Métodos con remuestreo

Ejemplo de solución ejercicio prático N°8

Como vimos, la Encuesta de Caracterización Socioeconómica Nacional (Casen) es realizada por el Ministerio de Desarrollo Social de forma periódica para conocer la situación de los hogares chilenos con relación a aspectos demográficos, de educación, salud, vivienda, trabajo e ingresos. Es la principal fuente de información para estimar la magnitud de la pobreza y la distribución del ingreso en el país.

Pregunta 1

Propongan una pregunta de investigación original, que involucre la comparación de una frecuencia de un evento o característica en dos grupos independientes. Fijando una semilla propia, seleccionen una muestra aleatoria de hogares (100 < n < 150) y respondan la pregunta propuesta utilizando el método Monte Carlo.

Pensando en la nueva política estatal sobre pensiones y el apoyo a personas con discapacidad, consideraremos la siguiente pregunta: *la proporción de entrevistados fuera de la Región Metropolitana que presentan problemas significativos de movilidad pero que están cotizando para su jubilación tiene igual proporción entre hombres y mujeres*.

Esta pregunta plantea la comparación de dos proporciones. Luego, el estadístico de interés que debemos utilizar en este caso es la **diferencia de las proporciones** entre dos grupos independiente de personas.

Comencemos cargando los paquetes que vamos a utilizar.

```
library(dplyr)
library(ez)
library(ggpubr)
library(tidyr)
```

Primero, carguemos los datos.

```
src_dir <- "~/Downloads"
src_basename <- "EP08 Datos CASEN 2017.csv"
src_file <- file.path(src_dir, src_basename)
datos <- read.csv(file = src_file, stringsAsFactors = TRUE)
```

Filtremos para obtener los datos de interés y obtengamos la muestra que se solicita.

```
set.seed(347)
n1 <- 125
muestra1 <- datos |> filter(region != "Región Metropolitana de Santiago") |>
  filter(h10e %in% c("No puede hacerlo", "Sí, mucha dificultad")) |>
  filter(!is.na(o29)) |>
  mutate(cotiza = ifelse(o29 == "No está cotizando", "No", "Sí")) |>
  select(sexo, cotiza) |>
  sample_n(n1)
```

Definamos una función que calcule la diferencia entre las proporciones de hombres y mujeres que cotizan para su jubilación.

```
get.prop.dif <- function(df, verbose = FALSE)
{
  tabla <- table(df)
  if(verbose)
    print(tabla)
  ph <- tabla[1, 2] / (tabla[1, 1] + tabla[1, 2])
  pm <- tabla[2, 2] / (tabla[2, 1] + tabla[2, 2])
  if(verbose)
  {
    cat("\n")
    cat("Proporción de personas que cotizan:\n")
    cat("Hombres:", round(ph, 4), "\n")
    cat("Mujeres:", round(pm, 4), "\n")
  }
  return(ph - pm)
}
```

Calculemos, usando esta función, la diferencia observada en la muestra obtenida.

```
dif.obs <- get.prop.dif(muestra1, TRUE)
cat("\nDiferencia de proporciones observada:", round(dif.obs, 3), "\n")
```

```
        cotiza
sexo     No Sí
  Hombre 46 30
  Mujer  30 19

Proporción de personas que cotizan:
Hombres: 0.3947
Mujeres: 0.3878

Diferencia de proporciones observada: 0.007
```

Vemos que la diferencia en las muestras es bastante pequeña: $0{,}7\%$.

Formulemos las hipótesis:

H$_0$: las proporciones de hombres (p_h) y mujeres (p_m) que viven en la Región Metropolitana de Santiago y que declaran tener problemas serios de movilidad pero que están cotizando para su jubilación son iguales: $p_h - p_m = 0$.

H$_A$: Por el contrario, estas proporciones son distintas: $p_h - p_m \neq 0$.

