# Primer uso de Bootstrapping - TP3


## Equipo:

- Galván, Hugo César
- García, José Manuel


**Fecha: 27/05/2024**

Bootstrapping es una técnica estadística que implica el remuestreo aleatorio con reemplazo de una muestra original para crear nuevas muestras simuladas. Su utilidad radica en estimar la distribución de un estadístico, como la media o la desviación estándar, sin hacer suposiciones sobre la distribución subyacente. Esto proporciona una distribución empírica de la estadística, que se puede utilizar para estimar su sesgo, error estándar y construir intervalos de confianza.

Un ejemplo de aplicación sería en un estudio de satisfacción del cliente. Supongamos que se encuestó a 500 clientes y se calculó una puntuación promedio de satisfacción de 4.2 en una escala de 1 a 5. Para estimar el intervalo de confianza de esta media sin asumir normalidad, se podría utilizar bootstrapping. Se tomarían miles de muestras de 500 clientes cada una, con reemplazo, de la muestra original, y se calcularía la media para cada una de estas muestras bootstrap. Luego, se utilizaría la distribución empírica de estas medias para construir el intervalo de confianza deseado.

Con una muestra limitada, el bootstrapping permitiría generar múltiples muestras bootstrap a partir de los datos originales y calcular la media en cada una de ellas. Esto daría una mejor idea de la variabilidad de la media y permitiría construir intervalos de confianza más precisos sin tener que confiar en supuestos teóricos.

---

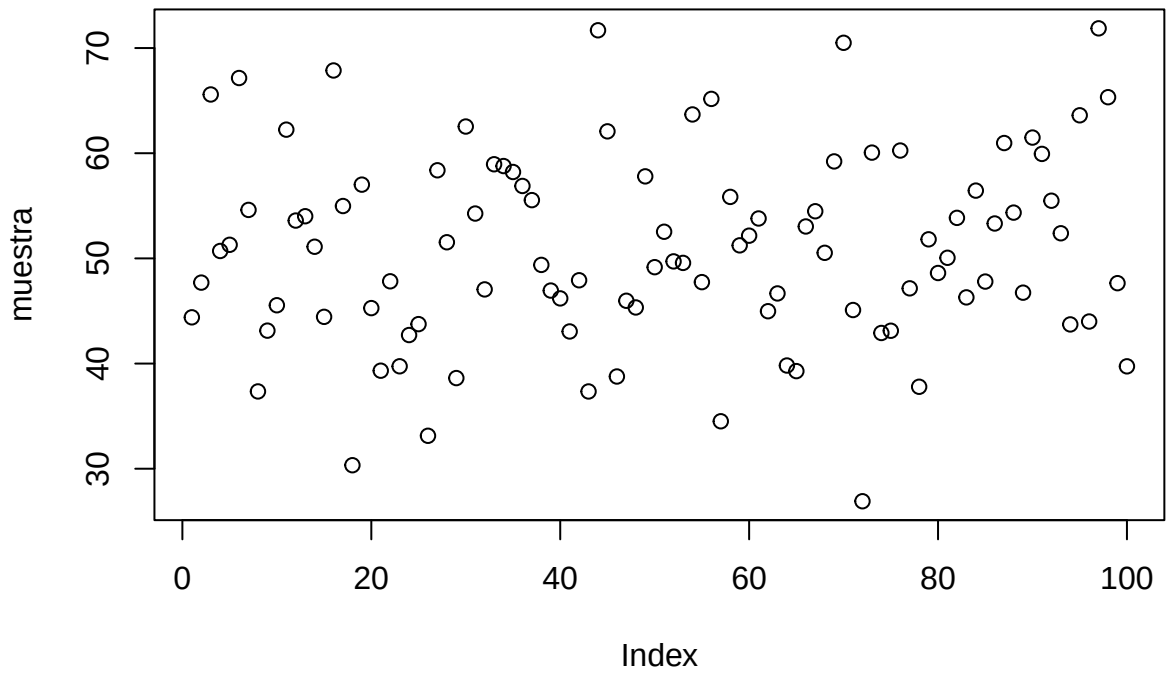
### Parte 1 - generar muestras de datos con distribución normal

