Tarea 1

Christian Francisco Badillo Hernández

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 gráfique 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%.

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

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

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

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

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

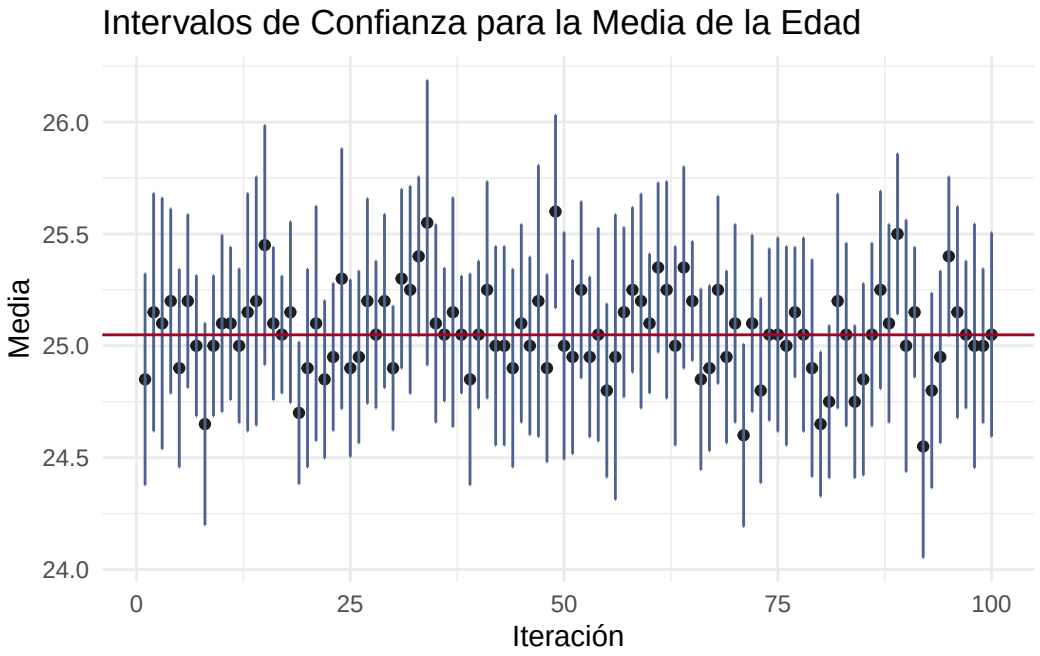

Figura 1: IC al 95% para la media de la edad con tamaño de muestra 20.

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%.

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

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

Gráficamos 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])
```

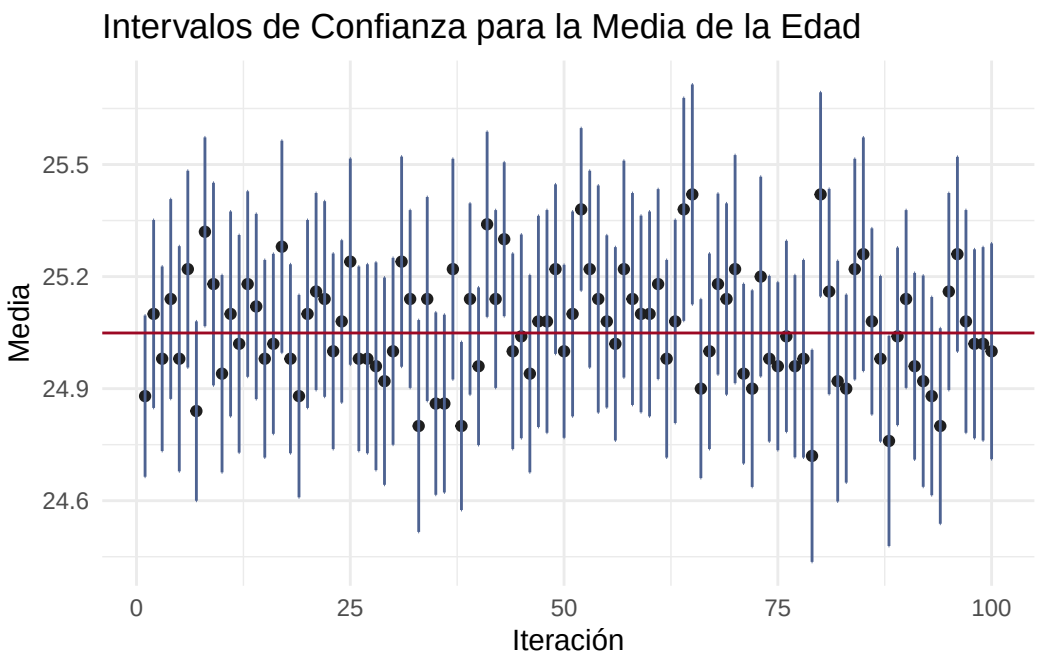

Figura 2: IC al 95% para la media de la edad con tamaño de muestra 50.

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)
```

