A2 - Análitica descriptiva e inferencial

Solución

Semestre 2021.2

Índice

Introducción

En esta actividad nos introducimos en la inferencia estadística. Para ello, usaremos el conjunto de datos `CensusIncome_clean.csv` que se ha preprocesado en la actividad anterior. Este conjunto de datos surge de un censo, donde se registra la información demográfica, personal y laboral de una muestra de la población. El conjunto de datos original se ha obtenido de la base de datos de la web Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Adult.

Las variables del conjunto de datos son:

- *CS_ID:* Identificador del individuo.
- *age:* Edad del individuo.
- *workclass*: Categorización del individuo en base al perfil laboral.
- *fnlwgt*: Número de unidades de la población objetivo que representa la unidad respondiente.
- *education_num*: Número de años de formación educativa del individuo.
- *marital_status*: Estado civil del individuo.
- *occupation*: Categorización del individuo en base a la tipología del trabajo.
- *relationship*: Parentesco de la unidad que responde de la familia.
- *race*: Grupo racial al que pertenece el individuo.
- *gender*: Género del individuo.
- *hours_per_week*: Horas por semana trabajadas por el individuo.
- *income*: Salario (anual) del individuo.
- *education_cat*: Nivel de educación.

Nota importante a tener en cuenta para entregar la actividad:

- Es necesario entregar el archivo Rmd y el fichero de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se debe respetar la misma numeración de los apartados que el enunciado.
- No se pueden realizar listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden usar las funciones **head** y **tail** que sólo muestran unas líneas del fichero de datos.
- Se valora la precisión de los términos utilizados (hay que usar de manera precisa la terminología de la estadística).
- Se valora también la concisión en la respuesta. No se trata de hacer explicaciones muy largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos de manera clara y concisa.

---

1. Lectura del fichero y preparación de los datos

Leed el fichero `CensusIncome_clean.csv` y guardad los datos en un objeto con identificador denominado *censo*. A continuación, verificad que los datos se han cargado correctamente.

Solución:

```
# Cargamos el conjunto de datos
censo<-read.csv("CensusIncome_clean.csv", sep=",", stringsAsFactors=TRUE)

# Obtenemos las dimensiones del conjunto de datos
dim(censo)
```

```
## [1] 32553    12
```

```
# Verificamos que no ha habido errores de lectura
problems(censo)

# Listado de variables con sus tipos de datos
str(censo)
```

```
## 'data.frame':    32553 obs. of  12 variables:
##  $ CS_ID         : Factor w/ 32553 levels "CS1","CS10","CS100",..: 1 11112 22221 25900 27011 28122 29
##  $ age           : int  50 38 53 28 37 49 52 31 42 37 ...
##  $ workclass     : Factor w/ 4 levels "Government","Other/Unknown",..: 4 3 3 3 3 3 4 3 3 3 ...
##  $ education_num : int  13 9 7 13 14 5 9 14 13 10 ...
##  $ marital_status: Factor w/ 5 levels "D","M","S","W",..: 2 1 2 2 2 2 2 3 2 2 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
##  $ occupation    : Factor w/ 6 levels "Blue-Collar",..: 6 1 1 3 6 5 6 3 6 6 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 3 3 5 3 5 5 5 3 ...
##  $ gender        : Factor w/ 2 levels "f","m": 2 2 2 1 1 1 2 1 2 2 ...
##  $ hours_per_week: num  13.5 40 40.5 40 40.5 16 45 50.5 40 50 ...
##  $ income        : num  52 51.7 50.1 44.2 0.1 ...
##  $ education_cat : Factor w/ 4 levels "Postuniversitaria",..: 4 3 3 4 1 2 3 1 4 4 ...
```

```
# Inspeccionamos 10 observaciones elegidas de manera aleatoria
set.seed(1)
censo[sample(nrow(censo), 10), ]
```

```
##         CS_ID age     workclass education_num marital_status  relationship
## 17401 CS17407  62       Private             6              D       Husband
## 24388 CS24395  76 Self-Employed            13              M       Husband
## 4775   CS4776  31       Private            10              M     Own-child
## 26753 CS26760  40       Private             9              M     Own-child
## 13218 CS13224  46       Private            13              D Not-in-family
## 26109 CS26116  26    Government             9              S     Unmarried
## 29143 CS29150  19       Private            10              S       Husband
## 10539 CS10545  76       Private             9              M Not-in-family
## 8462   CS8467  44    Government            14              M Not-in-family
## 4050   CS4051  29 Self-Employed            13              S       Husband
##         occupation  race gender hours_per_week income     education_cat
## 17401      Service White      f           40.0  32.08          Primaria
## 24388 White-Collar White      m           55.5  55.96     Universitaria
## 4775  Professional White      m           50.0  54.74     Universitaria
## 26753 White-Collar White      f           38.5  48.61        Secundaria
## 13218 Professional White      f           40.5  44.20     Universitaria
## 26109      Service White      m           16.0  51.27        Secundaria
## 29143        Sales White      f           25.5  36.08     Universitaria
## 10539  Blue-Collar White      m           50.5  56.37        Secundaria
## 8462  Professional White      m           40.5  58.70 Postuniversitaria
## 4050  Professional Black      m           40.0  38.05     Universitaria
```

```
# Nombres de las variables
names(censo)
```

```
##  [1] "CS_ID"          "age"            "workclass"      "education_num"
##  [5] "marital_status" "relationship"   "occupation"     "race"
##  [9] "gender"         "hours_per_week" "income"         "education_cat"
```