Definamos el número de permutaciones que vamos a trabajar (y si queremos mensajes).

```
R <- 2999
verbose <- FALSE
if(R < 10)
  verbose <- TRUE
```

Obtenemos las permutaciones, teniendo cuidado de usar una semilla, como indica el enunciado, que nos poermita obtener los mismos resultados cada que vez que ejecutemos el script.

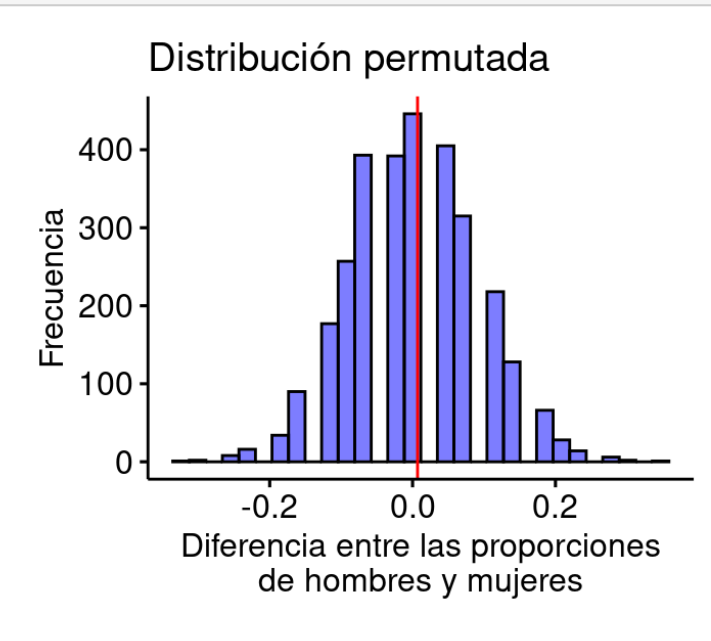
```
set.seed(349)
permutaciones <- lapply(1:R, function(i) sample(1:n1))
```

Obtenemos la distribución permutada de las diferencias de proporciones aplicando nuestra función `get.prop.dif()` a cada una de las permutaciones generadas. Notemos que **solo necesitamos permutar** la variable `sexo`, que es la que define en qué grupo está incluido cada caso.

```
get.prop.dif.perm <- function(indices, df, verbose = FALSE)
{
  df.nuevo <- data.frame(sexo = df[indices, "sexo"], df[["cotiza"]])
  get.prop.dif(df.nuevo, verbose)
}
distribucion <- sapply(permutaciones, get.prop.dif.perm, muestra1, verbose)
```

Revisemos cómo se ve esta distribución respecto del valor observado en la muestra original.

```
p1 <- gghistogram(data.frame(distribucion), "distribucion", bins = 30, fill = "blue",
                  title = "Distribución permutada",
                  xlab = "Diferencia entre las proporciones\nde hombres y mujeres",
                  ylab = "Frecuencia")
p1 <- p1 + geom_vline(xintercept = dif.obs, colour="red")
print(p1)
```



Podemos ver que la diferencia de las proporciones observada en la muestra original se ubica cerca del centro de distribución, no muy lejos del valor cero.

Calculemos el intervalo de 95% confianza y el valor p para una prueba bilateral (usando el valor absoluto de las diferencias) con 95% de confianza, es decir la probabilidad de encontrar diferencias al menos tan extremas como la diferencia observada.

```
ci1 <- quantile(distribucion, c(0.025, 0.975))
numerador1 <- sum(abs( distribucion) > abs(dif.obs))
valor_p1 <- (numerador1 + 1) / (R + 1)

cat("IC 95%: [", round(ci1[1], 3), ", ", round(ci1[2], 3), "]\n", sep = "")
cat("P-valor:", round(valor_p1, 3))
```

```
IC 95%: [-0.161, 0.175]
P-valor: 0.851
```

Lo que nos lleva a la siguiente conclusión:

Se puede concluir, con 95% confianza, que no hay evidencia suficiente para rechazar la hipótesis nula ($p = 0{,}851$) y tenemos que concluir que no es posible descartar que las proporciones de hombres y mujeres con problemas serios de movilidad que están cotizando para su jubilación son iguales (IC $95\% = [-0{,}161; 0{,}175]$).

