Tarea 1

Christian Badillo, Luis Nuñez, Luz Maria Santana, & Sealtiel Pichardo

Tabla de contenidos

Listado de Figuras

Listado de Tablas

1 Simulación de Intervalos de Confianza

Cargamos las librerías necesarias para la simulación y visualizaciones.

```
library(ggplot2)
library(knitr)
```

Importamos los datos.

```
edad.data <- read.csv("edad.csv")

set.seed(1)
```

Veamos las primeras observaciones.

```
knitr::kable(head(edad.data), caption = "Datos.")
```

Tabla 1: Datos.

id	edad
1	26
2	24
3	25
4	27
5	25
6	25

Vamos a definir una función que simule la estimación de la media y el intervalo de confianza a un nivel de confianza $100 * (1 - \alpha)\%$ especificado.

```
ic_estimation <- function(data = NULL, n=NULL, iter=1000, alpha = 0.05, seed = 1)
  {
    if(is.null(data))
    {
      stop("No se ha proporcionado un conjunto de datos.")
    }

    N <- nrow(data)

    if(is.null(n)){n <- N*0.1}

    mean.est <- numeric(iter)
    var.est <- numeric(iter)
    lower.ci <- numeric(iter)
```

```
    upper.ci <- numeric(iter)

    set.seed(seed)

    for(i in 1:iter){
      sample <- sample(data$edad, n)
      mean.est[i] <- mean(sample)
      var.est[i] <- (1 - n/N) * (var(sample)/n)
      lower.ci[i] <- mean.est[i] - qnorm(1 - alpha/2) * sqrt(var.est[i])
      upper.ci[i] <- mean.est[i] + qnorm(1 - alpha/2) * sqrt(var.est[i])
    }
    return(data.frame(mean.est, lower.ci, upper.ci))
}
```

También vamos a crear una función que grafique los intervalos de confianza y el valor real de la media.

```
plot_ic <- function(df=NULL, mean.est.col = "green", ci_color = "blue", real_col = "red")
{
    if(is.null(data))
        {
          stop("No se ha proporcionado un conjunto de datos.")
    }

    ggplot(df, aes(x = 1:iter, y = mean.est)) +
      geom_point(colour = mean.est.col) +
      geom_errorbar(aes(ymin = lower.ci, ymax = upper.ci), width = 0.1,
                    colour = ci_color) +
      geom_hline(yintercept = mean(edad.data$edad), color = real_col) +
      labs(title = "Intervalos de Confianza para la Media de la Edad",
           x = "Iteración",
           y = "Media") +
      theme_minimal()
}
```

1.1 Estimación de la Media de la Edad.

Estimamos la media de la edad y su intervalo de confianza para distintos tamaños de muestra.

1.1.1 Tamaño de Muestra 20.

Simulamos 100 muestras de tamaño 20 y estimamos la media y el intervalo de confianza al 95%.

```r
N <- nrow(edad.data)
n <- 20
iter <- 100
seed <- 1

df <- ic_estimation(data = edad.data, n = n, iter = iter, seed=seed)
```

Gráficamos los intervalos de confianza y el valor real de la media.

```r
colors <- c("#1E1E1E", "#4D6291", "#9C0824")

plot_ic(df, colors[1], colors[2], colors[3])
```

