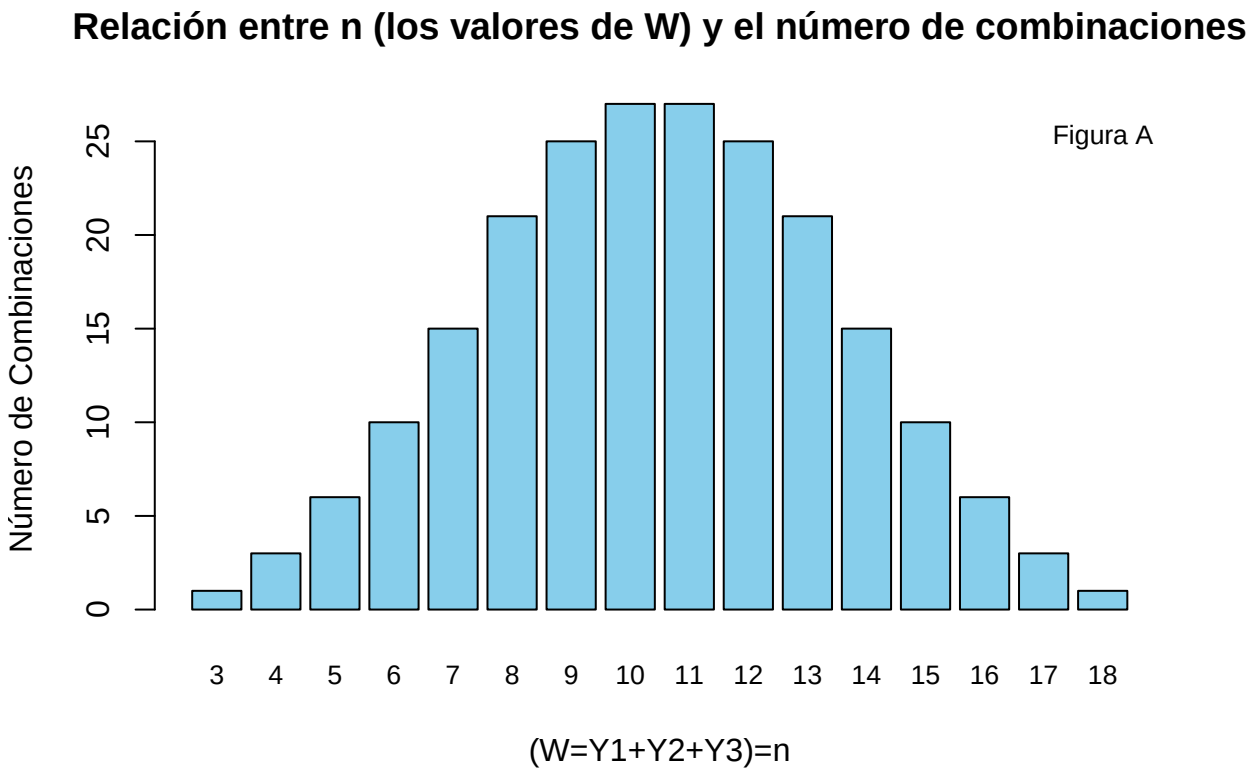
Calculando todas las combinaciones posibles de W ya que $P(W = n) = Y_1 + Y_2 + Y_3$ por código en R-software se tiene:

```
count_combinations <- function(n) {
  combinations <- 0

  for (Y1 in 1:6) {
    for (Y2 in 1:6) {
      for (Y3 in 1:6) {
        if (Y1 + Y2 + Y3 == n) {
          combinations <- combinations + 1
        }
      }
    }
  }

  return(combinations)
}

# Probamos la función para diferentes valores de n
for (n in 3:18) {
  cat("n =", n, ": Combinaciones =", count_combinations(n), "\n")
}
```

```
## n = 3 : Combinaciones = 1
## n = 4 : Combinaciones = 3
## n = 5 : Combinaciones = 6
## n = 6 : Combinaciones = 10
## n = 7 : Combinaciones = 15
## n = 8 : Combinaciones = 21
## n = 9 : Combinaciones = 25
## n = 10 : Combinaciones = 27
## n = 11 : Combinaciones = 27
## n = 12 : Combinaciones = 25
## n = 13 : Combinaciones = 21
## n = 14 : Combinaciones = 15
```

```
## n = 15 : Combinaciones = 10
## n = 16 : Combinaciones = 6
## n = 17 : Combinaciones = 3
## n = 18 : Combinaciones = 1
```

Realizamos un histograma de los valores de W con con sus combinaciones.



