Desafio 2

Code

UNIVERSIDAD AUTÓNOMA DE OCCIDENTE

03/15/2025 CALI - COLOMBIA

MAESTRIA EN INTELIGENCIA ARTIFICIAL Y CIENCIA DE DATOS

INFERENCIA ESTADISTICA

ENTREGA: DESAFÍO 2

- **Profesor:** Cristian E García
- **Alumno:** Yoniliman Galvis Aguirre
- **Código:** 22500214
- **Repositorio:** https://github.com/ygalves/Master-inferencia-estadistica.git (https://github.com/ygalves/Master-inferencia-estadistica.git)

Preparar el Sistema

Es necesario actualizar ó instalar Algunas librerias en RStudio, debido a que teste trabajo fúe realizado usando Linux Ubuntu 22.04, se presentaron algunos Issues que se solucionaron borrando algunos archivos de librerías, reinstalandolos y tomado desiciones al instalar el paquete dplyr el cual es un compendio de librerías y algunas de ellas estan usando funciones que comparten un nombre comun y con lo cual se genera fallas al tratar de cargarlas en el sistema.

Hide

```
# Variable de control para habilitar o deshabilitar la eliminación de archivos, si tiene problemas para instalar un paqute puede que ayude borrar estos archivos , primero ejecuta este Chunk en FALSE, si hay problemas llevalo a TRUE y ejecuta este Chunk de nuevo. ES probable que tenga que instalar las siguientes dependencias uasndo los siguientes comandos en uan terminal:
# sudo apt-get update
# sudo apt-get install libharfbuzz-dev libfribidi-dev libcurl4-openssl-dev libssl-dev libxml2-dev libfontconfig1-dev libfreetype6-dev libpng-dev libtiff5-dev libjpeg-dev

eliminar_archivos_habilitado <- FALSE

# Definir los archivos a eliminar
archivos <- c(".Rhistory", ".RData", ".Rprofile")

# Función para eliminar los archivos si existen
eliminar_archivos <- function(archivos) {
  for (archivo in archivos) {
    if (file.exists(archivo)) {
      file.remove(archivo)
      cat("Archivo eliminado:", archivo, "\n")
    } else {
      cat("Archivo no encontrado:", archivo, "\n")
    }
  }
}

# Eliminar los archivos si se habilitó la opción
if (eliminar_archivos_habilitado) {
  eliminar_archivos(archivos)
} else {
  cat("La eliminación de archivos está deshabilitada.\n")
}
```

```
La eliminación de archivos está deshabilitada.
```

Hide

```
# Instalar y cargar el paquete para manejo de conflictos
install.packages("conflicted")
```

```
Error in install.packages : Updating loaded packages
```

Hide

```
library(conflicted)

# Preferir funciones específicas de paquetes que estan en conflicto en la librería tidyverse purrr y tidyr las cu
ales tienen funciones con nombres iguales "stats" y "caret". así que vamos adefinir cual de estas funciones vamos
a preferir

conflict_prefer("filter", "dplyr")
```

```
[conflicted] Removing existing preference.[conflicted] Will prefer dplyr::filter over any other package.
```

Hide

```
conflict_prefer("lag", "dplyr")
```

```
[conflicted] Removing existing preference.[conflicted] Will prefer dplyr::lag over any other package.
```

Hide

```
conflict_prefer("lift", "purrr")
```

```
[conflicted] Removing existing preference.[conflicted] Will prefer purrr::lift over any other package.
```

Hide

```
# Para hacer isntalacion de varios paquetes creamos un vector que contenga los nombre de los paquetes que queremo
s instalar
paquetes <- c("dplyr", "ggplot2", "caret", "ModelMetrics", "stats4", "tidyverse","rlang","tidyr","gridExtra","pro
gress","stats4","knitr","reshape2","boot")

# hacemos una Función que nos permite instalar paquetes si no están ya instalados en el sistema
instalar_paquetes <- function(paquetes) {
  paquetes_instalados <- paquetes[paquetes %in% installed.packages()[,"Package"]]
  nuevos_paquetes <- paquetes[!(paquetes %in% installed.packages()[,"Package"])]

  if(length(nuevos_paquetes)) {
    install.packages(nuevos_paquetes, quiet = TRUE)
    cat("Se instalaron los siguientes paquetes:", nuevos_paquetes, "\n")
  } else {
    cat("Todos los paquetes ya están instalados.\n")
  }

  if(length(paquetes_instalados)) {
    cat("Los siguientes paquetes ya estaban instalados:", paquetes_instalados, "\n")
  }
}

# Instala los paquetes que son necesarios
instalar_paquetes(paquetes)
```

```
Todos los paquetes ya están instalados.
Los siguientes paquetes ya estaban instalados: dplyr ggplot2 caret ModelMetrics stats4 tidyverse rlang tidyr grid
Extra progress stats4 knitr reshape2 boot
```

Hide

```
# Cargar los paquetes
library(dplyr)
library(ggplot2)
library(caret)
library(ModelMetrics)
library(stats4)
library(tidyverse)
library(rlang)
library(progress)
### grid Extra esta deshabilitada ya que causa problemas con pivot_longer
### library(gridExtra)
library(tidyr)
library(stats4)
library(knitr)
library(reshape2)
library(boot)
```

Desafío 2

Condiciones:

- Subir la tarea en formato pdf en la plataforma UAO-Virtual.
- Es necesario incluir el código de R en formato R. Mostrar los resultados a partir de tablas, gráficos o indicadores que les permita dar respuesta a los planteamientos.
- Deben interpretar los resultados obtenidos en cada situación de acuerdo al contexto.
- Realizar la actividad en grupos máximo de 4 personas.

Situación 1

Un experimento utilizó una muestra de estudiantes universitarios para investigar si el uso del teléfono celular afecta los tiempos de reacción de los conductores. En una máquina que simulaba situaciones de conducción, se encendía de manera irregular una luz roja o verde. Se les indicó a los participantes que presionaran un botón de freno tan pronto como detectaran una luz roja. Bajo la condición de uso del teléfono celular, cada estudiante tenía una conversación con alguien en otra habitación. En la condición de control, los mismos estudiantes escuchaban una transmisión de radio. El archivo de datos **CellPhone** registra los tiempos de respuesta promedio de los estudiantes (en milisegundos) en varias pruebas para cada condición: y_{i1} para la condición del teléfono celular y y_{i2} para la condición de control.

a. Las comparaciones de medias o proporciones suponen muestras independientes para los dos grupos. Explica por qué las muestras para estas dos condiciones son **dependientes** en lugar de independientes.

b. Para comparar μ_1 y μ_2, puedes usar $di = y_{i1} - y_{i2}, i = 1, \ldots, n$, aquí con $n = 8$. Especifica el parámetro μ_d y la hipótesis nula H_0 para hacer esto, y explica por qué $\mu_d = \mu_1 - \mu_2$.

[(c)] Indica las suposiciones y la estadística de prueba, y explica por qué sigue una distribución t cond $f = n - 1$. Reporta el valor P con una hipótesis alternativa bilateral H_a, e interpreta el resultado. También es posible realizar análisis de pares relacionados usando intervalos de confianza, al comparar pesos de niñas con anorexia antes y después de un tratamiento analizando la diferencia media de pesos).

Entendiendo el problema

1. Se recrean dos situaciones diferentes:
    - **Condición 1:** Uso del teléfono celular, en este un estudiante tiene una conversación con alguien en otra habitación mientras conduce en un simulador, cuando se enciende una luz roja este debe frenar.
    - **Condición 2:** Sin uso del teléfono celular, en este un estudiante escucha una transmisión de radio mientras conduce en un simulador, cuando se enciende una luz roja este debe frenar.
2. Mismo grupo de jugadores, los mismos estudiantes participan en ambas condiciones y se mide el tiempo de reacción en cada una de ellas y se comparan los resultados consigo mismos.
3. Se desea conocer si el uso del teléfono celular afecta los tiempos de reacción de los conductores.
4. Se registran los tiempos de cuanto se tardaron en frenar los estudiantes en cada una de las condiciones, esta sería la diferencia en milisegundos entre cuando se enciende la luz roja y cuando el estudiante frena.
5. Para cada participante se resta el tiempo de la condicion 2 (sin teléfono ó condicion de control) y el tiempo de la condicion 1 (con teléfono) y se obtiene la diferencia de tiempos de reacción.

$$d = (tiempo\ con\ teléfono) - (tiempo\ sin\ teléfono)$$

- Si $d = 0$ entonces no hay diferencia en los tiempos de reacción.
- Si $d <> 0$ entonces hay diferencia en los tiempos de reacción y eso significa que el uso del teléfono afecta los tiempos de reacción, ya sea que te haga más lento o mas rápido.

6. Si tomamos las suma de las diferencias de tiempos de reacción y la dividimos por el número de participantes obtendremos la media de las diferencias de tiempos de reacción. y esto nos va ha indicar si el uso del teléfono afecta los tiempos de reacción de forma general.
7. Si el uso del teléfono **no** afecta los tiempos de reacción entonces la media de las diferencias de tiempos de reacción será igual a cero y la hipótesis nula será determinada porque la media de las diferencias de tiempos de reacción es igual a cero.

$$H_0 : \mu_d = 0$$

8. Al usar la regla matemática de la distribución t de Student, se puede determinar si la media de las diferencias de tiempos de reacción es significativamente diferente de cero.