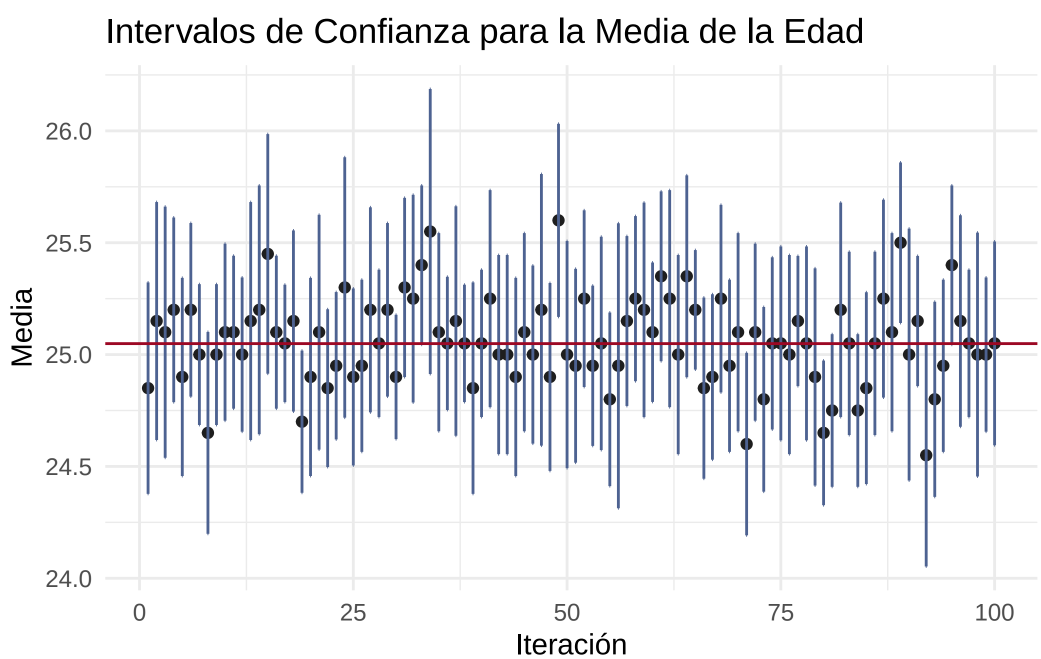

Figura 1: IC al 95% para la media de la edad con tamaño de muestra 20.

El porcentaje de los intervalos de confianza que contienen el valor real de la media es: 94%.

1.1.2 Tamaño de Muestra 50.

Simulamos 100 muestras de tamaño 50 y estimamos la media y el intervalo de confianza al 95%.

```r
n <- 50
seed <- 2
iter <- 100

df2 <- ic_estimation(data = edad.data, n = n, iter = iter, seed = seed)
```

Graficamos los intervalos de confianza y el valor real de la media.

```
colors <- c("#1E1E1E", "#4D6291", "#9C0824")

plot_ic(df2, colors[1], colors[2], colors[3])
```

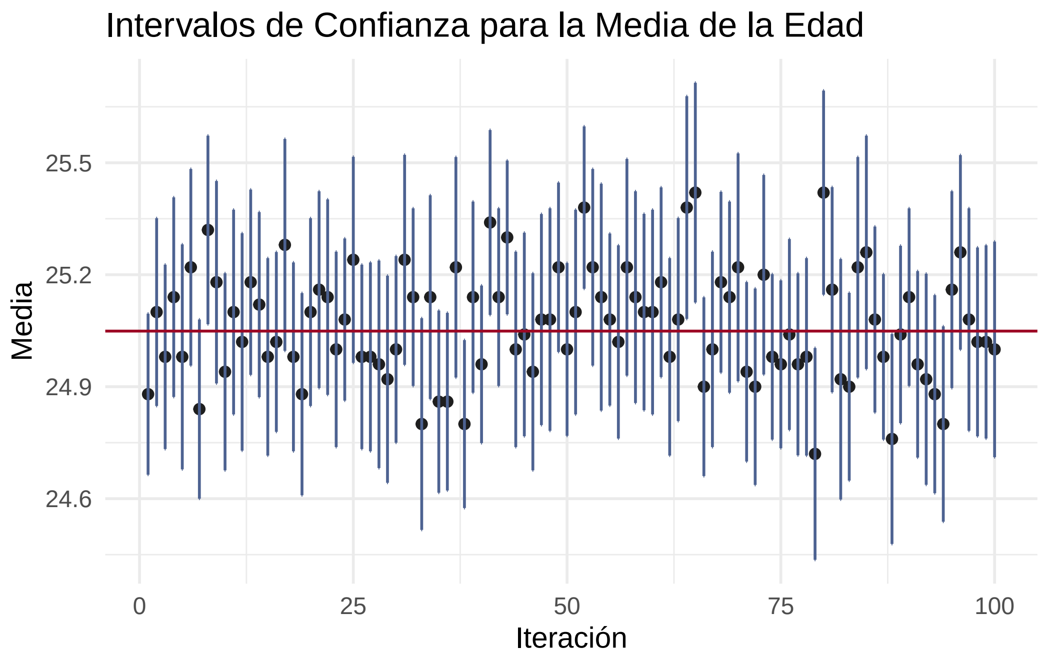

Figura 2: IC al 95% para la media de la edad con tamaño de muestra 50.