2. Edad

Para empezar el análisis, nos interesa conocer el valor medio de la edad del censo, a partir de los datos de la muestra. Para ello, calculad el intervalo de confianza de la media de edad. Seguid los pasos que se especifican a continuación.

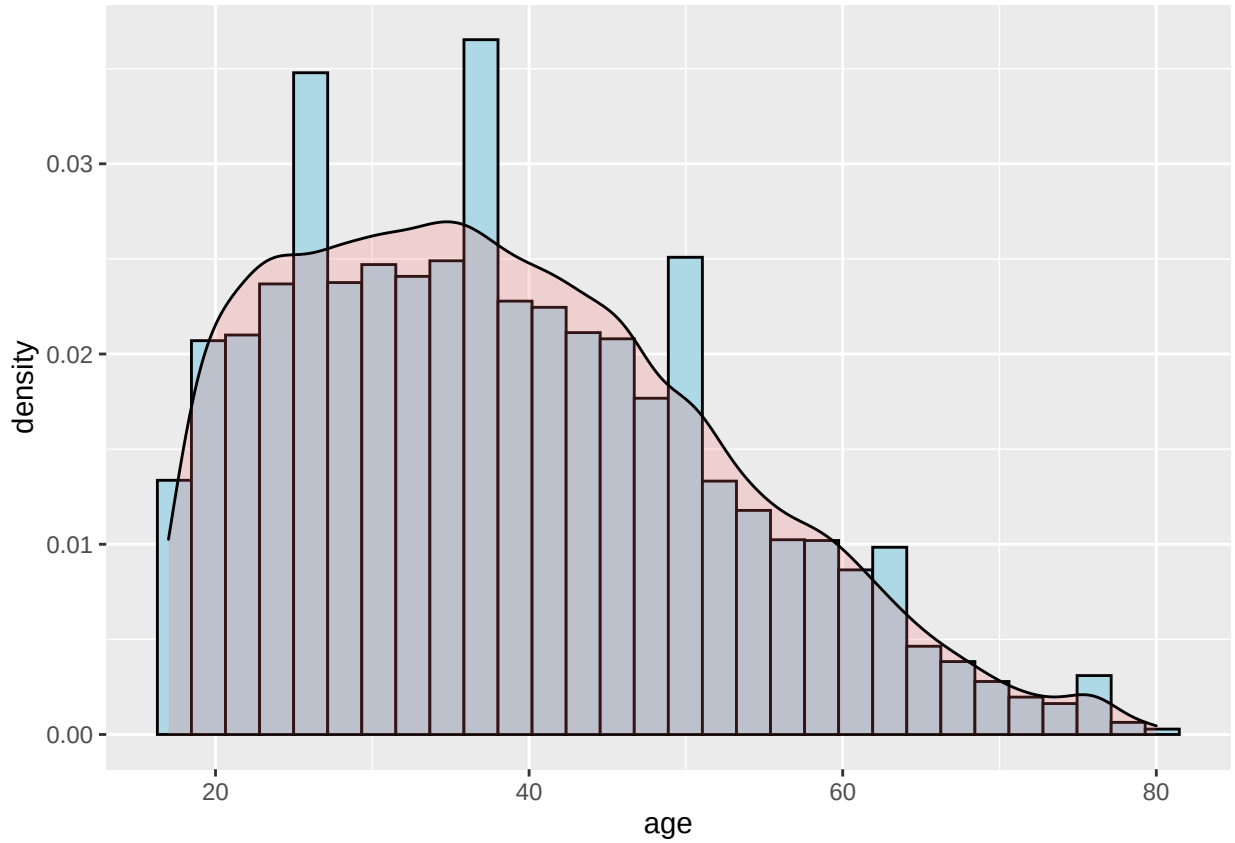
2.1. Distribución de edades

Visualizad gráficamente la distribución de la edad. Escoged el gráfico que sea más apropiado, considerando que se quiere conocer la distribución de la variable y si ésta sigue una distribución normal.

```
library(ggplot2)

ggplot( censo, aes(x=age)) +
geom_histogram( aes(y=..density..), colour="black", fill="lightblue")+
geom_density(alpha=.2, fill="#FF6666")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2.2. Normalidad

¿Podemos asumir normalidad para el cálculo del intervalo de confianza de la media de edad? Argumentad la respuesta.

Respuesta: Los datos, según muestra el diagrama anterior, no se distribuyen de forma normal. De todas formas, para calcular el intervalo de confianza de la media de edad, podemos aplicar el teorema del límite central (TLC). Según el TLC, la media de una variable de una muestra de tamaño grande (al menos, superior a 30) se comporta como una distribución normal.

2.3. Intervalo de confianza