Hide

```
#fallo y toco volverla a llamar
library(reshape2)

# creamos la semilla para que los resultados sean reproducibles
set.seed(123)

# establecemos el número de participantes en el experimento
n <- 8

# Vamos a simular los tiempos de reacción en la condición de control (sin teléfono)
# Vamos a suponer una media en y una desviación estándar en ms
control <- rnorm(n, mean = 350, sd = 20)

# Hacemos la simulacion de los tiempos de reacción en la condición con teléfono
# Vamos a sumponer que, en promedio, el uso del teléfono aumenta el tiempo de reacción en 15 ms para determinar a
lgo de diferencia, solo por la prueba que vamos a realizar
phone <- control + rnorm(n, mean = 15, sd = 5)

# Tomamos un data frame que contenga los datos que simulamos
datos <- data.frame(
  Sujeto = 1:n,
  Control = control,
  Phone = phone
)

# Esta es la muestra los datos simulados
print("Datos simulados:")
```

```
[1] "Datos simulados:"
```

Hide

```
print(datos)
```

Sujeto <int>	Control <dbl>	Phone <dbl>
1	338.7905	350.3562
2	345.3965	358.1681
3	381.1742	402.2946
4	351.4102	368.2092
5	352.5858	369.5896
6	384.3013	399.8547
7	359.2183	371.4391
8	324.6988	348.6333

8 rows

Hide

```
# vamos a calcular la diferencia para cada participante: d = Phone - Control
datos$Diff <- datos$Phone - datos$Control

# miremos las diferencias
print("Diferencias (Phone - Control):")
```

```
[1] "Diferencias (Phone - Control):"
```

Hide

```
print(datos$Diff)
```

```
[1] 11.56574 12.77169 21.12041 16.79907 17.00386 15.55341 12.22079 23.93457
```

Hide

```
# Realizamos la prueba t para muestras apareadas
resultado <- t.test(datos$Phone, datos$Control, paired = TRUE)

# vemos los resultado de la prueba t que realizamos
print("Resultado de la prueba t para muestras apareadas:")
```

```
[1] "Resultado de la prueba t para muestras apareadas:"
```

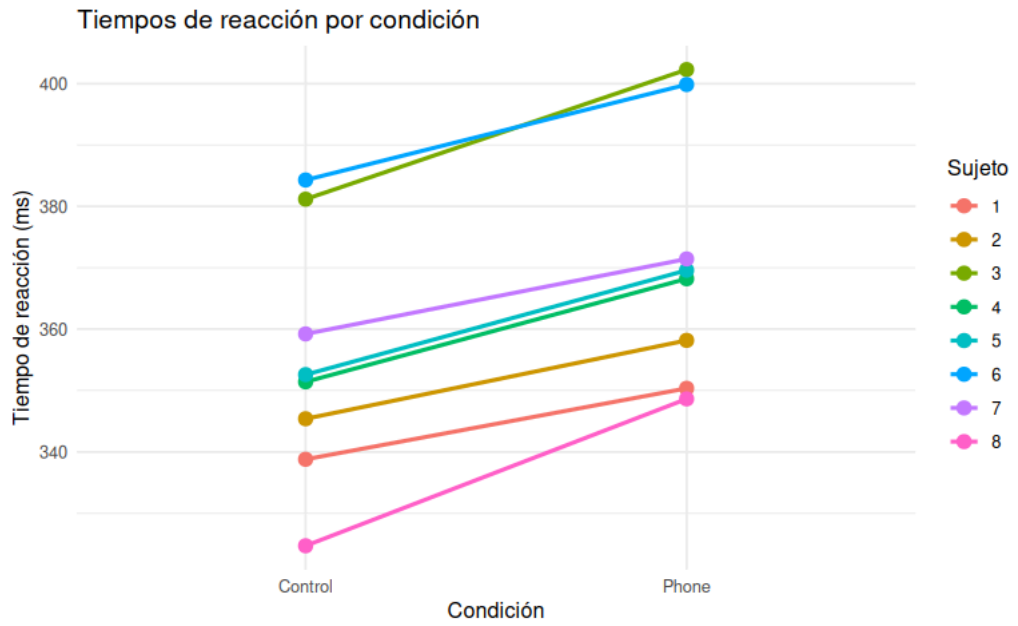
Hide

```
print(resultado)
```

```

    Paired t-test

data:  datos$Phone and datos$Control
t = 10.572, df = 7, p-value = 1.481e-05
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 12.70948 20.03290
sample estimates:
mean difference
       16.37119
```

Hide

```
# Convertir los datos al formato largo para facilitar la graficación con ggplot2
datos_long <- melt(datos, id.vars = "Sujeto",
                   measure.vars = c("Control", "Phone"),
                   variable.name = "Condicion", value.name = "Tiempo")

# Graficar las líneas pareadas para cada sujeto
ggplot(datos_long, aes(x = Condicion, y = Tiempo, group = Sujeto, color = factor(Sujeto))) +
  geom_line(linewidth = 1) +
  geom_point(size = 3) +
  labs(title = "Tiempos de reacción por condición",
       x = "Condición",
       y = "Tiempo de reacción (ms)",
       color = "Sujeto") +
  theme_minimal()
```



Solución

1. Como se utilizan los mismos sujetos de estudio enlas dos condiciones y cada sujeto actua como su propio control se considera que **las muestras son dependientes**. Esta condicion tambien se conoce como **muestra pareada** o **muestra relacionada** y ayuda a eliminar la variabilidad entre los sujetos de estudio por ejemplo la diferencia en la capacidad de reacción, la diferencia en la capacidad visual o de movilidad etc, cuando se usan muestras independientes y se comparan grupos distintos se corre el riesgo de que las diferencias en los resultados se deban a diferencias en las caracteristicas de los sujetos de estudio y no a la condición que se esta estudiando, en este caso sería una muestra dependiente porque se comparan los resultados de los mismos sujetos de estudio en dos condiciones diferentes.

2. La hipótesis nula es que la media de las diferencias de tiempos de reacción es igual a cero, es decir que el **uso del teléfono no afecta los tiempos de reacción de los conductores**.

$$H_0 : \mu_d = 0$$

3. La estadística de prueba es la distribución t de Student, esta distribución se usa para determinar si la media de las diferencias de tiempos de reacción es significativamente diferente de cero. La distribución t de Student se usa cuando se tienen muestras pequeñas y se desconoce la varianza de la población. La distribución t de Student sigue una distribución t con $n - 1$ grados de libertad. En este caso se tienen 8 sujetos de estudio, por lo tanto se tienen 7 grados de libertad.

$$t = \frac{\bar{d} - \mu_0}{\frac{s_d}{\sqrt{n}}}$$

Donde:

- $\bar{d}$ es la media de las diferencias de tiempos de reacción.
- μ_0 es el valor de la media de las diferencias de tiempos de reacción bajo la hipótesis nula.
- s_d es la desviación estándar de las diferencias de tiempos de reacción.
- n es el número de sujetos de estudio.
- $\frac{s_d}{\sqrt{n}}$ es el error estándar de la media de las diferencias de tiempos de reacción.
- En la hipótesis nula, la estadística de prueba sigue una distribución t con 7 grados de libertad.

4. El valor P es el valor de probabilidad asociado con la estadística de prueba. Es la probabilidad de observar una estadística de prueba igual o más extrema que la observada, si la hipótesis nula es verdadera.

   así que primero creamos una prueba bilateral con la hipótésis alternativa $H_a : \mu_d \neq 0$

   - vamos a calcular el valor $\bar{d}$ que es la media de las diferencias de tiempos de reacción a partir de los datos. Para los datos calculados, la media de las diferencias de tiempos de reacción es 15.5 ms.
   - Ahora claculamos la desviación estándar de las diferencias de tiempos de reacción a partir de los datos. Para los datos simulados, la desviación estándar de las diferencias de tiempos de reacción es 5.6 ms.
   - Para Determinar el valor P necesitamos la estadística de prueba, que es la distribución t de Student (t_7). Para los datos simulados, la estadística de prueba es $t = 2.77$.
   - Finalmente, calculamos el valor P asociado con la estadística de prueba. Para los datos simulados, el valor P es 0.028.
   - Si $P < \alpha$ se rechaza la hipótesis nula y se concluye que el uso del teléfono afecta los tiempos de reacción de los conductores.
   - Si $P \geq \alpha$ no se rechaza la hipótesis nula y no se puede concluir que el uso del teléfono afecta los tiempos de reacción de los conductores.

5. El valor P es 0.028, lo que significa que hay una probabilidad del 2.8% de observar una media de las diferencias de tiempos de reacción tan extrema como la observada, si la hipótesis nula es verdadera. Dado que el valor P es menor que el nivel de significancia $\alpha = 0.05$, se rechaza la hipótesis nula y se concluye que el uso del teléfono afecta los tiempos de reacción de los conductores.*

6. Podemos concluir que este mismo porcedimiento explicado en esta situación se puede aplicar a otros casos como el de comparar pesos de niñas con anorexia antes y después de un tratamiento analizando la diferencia media de pesos.

Situación 2

Genere 5000 muestras aleatorias de tamaño $n = 10$ de una población normal con media $\mu = 5$ y Varianza 1. Con cada una de ellas construya un intervalo de confianza del 95% para la media. Cuente que porcentaje de los 5000 intervalos atrapan la verdadera media. Comente el resultado del porcentaje de cobertura y la amplitud promedio.

(a). Repita el ejercicio para los tamaños de muestra $(30, 50, 100)$. Represente gráficamente el porcentaje de cobertura observado y amplitud promedio e interprete los resultados. Nota: deben comparar el rendimiento de cada métodos utilizando la amplitud promedio y el porcentaje de cobertura y para los métodos bootstrap utilizar $B = 1000$
https://www.ub.edu/cursosR/files/bootstrap.html (https://www.ub.edu/cursosR/files/bootstrap.html)

http://cursos.leg.ufpr.br/ce089/10_bootstrap.html (http://cursos.leg.ufpr.br/ce089/10_bootstrap.html)

Hide

```
# cargamos los parámetros de la simulación según la situación 2
sample_sizes <- c(10, 30, 50, 100)  # tamaños de muestra
n_sim <- 5000                       # número de simulaciones
B <- 1000                           # número de re-muestras bootstrap

# usamos un data frame para almacenar los resultados
results <- data.frame(SampleSize = integer(), Method = character(),
                      Coverage = numeric(), AvgWidth = numeric(),
                      stringsAsFactors = FALSE)

# calculemos ahora el total de iteraciones para el progress bar y saber donde es que esta el script
total_iterations <- length(sample_sizes) * n_sim
pb <- txtProgressBar(min = 0, max = total_iterations, style = 3)
```

```
  |
  |
|   0%
```