El porcentaje de los intervalos de confianza que contienen el valor real de la media es: 90%.

1.1.3 Tamaño de Muestra 50, IC al 90%.

Simulamos 100 muestras de tamaño 50 y estimamos la media y el intervalo de confianza al 90%.

```
n <- 50
seed <- 2
iter <- 100
alpha <- 0.1

df3 <- ic_estimation(data = edad.data, n = n, iter = iter, alpha = alpha, seed = seed)
```

Graficamos los intervalos de confianza y el valor real de la media.

```
colors <- c("#1E1E1E", "#4D6291", "#9C0824")
plot_ic(df3, colors[1], colors[2], colors[3])
```

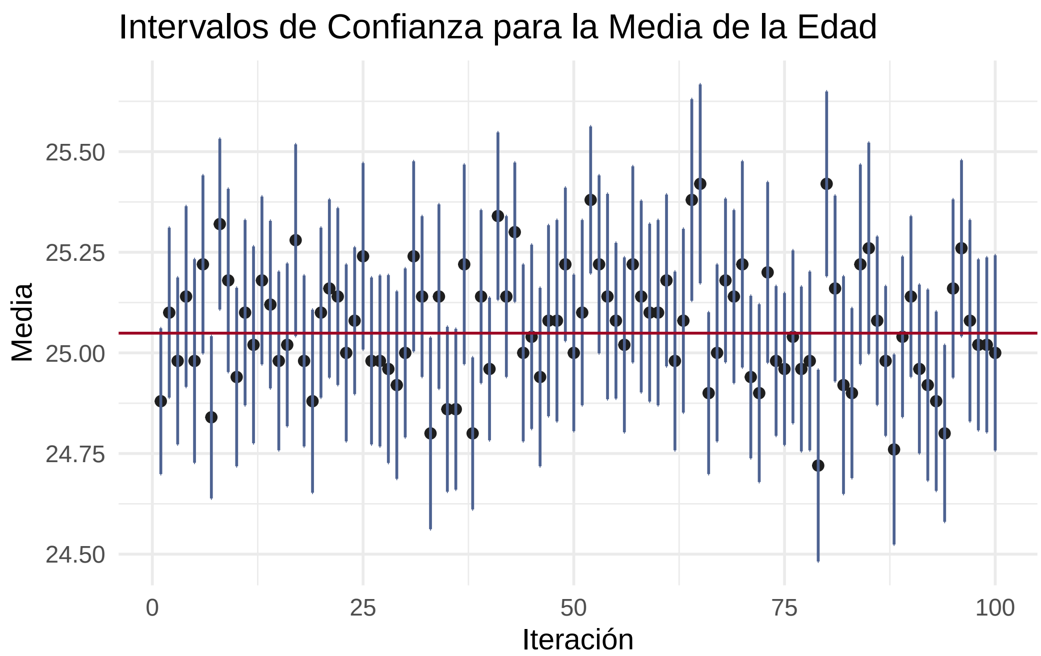

Figura 3: IC al 90% para la media de la edad con tamaño de muestra 50.

El porcentaje de los intervalos de confianza que contienen el valor real de la media es: 87%.