Calculad manualmente el intervalo de confianza de la media de la variable age. Para ello, definid una función IC que reciba la variable, la confianza, y que devuelva un vector con los valores del intervalo de confianza.

La cabecera de la función es:

```
   IC <- function( x, NC ){}
```

Nota: No se pueden usar funciones como t.test para el cálculo. Sí podéis usar otras funciones básicas de R como *mean*, *qnorm*, *qt*, *pnorm*, *pt*, etcétera.

Respuesta:

```
IC <- function( x, NC ){
  n <- length(x)
  alfa <- 1-(NC/100)
  sd <- sd(x)
  SE <- sd / sqrt(n)

  t <- qt( alfa/2, df=n-1, lower.tail=FALSE )
  L <- mean(x) - t*SE
  U <- mean(x) + t*SE
  return (c(L, U))
}
```

Nota: Calculamos la función del intervalo de confianza. Asumimos distribución normal por el TLC, con el matiz de que no conocemos la varianza de la población y por tanto usamos la varianza muestral para aproximar la varianza de la población. En este caso debemos aplicar la distribución t de Student con n-1 grados de libertad. A la práctica, como el tamaño de muestra es suficientemente grande, la distribución t de Student es muy similar a la distribución normal. Por tanto, se aceptaría también que se use qnorm en lugar de qt.

2.4. Cálculos

Calculad el intervalo de confianza al 90 % y 95 %. Comparad los resultados.