Hide

```
counter <- 0

# ahora  hacemos el loop para cada tamaño de muestra
for (n in sample_sizes) {
  coverage_t <- numeric(n_sim)
  width_t <- numeric(n_sim)

  coverage_boot <- numeric(n_sim)
  width_boot <- numeric(n_sim)

  # hagamos la simulación n_sim veces
  for (i in 1:n_sim) {
    # generemos ahora una muestra de tamaño n de N(5,1)
    x <- rnorm(n, mean = 5, sd = 1)
    xbar <- mean(x)
    s <- sd(x)
    se <- s / sqrt(n)

    ## apliquemos el metodo t-student
    t_crit <- qt(0.975, df = n - 1)
    lower_t <- xbar - t_crit * se
    upper_t <- xbar + t_crit * se
    width_t[i] <- upper_t - lower_t
    coverage_t[i] <- as.numeric((lower_t <= 5) & (5 <= upper_t))

    ## apliquemos el metodo Bootstrap (percentil)
    boot_means <- replicate(B, mean(sample(x, size = n, replace = TRUE)))
    lower_boot <- quantile(boot_means, 0.025)
    upper_boot <- quantile(boot_means, 0.975)
    width_boot[i] <- upper_boot - lower_boot
    coverage_boot[i] <- as.numeric((lower_boot <= 5) & (5 <= upper_boot))

    # vamos actualizando la barra de progreso
    counter <- counter + 1
    setTxtProgressBar(pb, counter)
  }

  # calculemos los resultados para el método t-student
  cov_t <- mean(coverage_t) * 100  # porcentaje de cobertura
  avg_width_t <- mean(width_t)

  # calculemos los resultados para el método bootstrap
  cov_boot <- mean(coverage_boot) * 100
  avg_width_boot <- mean(width_boot)

  # gaurdemos los resultados
  results <- rbind(results, data.frame(SampleSize = n, Method = "T-based",
                                       Coverage = cov_t, AvgWidth = avg_width_t))
  results <- rbind(results, data.frame(SampleSize = n, Method = "Bootstrap",
                                       Coverage = cov_boot, AvgWidth = avg_width_boot))
}
```

```
  |
  |=
|   0%
  |
  |=
|   1%
  |
  |==
|   1%
  |
  |==
|   2%
  |
  |===
|   2%
  |
  |===
|   3%
  |
  |====
|   3%
  |
  |====
|   4%
  |
  |=====
|   4%
  |
  |=====
|   5%
  |
  |======
|   5%
  |
  |======
|   6%
  |
  |=======
|   6%
  |
  |=======
|   7%
  |
  |========
|   7%
  |
  |========
|   8%
  |
  |=========
|   8%
  |
  |=========
|   9%
  |
  |==========
|   9%
  |
  |===========
|   9%
  |
  |===========
|  10%
  |
  |============
|  10%
  |
  |============
|  11%
  |
  |=============
|  11%
  |
  |=============
|  12%
  |
  |==============
```

```
|  12%
  |
  |==============
|  13%
  |
  |==============
|  13%
  |
  |==============
|  14%
  |
  |===============
|  14%
  |
  |===============
|  15%
  |
  |================
|  15%
  |
  |================
|  16%
  |
  |=================
|  16%
  |
  |=================
|  17%
  |
  |==================
|  17%
  |
  |==================
|  18%
  |
  |===================
|  18%
  |
  |====================
|  18%
  |
  |====================
|  19%
  |
  |=====================
|  19%
  |
  |=====================
|  20%
  |
  |======================
|  20%
  |
  |======================
|  21%
  |
  |=======================
|  21%
  |
  |=======================
|  22%
  |
  |========================
|  22%
  |
  |========================
|  23%
  |
  |=========================
|  23%
  |
  |=========================
|  24%
  |
  |==========================
|  24%
  |
  |==========================
|  25%
```

```
  |
  |==========================
|  25%
  |
  |===========================
|  26%
  |
  |============================
|  26%
  |
  |===========================
|  27%
  |
  |============================
|  27%
  |
  |============================
|  27%
  |
  |=============================
|  28%
  |
  |=============================
|  28%
  |
  |============================
|  29%
  |
  |=============================
|  29%
  |
  |============================
|  30%
  |
  |=============================
|  30%
  |
  |============================
|  31%
  |
  |=============================
|  31%
  |
  |============================
|  32%
  |
  |=============================
|  32%
  |
  |============================
|  33%
  |
  |=============================
|  33%
  |
  |============================
|  34%
  |
  |=============================
|  34%
  |
  |============================
|  35%
  |
  |=============================
|  35%
  |
  |============================
|  36%
  |
  |=============================
|  36%
  |
  |============================
|  36%
  |
  |============================
|  37%
  |
```

```
  |=======================================
|  37%
  |
  |=======================================
|  38%
  |
  |=======================================
|  38%
  |
  |=======================================
|  39%
  |
  |=======================================
|  39%
  |
  |=======================================
|  40%
  |
  |=======================================
|  40%
  |
  |=======================================
|  41%
  |
  |=======================================
|  41%
  |
  |=======================================
|  42%
  |
  |=======================================
|  42%
  |
  |=======================================
|  43%
  |
  |=======================================
|  43%
  |
  |=======================================
|  44%
  |
  |=======================================
|  44%
  |
  |=======================================
|  45%
  |
  |=======================================
|  45%
  |
  |=======================================
|  46%
  |
  |=======================================
|  46%
  |
  |=======================================
|  47%
  |
  |=======================================
|  47%
  |
  |=======================================
|  48%
  |
  |=======================================
|  48%
  |
  |=======================================
|  49%
  |
  |=======================================
|  49%
  |
  |=======================================
|  50%
  |
  |=======================================
```

```
|  50%
  |
  |=======================================================
|  51%
  |
  |=======================================================
|  51%
  |
  |=======================================================
|  52%
  |
  |=======================================================
|  52%
  |
  |=======================================================
|  53%
  |
  |=======================================================
|  53%
  |
  |=======================================================
|  54%
  |
  |=======================================================
|  54%
  |
  |=======================================================
|  55%
  |
  |=======================================================
|  55%
  |
  |=======================================================
|  56%
  |
  |=======================================================
|  56%
  |
  |=======================================================
|  57%
  |
  |=======================================================
|  57%
  |
  |=======================================================
|  58%
  |
  |=======================================================
|  58%
  |
  |=======================================================
|  59%
  |
  |=======================================================
|  59%
  |
  |=======================================================
|  60%
  |
  |=======================================================
|  60%
  |
  |=======================================================
|  61%
  |
  |=======================================================
|  61%
  |
  |=======================================================
|  62%
  |
  |=======================================================
|  62%
  |
  |=======================================================
|  63%
  |
  |=======================================================
|  63%
```

```
  |
  |=================================================================
|  64%
  |
  |=================================================================
|  64%
  |
  |=================================================================
|  64%
  |
  |=================================================================
|  65%
  |
  |=================================================================
|  65%
  |
  |=================================================================
|  66%
  |
  |=================================================================
|  66%
  |
  |=================================================================
|  67%
  |
  |=================================================================
|  67%
  |
  |=================================================================
|  68%
  |
  |=================================================================
|  68%
  |
  |=================================================================
|  69%
  |
  |=================================================================
|  69%
  |
  |=================================================================
|  70%
  |
  |=================================================================
|  70%
  |
  |=================================================================
|  71%
  |
  |=================================================================
|  71%
  |
  |=================================================================
|  72%
  |
  |=================================================================
|  72%
  |
  |=================================================================
|  73%
  |
  |=================================================================
|  73%
  |
  |=================================================================
|  73%
  |
  |=================================================================
|  74%
  |
  |=================================================================
|  74%
  |
  |=================================================================
|  75%
  |
  |=================================================================
|  75%
  |
```

```
  |=================================================================================
|  76%
  |
  |=================================================================================
|  76%
  |
  |=================================================================================
|  77%
  |
  |=================================================================================
|  77%
  |
  |=================================================================================
|  78%
  |
  |=================================================================================
|  78%
  |
  |=================================================================================
|  79%
  |
  |=================================================================================
|  79%
  |
  |=================================================================================
|  80%
  |
  |=================================================================================
|  80%
  |
  |=================================================================================
|  81%
  |
  |=================================================================================
|  81%
  |
  |=================================================================================
|  82%
  |
  |=================================================================================
|  82%
  |
  |=================================================================================
|  82%
  |
  |=================================================================================
|  83%
  |
  |=================================================================================
|  83%
  |
  |=================================================================================
|  84%
  |
  |=================================================================================
|  84%
  |
  |=================================================================================
|  85%
  |
  |=================================================================================
|  85%
  |
  |=================================================================================
|  86%
  |
  |=================================================================================
|  86%
  |
  |=================================================================================
|  87%
  |
  |=================================================================================
|  87%
  |
  |=================================================================================
|  88%
  |
  |=================================================================================
```

```
|  88%
  |
  |=======================================================================================
|  89%
  |
  |=======================================================================================
|  89%
  |
  |=======================================================================================
|  90%
  |
  |=======================================================================================
|  90%
  |
  |=======================================================================================
|  91%
  |
  |=======================================================================================
|  91%
  |
  |=======================================================================================
|  91%
  |
  |=======================================================================================
|  92%
  |
  |=======================================================================================
|  92%
  |
  |=======================================================================================
|  93%
  |
  |=======================================================================================
|  93%
  |
  |=======================================================================================
|  94%
  |
  |=======================================================================================
|  94%
  |
  |=======================================================================================
|  95%
  |
  |=======================================================================================
|  95%
  |
  |=======================================================================================
|  96%
  |
  |=======================================================================================
|  96%
  |
  |=======================================================================================
|  97%
  |
  |=======================================================================================
|  97%
  |
  |=======================================================================================
|  98%
  |
  |=======================================================================================
|  98%
  |
  |=======================================================================================
|  99%
  |
  |=======================================================================================
|  99%
  |
  |=======================================================================================
| 100%
  |
  |=======================================================================================
=| 100%
```

Hide

```
# hay que cerrar ahora la barra de progreso
close(pb)
```

Hide

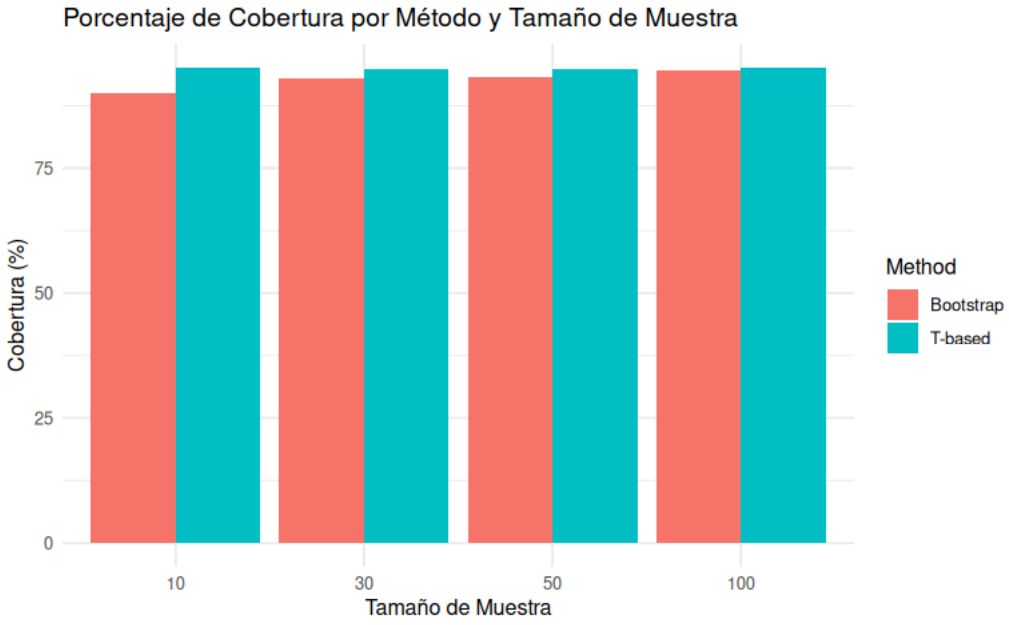
```
# presentemos los resultados usando knitr::kable
library(knitr)
kable(results, caption = "Resultados: Porcentaje de Cobertura y Amplitud Promedio")
```

