Segundo examen parcial

Simulación

Instrucciones:

Resuelve los siguientes ejercicios. Una entrega exitosa será la que pueda compilar y generar un pdf sin problemas, y aún mas importante, que incluya todas las respuestas del examen.

```
semilla <- 0 # Tu clave unica
set.seed(semilla)

library(dplyr)    # Para manipular y arreglar los datos ------------------------
library(tidyr)
library(purrr)
library(rsample)  # Para hacer el remuestreo -----------------------------------
library(ggplot2)  # Para graficar ----------------------------------------------
```

1. Tráfico

La base de datos *amis* (publicada por G. Amis) contiene información de velocidades de coches en millas por hora, las mediciones se realizaron en caminos de Cambridgeshire, y para cada ubicación se realizan mediciones en dos sitios, en uno de estos sitios se situó una señal de alerta (de dismunición de velocidad). Mas aún, las mediciones se realizaron en dos ocasiones, antes y después de que se instalara la señal de alerta. La cantidad de interés es el cambio medio relativo de velocidad en el cuantil 0.85. Se eligió esta cantidad porque el objetivo de la señal de alerta es disminuir la velocidad de los conductores más veloces.

Variables con las que cuenta este conjunto de datos son:

- `speed`: velocidad de los autos en mph,
- `period`: periodo en que se hicieron las mediciones. Es decir, 1 indica antes de la señal, 2 cuando ya había señal,
- `pair`: carretera en que se hizo la medición (1,2,5,7,8,9,10,11,12,13,14),
- `warning`: si se colocó señal de alerta en el sitio. Es decir, 1 indica que si había señal, 2 que no había.

Por interpretabilidad haremos un cambio de los valores de la variables con el bloque de abajo.

```
data <- read.csv("datos/amis.csv") |> tibble() |>
  mutate(
    period  = ifelse(period == 1, "antes", "despues"),
    warning = ifelse(warning ==  1, "conalerta", "sinalerta")
  )
```

a) ¿Las observaciones conforman una muestra aleatoria? Explica tu respuesta y en caso de ser negativa explica la estructura de los datos.

b) El estadístico de interés se puede escribir como

$$\theta = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\eta_{i,a}^{(1)} - \eta_{i,b}^{(1)}\right) - \left(\eta_{i,a}^{(0)} - \eta_{i,b}^{(0)}\right)\right],$$

donde $\eta_{i,1}^{(1)}$, $\eta_{i,2}^{(1)}$ corresponden a los cuartiles 0.85 de la distribución de velocidad en los sitios en los que se colocó la señal de alerta, (1 corresponde a las mediciones antes de la señal y 2 después) y $\eta_{i,1}^{(0)}$, $\eta_{i,2}^{(0)}$ son los correspondientes para los sitios sin alerta, N denota el número de carreteras. Es decir, denotamos por

$$\eta_{i,\text{period}}^{(\text{warning})},$$

donde η_i es el percentil .85 de la carretera i para cuando se establece la señal $\text{warning} \in 1, 2$.

Calcula el estimador *plug-in* de θ.

```
data |>
  # Rellena el código para realizar el agrupado. Hint: puedes usar col1,..., colp
  group_by( ... ) |>
  # Rellena el código para calcular el cuantil
  summarise(qspeed = quantile(speed, ... ),
            .groups = "drop")

calcula_estimador <- function(muestra, ...){
  muestra |>
    analysis() |>
    # Copia el código necesario para calcular el estadistico que nos interesa --

    # ------------------------------------------------------------------------
    pivot_wider(
      id_cols = pair,
      values_from = qspeed,
      names_from = c(warning, period)
    ) |>
    mutate(
      theta = (conalerta_antes - conalerta_despues) - (sinalerta_antes - sinalerta_despues)
    ) |>
    summarise(estimate = mean(theta),
              term = "theta")
}
```

En el bloque anterior ¿qué es lo que hace la función `pivot_wider`?

c) Genera $B = 1,500$ replicaciones bootstrap de θ y realiza un histograma.

```
boots <- data |>
  # Para poder usar el estrato necesitamos qué este sea una columna. Actualmente
  # son varias. En la funcion _paste_ de abajo incorpora las columnas que se
  # tienen que concatenar para crear /una columna/ de estrato para la función
  # _bootstraps_.
  mutate(estrato = paste(...)) |>
  bootstraps(strata = estrato, times = ...)
```

El siguiente bloque debería de generar el histograma de nuestro estimador `estimate`.

```
boots <- boots |>
  mutate(resultados = map(splits, calcula_estimador))

boots |> unnest(resultados) |>
```

```
ggplot(aes(estimate)) +
geom_histogram()
```

d) Genera intervalos de confianza usando la aproximación normal y percentiles. Comparalos y en caso de encontrar diferencias explica a que se deben.

2. Cobertura de intervalos

En este problema realizarás un ejercicio de simulación para comparar la exactitud de distintos intervalos de confianza. Simularás muestras de
una distribución Poisson con parámetro $\lambda = 2.5$ y el estadístico de interés
es $\theta = \exp(-2\lambda)$.

Sigue el siguiente proceso:

i) Genera una muestra aleatoria de tamaño $n = 60$ con distribución $Poisson(\lambda)$, parámetro $\lambda = 2.5$ (en R usa la función `rpois()`).

ii) Genera $2,500$ muestras bootstrap y calcula intervalos de confianza del 95% para $\hat{\theta}$ usando 1) el método normal y 2) percentiles.

iii) Revisa si el intervalo de confianza contiene el verdadero valor del parámetro ($\theta = \exp(-2 \cdot 2.5)$), en caso de que no lo contenga registra si falló por la izquierda (el límite inferior mayor $2\exp(-2 * \lambda)$) o falló por la derecha (el límite superior menor $2\exp(-2 * \lambda)$).

a) Repite el proceso descrito 1000 veces y llena la siguiente tabla:

Método	% fallo izquierda	% fallo derecha	Cobertura	Longitud promedio
Normal				
Percentiles				

La columna cobertura es una estimación de la cobertura del intervalo basada en las simulaciones, para calcularla simplemente escribe el porcentaje de los intervalos que incluyeron el verdadero valor del parámetro. La longitud promedio es la longitud promedio de los intervalos de confianza bajo cada método.

b) Repite el inciso a) seleccionando muestras de tamaño 300.

Método	% fallo izquierda	% fallo derecha	Cobertura	Longitud promedio
Normal				
Percentiles				