```
  ic95<-IC(censo$age, 95)
  ic90<-IC(censo$age, 90)

  ic95; ic90
```

```
## [1] 38.40276 38.69698
```

```
## [1] 38.42642 38.67333
```

```
  #Comprobación
  t.test( censo$age )
```

```
##
```

```
##  One Sample t-test
##
## data:  censo$age
## t = 513.63, df = 32552, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  38.40276 38.69698
## sample estimates:
## mean of x
##  38.54987
```

Como se puede observar, el intervalo de confianza 95 % es más amplio que el intervalo de confianza al 90 %. Ver explicación en el siguiente subapartado.

2.5. Interpretación

Explicad cómo se interpreta el intervalo de confianza a partir de los resultados obtenidos.

Respuesta: La interpretación del intervalo de confianza es que si realizamos un muestreo elevado de muestras de la población, el C % (95 % o 90 %) de los intervalos de confianza obtenidos de estas muestras contienen el valor de la media poblacional de edad.

---

3. Salario

Vamos a investigar ahora el salario de la población. En particular, nos preguntamos si en media, el salario de las personas Self-Employed es inferior al del resto de modalidades. Seguid los pasos que se especifican a continuación.

3.1. Pregunta de investigación

Formulad la pregunta de investigación.

Respuesta: ¿El salario medio poblacional de las personas Self-Employed es inferior al salario medio poblacional del resto de perfiles?

3.2. Hipótesis

Escribid las hipótesis (hipótesis nula e hipótesis alternativa).

Respuesta:

$H_0 : \mu_{Self-Employed} = \mu_{Other}$

$H_1 : \mu_{Self-Employed} < \mu_{Other}$

3.3. Test a aplicar

Explicad qué tipo de test podéis aplicar dada la pregunta de investigación planteada y las características de la muestra. Justificad vuestra elección.

Nota: Podéis usar las funciones de R que consideréis necesarias para responder esta pregunta.

Respuesta:

Es un test de dos muestras sobre la media con varianzas desconocidas. Por el teorema del límite central, podemos asumir normalidad. Comprobamos igualdad de varianzas:

```r
SE <- censo[censo$workclass=="Self-Employed",]
nSE <- censo[censo$workclass!="Self-Employed",]
var.test( SE$income, nSE$income )
```

```
##
##  F test to compare two variances
##
## data:  SE$income and nSE$income
## F = 0.63242, num df = 3656, denom df = 28895, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6026580 0.6642595
## sample estimates:
## ratio of variances
##          0.6324152
```

El resultado del test es un valor p<0.05. Por tanto, rechazamos la hipótesis nula de igualdad de varianzas. En consecuencia, el test se corresponde con un test de dos muestras independientes sobre la media con varianzas desconocidas diferentes. El test es unilateral.

3.4. Cálculo

Calculad el test usando una función propia. Implementad una función que realice el cálculo del test y que podáis usar con distintos valores de nivel de confianza.

Calculad el contraste para un nivel de confianza del 95 % y del 90 %. Mostrad los resultados (valor observado, crítico y valor p) en una tabla.

Nota: No se pueden usar funciones como *t.test* para el cálculo. Sí podéis usar otras funciones básicas de R como *mean*, *qnorm*, *qt*, *pnorm*, *pt*, etcétera.

```r
my.ttest <- function( x1, x2, alternative="bilateral", CL=95 ){
mean1<-mean(x1); n1<-length(x1); sd1<-sd(x1)
mean2<-mean(x2); n2<-length(x2); sd2<-sd(x2)
alfa <- (1-CL/100)

#varianzas diferentes
Sb <- sqrt( sd1^2/n1 + sd2^2/n2 )
denom <- ( (sd1^2/n1)^2/(n1-1) + (sd2^2/n2)^2/(n2-1))
df <- ( (sd1^2/n1 + sd2^2/n2)^2 ) / denom
#valor observado
t<- (mean1-mean2) / Sb

if (alternative=="bilateral"){
tcritical <- qt( alfa/2, df, lower.tail=FALSE ) #two sided
pvalue<-pt( abs(t), df, lower.tail=FALSE )*2 #two sided
}
else if (alternative=="<"){
tcritical <- qt( alfa, df, lower.tail=TRUE )
pvalue<-pt( t, df, lower.tail=TRUE )
}
else{ #(alternative==">")
tcritical <- qt( alfa, df, lower.tail=FALSE )
pvalue<-pt( t, df, lower.tail=FALSE )
}
#Guardamos el resultado en un named vector
```

Cuadro 1: ¿El salario medio de Self-Employed es menor que el del resto? NC=95

	x
mean1	49.391225
mean2	48.671835
t	6.889720
tcritical(95)	-1.645145
pvalue	1.000000
df	5239.123790

Cuadro 2: ¿El salario medio de Self-Employed es menor que el del resto? NC=90

	x
mean1	49.391225
mean2	48.671835
t	6.889720
tcritical(90)	-1.281713
pvalue	1.000000
df	5239.123790