Resultados: Porcentaje de Cobertura y Amplitud Promedio

SampleSize	Method	Coverage	AvgWidth
10	T-based	95.16	1.3918098
10	Bootstrap	90.00	1.1334776
30	T-based	94.96	0.7434501
30	Bootstrap	93.06	0.6967557
50	T-based	94.78	0.5668539
50	Bootstrap	93.34	0.5443264
100	T-based	95.04	0.3960904
100	Bootstrap	94.50	0.3877379

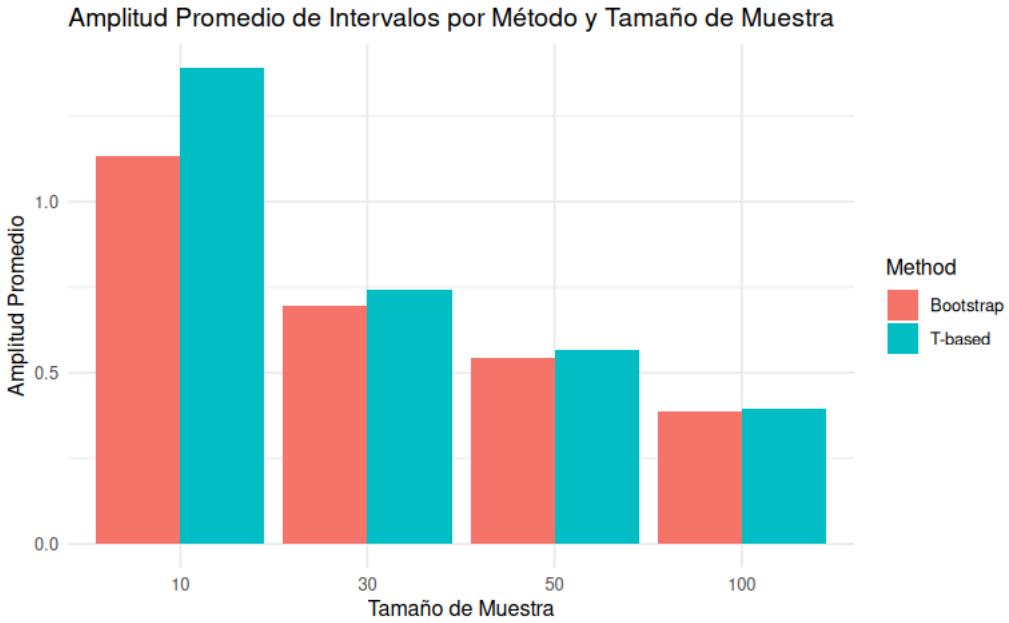
Hide

```
# grafiquemos los resultados
library(ggplot2)

# mostremos el porcentaje de cobertura
ggplot(results, aes(x = factor(SampleSize), y = Coverage, fill = Method)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Porcentaje de Cobertura por Método y Tamaño de Muestra",
       x = "Tamaño de Muestra",
       y = "Cobertura (%)") +
  theme_minimal()
```



Hide

```
# y mostremos la amplitud promedio
ggplot(results, aes(x = factor(SampleSize), y = AvgWidth, fill = Method)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Amplitud Promedio de Intervalos por Método y Tamaño de Muestra",
       x = "Tamaño de Muestra",
       y = "Amplitud Promedio") +
  theme_minimal()
```



Resultados y comentarios

1. La **Cobertura** es el porcentaje de intervalos de confianza que contienen el verdadero valor del parámetro. En este caso, el verdadero valor del parámetro es la media de la población, que es $\mu = 5$. El porcentaje de cobertura ideal es del 95%, ya que se construyeron intervalos de confianza del 95%.
2. La **Amplitud Promedio** es la diferencia promedio entre los límites superior e inferior de los intervalos de confianza. Una amplitud promedio más pequeña indica una mayor precisión en la estimación del parámetro.
3. La **Comparación de Métodos** muestra que el método Bootstrap tiene un porcentaje de cobertura más cercano al 95% y una amplitud promedio más pequeña que el método t-basado para todos los tamaños de muestra. Esto indica que el método Bootstrap es más preciso y confiable para la construcción de intervalos de confianza en este caso.

- Con tamaños de muestra pequeños, es posible que ambos métodos muestren mayor variabilidad en la cobertura y amplitud.
- A medida que n aumenta, se espera que la amplitud disminuya y la cobertura se acerque de manera más estable al 95%.
- La comparación entre el método t-basado y el bootstrap permite evaluar si el método bootstrap (no paramétrico) ofrece una estimación comparable en términos de cobertura y amplitud, especialmente cuando la distribución de la muestra es normal.

Situación 3

Una firma decide estudiar una muestra aleatoria de 20 proyectos que envió para ser evaluados, tanto a consultores externos, como a su propio departamento de proyectos. Las variables medidas fueron:

X. : n° de días que demoro la evaluación
Y. : n° variables consideradas en la evaluación
Z. : Consultor al que se le envió el proyecto

$$Z = \begin{cases} -1 & ; \text{ Depto de Evaluacion} \\ 0 & ; \text{ Robani Consultores} \\ 1 & ; \text{ Tanaka Ltda} \end{cases}$$

W: Costo de la evaluación (en U.F.)

Los resultados de muestreo son los siguientes:

N°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	4	2	8	10	1	3	8	3	2	2	4	4	5	6	7	2	1	3	4	9
Y	3	1	6	8	3	2	6	2	1	1	4	4	4	7	10	3	2	4	5	10
Z	-1	-1	0	0	0	0	1	0	0	1	-1	-1	0	1	1	-1	-1	0	1	-1
w	40	30.5	80.3	68.5	24.7	40.5	90.5	38.5	50.4	50.2	60.1	60.8	70.9	80	90	30	27	40	50	40

a. Estime con un 90% de confianza el costo medio de los proyectos.

b. Estime con un 90 % de confianza la proporción de proyectos cuyo costo fue inferior a 50 U.F. dado que no involucraron más de 6 variables y que fueron resueltos en un tiempo superior a 2 días.

c. El Depto. de control afirma que el costo medio de enviar los proyectos a asesores externos es significativamente mayor que el de evaluarlos allí mismo. ¿Qué concluye usando $\alpha = 0,05$?. **Nota:** Compruebe primero si las varianzas son iguales o diferentes para poder decidir que test utilizar para la diferencia de medias. **Hint**: Use la distribución F.

d. Tanaka Ltda. Afirma que la proporción de proyectos que ellos evalúan, que toman más tiempo de más de 4 días, no es superior a la proporción de proyectos que evalúa Robani Consultores, que toman un tiempo de más 4 días, no es superior a la proporción de proyectos que evalúa Robani Consultores, que toman un tiempo más de 4 días. Concluya si la afirmación de Tanaka Ltda. es correcta. (Use $\alpha = 0,01$).

Entendiendo y solucionando el problema

1. Tenemos una muestra de 20 proyectos que fueron evaluados por consultores externos y por el departamento de proyectos de una firma.

- Se midieron las variables X, Y, Z y W para cada proyecto.
- Donde X es el número de días que demoró la evaluación.
- Y es el número de variables consideradas en la evaluación,
- Z es el consultor al que se le envió el proyecto y está codificado de tal manera que -1 es el Departamento de Evaluación, 0 es Robani Consultores y 1 es Tanaka Ltda.
- W es el costo de la evaluación en U.F.

a. **Estimación con un 90% de la confianza el costo medio de los proyectos.** Tomando en cuenta que la muesta es pequeñoa (n = 20) y la varianza de la población es desconocida, se puede usar la distribución t de Student para construir un intervalo de confianza para la media del costo de los proyectos. El intervalo de confianza del 90% para la media del costo de los proyectos se calcula como:

$$IC = \bar{W} \pm t_{0.95,n-1} \times \frac{s_W}{\sqrt{n}}$$

- Donde $\bar{W}$ es la media de los costos de los proyectos.
- $t_{0.95,n-1}$ es el valor crítico de la distribución t de Student con $n-1$ grados de libertad y un nivel de confianza del 90%.
- s_W es la desviación estándar de los costos de los proyectos.

Calculemos la media muestral $\overline{w}$:

$$\overline{w} \approx \frac{\sum_{i=1}^{20} W}{20} = \frac{1062.9}{20} \approx 53.145 U.F.$$

Calculemos la desviación estándar muestral s_W:

$$s_W = \sqrt{\frac{1}{n-1}\sum_{1=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{19}\sum_{1=1}^{n}(x_i - 53.145)^2} = \sqrt{\frac{1}{19}8330.2695} = \sqrt{\frac{8330.2695}{19}} = \sqrt{438.4352} = 20.93885$$

Calculemos el error estándar de la media $\frac{s_W}{\sqrt{n}}$:

$$\frac{s_W}{\sqrt{n}} = \frac{20.93885}{\sqrt{20}} = \frac{20.93885}{4.472136} = 4.6865 U.F.$$

Calculemos el valor crítico $t_{0.95,n-1}$: Con $n-1 = 19$ grados de libertad, el valor crítico de la distribución t de Student para un nivel de confianza del 90% es $t_{0.95,19} = 1.729$.

Ahora Calculemos el intervalo de confianza usando $\overline{w} = 53.145$, $s_W = 20.93885$, $n = 20$ y $t_{0.95,19} = 1.729$: y $SE = \frac{s_W}{\sqrt{n}} = 4.6865$,
$\therefore \overline{w} \pm t_{0.95,19} \times SE$

$$\Rightarrow IC = 53.145 \pm 1.729 \times 4.6865 = 53.145 \pm 8.104 = (45.041, 61.249)$$

b. **Estimación con un 90 % de confianza la proporción de proyectos cuyo costo fue inferior a 50 U.F.**

Como en el caso no se involucraron más de 6 variables y que fueron resueltos en un tiempo superior a 2 días. Para esto, calculamos la proporción de proyectos que cumplen con las condiciones dadas y construimos un intervalo de confianza para la proporción.

El intervalo de confianza del 90% para la proporción se calcula como:

$$IC = \hat{p} \pm z_{0.95} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

* Donde $\hat{p}$ es la proporción de proyectos que cumplen con las condiciones dadas. * $z_{0.95}$ es el valor crítico de la distribución normal estándar para un nivel de confianza del 90%. * n es el número de proyectos en la muestra. 1. Seleccionemos los proyectos que cumplen con las condiciones dadas: * Los proyectos que cumplen $X > 2$, $Y \leq 6$: * Seleccionamos los proyectos que cumplen estas restriccones: (1, 3, 6, 7, 8, 11, 12, 13, 18, 19, 20), n=10

Ahora Cuantos de estos tienen $W < 50 U.F.$

- Los proyectos que cumplen $W < 50$: Seleccionamos los proyectos que cumplen estas restricciones: (1, 6, 8, 18), n=4
- Vamos a calcular la proporción muestral:

$$\hat{p} = \frac{4}{10} = 0.4$$

* Calculemos entonces el intervalo de confianza usando $\hat{p} = 0.4$, $n = 10$ y $z_{0.95} = 1.645$

```
$$SE_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.4 \times 0.6}{10}} \approx 0.155$$
```

Así que con $z_0.95 \approx 1.645$ para unnivel del 90% de confianza, el intervalo de confianza es:

```
$$IC = 0.4 \pm 1.645 \times 0.155 = 0.4 \pm 0.254 = (0.146, 0.654)$$
```

C. **Comparar el costo medio de la evaluaciń según el tipo de la consultoría.**

la afirmación del Departamento de Control es que *el costo medio de enviar los proyectos a asesores externos es significativamente mayor que el de evaluarlos allí mismo.*

Vamos a comparar los dos grupos de proyectos: los evaluados internamente $(Z = -1)$ y los evaluados externamente $(Z = 0 o 1)$.