1.1.4 Tamaño de Muestra 50, IC al 99%.

Simulamos 100 muestras de tamaño 50 y estimamos la media y el intervalo de confianza al 99%.

```
n <- 50
seed <- 2
iter <- 100
alpha <- 0.01

df4 <- ic_estimation(data = edad.data, n = n, iter = iter, alpha = alpha, seed = seed)
```

Graficamos los intervalos de confianza y el valor real de la media.

```
colors <- c("#1E1E1E", "#4D6291", "#9C0824")
plot_ic(df4, colors[1], colors[2], colors[3])
```

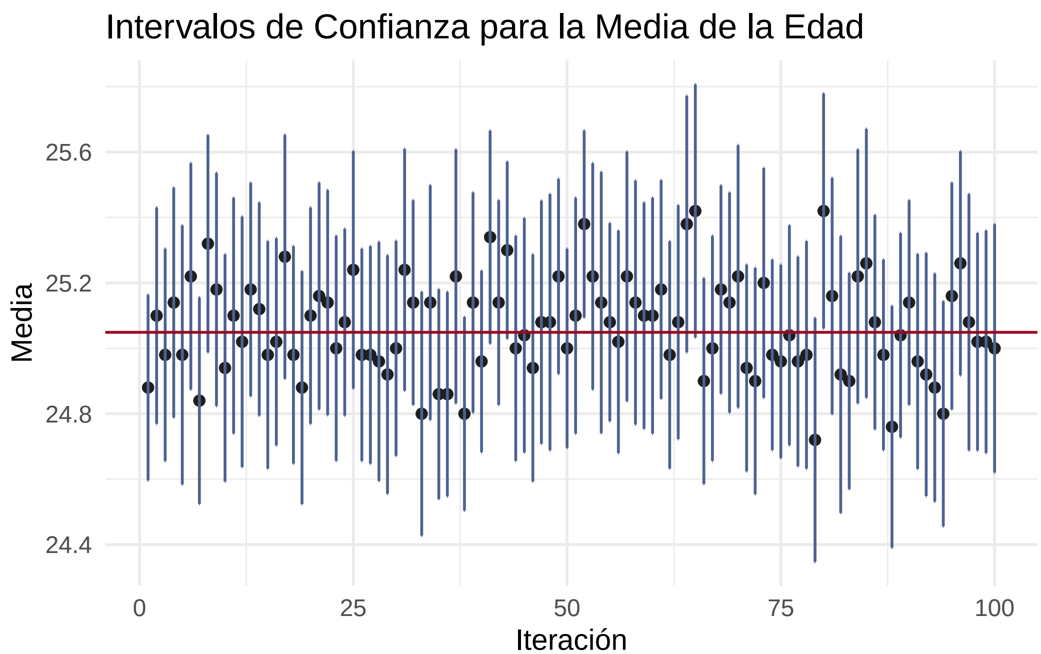

Figura 4: IC al 99% para la media de la edad con tamaño de muestra 50.

El porcentaje de los intervalos de confianza que contienen el valor real de la media es: 98%.

1.1.5 Tamaño de Muestra 50, IC al 87%.

Simulamos 100 muestras de tamaño 50 y estimamos la media y el intervalo de confianza al 87%.

```
n <- 50
seed <- 2
iter <- 100
alpha <- 0.13

df5 <- ic_estimation(data = edad.data, n = n, iter = iter, alpha = alpha, seed = seed)
```

Gráficamos los intervalos de confianza y el valor real de la media.

```
colors <- c("#1E1E1E", "#4D6291", "#9C0824")
plot_ic(df5, colors[1], colors[2], colors[3])
```