```
info<-c(mean1, mean2, t,tcritical,pvalue,df)
names(info)<-c("mean1", "mean2", "t","tcritical", "pvalue", "df")
return (info)
}

test.income.95<-my.ttest( SE$income, nSE$income, "<", 95)
names(test.income.95)<-c("mean1", "mean2", "t","tcritical(95)", "pvalue", "df")
test.income.95 %>% kable( caption="¿El salario medio de Self-Employed es menor que el del resto? NC=95"
```

```
test.income.90<-my.ttest( SE$income, nSE$income, "<", 90)
names(test.income.90)<-c("mean1", "mean2", "t","tcritical(90)", "pvalue", "df")

test.income.90 %>% kable( caption="¿El salario medio de Self-Employed es menor que el del resto? NC=90"
```

```
#Comprobación
t.test( SE$income, nSE$income, alternative="less", conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  SE$income and nSE$income
## t = 6.8897, df = 5239.1, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.8911684
## sample estimates:
## mean of x mean of y
##  49.39123  48.67183
```

3.5. Conclusión

A partir de los resultados obtenidos, dad respuesta a la pregunta de investigación.

Respuesta: El valor p del test es 1. No podemos rechazar la hipótesis nula. Por tanto, no podemos afirmar que el salario medio de los profesionales autónomos sea inferior al del resto.

A partir del valor observado, se llega a la misma conclusión. El valor observado es 6.8897204 y el valor crítico para un nivel de confianza del 95 % es -1.6451445. Por tanto, no podemos rechazar la hipótesis nula de igualdad de salarios medios, con un nivel de confianza del 95 %.

Con un nivel de confianza del 90 %, la conclusión es la misma.

---

4. Proporción de Self-Employed

Nos preguntamos si el porcentaje de Self-Employed en la población es superior al 10 %. Aplicad el test necesario para dar respuesta a esta pregunta. Seguid los pasos que se indican a continuación.

4.1. Pregunta

Formulad la pregunta de investigación que se plantea en esta sección.

¿La proporción de Self-Employed en la población es superior a 0.1?

4.2. Hipótesis

Escribid la hipótesis nula y la hipótesis alternativa.

$H_0 : p_{SelfEmployed} = 0.1$

$H_1 : p_{SelfEmployed} > 0.1$

4.3. Análisis visual

Representad de forma gráfica la proporción de Self-Employed en la muestra.

```
df <- data.frame( se=c("SE","NSE"), n=c(nrow(SE), nrow(nSE)))
ggplot(df) +
aes( x="", y=n, fill= se) +
geom_bar(stat="identity", width=1) +
coord_polar("y", start=0)+
labs(title = "Proporción de Self-Employed en la muestra")
```

4.4. Contraste

Explicad qué tipo de contraste podéis aplicar dada la pregunta de investigación planteada y las características de la muestra. Justificad vuestra elección.

Es un test de la proporción sobre una muestra. Aplicamos el contraste de una muestra sobre la proporción, asumiendo la aproximación de la distribución binomial a una distribución normal para muestras grandes. El contraste es unilateral.

4.5. Cálculo

Calculad el test usando una función propia. Podéis crear una función que reciba los parámetros necesarios y el nivel de confianza. Luego, calculad el contraste, llamando esta función, con nivel de confianza del 95 %.

Mostrad los resultados (valor observado, crítico y valor p) en una tabla.

Nota: No podéis usar *prop.test* o funciones ya implementadas en R. Sí podéis usar *qnorm*, *qt*, etcétera.