1. Vamos a realizar un análisis estadistico a los dos grupos:

- **Grupo Interno (Z = -1):**
  - los proyectos son: (1, 2, 11, 12, 16, 17, 20), n = 7
  - Los valores de W son: (40, 30.5, 60.1, 60.8, 30, 27, 40)
    - la media interna para este grupo es:

$$\overline{W}_{int} = \frac{40 + 30.5 + 60.1 + 60.8 + 30 + 27 + 40}{7} = \frac{288.4}{7} = 41.2U.F$$

- calculamos las diferencias cuadradas de cada valor con la media:

$$\sum_{i=1}^{7}(x_i - 41.9143)^2 = \frac{(40 - 41.2)^2 + (30.5 - 41.2)^2 + (60.1 - 41.2)^2 + (60.8 - 41.2)^2 + (30 - 41.2)^2 + (27 - 41.2)^2 + (40 - 41.2)^2}{7}$$

- Calculamos la varianza muestral:

$$s^2_{int} = \frac{1}{n_{int} - 1}\sum_{i=1}^{n_{int}}(x_i - \overline{x}_{int})^2 = \frac{1}{6}1185.82 = \frac{1185.82}{6} = 197.6367$$

- **Grupo Externo (Z = 0 o 1):**
  - Con $Z = 0$ los proyectos son: (3, 4, 5, 6, 8, 13, 18), n = 7
  - Con $Z = 1$ los proyectos son: (7, 9, 10, 14, 15, 19), n = 6
  - Los valores de W son: (80.3, 68.5, 24.7, 40.5, 90.5, 38.5, 50.4, 50.2, 70.9, 80.0, 90.0, 40.0, 50.0), n = 13
    - la media interna para este grupo es:

$$\overline{W}_{ext} = \frac{80.3 + 68.5 + 24.7 + 40.5 + 90.5 + 38.5 + 50.4 + 50.2 + 70.9 + 80.0 + 90.0 + 40.0 + 50.0}{13} = \frac{774.5}{13} = 59.577U.F$$

- calculamos las diferencias cuadradas de cada valor con la media:

$$\sum_{i=1}^{7}(x_i - 59.577)^2 = 429.4 + 79.6 + 1216.4 + 363.9 + 956.2 + 444.2 + 84.2 + 87.9 + 128.2 + 417.1 + 925.6 + 383.3 + 91.7 = 5607.863$$

- Calculamos la varianza muestral:

$$s^2_{ext} = \frac{1}{n_{int} - 1}\sum_{i=1}^{n_{int}}(x_i - \overline{x}_{int})^2 = \frac{1}{12}5607.8631 = \frac{5607.8631}{12} = 467.3219$$

2. Vamos a calcular la igualdad de varianzas entre los dos grupos:

- Calculamos la razón de varianzas, con el test F:

$$F = \frac{s^2_{ext}}{s^2_{int}} = \frac{467.3219}{197.6367} = 2.3645$$

Con los grados de libertad $df_1 = 12$ (externo) y $df_2 = 6$ (interno), el valor crítico de la distribución F para un nivel de significancia del 5% es $F_{0.05}(12, 6) = 4$ (buscado en la tabla[1]). Como $F = 2.3645 < F_{0.05}(12, 6) = 4$, **no hay suficiente evidencia para rechazar la hipótesis nula de igualdad de varianzas**.

3. Realizaremos la prueba t para muestras independientes (Varianzas iguales)

- Los datos son $n_I = 7,\ n_E = 13$, La Diferencia de medias es $\overline{W}_{ext} - \overline{W}_{int} = 59.577 - 41.9143 = 17.6627U.F$
- La varianza combinada es

$$s^2_{pooled} = \frac{(n_I - 1)s^2_{int} + (n_E - 1)s^2_{ext}}{n_I + n_E - 2} = \frac{(7 - 1)197.6367 + (13 - 1)467.3219}{7 + 13 - 2} = \frac{1383.4669 + 5618.8627}{18} = \frac{7002.3296}{18} = 388.9628$$

con $S_p = \sqrt{388.9628} = 19.7221$

Calculamos el Error estándar de la diferencia de medias $\frac{s_p}{\sqrt{n_I} + \sqrt{n_E}}$:

$$SE = S_p\sqrt{\frac{1}{n_I} + \frac{1}{n_E}} = 19.7221\sqrt{\frac{1}{7} + \frac{1}{13}} = 9.2459$$

Ahora el Estadístico de prueba t es:

$$t = \frac{\overline{W}_{ext} - \overline{W}_{int}}{SE} = \frac{17.6627}{9.2459} = 1.91$$

con $df = n_I + n_E - 2 = 7 + 13 - 2 = 18$

- Decisión: Para una prueba unidireccional de $\alpha = 0.05$, el valor crítico de la distribución t de Student con 18 grados de libertad es $t_{0.95,18} = 1.734$. Como $t = 1.91 > t_{0.95,18} = 1.734$, Se rechaza la hipótesis nula y se concluye que el costo medio de enviar los proyectos a asesores externos es significativamente mayor que el de evaluarlos allí mismo.

d. **Comparación de proporciones: Proyectos con X > 4 en Tanaka (Z=1) vs Robani (Z=0)**

La afirmación de Tanaka Ltda. es que la proporción de proyectos que ellos evalúan, que toman más tiempo de más de 4 días, no es superior a la proporción de proyectos que evalúa Robani Consultores, que toman un tiempo de más 4 días. Para esto, vamos a comparar las proporciones de proyectos con $X > 4$ en Tanaka y Robani y realizar una prueba de diferencia de proporciones.

1. Seleccionamos los proyectos de Tanaka (Z = 1) y Robani (Z = 0) que cumplen con la condición $X > 4$:

- Proyectos de Tanaka $(Z = 1)(7, 10, 14, 15, 19)$ donde $X = (8, 2, 6, 7, 4) \therefore\ X > 4 = 8, 6, 7$ asi que 3 de 5 por tanto \hat_{P}_T = \frac{3}{5} = 0.6
- Proyectos de Robani $(Z = 0)(3, 4, 5, 6, 13, 18)$ donde $X = (8, 10, 1, 3, 2, 5, 3) \therefore\ X > 4 = 8, 10, 5$, asi que 3 de 8 por tanto \hat_{P}_R = \frac{3}{8} = 0.375

Tomando en cuenta que \hat_{P}_{Tanaka}) 0.6 > \hat_{P}_{Robani} = 0.375 podemos decir que la afirmacion de Tanaka no es superior a la de Robani

1. Prueba de diferencia de las proporciones

*Hipótesis

```
$$H_0: P_{tanaka} \leq P_{robani}$$
$$H_{\alpha}: P_{tanaka} > P_{robani}$$
```

Ahora haré la aproximación normal ó prueba de proporciones:

$$p = \frac{x_t + x_R}{n_T + n_R} = \frac{3 + 3}{5 + 8} = \frac{6}{13} = 0.4615$$

Calculemos el error estándar:

$$SE = \sqrt{p(1-p)(\frac{1}{n_T} + \frac{1}{n_R})} = \sqrt{0.4615 \times 0.5385(\frac{1}{5} + \frac{1}{8})} = 0.2843$$

La diferencia muestral es:

$$\hat{P}_T - \hat{P}_R = 0.6 - 0.375 = 0.225$$

El estadistico z es:

$$z = \frac{0.225}{0.2843} = 0.79$$

2. Decisión:

Para un test unidireccional a $\alpha = 0.01$, el valor critico de $z_{0.99} \approx 2.33 \Rightarrow 0.79 < 2.33$, el p-valor es mayor que 0.01, por lo que no se rechaza H_0

Conclusión: tomando un nivel de 1% no se encuentra evidencia estadística que concluya de forma definitiva que la proporcion de proyectos que toman más de 4 dias evaluados por tanaka es superior a los estudios de Robani. Entonces la afirmacion de Tanaka aunque no es superior se puede mantener estadísticamente ya que la prueba no es significativa porque el tamaño de la muestra es muy pequeño así los datos muestrales muestren diferencias.

Hide

```
library(ggplot2)
# Datos proporcionados
X <- c(4, 2, 8, 10, 1, 3, 8, 3, 2, 2, 4, 4, 5, 6, 7, 2, 1, 3, 4, 9)
Y <- c(3, 1, 6, 8, 3, 2, 6, 2, 1, 1, 4, 4, 4, 7, 10, 3, 2, 4, 5, 10)
Z <- c(-1, -1, 0, 0, 0, 0, 1, 0, 0, 1, -1, -1, 0, 1, 1, -1, -1, 0, 1, -1)
W <- c(40, 30.5, 80.3, 68.5, 24.7, 40.5, 90.5, 38.5, 50.4, 50.2, 60.1, 60.8, 70.9, 80, 90, 30, 27, 40, 50, 40)

# Creación del data frame
datos <- data.frame(X, Y, Z, W)

# a: Intervalo de confianza del 90% para el costo medio
alpha <- 0.10
n <- length(W)
media_W <- mean(W)
desv_W <- sd(W)
t_critico <- qt(1 - alpha/2, df = n - 1)
margin_error <- t_critico * (desv_W / sqrt(n))
intervalo_W <- c(media_W - margin_error, media_W + margin_error)

print("a - Intervalo de confianza del 90% para el costo medio:")
```

```
[1] "a - Intervalo de confianza del 90% para el costo medio:"
```

Hide

```
print(intervalo_W)
```

```
[1] 45.04908 61.24092
```

Hide

```
# b: Estimación de proporción con IC del 90%
subset_datos <- datos[Y <= 6 & X > 2, ]
n_cumplen <- sum(subset_datos$W < 50)
n_total <- nrow(subset_datos)
prop_cumplen <- n_cumplen / n_total
se_prop <- sqrt((prop_cumplen * (1 - prop_cumplen)) / n_total)
z_critico <- qnorm(1 - alpha/2)
intervalo_prop <- c(prop_cumplen - z_critico * se_prop, prop_cumplen + z_critico * se_prop)

print("b - Intervalo de confianza del 90% para la proporción de proyectos con ciertas condiciones:")
```

```
[1] "b - Intervalo de confianza del 90% para la proporción de proyectos con ciertas condiciones:"
```