Genere una muestra con 100 datos, y que tenga una media de X y desvio estandar Z.. X y Z son valores que debe elegir

```
# Establecer la semilla para reproducibilidad
set.seed(123)

# Definir la media y el desvío estándar deseados
X <- 50
Z <- 10

# Generar una muestra de 100 datos con distribución normal
muestra <- rnorm(n = 100, mean = X, sd = Z)

# Imprimir los primeros 10 elementos de la muestra
head(muestra, 10)
```

```
##  [1] 44.39524 47.69823 65.58708 50.70508 51.29288 67.15065 54.60916 37.34939
##  [9] 43.13147 45.54338
```

```
plot(muestra)
```




**Detalle del script anterior:**
Este código generará una muestra de 100 números aleatorios que siguen una distribución normal con media 50 y desviación estándar 10. La función `rnorm()` se utiliza para generar números aleatorios con distribución normal, donde `n` especifica el número de observaciones, `mean` es la media deseada y `sd` es el desvío estándar deseado.

La línea `set.seed(123)` establece una semilla para el generador de números aleatorios, lo que asegura que los resultados sean reproducibles.

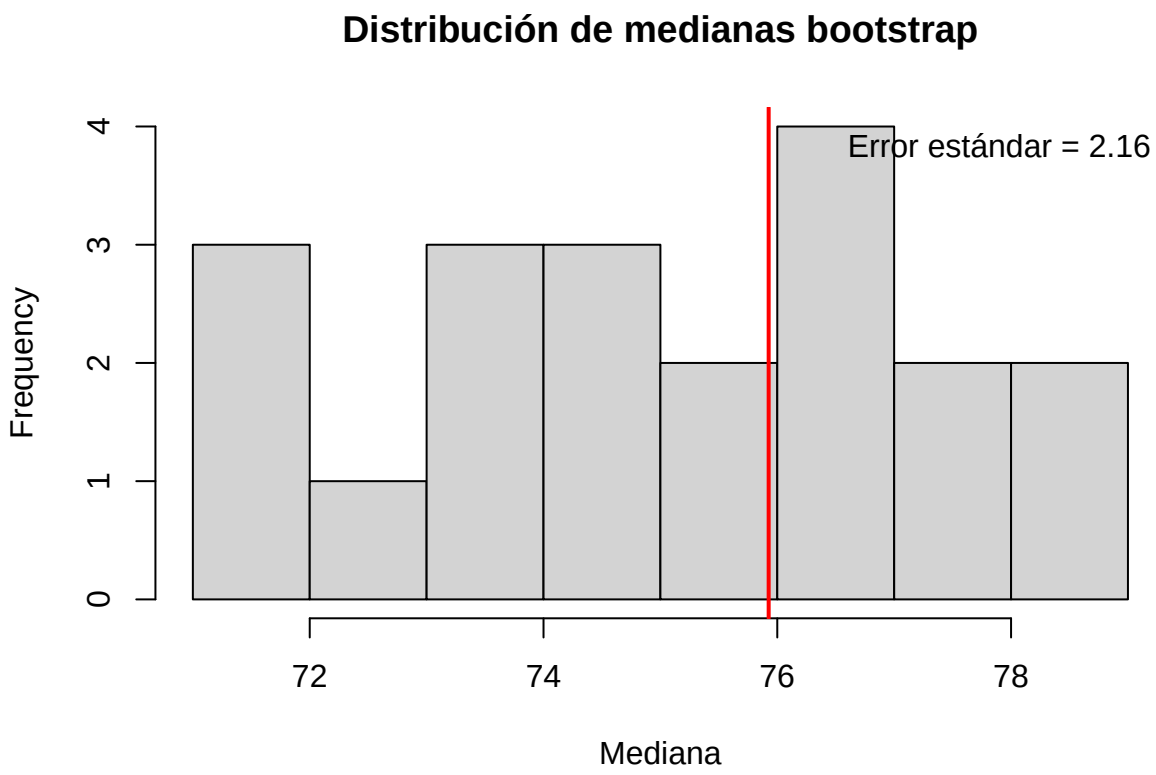
Finalmente, `head(muestra, 10)` imprime los primeros 10 elementos de la muestra generada.