```
my.proptest<-function( p, p0, n, alternative="bilateral", CL=95 ){
  z <- (p-p0)/sqrt( (p0*(1-p0)/n))
  alfa <- 1 - CL/100
  if (alternative=="bilateral"){
    pvalue <- pnorm( abs(z), lower.tail=FALSE)*2
    zcritical <- qnorm( alfa/2, lower.tail=FALSE )
  }
  else if (alternative==">"){
    pvalue <- pnorm( z, lower.tail=FALSE)
    zcritical <- qnorm( alfa, lower.tail=FALSE )
```

Cuadro 3: ¿La proporción de Self-Employed en la población es superior a 0.1?

	x
p	0.1123399
p0	0.1000000
zobs	7.4213868
zcrit(95)	1.6448536
valor p	0.0000000

```
  }
  else{ #alternative is less
    pvalue <- pnorm(z, lower.tail=TRUE)
    zcritical <- qnorm( alfa, lower.tail=TRUE )
  }

  info<-c(p,p0,z, zcritical, pvalue)
  names(info)<-c("p","p0","z", "zcritical", "pvalue")
  return (info)
}

test.pr <- my.proptest( nrow(SE)/nrow(censo), p0=0.1, nrow(censo), alternative=">")
names(test.pr)<-c("p", "p0", "zobs", "zcrit(95)", "valor p")
test.pr %>% kable(caption="¿La proporción de Self-Employed en la población es superior a 0.1?") %>% ka
```

```
#Comprobación
prop.test(x=nrow(SE), n=nrow(censo), p=0.1, alternative="greater", correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  nrow(SE) out of nrow(censo), null probability 0.1
## X-squared = 55.077, df = 1, p-value = 5.795e-14
## alternative hypothesis: true p is greater than 0.1
## 95 percent confidence interval:
##  0.1094932 1.0000000
## sample estimates:
##         p
## 0.1123399
```

4.6. Conclusión

A partir de los resultados obtenidos, dad respuesta a la pregunta de investigación.

El valor p obtenido es 0. Por tanto, podemos rechazar la hipótesis nula a favor de la alternativa, concluyendo que la proporción de Self-Employed en la población es superior a 0.1 con un nivel de confianza del 95 %.

---

5. Proporción de Self-Employed en mujeres y hombres

Nos preguntamos si la proporción de Self-Employed es menor entre las mujeres que entre los hombres en la población. Para dar respuesta a esta pregunta, seguid los pasos que se indican a continuación.

5.1. Pregunta de investigación

Formulad la pregunta de investigación que se plantea en esta sección.

¿La proporción de Self-Employed entre las mujeres es inferior que entre los hombres en la población?

5.2. Análisis visual

Representad de forma gráfica la proporción de Self-Employed en la muestra de hombres y mujeres respectivamente.

```
df <- data.frame(SE=c("SE", "SE", "NSE", "NSE"),
                 gender=rep(c("F","M"),2),
                 Q=c( nrow( SE[SE$gender=="f",] ),
                        nrow( SE[SE$gender=="m",] ),
                        nrow( nSE[nSE$gender=="f",]),
                        nrow( nSE[nSE$gender=="m",])
                 )
)

df
```

```
##    SE gender     Q
## 1  SE      F   534
## 2  SE      M  3123
## 3 NSE      F 10233
## 4 NSE      M 18663
```

```
ggplot( df ) +
aes(x=gender, y=Q, fill=SE) +
geom_bar(stat="identity")+
labs(title = "Proporción de Self-Employed según el género")
```

5.3. Hipótesis

Escribid la hipótesis nula y la hipótesis alternativa.

$H_0 : p_{SelfEmployedF} = p_{SelfEmployedM}$

$H_1 : p_{SelfEmployedF} < p_{SelfEmployedM}$

5.4. Test

Explicad qué tipo de test podéis aplicar dada la pregunta de investigación planteada y las características de la muestra. Justificad vuestra elección.

Aplicamos un contraste sobre la diferencia de proporciones, asumiendo la aproximación de la distribución binomial a una normal para muestras grandes. El contraste es unilateral.

5.5. Cálculo

Calculad el test usando una función propia. Al igual que en apartados anteriores, se recomienda definir una función que realice el cálculo y que reciba los parámetros necesarios.

Calculad el contraste para un nivel de confianza del 97 %. Mostrad los resultados (valor observado, crítico y valor p) en una tabla.

Nota: No podéis usar funciones como *prop.test* o funciones ya implementadas en R para el contraste. Sí podéis usar funciones básicas como *qnorm*, *qt*, etcétera.