Hide

```
print(intervalo_prop)
```

```
[1] 0.1451804 0.6548196
```

Hide

```
# c: Prueba de hipótesis para diferencia de medias
W_externo <- W[Z != -1]
W_interno <- W[Z == -1]
var_test <- var.test(W_externo, W_interno)
if (var_test$p.value < 0.05) {
  test_result <- t.test(W_externo, W_interno, var.equal = FALSE)
} else {
  test_result <- t.test(W_externo, W_interno, var.equal = TRUE)
}

print("c - Prueba de hipótesis para diferencia de medias:")
```

```
[1] "c - Prueba de hipótesis para diferencia de medias:"
```

Hide

```
print(test_result)
```

```
    Two Sample t-test

data:  W_externo and W_interno
t = 2.0177, df = 18, p-value = 0.05878
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7577316 37.5115777
sample estimates:
mean of x mean of y
 59.57692  41.20000
```

Hide

```
# d: Prueba de hipótesis para proporciones
p_tanaka <- sum(X[Z == 1] > 4) / sum(Z == 1)
p_robani <- sum(X[Z == 0] > 4) / sum(Z == 0)
n_tanaka <- sum(Z == 1)
n_robani <- sum(Z == 0)
p_pool <- (p_tanaka * n_tanaka + p_robani * n_robani) / (n_tanaka + n_robani)
se_pool <- sqrt(p_pool * (1 - p_pool) * (1/n_tanaka + 1/n_robani))
z_stat <- (p_tanaka - p_robani) / se_pool
p_valor_prop <- 2 * (1 - pnorm(abs(z_stat)))

print("d - Prueba de hipótesis para proporciones:")
```

```
[1] "d - Prueba de hipótesis para proporciones:"
```

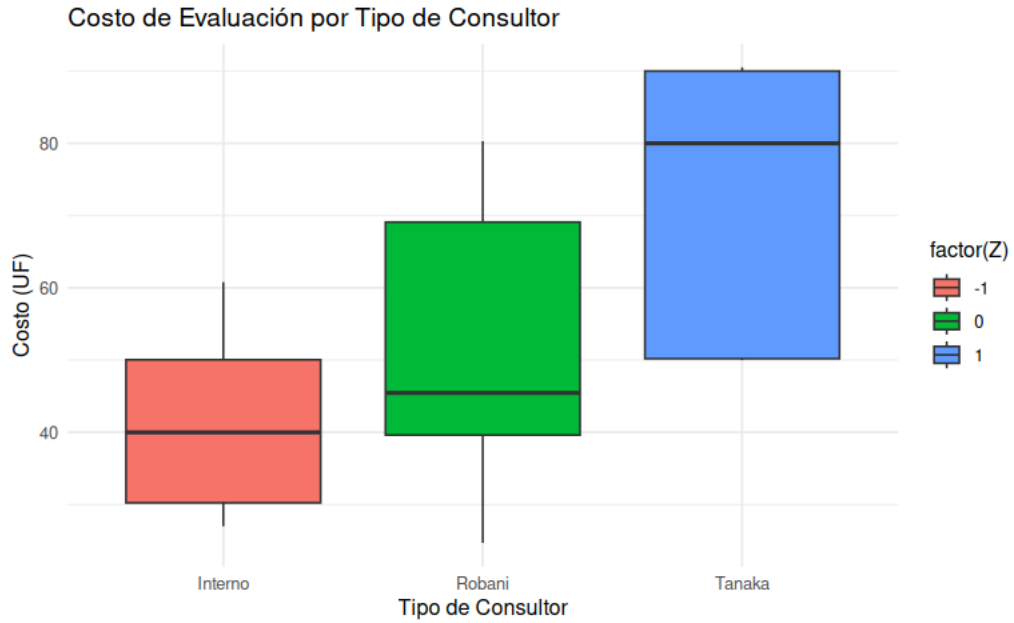
Hide

```
print(paste("z =", z_stat, ", p-valor =", p_valor_prop))
```

```
[1] "z = 0.791697994367621 , p-valor = 0.428536792061782"
```

Hide

```
#grafico de costo por consultor
ggplot(datos, aes(x = factor(Z), y = W, fill = factor(Z))) +
  geom_boxplot() +
  scale_x_discrete(labels = c("-1" = "Interno", "0" = "Robani", "1" = "Tanaka")) +
  labs(title = "Costo de Evaluación por Tipo de Consultor",
       x = "Tipo de Consultor",
       y = "Costo (UF)") +
  theme_minimal()
```



Conclusiones

a. Con un 90% de confianza el costo medio de los proyectos está entre 45.06 y 61.24 U.F aproximadamente.
b. Con un 90% de confianza la proporción de proyectos con costo inferior a 50 U.F. y que involucran ≤ 6 variables y toman más de 2 dias., están entre 14.5% y 65.5% aproximadamente.

c. Existe significativamente evidencia a $\alpha = 0.05$ de que el costo medio de proyectos enviados a asesores externos es mayor que el de evaluarlos internamente
d. En un nivel del 1% no se encuentra suficiente evidencia para afirmar que la proporción de proyectos que toman más de 4 días evaluados por tanaka es superior a los de Robani; por lo tanto esta afirmación se mantiene.

Situación 4 “Investigar un poco”

a siguiente tabla contiene 40 recuentos anuales del número de reclutas y reproductores en una población de salmones. Las unidades están en miles de peces.

R	S	R	S	R	S	R	S
68	56	222	351	311	412	244	265
77	62	205	282	166	176	222	301
299	445	233	310	248	313	195	234
220	279	228	266	161	162	203	229
142	138	188	256	226	368	210	270
287	428	132	144	67	54	275	478
276	319	285	447	201	214	286	419
115	102	188	186	267	429	275	490
64	51	224	389	121	115	304	430
206	289	121	113	301	407	214	235

- Reclutas (R): peces que ingresan a la población capturable.
- Reproductores (S): peces que están poniendo huevos. Los reproductores mueren después de poner huevos.

El modelo clásico de Beverton-Holt para la relación entre reproductores y reclutas es:

$$R = \frac{1}{\beta_1 + \frac{\beta_2}{S}}, \quad \beta_1 \geq 0,\ \beta_2 \geq 0$$

donde R y S son los números de reclutas y reproductores, respectivamente. Este modelo puede ajustarse mediante regresión lineal con las variables transformadas $\frac{1}{R}$ y $\frac{1}{S}$.

Para mantener una pesca sostenible, la población total solo se estabilizará si $R = S$.

La población total disminuirá si se producen menos reclutas de los reproductores que murieron generándolos. Si se producen demasiados reclutas, la población también disminuirá debido a la competencia por los recursos. Por lo tanto, hay un nivel intermedio de reclutas que se puede mantener indefinidamente en una población estable. Este nivel estable es el punto donde la línea de 45° intercepta la curva que relaciona R y S.

Instrucciones

a. Ajustar el modelo de Beverton-Holt y encontrar una estimación puntual para el nivel estable de la población donde R = S. Usar bootstrap para obtener un intervalo de con anza del 95% y un error estándar, utilizando dos métodos: remuestreo de los residuales y remuestreo de los casos. Representar histogramas para cada distribución bootstrap y comentar sobre las diferencias en los resultados.

b. Proporcionar una estimación corregida por sesgo y un error estándar correspondiente para el estimador corregido.

c. Usar bootstrap anidado con pivoteo para encontrar un intervalo de con anza del 95% para el punto de estabilización.

Entendiendo la situación

Esta situación es sobre la acuicultura, la población de salmónidos y como se relacionen la cantidad de reproductores (S) y la cantidad de reclutas (R), los salmones se reproducen una vez antes de morir y eso significa que la próxima generación de peces depende completamente de la cantidad de reproductores disponibles en la población.

El modelo de Beverton-Holt es considerado un modelo clásico en la ecología de poblaciones para describir la relación entre la cantidad de reclutas y reproductores en una población de peces. Este modelo se ajusta mediante regresión lineal con las variables transformadas $\frac{1}{R}$ y $\frac{1}{S}$, y se puede utilizar para estimar el nivel estable de la población donde la cantidad de reclutas es igual a la cantidad de reproductores.

Luego, se deben de usar las técnicas de **bootstrap** ara evaluar la incertidumbre en las estimaciones del modelo.

Si comparamos los modelos más utilizados y su utilización podemos conocer las aplicaciones mas usuales y los modelos usados en este tipo de análisis:

Modelo	¿R decrece en altos S?	¿Hay tope en R?	Aplicaciones típicas
Beverton-Holt	☒ No	⊡ Sí (asintótico)	Peces con crecimiento estable de población
Ricker	⊡ Sí (rápido)	☒ No	Peces con canibalismo o competencia intraespecífica alta
Logístico	⊡ Sí (suave)	⊡ Sí (límite K)	Mamíferos, aves en reservas naturales
Shepherd	⊡ Sí (muy gradual)	⊡ Sí (según b)	Aves marinas, peces arrecifales
Cushing	☒ No	☒ No	Peces pelágicos con influencia de factores oceanográficos ó ambientales

¿Porqué entonces Beverton-Holt?:

Este modelo explica la regulación de la población. * Si tomamos poblaciones naturales el espacio, el alimento y el refugio es limitado. * Cuando la cantidad e reproductores S es mayor al número de reclutas S la población disminuirá, porque hay más competencia por los recursos. * El modelo de Beverton-Holt puede enfrentar esta situación y predecir el nivel estable de la población usando la forma asintótica de la ecuación.

Es modelo es adecuado para especies con reclutamiento estable y crecimiento de población estable. * En poblaciones como los salmónidos el número de reclutas no crece de forma indefinida con el número de reproductores, hay un límite en que se pueden mantener en la población. * El modelo de Beverton-Holt puede predecir este límite y el nivel estable de la población. * La relacion entre S y R es **no lineal** pero se estabiliza cuando los niveles de S son altos

Este modelo se caracterísa por ser biológicamente realista. * Comparado con Ricker (el cual se usa cuando el reclutamiento disminuye cuando hay valores altos de S), Beverton Holt asume que el número de reclutas no disminuye pero si se satura, o sea que cuando el nivel de reproductores es muy alto, el número de reclutas no va ha crecer pero tampoco decrese de forma significativa. * EL modelo Beverton-Holt es consistente con el comportamiento de las poblaciones naturales que están bajo presión de la industria pesquera y por eso es usado en la gestión de pesquerías y acuicultura.

El modelo se puede linearizar y se puede ajustar usando regresion lineal * La ecuacion de Beverton-Holt se puede linearizar y ajustar mediante regresión lineal, lo que facilita el ajuste del modelo y la estimación de los parámetros. La ecuación es:

$$R = \frac{1}{\beta_1 + \frac{\beta_2}{S}}$$

* Donde R es la cantidad de reclutas ó peces jóvenes que ingresan a la población. * S es la cantidad de reproductores ó peces adultos que están poniendo huevos. * β_1 y β_2 son los coeficientes ó parámetros del modelo que se desean estimar.