Pregunta 2

Propongan una pregunta de investigación original, que involucre la comparación de las medias de más de dos grupos independientes. Fijando una semilla distinta a la anterior, seleccionen una muestra aleatoria de hogares (200 < n < 300) y respondan la pregunta propuesta utilizando *bootstrapping*. Solo por ejercicio académico, aplique un análisis *post-hoc con bootstrapping* aunque este no sea necesario.

En este caso, consideraremos la pregunta: *¿es igual el ingreso per cápita en las regiones de Atacama, Coquimbo y del Maule?*

Esta pregunta requiere contrastar las medias de 3 grupos independientes. Al haber más de dos medias, no es posible comparar directamente sus diferencias. Así, lo más conveniente en este caso es utilizar el **estadístico F** para evaluar su igualdad.

Filtremos los datos y obtengamos la muestra como se solicita.

```
n2 <- 275
set.seed(572)
regiones <- c("Región de Coquimbo", "Región de Atacama", "Región del Maule")
muestra2 <- datos |> filter(region %in% regiones) |> droplevels() |>
  mutate(region = recode(region, "Región de Atacama" = "Atacama")) |>
  mutate(region = recode(region, "Región de Coquimbo" = "Coquimbo")) |>
  mutate(region = recode(region, "Región del Maule" = "Maule")) |>
  select(ytotcorh, numper, region) |>
  mutate(ypercap = ytotcorh/numper, .keep = "unused") |>
  sample_n(n2)
```

Formulamos las hipótesis:

H$_0$: El ingreso per cápita promedio es igual en las regiones de Atacama (μ_A), Coquimbo (μ_C) y el Maule (μ_M): $\mu_A = \mu_C = \mu_M$.

H$_A$: El ingreso per cápita promedio es distinto en las regiones de Atacama, Coquimbo y el Maule:
$\exists (a, b) \in \{A, C, M\} \mid \mu_a \neq \mu_b$

Revisemos el tamaño de las muestras de observaciones en cada grupo.

```
print(summary(muestra2[["region"]]))
```

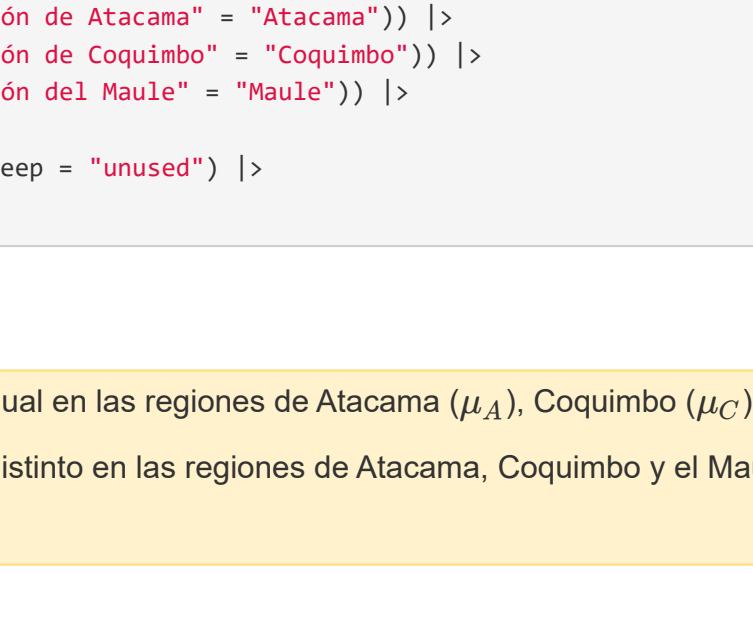
```
 Atacama Coquimbo    Maule 
      58       74      143 
```

Ooops! Una complicación extra, puesto que las muestras tienen tamaños distintos y debemos **mantener esos tamaños** en el remuestreo. Separamos los índices de cada región con este objetivo en mente.

```
iAtacama <- which(muestra2[["region"]] == "Atacama")
iCoquimbo <- which(muestra2[["region"]] == "Coquimbo")
iMaule <- which(muestra2[["region"]] == "Maule")
```