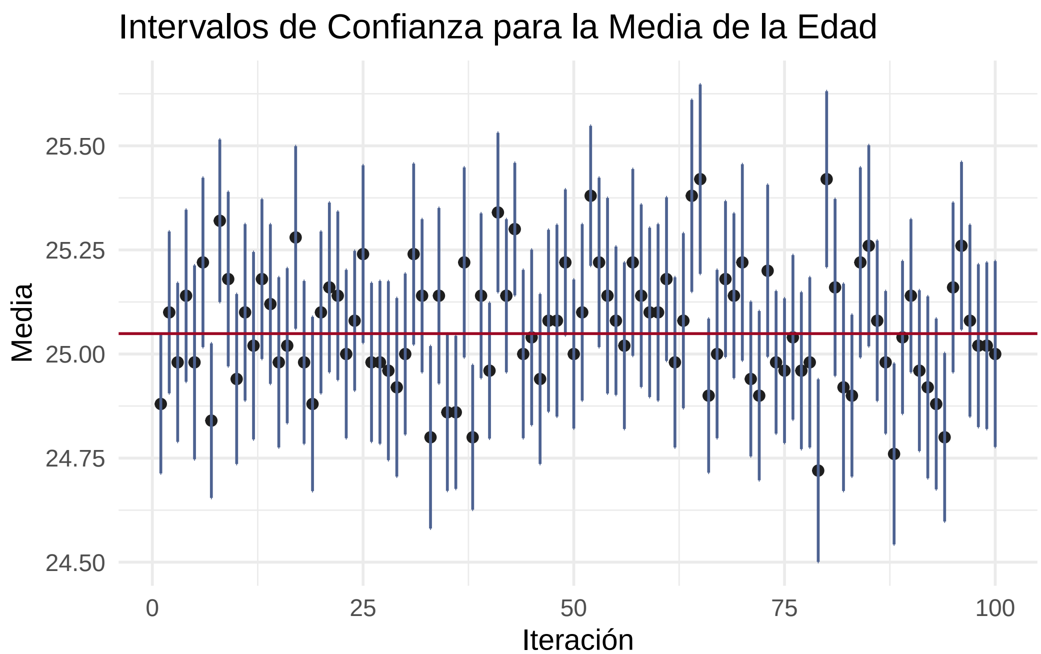

Figura 5: IC al 87% para la media de la edad con tamaño de muestra 50.

El porcentaje de los intervalos de confianza que contienen el valor real de la media es: 82%.

1.2 Conclusiones.

Los intervalos de confianza se pueden definir como el rango de valores en el que se espera que se encuentre el valor verdadero de un parámetro en un porcentaje de $(1-\alpha)*100$ veces de las que se repite la estimación del parámetro con muestras distintas de tamaño n. Como se observó en la Figura 1, los intervalos de confianza al 95% para la media de la edad con tamaño de muestra 20 son más amplios que con un tamaño de 50, esto se debe a que la varianza de la media estimada $\widehat{V}(\hat{\bar{y}}_{\text{Edad}})$, es mayor con un tamaño de muestra menor dado que conforme $n \to N$, $\widehat{V}(\hat{\bar{y}}_{\text{Edad}}) \to 0$.

Al aumentar el tamaño de muestra, la varianza de la media estimada disminuye y por lo tanto el intervalo de confianza es menor lo que conlleva a una estimación más precisa, como lo muestra la Figura 2.

Al variar el nivel de confianza, se observa que al disminuir el nivel de confianza, el intervalo de confianza es menor, como se observa en la Figura 3 y Figura 5. Por otro lado, al aumentar el nivel de confianza, el intervalo de confianza es mayor, como se observa en la Figura 4, dado que el cuantil de la distribución normal es mayor y por ende el intervalo de confianza es mayor a un nivel fijo de n.

	Semilla	Tamaño	% de confianza	# de intervalos
1	1	20	95%	94
2	2	50	95%	90
3	2	50	90%	87
4	2	50	99%	98
5	2	50	87%	82

1.3 Estimación de la Proporción de Familias

Generamos los datos

```
familias.data <- cbind(c(rep(1, 468), rep(0, 117)))

colnames(familias.data)[1]<- "Baño Interior"  #1 Sí; 0 No
```

Veamos las primeras observaciones

```
knitr::kable(head(familias.data), caption = "Datos")
```

Tabla 3: Datos

Baño Interior
1
1
1
1
1
1

1. Con una muestra de 585 familias, de una población de 29,661, estimamos el porcentaje de familias que cuentan con un baño interior de uso exclusivo a un intervalo de confianza del 95%

```
N <- 29661
n <- length(familias.data)
alpha <- 0.05

p_mean.est <- mean(familias.data)
p_var.est <- (1-n/N)*(p_mean.est*((1-p_mean.est))/(n-1))
p_lower.ci <- p_mean.est - qnorm(1 - alpha/2) * sqrt(p_var.est)
p_upper.ci <- p_mean.est + qnorm(1 - alpha/2) * sqrt(p_var.est)
```

```
p_mean.est
```