La probabilidad $P(\bar{Y} = \frac{n}{3})$ está determinada por la suma de las combinaciones posibles determinadas en R-software para que un n tenga como resultado con $n \in \{3, \cdots, 18\}$, lo cuál describe que es simétrica y la mayor concentración de los datos está alrededor de la media de W y tiene forma de campana, por lo cuál es una distribución normal con las variables aleatorias. Y_1, Y_2, Y_3.

4. Implementación en R mediante simulación

En el siguiente espacio de código en R-Software se toma la muestra de las caras del dado (que es la población a estudiar) y con la función sample, se toman 3 valores arbitrarios de la muestra.

```r
muestra <- c(1,2,3,4,5,6)
tres_valores_arb <- sample(muestra, 3, replace = FALSE)
```

Ahora, tomamos un espacio muestral de 100 repeticiones del experimento y calculamos su media definida de la forma:

$$\bar{Y} = \frac{Y_1 + Y_2 + Y_3}{3} \qquad (1)$$

En el siguiente código, realizamos una matriz de 4000 filas con 5 columnas y mostramos en el documento solo las primeras 6 filas donde las columnas Y_1, Y_2, Y_3 son las variables aleatorias, W que es la suma $W = Y_1 + Y_2 + Y_3$ y la columna $Media$ es el cálculo de la media definida con la ecuación (3.1) en cada una de las filas.

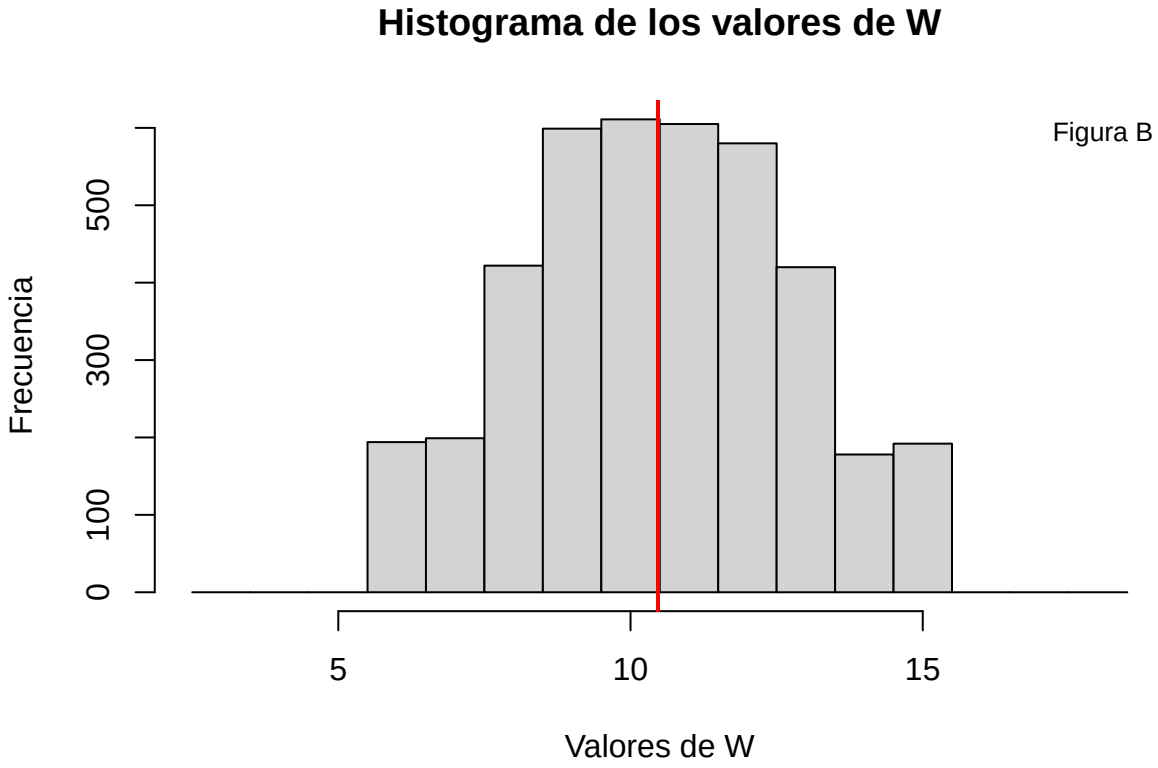
```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Crear un vector de nombres para las columnas
nombres_columnas <- c("Y1", "Y2", "Y3", "W", "Media")

# Crear el dataframe vacío con los nombres de columnas adecuados
espacio_muestral <- data.frame(matrix(numeric(0), nrow = 0, ncol = length(nombres_columnas)))
names(espacio_muestral) <- nombres_columnas

# Generar los datos y agregarlos al dataframe
for (i in 1:4000) {
  tres_valores_arb <- sample(muestra, 3, replace = FALSE)
  w <- sum(tres_valores_arb)  # Calculamos W = Y1 + Y2 + Y3
  media_definida <- mean(tres_valores_arb)
  espacio_muestral <- rbind(espacio_muestral, c(tres_valores_arb, w, media_definida))
}

# Renombrar las columnas del dataframe
espacio_muestral <- espacio_muestral %>%
  setNames(nombres_columnas)
head(espacio_muestral)
```

```
##   Y1 Y2 Y3  W    Media
## 1  5  1  4 10 3.333333
## 2  5  6  3 14 4.666667
## 3  5  6  1 12 4.000000
## 4  4  3  2  9 3.000000
## 5  4  5  2 11 3.666667
## 6  4  2  6 12 4.000000
```