Gráficamos 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])
```

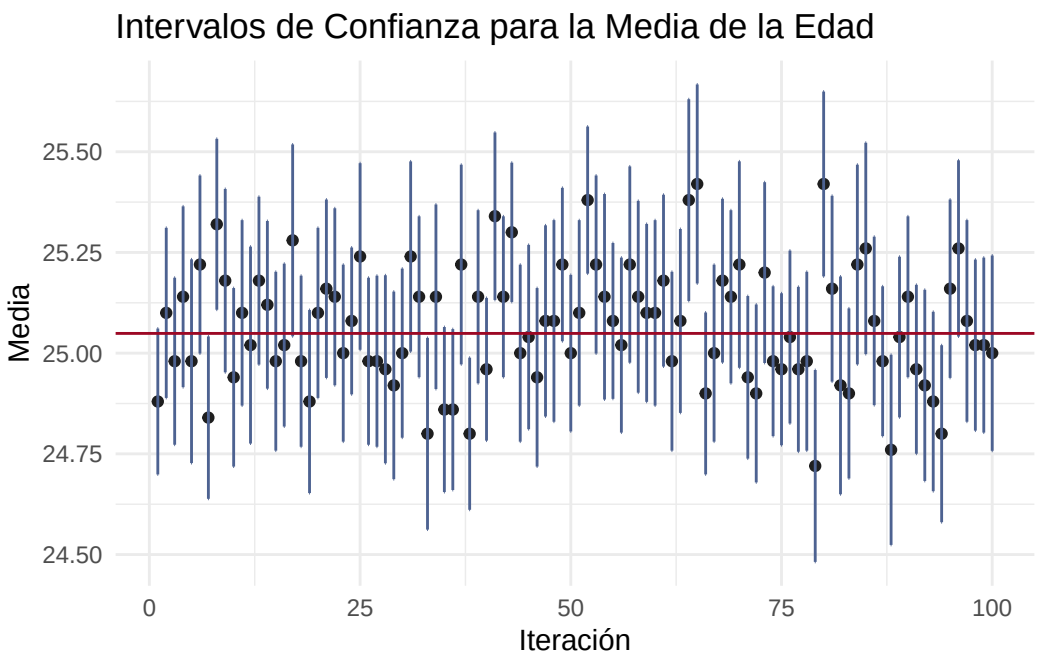

Figura 3: IC al 90% para la media de la edad con tamaño de muestra 50.

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)
```

Gráficamos 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])
```

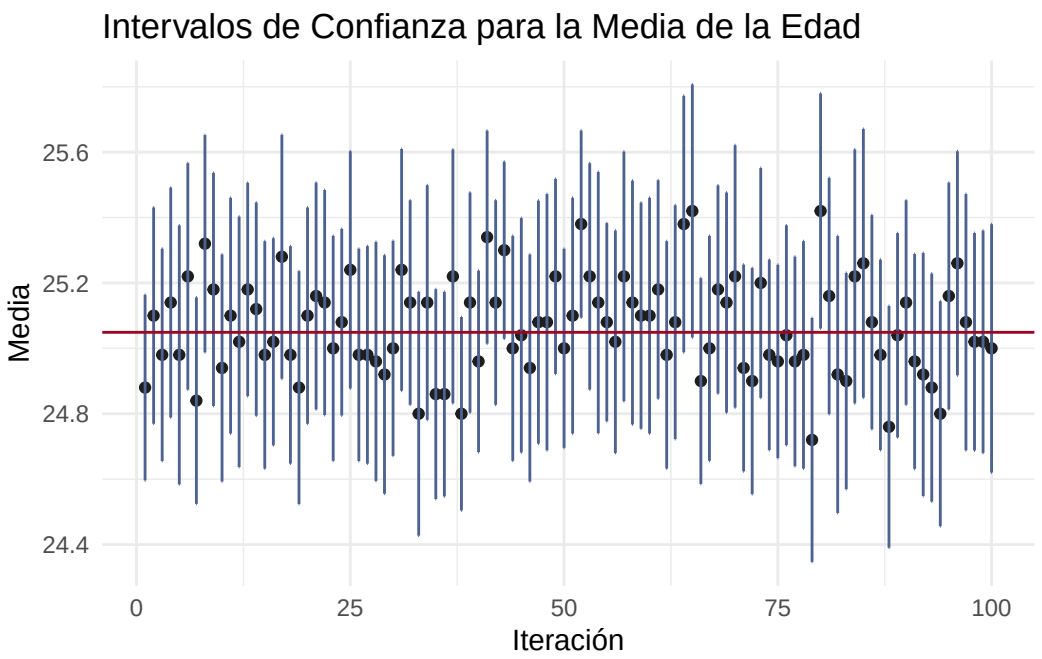

Figura 4: IC al 99% para la media de la edad con tamaño de muestra 50.

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])
```