Revisemos los datos.

```
muestra2[["ypercap_miles"]] <- muestra2[["ypercap"]] / 1000
p2 <- ggboxplot(muestra2, x = "region", y = "ypercap_miles", fill = "region")
p2 <- p2 + xlab("Región") + ylab("Ingreso per cápita (miles de pesos)")
print(p2)
```



Podemos ver que los datos, además de estar desbalanceados, parecen no cumplir con la condición de heterocedasticidad y presentan numerosos valores atípicos. Así, no podemos usar un método de análisis clásico y debemos recurrir a métodos apropiados, en este caso **remuestreo con bootstrapping**.

Definimos una función que calcula el estadístico de interés que, como se dijo, corresponde al estadístico F usado por ANOVA para muestras independientes.

```
get.F <- function(df, iA, iC, iM, verbose = FALSE)
{
  # Armamos la matriz de datos con los índices recibidos
  i <- c(iA, iC, iM)
  ids <- factor(1:length(i))
  datos <- cbind(id = ids, df[i, ])
  dd <- datos

  ez <- ezANOVA(datos, ypercap, id, between = region, type = 2)
  if(verbose)
    print(ez)
  return(ez[["ANOVA"]][["F"]])
}
```

Obtenemos el estadístico para la muestra original.

```
F.obs <- get.F(muestra2, iAtacama, iCoquimbo, iMaule, TRUE)
```

```
Warning: Data is unbalanced (unequal N per group). Make sure you specified a
well-considered value for the type argument to ezANOVA().
```

```
Coefficient covariances computed by hccm()
```

```
$ANOVA
  Effect DFn DFd        F        p p<.05        ges
1 region   2 272 4.557016 0.011307     * 0.03242112

$`Levene's Test for Homogeneity of Variance`
  DFn DFd          SSn          SSd        F         p p<.05
1   2 272 255519071781 1.913278e+13 1.816286 0.1645955      
```

Notemos que la llamada a `ezANOVA()` genera un mensaje (inútil y molesto a estas alturas) y una advertencia relacionada al desbalance en los datos. Como se comentó en el apunte, especificando `type = 2` resuelve este inconveniente en la mayoría de los casos.

Debemos recordar que, a diferencia del caso anterior con permutaciones, con bootstrapping (al remuestrear con reemplazo la muestra original) las remuestras **no representan la hipótesis nula**. Cuando teníamos dos muestras, podíamos **recentrar** la distribución bootstrap en el valor nulo. Pero ¿cómo hacemos esto con el estadístico F?

Para ello debemos considerar que si la hipótesis nula se cumple, los tres grupos tienen las mismas medias y varianzas. Con más de dos grupos, es más fácil hacer estos ajustes **antes del remuestreo**. Para ello, nos basamos en las ideas propuestas por Fisher & Hall (1990), Hall & Wilson (1991) y Martin (2007).

Primero vamos a obtener las medidas generales (*pooled*).

```
media.gral <- mean(muestra2[["ypercap"]])
sd.gral <- sd(muestra2[["ypercap"]])
```

Luego obtenemos las medidas por grupo (por región en este caso):

```
grupos <- muestra2 |>
  group_by(region) |>
  summarise(media = mean(ypercap), sd = sd(ypercap)) |>
  as.data.frame()
```

Ahora **desplazamos** los valores vistos para que los tres grupos tengan la misma media e igual varianza (en rigor, desviación estándar).

```
muestra2b <- muestra2
muestra2b[iAtacama, "ypercap"] <- media.gral +
  (muestra2b[iAtacama, "ypercap"] - grupos[1, "media"]) *
  (sd.gral / grupos[1, "sd"])
muestra2b[iCoquimbo, "ypercap"] <- media.gral +
  (muestra2b[iCoquimbo, "ypercap"] - grupos[2, "media"]) *
  (sd.gral / grupos[2, "sd"])
muestra2b[iMaule, "ypercap"] <- media.gral +
  (muestra2b[iMaule, "ypercap"] - grupos[3, "media"]) *
  (sd.gral / grupos[3, "sd"])
```