[1] 0.8

```
p_lower.ci;p_upper.ci
```

[1] 0.76788

[1] 0.83212

La estimación puntual del porcentaje de familias en el área de la ciudad que tienen un baño interior en casa es del 80% y el intervalo de confianza al 95% es de [76.788, 83.212].

1.3.1 Tamaño de Muestra

2. Estimamos el tamaño de muestra que necesitaríamos si obtuvimos una proporción de 0.75 en un estudio piloto, queremos una confianza al 95% y una precisión de 0.03

```
p_piloto <- 0.75
alpha <- 0.05
sigma_precision <- 0.03

n0 <- (qnorm(1 - alpha/2)^2)*(p_piloto)*(1-p_piloto)/(sigma_precision^2)
```

```
n0
```

[1] 800.3039

Ajustamos el tamaño de muestra al de la población:

```
n_est <- n0/(1+ n0/N)
```

```
n_est
```

[1] 779.2777

El tamaño de muestra estimado y ajustado es de aproximadamente 779. Notamos que es mayor al del inciso anterior(585), esto nos habla que este estudio tendrá un precisión mayor.

1.4 Estimación del Total de Mercancía

Para reducir el trabajo de inventario de existencias en una bodega donde existen 226 estantes, se tomó una muestra aleatoria simple de 26 estantes en los cuales se hizo un conteo del valor ($) de los artículos en cada uno de los estantes.

Generamos los datos

```
df_estantes <- data.frame(valor = c(40, 42, 44, 45, 49, 51, 54, 53, 57, 58, 60, 61, 61, 61, 62, 61, 63, 63, 66
```

1. Estimamos el valor total de la mercancía que se encuentra en la bodega y calculamos un intervalo de 95% de confianza.

```
N <- 226
n <- length(df_estantes$valor)
alpha <- 0.05

t_mean.est <- (N)* mean(df_estantes$valor)

t_var.est <- N^2*(1- n/N)*(var(df_estantes$valor)/n)
t_lower.ci <- t_mean.est - qnorm(1 - alpha/2) * sqrt(t_var.est)
t_upper.ci <- t_mean.est + qnorm(1 - alpha/2) * sqrt(t_var.est)

t_mean.est
```

```
[1] 13699.08
```

```
t_lower.ci; t_upper.ci
```

```
[1] 12738.22
```

```
[1] 14659.93
```

La estimación puntual del valor total de la mercancía en la bodega es de $13,699.08 y el intervalo de 95% de confianza donde se encuantra el valor total de la bodega es [$12,738.22, $14,659.93].Podemos asegurar (con el 95%) que la bodega tiene un valor mínimo de $12,000 aproximadamente y un valor máximo aproximado de $15,000.

1.4.1 Tamaño de Muestra

Estimamos el tamaño de muestra que necesitaríamos si quisieramos estimar el valor total de la mercancía con un nivel de confianza del 95%, una precisión de $750 y sabiendo que la S es aproximadamente $10.35.

```
alpha <- 0.05
t_precision <- 750
t_var.est <- 10.35

n0 <- (qnorm(1 - alpha/2)^2) * N^2 * t_var.est^2 / t_precision^2

n0
```

[1] 37.36554

Ajustamos el tamaño de muestra al de la población:

```
t_n <- n0/(1+ n0/N)

t_n
```

[1] 32.06422

El tamaño de muestra estimado y ajustado es de aproximadamente 32.Este tamaño de muestra nos daría con el 95% el valor total de la bodega, notemos que sería muy cercana al valor real debido a que al buscar una precision de $750, comparada con la media estimada, es mínima la variación que buscamos; de ahí que la muestra sea mayor al inciso anterior.