```
my.proptest2 <-function ( x1,x2,n1,n2, alternative="bilateral", CL=95){
 p1 <- x1/n1
```

Cuadro 4: ¿La proporción de Self-Employed es inferior en mujeres que en hombres en la población?

	x
p Fem	0.0495960
p Male	0.1433489
zobs	-25.2020125
zcrit(97)	-1.8807936
valor p	0.0000000

```
 p2 <- x2/n2
 alfa <- 1 - CL/100
 p<-(n1*p1 + n2*p2) / (n1+n2)
 zobs <- (p1-p2)/( sqrt(p*(1-p)*(1/n1+1/n2)) )
 if (alternative=="bilateral"){
    pvalue <- pnorm( abs(zobs), lower.tail=FALSE)*2
    zcrit <- qnorm( alfa/2, lower.tail=FALSE )
  }
  else if (alternative==">"){
    pvalue <- pnorm( zobs, lower.tail=FALSE)
    zcrit <- qnorm( alfa, lower.tail=FALSE )
  }
  else{ #alternative is less
    pvalue <- pnorm(zobs, lower.tail=TRUE)
    zcrit <- qnorm( alfa, lower.tail=TRUE )
  }
 result <-  c(p1,p2,zobs, zcrit, pvalue)
 names(result) <- c("p1", "p2", "zobs","zcrit", "pvalue")
 return (result)
}

nFSE <- nrow(SE[SE$gender=="f",])
nF   <- nrow( censo[censo$gender=="f",])
nMSE <- nrow(SE[SE$gender=="m",])
nM   <- nrow( censo[censo$gender=="m",])

test.pr2 <- my.proptest2(nFSE,nMSE,nF,nM, "<",97)
names(test.pr2)<-c("p Fem", "p Male", "zobs", "zcrit(97)", "valor p")
test.pr2 %>% kable(caption="¿La proporción de Self-Employed es inferior en mujeres que en hombres en la
```

```
#Validación con prop.test
success<-c( nFSE, nMSE)
nn<-c(nF,nM)
prop.test(success, nn, alternative="less", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 635.14, df = 1, p-value < 2.2e-16
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000 -0.0885477
```

```
## sample estimates:
##     prop 1     prop 2
## 0.04959599 0.14334894
```

5.6. Conclusión

A partir de los resultados obtenidos, proporcionad una respuesta a la pregunta de investigación.

El valor p obtenido es 0. Por tanto, podemos rechazar la hipótesis nula a favor de la alternativa, concluyendo que la proporción de Self-Employed en mujeres es inferior a la proporción de Self-Employed en hombres en la población, con un nivel de confianza del 97 %.

---

6. Dependencia Género - Self-Employed

Otra forma de abordar si existen diferencias en la proporción de Self-Employed según el género es realizando un test de independencia de dos variables cualitativas. Concretamente, nos preguntamos si el género y ser Self-Employed están relacionadas o se pueden considerar variables independientes. Las variables serían independientes si el género no influye en la proporción de Self-Employed, es decir, si no hay diferencias en las proporciones de Self-Employed según el género.

En esta sección se pide aplicar el test de independencia Chi cuadrado para evaluar si las variables género y Self-Employed son independientes. Seguid los pasos que se indican a continuación.

6.1. Pregunta de investigación

Formulad la pregunta de investigación.

¿Las variables género y Self-Employed están relacionadas?

6.2. Hipótesis

Escribid la hipótesis nula y alternativa.

H_0 : las variables género y Self-Employed son independientes.

H_1 : las variables género y Self-Employed están relacionadas.

6.3. Test

Describid brevemente en qué consiste el test chi cuadrado. Calculad la matriz de contingencia y mostrad sus valores.

El test Chi cuadrado se usa para evaluar la independencia (o dependencia) entre dos variables. Se basa en el cálculo de las frecuencias observadas en cada categoría de las variables analizadas. Estas frecuencias se comparan con los valores esperados de las variables si éstas fueran independientes. Si la diferencia entre las frecuencias observadas y las esperadas es suficientemente grande, entonces se considera que las variables no son independientes.