Nota: Los valores exactos de la muestra generada variarán cada vez que se ejecute el código debido a la naturaleza aleatoria del proceso, pero la media y el desvío estándar de la muestra completa se acercarán a los valores especificados (50 y 10, respectivamente).

**Parte 2 - integramos el paso anterior con el resto del procedimiento solicitado**

```
# Paso 1: Generar una muestra con 100 datos
set.seed(123) #
X <- 75
Z <- 15
muestra <- rnorm(100, mean = X, sd = Z) #rnorm genera muestras aleatorias con distribución normal, y no

# Paso 2: Obtener 20 muestras bootstrap
n_bootstrap <- 20
muestras_bootstrap <- lapply(1:n_bootstrap, function(i) sample(muestra, replace = TRUE)) # lapply gener
```

```r
# Paso 3: Calcular la mediana para cada muestra bootstrap
medianas_bootstrap <- sapply(muestras_bootstrap, median) # sapply es función vectorizada que permite it

# Paso 4: Calcular los errores estándar de la distribución de medianas y graficar
error_estandar <- sd(medianas_bootstrap) #sd nos devuelve el error estandar de los valores guardados en
hist(medianas_bootstrap, main = "Distribución de medianas bootstrap", xlab = "Mediana")
abline(v = median(muestra), col = "red", lwd = 2)
legend("topright", legend = paste("Error estándar =", round(error_estandar, 2)), bty = "n")
```




```r
# Se observa que las medianas de las muestras bootstrap se distribuyen alrededor de la mediana de la mu

# Paso 5: Generar una función para calcular los errores estándar por bootstrap
# La función "calcular_error_estandar_bootstrap" devuelve el error estandar a partir de la entrada de u
calcular_error_estandar_bootstrap <- function(datos, n_bootstrap) {
 muestras_bootstrap <- lapply(1:n_bootstrap, function(i) sample(datos, replace = TRUE))
 medianas_bootstrap <- sapply(muestras_bootstrap, median)
 error_estandar <- sd(medianas_bootstrap)
 return(error_estandar)
}

# Calcular el error estándar con la función
error_estandar_bootstrap <- calcular_error_estandar_bootstrap(muestra, n_bootstrap)
print(error_estandar_bootstrap) #imprime el error estandard del boostrap
```

```
## [1] 2.371709
```

**Explicación del ejercicio.**

1. Se genera una muestra de 100 datos con media 75 y desviación estándar 15 utilizando `rnorm`.
2. Se obtienen 20 muestras bootstrap a partir de la muestra original utilizando `sample` con `replace = TRUE`.
3. Se calcula la mediana para cada muestra bootstrap utilizando `sapply` y `median`.
4. Se calcula el error estándar de las medianas bootstrap utilizando `sd` y se grafica un histograma de las medianas junto con la mediana de la muestra original y el error estándar.

   Se observa que las medianas de las muestras bootstrap se distribuyen alrededor de la mediana de la muestra original.
   El error estándar proporciona una estimación de la variabilidad de la mediana.
5. Se crea una función `calcular_error_estandar_bootstrap` que toma los datos originales y el número de muestras bootstrap, y devuelve el error estándar de las medianas bootstrap.
6. Se utiliza la función para calcular el error estándar de las medianas bootstrap y se imprime el resultado.

La salida del código mostrará el histograma de las medianas bootstrap y el error estándar calculado. También se imprimirá el error estándar calculado por la función `calcular_error_estandar_bootstrap`.

Observación: Las medianas de las muestras bootstrap se distribuyen alrededor de la mediana de la muestra original, y el error estándar proporciona una estimación de la variabilidad de la mediana.