Ahora, por R-software, realizamos el histograma de los 4000 datos donde relacionaremos la frecuencia de los valores repetivos en $W = Y_1 + Y_2 + Y_3$ con los posibles valores que están en el cojunto $\{1, \cdots, 18\}$ y observamos la $\mu_{\bar{Y}}$ que es la media de las medias.



Lo que coincide el estudio teórico de la **Figura A** a la simulación de la **Figura B** con valores de las variables aleatorias Y_1, Y_2, Y_3

5. Conclusiones

Al ser un dado justo (que no está cargado), lanzado 3 veces tal que

$$\bar{Y} = \frac{W}{3} \text{ con } W = Y_1 + Y_2 + Y_3$$

se logró responder las siguientes preguntas:

- ¿Cuáles son la media $\mu_{\bar{Y}}$ y la desviación estándar $\sigma_{\bar{Y}}$, de $\bar{Y}$?

  Con valores $Y_1 = 1, Y_2 = 2, Y_3 = 3$ el valor de la media es $\mu_{\bar{Y}} = 2$ y su desviación estándar $\sigma_{\bar{Y}} \approx .986013$. La desviación es una medida de dispersión que nos indica cuánto varían estos valores promedio alrededor de la media, lo cuál es baja su variación.

- ¿Cómo podemos determinar la distribución muestral de $\bar{Y}$?

  Como la probabilidad $P(W = n)$ está determinada por la suma de las combinaciones posibles determinadas en R-software para que un n tenga como resultado con $n \in \{3, \cdots, 18\}$, lo que describe que es simétrica y la mayor concentración de los datos está alrededor de la media de W ya que tiene una forma de campana, por lo cuál es una distribución normal con las variables aleatorias. Y_1, Y_2, Y_3.

- Además, se simuló con valores aleatorios el lanzamiento de un dado, con R-Software comparando la frecuencia de W con los valores esperados del conjunto $\{3, \cdots, 18\}$ que fue el estudio teórico de las probabilidades que saliera un elemento del conjunto. Ambas figuras (A y B) coinciden dada la comparación.

6. Referencias

- **Wackerly, Dennis D., William Mendenhall III, y Richard L. Scheaffer. Inferencia Estadística con aplicaciones. Séptima edición. Cengage Learning.**
- **Walpole, Ronald E., Raymond H. Myers, Sharon L. Myers, y Keying Ye. 2006. Probability and Statistics for Engineers and Scientists. 8ª edición. Prentice Hall.**