Definamos el número de remuestreos que vamos a utilizar (y si queremos mensajes a pantalla).

```
B <- 2999
verbose <- FALSE
if(B < 10)
  verbose <- TRUE
```

Generamos las remuestras definiendo una semilla adecuada y muestreando con reemplazo los índices de los datos de cada grupo.

```
set.seed(573)
re.iAtacama <- lapply(1:B, function(i) sample(iAtacama, replace = TRUE))
re.iCoquimbo <- lapply(1:B, function(i) sample(iCoquimbo, replace = TRUE))
re.iMaule <- lapply(1:B, function(i) sample(iMaule, replace = TRUE))
```

Obtenemos la distribución bootstrapping remuestreando cada región por separado (evitando mensajes y advertencias)

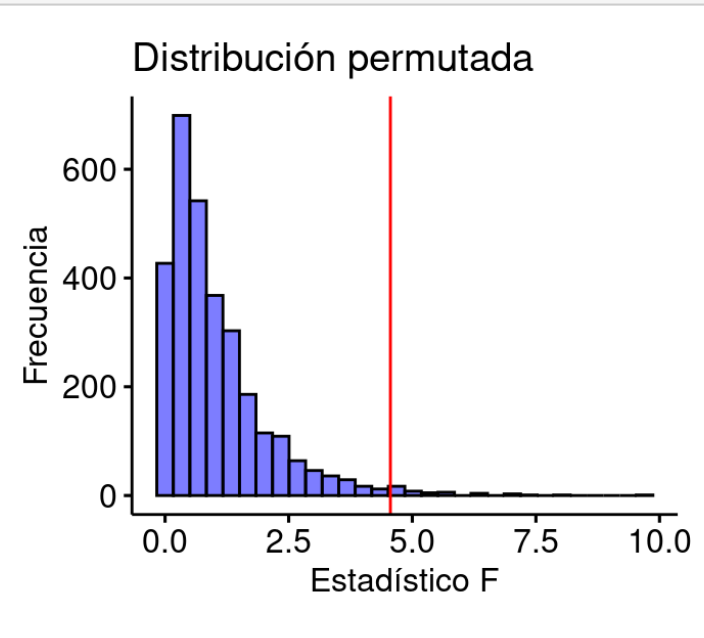
```
cat("Remuestreando, por favor espere...\n")
get.F.boot <- function(i, df, verbose = FALSE)
  get.F(df, re.iAtacama[[i]], re.iCoquimbo[[i]], re.iMaule[[i]], verbose)

distribucion <- suppressMessages(suppressWarnings(
  sapply(1:B, function(i) get.F.boot(i, muestra2b, verbose))
))
```

```
Remuestreando, por favor espere...
```

Revisemos cómo se ve esta distribución respecto del valor observado en la muestra original.

```
p2 <- gghistogram(data.frame(distribucion), x = "distribucion",
                  title = "Distribución permutada",
                  xlab = "Estadístico F", ylab = "Frecuencia",
                  bins = 30, fill = "blue")
p2 <- p2 + geom_vline(xintercept = F.obs, colour="red")
print(p2)
```



Vemos que el valor F observado parece estar bastante alejado de lo esperado si la hipótesis nula fuera cierta. Calculemos el valor crítico de F con 95% confianza en esta distribución empírica y estimemos el valor p correspondiente para el valor F observado.

```
F_crit <- quantile(distribucion, 0.95)
cat("F crítico con 95% de confianza:", round(F_crit, 3), "\n")

numerador2 <- sum(distribucion > F.obs)
valor_p2 <- (numerador2 + 1) / (B + 1)
cat("P-valor:", round(valor_p2, 3))
```

```
F crítico con 95% de confianza: 3.081
P-valor: 0.015
```