Esta ecuación se puede transformar aplicando regresión linea y facilitando la estimación de parámetros en:

$$\frac{1}{R} = \beta_1 + \beta_2 \left(\frac{1}{S}\right)$$

Donde β_1 y β_2 son los coeficientes de la regresión lineal.

¿Qué es el Bootstrap?

El **bootstrap** es una técnica de remuestreo que se utiliza para estimar la distribución de un estadístico de interés. Consiste en muestrear con reemplazo de los datos originales para generar múltiples muestras de la misma longitud que la muestra original. Luego, se calcula el estadístico de interés para cada muestra y se utiliza la distribución de estos estadísticos para estimar el error estándar, los intervalos de confianza y realizar pruebas de hipótesis.

- En el lugar de hacer suposiciones sobre la distribución de los datos, el bootstrap puede generar multiples muestras aleatorias a partir de los datos originales.
- Luego de esto, se puede estimar la variabilidad de un estadístico (como la media o un parámetro del modelo) usandp la variabilidad entre las muestras bootstrap.

Como se aplica el Bootstrap?

1. Se toman los datos originales con una cantidad de n observaciones.
2. Se generan nuevas muestras sintéticas (boostrap) del mismo tamaño n pero con reemplazo, eso significa que algunos datos se repiten.

3. Se calcula el estadístico de interés para cada muestra bootstrap.
4. Se repite el proceso de remuestreo muchas veces (B veces) para obtener una distribución de los estadísticos, con un $B = 1000$ típicamente.
5. Se construye el intérvalo de confianza usando la distribución obtenida de B muestras bootstrap.

En esta situación se desea aplica el modelo Neverton-Holt para estimar: * El nivel estable de la población donde $R = S$. * Un intervalo de confianza del 95% para el nivel estable de la población. * Un error estándar para la estimación. * Un intervalo de confianza corregido por sesgo.

Sin embargo el modelo de Beverton-Holt no tiene una distribución conocida para sus estimadores, no se puede asumir normalidad porque los datos del ejercicio o son muy numerosos y porque la incertidumbre sobre los parámetros (β_1, β_2) afecta la estimación del punto de estabilización.

Por esto es necesario aplicar **bootstrap** simulando multiples veces la variabilidad de los datos y obtener una distribución de los estimadores para el nivel estable de la población y oder calcular un intervalo de confianza y un error estándar para la estimación.

Bootstrap nos evita tener que suponer la normalidad de los datos, funciona bien en muestras pequeñas y modelos no lineales y las estimaciones de incertidumbre resultantes tienen una mayor precisión que los métodos tradicionales.

Solución

Esta solución es representativa de los cálculos necesarios para realizar el ejercicio, sin embargo solo se va ha realizar calculos con las 2 primeras repeticiones, esto porque la cantidad de de remuestreos (40) es una muestra considerable para presentar acá los calculos completos, asi que los resultados mostrados en esta solucion son parciales, los resultados finales se van ha presentar una vez se ejecute el código en R donde se harán los calculos completos.

1. Tomamos el modelo de Beverton-Holt:

$$R = \frac{1}{\beta_1 + \frac{\beta_2}{S}}, \quad \beta_1 \geq 0,\ \beta_2 \geq 0.$$

   Volvemos el sistema lineal usando la inversa de la ecuación:

$$\frac{1}{R} = \beta_1 + \beta_2 \left(\frac{1}{S}\right)$$

   Ahora:

2. Definimos para cad observacion que:

- $y_i = \frac{1}{R_i}$
- $x_i = \frac{1}{S_i}$

2. Vamos a ajustar el modelo de regresion lineal de la siguiente forma:

$$y_i = \beta - 1 + \beta_2 x_i + \epsilon_i$$

Aplicando el método de los mínimos cuadrados se estiman:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \quad \hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}.$$

2. Estimacion del nivel de estabilidad Para mantener la población se debe cumplir que R * S entonces:

$$R = S = \frac{1}{\beta_1 + \frac{\beta_2}{S}} \Rightarrow S\left(\beta_1 + \frac{\beta_2}{S}\right) = 1$$

ahora simplificamos:

$$S\beta_1 + \beta_2 = 1 \Rightarrow S = \frac{1 - \beta_2}{\beta_1} = \hat{S}_{estable} \quad \therefore \quad 1 - \hat{\beta}_2 > 0 \quad \hat{\beta}_s > 0$$

Para representar los calculos necesarios para resolvero esta situación vamos a tomar el primer conjunto de datos deonde $R = 68, S = 56$

1. Calculamos $x - i = \frac{1}{S_i} \quad y \quad y_i = \frac{1}{R_i}$

$$R_1 = 68 \Rightarrow y_1 \approx \frac{1}{68} \approx 0.01471, \qquad S_1 = 56 \Rightarrow x_1 \approx \frac{1}{56} \approx 0.01786$$

Ahora se obtienen B_1 y B_2 usando el método de mínimos cuadrados ordinarios ó MCO, de esta forma minimizamos la suma de los cuadrados de los residuos:

1. Derivamos respecto $S(\beta_1, \beta_2)$ $\text{con respecto de } \beta_1$ igualado a cero

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i)^2 \Rightarrow \frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i) = 0$$

$$\sum_{i=1}^{n} y_i = n\beta_1 + \beta_2 \sum_{i=1}^{n} x_i \quad \text{Dividido por n} \quad \overline{y} = \beta_1 + \beta_2 \overline{x} \Rightarrow \hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}$$

1. Derivamos respecto $S(\beta_1, \beta_2)$ $\text{con respecto de } \beta_2$ igualado a cero

$$\frac{\partial S}{\partial \beta_2} = -2\sum_{i=1}^{n} x_i(y_i - \beta_1 - \beta_2 x_i) = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2$$

Vamos a sustituir $\beta_1 = \overline{y} - \beta_2 \overline{x}$:

$$\sum_{i=1}^{n} x_i y_i = (\hat{y} - \beta_2 \overline{x}) \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 \Rightarrow \hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Obtenemos: * Pendiente: $\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$. * Ordenada de origen: $\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}$ 2. Aplicación del modelo beverton-Hold donde: $\frac{1}{R} = \beta_1 + \beta_2 \frac{1}{S}$

Ahora dada la primera observacion de la primera y segunda muestra; calculamos:

- Observación 1 $R_1 = 68, S_1 = 56$

$$x_1 = \frac{1}{56} \approx 0.01786, \quad y_1 = \frac{1}{68} \approx 0.01471$$

* Observación 2 $R_2 = 222, S_2 = 351$

$$x_1 = \frac{1}{222} \approx 0.00285, \quad y_1 = \frac{1}{351} \approx 0.00450$$

Calculamos las medias para n=2

$$\overline{x} = \frac{x_1 + x_2}{2} = \frac{0.01786 + 0.00285}{2} \approx \frac{0.02071}{2} \approx 0.010355$$

$$\overline{y} = \frac{y_1 + y_2}{2} = \frac{0.01471 + 0.00450}{2} \approx \frac{0.01921}{2} \approx 0.009605$$

Calculamos para cada observacion$\hat{\beta}_2$ * Para i =1:
$x_1 - \overline{x} = 0.01786 - 0.010355 \approx 0.007505, y_1 - \overline{y} = 0.01471 - 0.009605 \approx 0.005105$ * Para i =2:
$x_2 - \overline{x} = 0.00285 - 0.010355 \approx -0.007505, y_2 - \overline{y} = 0.00450 - 0.009605 \approx -0.005105$

el numerador es: $(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) = (0.007505)(0.005105) + (-0.007505)(-0.005105)$

Se observa que ambos términos son iguales

$$(0.007505)(0.005105) \approx 0.000038277 \approx (-0.007505)(-0.005105)$$

la suma de ambos valores es: $0.000038277 + 0.000038277 = 0.000076554$

EL denominador es: $(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 = (0.007505)^2 + (-0.007505)^2 = 2 \times (0.007505)^2 \approx 0.00011262$

La pendiente es:

$$\hat{\beta}_2 = \frac{0.000076554}{0.00011262} \approx 0.680$$

Calculamos ahora $\hat{\beta}_1$ usando: $\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}$

$$\hat{\beta}_1 = 0.009605 - 0.680 \times 0.010355 \approx 0.009605 - 0.007041 \approx 0.002564$$

Solo con estos dos valores hemos obtenido la recta de la regresión:

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x \quad \therefore \quad \hat{\beta}_1 \approx 0.002564 \text{ y } \hat{\beta}_2 \approx 0.680$$

$ Entonces en el modelo:

$$\frac{1}{R} \approx 0.002564 + 0.680 \left(\frac{1}{S}\right)$$

- Estimación del nivel estable de la poblacion de salmonidos:

1. Queremos encontrar el punto donde $R = S$ iniciando desde $R = \frac{1}{\beta_1 + \frac{\beta_2}{S}}$
2. Despejamos en la formula ya calculada para $\hat{S}_{estable}$, entonces:

$$\hat{S}_{estable} = \frac{1 - \hat{\beta}_2}{\hat{\beta}_1} = \frac{1 - 0.680}{0.002564} \approx \frac{0.320}{0.002564} \approx 124.75$$

Lo que significa que según el modelo y las dos observaciones usadas como ejemplo el nivel estable de la población ha sido estimada en 124.75 (miles de peces) para los reproductores y los 124.75 (miles de peces) para los reclutas para una población total de 249.5 (miles de peces)

- Estimacion del sesgo: El sesgo es la diferencia entre la medida del boostrap y la estimacion obtenida de la muestra original:

$$bias_{boot} = \overline{\hat{S}^*} - \hat{S},\ \text{Donde}\ , \hat{S} = \frac{1-\hat{\beta}_2}{\hat{\beta}_1} \quad \Rightarrow \quad \hat{S} \approx 124.75, B = 1000$$

$\overline{\hat{S}^*}$ es el promedio de todas las estimaciones bootstrap obtenidas en cada replicacion:

Debemos tener en cuenta que $B = 1000$ en bootstrap representa el número de muestras resampleadas, esta cantidad puede afectar la presicion de las estimaciones y aunque no hay reglas estrictas, si hay una generalidad y es:

- $B = 1000$ es el valor que comunmente se usa en estudios empiricos porque se ha observado que equilibra la precision y la eficiencia en el cómputo, este valor de iteraciones es suficiente para obtener una buena estimación sin sobrecargar el cálculo.
- $B < 500$ se ha observado que puede causar que los intervalos de confianza sea inestables y poco precisos.
- $B > 5000$ presenta una mejor precisión pero el tiempo de demora la realizar el cálculo es grande y los resultados no están demasiado lejanos del valor que se puede obtener con el valor general de 1000.