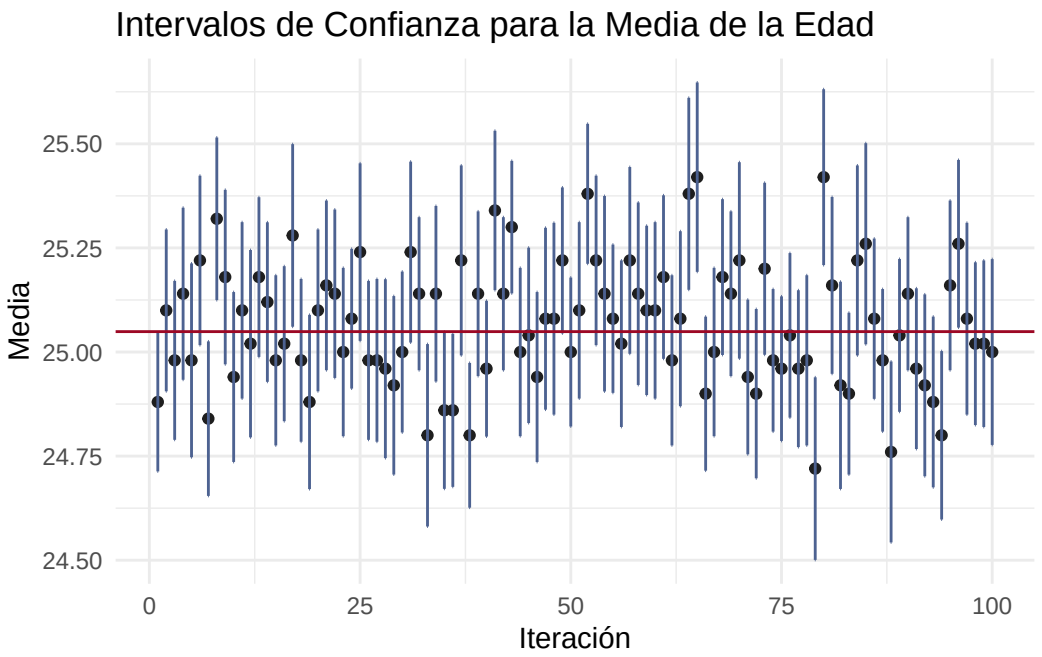

Figura 5: IC al 87% para la media de la edad con tamaño de muestra 50.

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 observo 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 $\hat{V}(\hat{\bar{y}}_{\text{Edad}})$, es mayor con un tamaño de muestra menor dado que conforme $n \to N$, $\hat{V}(\hat{\bar{y}}_{\text{Edad}}) \to 0$.

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.

2 Estimación de proporción.

1. Problema: Se eligió una muestra aleatoria simple de 585 familias de un área de la ciudad conteniendo 29661 familias. A cada familia se le preguntó si tenían el uso exclusivo de un baño interior en su casa. Los resultados fueron los siguientes:

Tabla 2: Uso exclusivo de baño interior en familias encuestadas.

	Sí	No	Total
Uso de un baño	468	117	585

- A). Se estima el porcentaje de familias que tienen el uso exclusivo de un baño interior en su casa.

```
N <- 29661
n <- 585

datos.prop <- data.frame(toilet = c("Sí", "No"), frecuencia = c(468, 117))

prop.est <- datos.prop$frecuencia[1] / sum(datos.prop$frecuencia)

finitud <- (1 - n/N)
var.prop.est <- (finitud * 1/n * N/(N-1)) * prop.est * (1 - prop.est)
ci <- prop.est + c(-1, 1) * qnorm(1 - 0.05/2) * sqrt(var.prop.est)
```

La estimación del porcentaje de familias que tienen el uso exclusivo de un baño interior en su casa es del 80% con un intervalo de confianza al 95% de (76.79, 83.21).

- B). EL tamaño de muestra necesario para una prueba piloto con una precisión del 3% y un nivel de confianza del 95%, sabiendo que la proporción de familias que tienen el uso exclusivo de un baño interior en su casa es del 75% en el área de interés de la ciudad, estimamos el tamaño de muestra necesario.

```
p.est <- 0.75
accuracy <- 0.03
confidence <- qnorm(1 - 0.05/2)

## Dado que la N se puede considerar como grande se usa n0
n0 <- (confidence ^2 * p.est * (1 - p.est)) / accuracy ^ 2
n0 <- ceiling(n0) # Función techo
```

Dado que el tamaño de la población es grande, se puede usar la fórmula para el tamaño de muestra sin corregir, por lo que el tamaño de muestra necesario es de 801.