Estamos en condiciones de concluir respecto a la prueba ómnibus:

Observamos que si la hipótesis nula es correcta, el estadístico $F(2, 272)$ no debiera ser superior a $3{,}08$ con 95% confianza. Como el estadístico F observado (en la muestra original) fue $F(2, 272) = 4{,}56$, que tiene una baja probabilidad de ser encontrado ($p = 0{,}015$), se rechaza la hipótesis nula en favor de la alternativa. Concluimos entonces, con 95% confianza, que el ingreso per cápita promedio no es igual en las regiones estudiadas. Corresponde, entonces, hacer un análisis post-hoc.

Para el análisis post-hoc, por conveniencia haremos comparaciones entre pares de regiones, teniendo cuidado de utilizar las **mismas remuestras** que usamos en la prueba ómnibus. Al analizar pares de regiones, podemos usar la **diferencia de las medias** como estadístico de interés.

Escribamos la típica función que calcula esta diferencia.

```
get.dif.medias <- function(df, i1, i2)
{
  media1 <- mean(df[i1, "ypercap"])
  media2 <- mean(df[i2, "ypercap"])
  return(media1 - media2)
}
```

Obtenemos las diferencias observadas entre cada par de regiones.

```
dif.obs.A.C <- get.dif.medias(muestra2, iAtacama, iCoquimbo)
dif.obs.A.M <- get.dif.medias(muestra2, iAtacama, iMaule)
dif.obs.C.M <- get.dif.medias(muestra2, iCoquimbo, iMaule)

cat("Atacama - Coquimbo:", round(dif.obs.A.C), "\n")
cat("Atacama - Maule:", round(dif.obs.A.M), "\n")
cat("Coquimbo - Maule:", round(dif.obs.C.M), "\n")
```

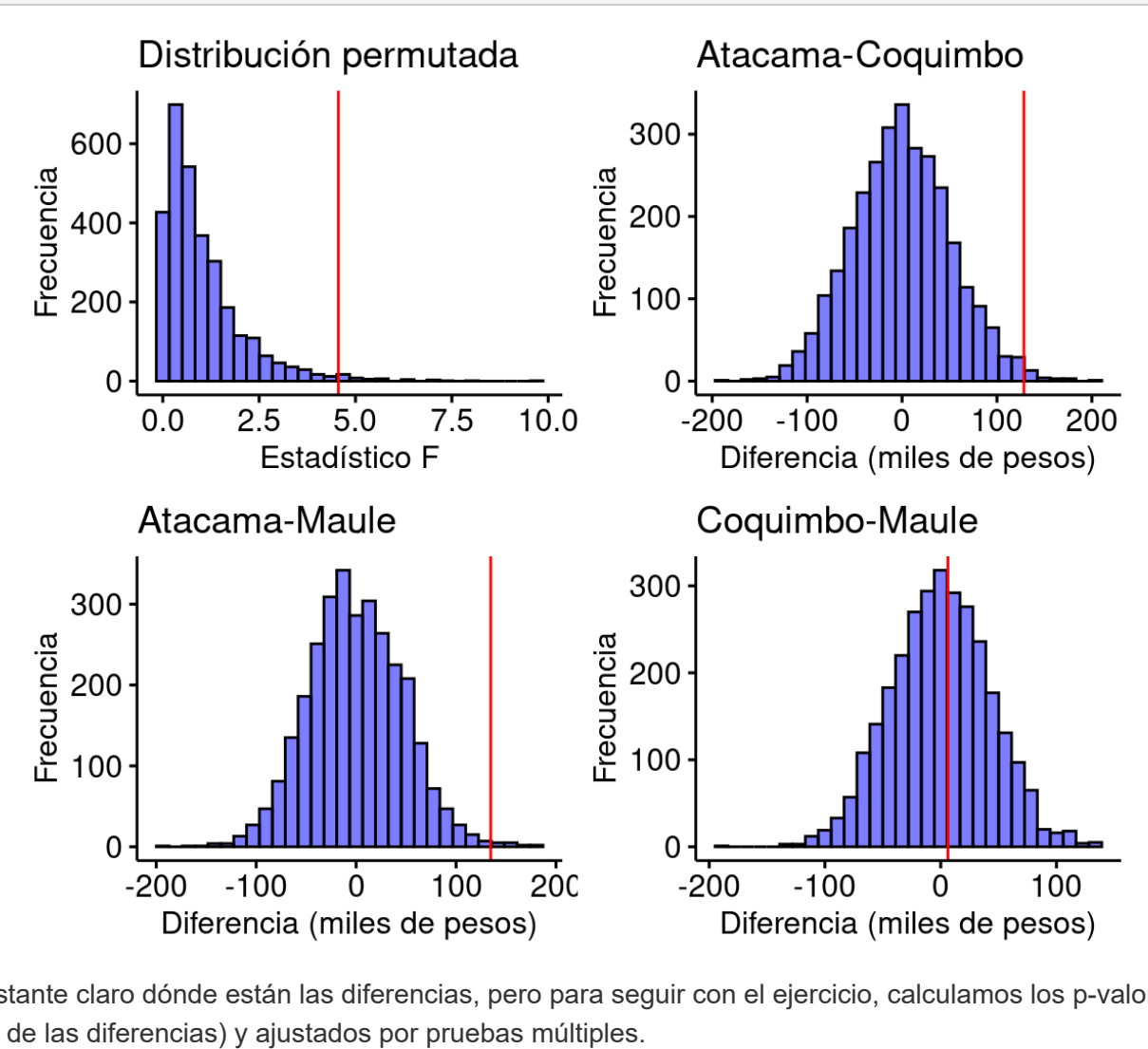
```
Atacama - Coquimbo: 128427
Atacama - Maule: 134652
Coquimbo - Maule: 6225
```