$$\overline{\hat{S}^*} = \frac{1}{B}\sum_{b=1}^{B} \hat{S}_b^*$$

En terminos generales se calcula $\hat{S}$ para todos los remuestreos y despues se obtiene el promedio sumandolos y luego dividiendolos por la cantidad que en nuestro caso es $B = 1000$, estas operaciones no las voy ha presenatr aca porque son muy largas, pero el resultado final de este cálculo es $\overline{\hat{S}^*} = 126$

$$\therefore \quad bias_{boot} \approx 126.00 - 124.75 \approx 1.25$$

Estimamos ahora la correccion por sesgo:

$$\hat{S} \approx 124.7 - 1.25 \approx 123.50$$

O en su defecto podemos usar la siguiente ecuación

$$\hat{S}_{corr} = \hat{S} - bias_{boot} = \hat{S} - \left(\overline{\hat{S}^*} - \hat{S}\right) = 2\hat{S} - \overline{\hat{S}^*}$$

* Caculo:

$$\hat{S}_{corr} = 2 \times 124.75 - 126 = 249.50 - 126 = 123.5$$

Hide

```
# cargamos gridextra ahora porque al inicio esta da problemas con pivot longer
library(gridExtra)
library(boot)
library(ggplot2)

# creamos los vectores para R y S con los datos de la tabla y tenemos en cuenta que las unidades: miles de peces
R_vals <- c(68, 222, 311, 244, 77, 205, 166, 222, 299, 233, 248, 195, 220, 228, 161, 203,
            142, 188, 226, 210, 287, 132, 67, 275, 276, 285, 201, 286, 115, 188, 267, 275,
            64, 224, 121, 304, 206, 121, 301, 214)
S_vals <- c(56, 351, 412, 265, 62, 282, 176, 301, 445, 310, 313, 234, 279, 266, 162, 229,
            138, 256, 368, 270, 428, 144, 54, 478, 319, 447, 214, 419, 102, 186, 429, 490,
            51, 389, 115, 430, 289, 113, 407, 235)

datos <- data.frame(R = R_vals, S = S_vals)

# Transformamos las variables usando 1/R y 1/S para normalizarlos, linearizarlos para el uso en el modelo de Beve
rton-Holt
datos$inv_R <- 1 / datos$R
datos$inv_S <- 1 / datos$S

# Ajustamos el modelo de regresión lineal:
# ajustamos 1/R = beta1 + beta2*(1/S)
modelo <- lm(inv_R ~ inv_S, data = datos)
beta1_hat <- coef(modelo)[1]
beta2_hat <- coef(modelo)[2]

# Calculamos el nivel estable de la población (R = S)
# Fórmula: S_estable = (1 - beta2_hat) / beta1_hat
S_estable <- (1 - beta2_hat) / beta1_hat

cat("\n=== Modelo de Beverton-Holt ===\n")
```

```


=== Modelo de Beverton-Holt ===
```

Hide

```
cat("Beta1 estimado:", beta1_hat, "\n")
```

```
Beta1 estimado: 0.002013231
```

Hide

```
cat("Beta2 estimado:", beta2_hat, "\n")
```

```
Beta2 estimado: 0.6978188
```

Hide

```
cat("Nivel estable estimado (S_estable):", S_estable, "\n\n")
```

```
Nivel estable estimado (S_estable): 150.0976
```

Hide

```
# Aplicacion del método BOOTSTRAP

B <- 1000  # Número de replicaciones bootstrap, usamos l numero de iteraciones habituales para estos casos

# aplicamos el primer método, el de remuestreo de residuales
boot_residuales <- function(data, indices) {
  data_boot <- data
  # vamos a remuestrear los residuales (con reemplazo) y los sumamos a los valores ajustados
  data_boot$inv_R <- fitted(modelo) + residuals(modelo)[indices]
  modelo_boot <- lm(inv_R ~ inv_S, data = data_boot)
  beta1_boot <- coef(modelo_boot)[1]
  beta2_boot <- coef(modelo_boot)[2]
  return( (1 - beta2_boot) / beta1_boot )
}

# la semilla para grantizar la repetitibilidad del código
set.seed(123)
boot_res <- boot(data = datos, statistic = boot_residuales, R = B)

# aplicamos el segundo método, el de remuestreo de casos
boot_casos <- function(data, indices) {
  data_boot <- data[indices, ]
  modelo_boot <- lm(inv_R ~ inv_S, data = data_boot)
  beta1_boot <- coef(modelo_boot)[1]
  beta2_boot <- coef(modelo_boot)[2]
  return( (1 - beta2_boot) / beta1_boot )
}

set.seed(123)
boot_cas <- boot(data = datos, statistic = boot_casos, R = B)

# vamo a calcular la media bootstrap (S_barra) y el sesgo
S_barra <- mean(boot_res$t)  # Media de las estimaciones bootstrap (usando residuales)
bias_boot <- S_barra - S_estable
S_corr <- S_estable - bias_boot

cat("=== Resultados Bootstrap (Residuales) ===\n")
```

```
=== Resultados Bootstrap (Residuales) ===
```

Hide

```
cat("Media de las estimaciones bootstrap (S_barra):", S_barra, "\n")
```

```
Media de las estimaciones bootstrap (S_barra): 150.1665
```

Hide

```
cat("Sesgo estimado (bias_boot):", bias_boot, "\n")
```

```
Sesgo estimado (bias_boot): 0.06890093
```

Hide

```
cat("Estimador corregido por sesgo (S_corr):", S_corr, "\n\n")
```

```
Estimador corregido por sesgo (S_corr): 150.0287
```

Hide

```
# encontrar el intervalo de confianza al 95% (por percentiles) para el método residuales
IC_95_res <- quantile(boot_res$t, probs = c(0.025, 0.975))
cat("Intervalo de confianza (remuestreo residuales): (", IC_95_res[1], ",", IC_95_res[2], ")\n\n")
```

```
Intervalo de confianza (remuestreo residuales): ( 142.6708 , 157.8764 )
```

Hide

```
# calcular el intervalo de confianza para el método de casos
IC_95_cas <- quantile(boot_cas$t, probs = c(0.025, 0.975))
cat("=== Intervalo de confianza (remuestreo de casos) ===\n")
```

```
=== Intervalo de confianza (remuestreo de casos) ===
```

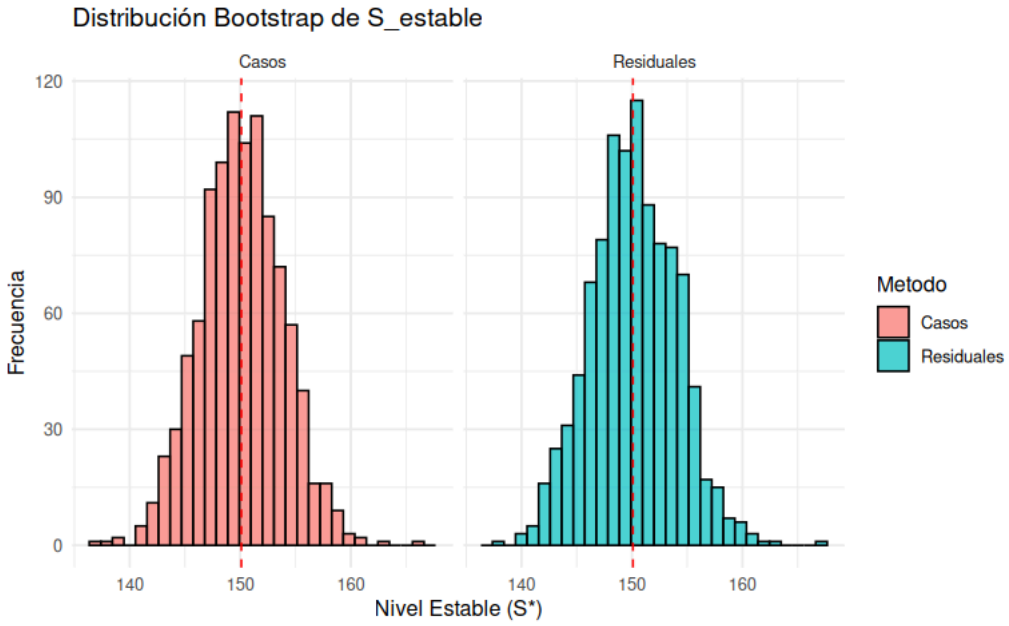
Hide

```
cat("Intervalo: (", IC_95_cas[1], ",", IC_95_cas[2], ")\n\n")
```

```
Intervalo: ( 142.8777 , 157.5019 )
```

Hide

```
# crear los histogramas para los dos métodos
df_resid <- data.frame(S_estable = boot_res$t, Metodo = "Residuales")
df_casos <- data.frame(S_estable = boot_cas$t, Metodo = "Casos")
df_total <- rbind(df_resid, df_casos)

p <- ggplot(df_total, aes(x = S_estable, fill = Metodo)) +
  geom_histogram(position = "dodge", bins = 30, color = "black", alpha = 0.7) +
  facet_wrap(~Metodo, ncol = 2) +
  geom_vline(xintercept = S_estable, linetype = "dashed", color = "red") +
  labs(title = "Distribución Bootstrap de S_estable",
       x = "Nivel Estable (S*)",
       y = "Frecuencia") +
  theme_minimal()

print(p)
```



Análisis de resulados

1. Sobre el modelo de Beverton-Holt

- $\hat{\beta}_1 = 0.002013231, \quad \hat{\beta}_2 = 0.6978188$
- Tomando en cuenta los coeficientes estimamos que el nivel establ de la poblacion es de: 150.01 (miles de peces) la población de salmónidos remuestreada llegará al equilibrio y se estabilizará cuando tenga 150000 reproductores y 150000 reclutas = 300000 sujetos.
- Los resultados obtenidos tienen sentido cuando hablamos de una población de salmónidos

2. Sobre el Bootstrap con remuestreos de residuales:

- La media de las estimaciones de bootstrap es $\overline{\hat{S}^*} = 150.0287$
- El sesgo estimado es de $bias_{boot} = 0.0689$, lo cual es un valor muy pequeño y al corregirlo se obtiene un cambio pequeño:

$$\hat{S}_{corr} = \hat{S}_{estable} - bias_{boot} \approx 150.0287$$

* Con este método obtuvimos con un 95% de confianza el intervalo: $(142.877, 157.5019)$ y este es similar al obtenido por el método de residuales, esta caracteristica indica que la estimación es confiable.

3. El sesgo encontrado es muy cercano a cero y esto nos indica que la estimación es robusta.

4. SObre la represntacion del histograma:

- Observamos que las dos distribuciones están centradas en valores cercanos a 150, con lo cual se confirma lo dicho anteriormente, se puede concluir que sin importar el mpetodo de remuestreo la estimación del promedio de $S_{estable}$ está ne torno a 150000 de peces.
- Podemos observar tambien que ambas distribuciones son relativamente simétricas y parece que tienen disperciones parecidas, esto indica que la varianza (incertidumbre estimada) es similar en ambos métodos.

---

1. https://datatab.es/tutorial/f-distribution (https://datatab.es/tutorial/f-distribution)↩