Ep08

Grupo 7

2024-11-19

¿Existen diferencias significativas entre la cantidad de personas asalariadas y casadas entre las comunas de maipú y puente alto?

```
library(dplyr)
```

```
##
## Adjuntando el paquete: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
set.seed(10)
datos <- read.csv("EP08 Datos CASEN 2017.csv")

datosP1 <- datos %>% filter(ch1=="Asalariado", comuna == c("Puente Alto","Maipú"), ecivil=="Casado(a)")


datosP1Final <- datos %>% filter(comuna == c("Puente Alto","Maipú"))  %>% mutate(ecivil = (ecivil=="Casa

diferenciaProporciones <- 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 son Asalariado(a) y Casado(a):\n")
    cat("Maipu:", round(ph, 4), "\n")
    cat("Puente Alto:", round(pm, 4), "\n")
  }
  return(ph - pm)
}
dif <- diferenciaProporciones(datosP1Final,TRUE)
```

```
##              ecivil
## comuna        FALSE TRUE
##   Maipú          51    8
##   Puente Alto    56   10
##
## Proporción de personas que son Asalariado(a) y Casado(a):
## Maipu: 0.1356
## Puente Alto: 0.1515
```

Se proponen las siguientes hipótesis:

H0: Las proporciones de personas Asalariadas y Casadas que son de las comunas de Maipú y Puente alto son iguales: $p_M - p_P = 0$

HA: Las proporciones de personas Asalariadas y Casadas que son de las comunas de Maipú y Puente alto son diferentes: $p_M - p_P \neq 0$

Seteamos la cantidad de remuestreos en 4000, y definimos una funcion que calcula la diferencia de las proporciones.

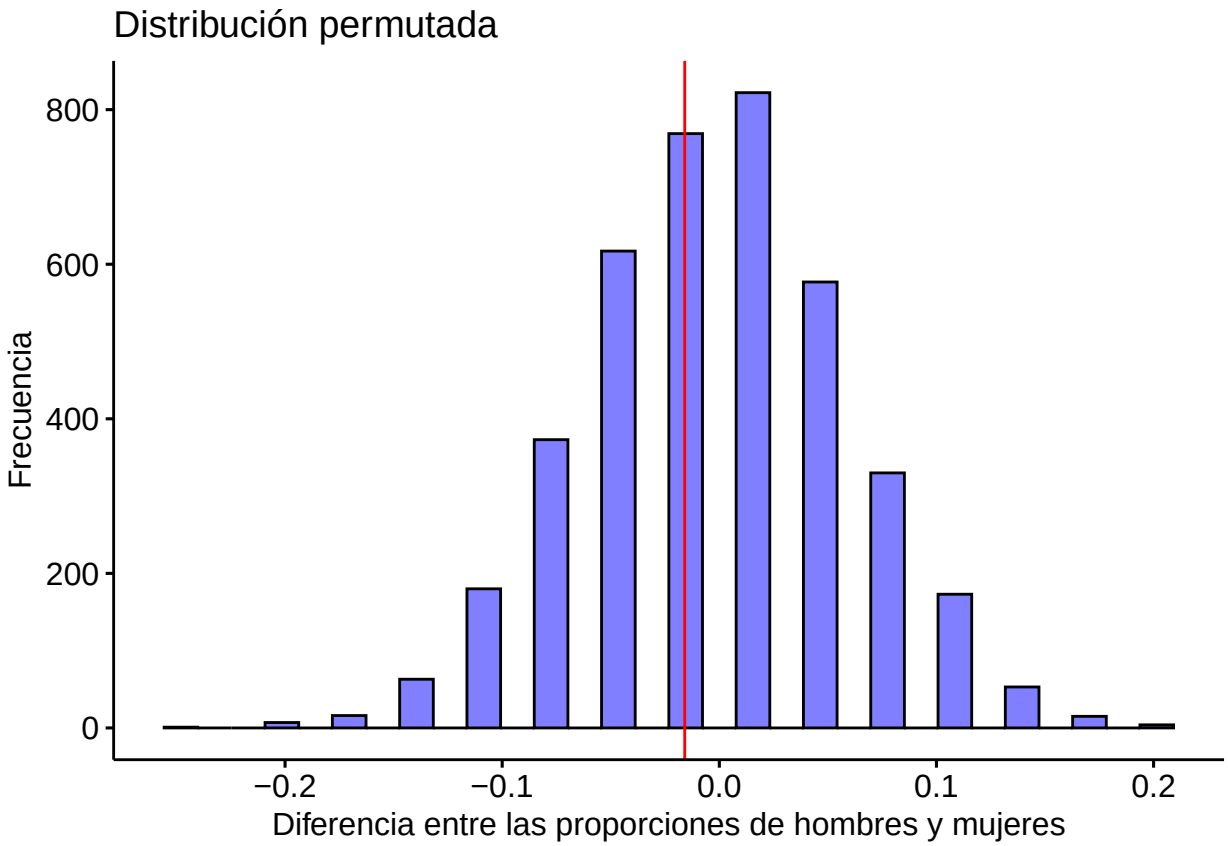
```
R <- 4000
set.seed(10)
permutaciones <- lapply(1:R,function(i) sample(1:125))

get.prop.dif.perm <- function(indices, df, verbose = FALSE)
{
  df.nuevo <- data.frame(comuna = df[indices, "comuna"], df[["ecivil"]])
  diferenciaProporciones(df.nuevo, verbose)
}
distribucion <- sapply(permutaciones, get.prop.dif.perm, datosP1Final)
```

Revisamos el resultado de la distribución resultante.

```
library(ggpubr)
```

```
## Cargando paquete requerido: ggplot2
```

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



Podemos ver luego de graficar, que la diferencia de las proporciones observada se encuentra cercana a 0. Ahora calcularemos el intervalo de confianza de 95% y el valor de p.