Obtenemos las distribuciones bootstrap para cada diferencia.

```
dist.boot.dif.A.C <- sapply(1:B,
                            function(i) get.dif.medias(muestra2b,
                                                       re.iAtacama[[i]],
                                                       re.iCoquimbo[[i]]))
dist.boot.dif.A.M <- sapply(1:B,
                            function(i) get.dif.medias(muestra2b,
                                                       re.iAtacama[[i]],
                                                       re.iMaule[[i]]))
dist.boot.dif.C.M <- sapply(1:B,
                            function(i) get.dif.medias(muestra2b,
                                                       re.iCoquimbo[[i]],
                                                       re.iMaule[[i]]))
```

Y las graficamos (en miles de pesos).

```
p3a <- gghistogram(data.frame(Diferencia = dist.boot.dif.A.C / 1000), x = "Diferencia",
                   title = "Atacama-Coquimbo",
                   xlab = "Diferencia (miles de pesos)", ylab = "Frecuencia",
                   bins = 30, fill = "blue")
p3a <- p3a + geom_vline(xintercept = dif.obs.A.C / 1000, colour="red")
p3b <- gghistogram(data.frame(Diferencia = dist.boot.dif.A.M / 1000), x = "Diferencia",
                   title = "Atacama-Maule",
                   xlab = "Diferencia (miles de pesos)", ylab = "Frecuencia",
                   bins = 30, fill = "blue")
p3b <- p3b + geom_vline(xintercept = dif.obs.A.M / 1000, colour="red")
p3c <- gghistogram(data.frame(Diferencia = dist.boot.dif.C.M / 1000), x = "Diferencia",
                   title = "Coquimbo-Maule",
                   xlab = "Diferencia (miles de pesos)", ylab = "Frecuencia",
                   bins = 30, fill = "blue")
p3c <- p3c + geom_vline(xintercept = dif.obs.C.M / 1000, colour="red")
p3 <- ggarrange(p2, p3a, p3b, p3c, nrow = 2, ncol = 2)
print(p3)
```



En los gráficos queda bastante claro dónde están las diferencias, pero para seguir con el ejercicio, calculamos los p-valores de pruebas bilaterales (usando el valor absoluto de las diferencias) y ajustados por pruebas múltiples.

```
valor_p.A.C <- (sum(abs(dist.boot.dif.A.C) > abs(dif.obs.A.C)) + 1) / (B + 1)
valor_p.A.M <- (sum(abs(dist.boot.dif.A.M) > abs(dif.obs.A.M)) + 1) / (B + 1)
valor_p.C.M <- (sum(abs(dist.boot.dif.C.M) > abs(dif.obs.C.M)) + 1) / (B + 1)
valores_p.adj <- p.adjust(c(valor_p.A.C, valor_p.A.M, valor_p.C.M), method = "BH")

cat("Valores p de pruebas bilaterales:\n")
cat("Atacama - Coquimbo:", round(valores_p.adj[1], 3), "\n")
cat("Atacama - Maule   :", round(valores_p.adj[2],3), "\n")
cat("Coquimbo - Maule  :", round(valores_p.adj[3], 3), "\n")
```

```
Valores p de pruebas bilaterales:
Atacama - Coquimbo: 0.02 
Atacama - Maule   : 0.02 
Coquimbo - Maule  : 0.885 
```

También podemos usar las remuestras para estimar los intervalos de confianza de las diferencias de las medias de ingresos per cápita entre las regiones observadas en en estudio.

```
dist.boot.dif.obs.A.C <- sapply(1:B,
                                function(i) get.dif.medias(muestra2,
                                                           re.iAtacama[[i]],
                                                           re.iCoquimbo[[i]]))
dist.boot.dif.obs.A.M <- sapply(1:B,
                                function(i) get.dif.medias(muestra2,
                                                           re.iAtacama[[i]],
                                                           re.iMaule[[i]]))
dist.boot.dif.obs.C.M <- sapply(1:B,
                                function(i) get.dif.medias(muestra2,
                                                           re.iCoquimbo[[i]],
                                                           re.iMaule[[i]]))

ci.dif.obs.A.C <- quantile(dist.boot.dif.obs.A.C, c(0.025, 0.975))
ci.dif.obs.A.M <- quantile(dist.boot.dif.obs.A.M, c(0.025, 0.975))
ci.dif.obs.C.M <- quantile(dist.boot.dif.obs.C.M, c(0.025, 0.975))

cat("Intervalos de 95% confianza:\n")
cat("Atacama - Coquimbo: [", round(ci.dif.obs.A.C[1], 3), ", ",
                          round(ci.dif.obs.A.C[2], 3), "]\n", sep = "")
cat("Atacama - Maule   : [", round(ci.dif.obs.A.M[1], 3), ", ",
                          round(ci.dif.obs.A.M[2], 3), "]\n", sep = "")
cat("Coquimbo - Maule  : [", round(ci.dif.obs.C.M[1], 3), ", ",
                          round(ci.dif.obs.C.M[2], 3), "]\n", sep = "")
```

```
Intervalos de 95% confianza:
Atacama - Coquimbo: [26630.42, 242873.9]
Atacama - Maule   : [29800.36, 248522.9]
Coquimbo - Maule  : [-61347.28, 68950.19]
```

Así, llegamos a la siguiente conclusión.

En base al procedimiento post-hoc, podemos concluir con 95% de confianza que no hay diferencias significativas en el ingreso per cápita promedio de las regiones de Coquimbo y del Maule (IC $95\% = [-\$61.347, \$68.950]$; $p = 0{,}885$), pero que estos son significativamente menores al ingreso per cápita promedio de la región de Atacama (Coquimbo: IC $95\% = [\$26.630, \$242.874]$; $p = 0{,}020$; Maule: IC $95\% = [\$29.800, \$248.523]$; $p = 0{,}020$).

Referencias

Fisher, N. I., & Hall, P. (1990). On bootstrap hypothesis testing. Australian Journal of Statistics, 32(2), 177-190.

Hall, P., & Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. Biometrics, 757-762.

Martin, M. A. (2007). Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties. Computational Statistics & Data Analysis, 51(12), 6321-6342.