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

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

```
## IC 95%: [-0.112, 0.112]
```

```
cat("P-valor:", round(valor_p1, 3))
```

```
## P-valor: 0.808
```

R: Entonces podemos concluir con un 95% de confianza que no existe suficiente evidencia para rechazar H0. Por lo tanto no es posible descartar que las proporciones de personas asalariadas y casadas en las comunas de maipu y puente alto sean distintas.

Pregunta 2: ¿Existen diferencias significativas en los ingresos entre los hombres Casados, Solteros o Conviviente de la provincia de Santiago, de entre 25 a 40 años?

Se proponen las siguientes Hipotesis:

H0: No existen diferencias significativas en el ingreso promedio entre hombres, de entre 25 y 40 años de la provincia de Santiago, de estado civil casado, soltero o conviviente

HA: Al menos uno de los grupos tiene diferencias significativas en el ingreso promedio respecto a los otros grupos.

Primero se filtran los datos y se verifica si la condición de homocedasticidad se cumple.

```
library(dplyr)
library(WRS2)
```

```
## Warning: package 'WRS2' was built under R version 4.4.2
```

```
library(car)
```

```
## Cargando paquete requerido: carData
```

```
##
## Adjuntando el paquete: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
# Filtrar y limpiar los datos
datos_filt <- datos %>%
  filter(provincia == "Santiago",
         sexo == "Hombre",
         edad >= 25, edad <= 40,
         ecivil %in% c("Casado(a)", "Soltero(a)", "Conviviente o pareja sin acuerdo de unión civil")) %
  select(ecivil, ytot)

set.seed(123)
muestra <- datos_filt %>%
  sample_n(230)

# Verificar homocedasticidad
levene <- car::leveneTest(ytot ~ ecivil, data = muestra, center = "median")
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
cat("Test de Levene para homocedasticidad:\n")
```

```
## Test de Levene para homocedasticidad:
```

```
print(levene)
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value   Pr(>F)
## group   2  4.9926 0.007553 **
##       227
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se observa que p < 0.05 en el test de Levene, por lo que no se cumple con la condición de homocedasticidad. Además de esto, se tiene que los tamaños muestrales son bastante diferentes, por lo que se procede a utilizar el método de ANOVA rubusta con bootstrapping. Se realizan 3999 repeticiones bootstrapping y se fija un alpha $= 0.05$:

```
# ANOVA robusta con bootstrapping
set.seed(568)
anova <- t1waybt(ytot ~ ecivil, data = muestra, tr = 0.2, nboot = 3999)
print(anova)
```

```
## Call:
## t1waybt(formula = ytot ~ ecivil, data = muestra, tr = 0.2, nboot = 3999)
##
## Effective number of bootstrap samples was 3999.
##
## Test statistic: 5.5513
## p-value: 0.01725
## Variance explained: 0.108
## Effect size: 0.328
```

```
cat("\nResultados del ANOVA robusta con bootstrapping:\n")
```

```
##
## Resultados del ANOVA robusta con bootstrapping:
```

```
print(anova)
```

```
## Call:
## t1waybt(formula = ytot ~ ecivil, data = muestra, tr = 0.2, nboot = 3999)
##
## Effective number of bootstrap samples was 3999.
##
## Test statistic: 5.5513
## p-value: 0.01725
## Variance explained: 0.108
## Effect size: 0.328
```

Vemos que p < 0.05, por lo que se rechaza la hipótesis nula en favor de la alternativa, y se concluye que hay diferencias significativas en el ingreso promedio entre hombres, de entre 25 y 40 años de la provincia de Santiago, entre al menos un par de los grupos elegidos.

Para poder hallar cuáles son los grupos diferentes, se procede a realizar un análisis Post-hoc:

```
# Post-hoc con bootstrapping para diferencias entre pares de grupos
post_hoc <- lincon(ytot ~ ecivil, data = muestra, tr = 0.2, alpha = 0.05)

cat("\nResultados del análisis post-hoc con bootstrapping:\n")
```

```
##
## Resultados del análisis post-hoc con bootstrapping:
```

```r
print(post_hoc)
```

```
## Call:
## lincon(formula = ytot ~ ecivil, data = muestra, tr = 0.2, alpha = 0.05)
##
##                                                                  psihat
## Casado(a) vs. Conviviente o pareja sin acuerdo de unión civil  235981.7
## Casado(a) vs. Soltero(a)                                       416210.4
## Conviviente o pareja sin acuerdo de unión civil vs. Soltero(a) 180228.7
##                                                                  ci.lower
## Casado(a) vs. Conviviente o pareja sin acuerdo de unión civil  -171708.78
## Casado(a) vs. Soltero(a)                                         40018.60
## Conviviente o pareja sin acuerdo de unión civil vs. Soltero(a)  -26141.11
##                                                                ci.upper p.value
## Casado(a) vs. Conviviente o pareja sin acuerdo de unión civil  643672.3 0.15862
## Casado(a) vs. Soltero(a)                                       792402.3 0.02705
## Conviviente o pareja sin acuerdo de unión civil vs. Soltero(a) 386598.5 0.07309
```

Vemos que el valor p es menor a 0.05 si comparamos a las personas casadas y a las solteras, esto significa que ambos grupos tienen diferencias significativas en sus ingresos. Mientras que no hay evidencia para afirmar que existen diferencias significativas entre los otros grupos, pues el valor p es mayor que 0.05.