```
censo2<-censo
censo2$workclass <- ifelse( censo2$workclass=="Self-Employed", "Self-Emmployed", "other")
tab1 <- table(censo2$gender, censo2$workclass); tab1
```

```
##
##     other Self-Emmployed
##   f 10233            534
##   m 18663           3123
```

6.4. Cálculos

Realizad los cálculos del test Chi cuadrado, implementando una función propia. Calculad el contraste para un nivel de confianza de 97 %.

Nota: No podéis usar la función *chisq.test* de R. Sí podéis usar *pchisq* para consultar los valores de la distribución.

```
#cálculos propios
my.chisq <- function( tab, CL=95 ){
  N <- sum(tab)
  E11 <- sum(tab[1,])* sum(tab[,1]) / N
  E12 <- sum(tab[1,])* sum(tab[,2]) / N
  E21 <- sum(tab[2,])* sum(tab[,1]) / N
  E22 <- sum(tab[2,])* sum(tab[,2]) / N
  chi <-
  (tab[1,1] - E11)^2/E11 +
  (tab[1,2] - E12)^2/E12 +
  (tab[2,1] - E21)^2/E21 +
  (tab[2,2] - E22)^2/E22

  pvalue <- pchisq(chi, df=1, lower.tail=FALSE)

  result <- c(chi, pvalue)
  names(result)<-c("chi","pvalue")
  return (result)
}

test.ind <- my.chisq(tab1, CL=97); test.ind
```

```
##           chi        pvalue
##  6.351414e+02 3.807275e-140
```

```
#comprobación
chisq.test( tab1,correct=FALSE )
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 635.14, df = 1, p-value < 2.2e-16
```

6.5. Conclusión

Responded la pregunta de investigación planteada en este apartado. Relacionad el resultado con la aproximación de la sección anterior, donde se realiza un test sobre las proporciones.

El test obtiene un valor p<0.001. Por lo tanto, rechazamos la hipótesis nula y concluimos que se observa evidencia suficiente para afirmar que existe dependencia entre género y Self-Employed. El resultado es coherente con el obtenido en la sección anterior, donde se ha observado que la proporción de Self Employed entre las mujeres es inferior que entre los hombres a un nivel de confianza del 97 %. Por tanto, el género influye en la elección de autoempleo.

7. Resumen y conclusiones

Presentad una tabla con los resultados principales de cada sección: la pregunta de investigación planteada, los valores obtenidos del contraste y la conclusión obtenida en cada apartado. La tabla puede tener un formato

como el que se muestra a continuación (se aporta un ejemplo para la primera fila de datos).

N	Pregunta	Resultado (valor observado, crítico, valor p...)	Conclusión
2	Intervalo de confianza de edad al 95 %	(25.22,26.24)	El intervalo de confianza al 95 % es...
3	texto	valores	texto
4	texto	valores	texto
5	texto	valores	texto
6	texto	valores	texto

Respuesta:

N	Pregunta	Resultado (valor observado, crítico, valor p...)	Conclusión
2a	IC 95 % de edad	(38.4027638, 38.6969812)	El 95 % de los IC de infinitas muestras contienen el valor del parámetro.
2b	IC 90 % de edad	(38.4264161, 38.6733289)	El 90 % de los IC de infinitas muestras contienen el valor del parámetro.
3	¿El salario de las personas Self Employed es inferior al salario del resto de perfiles?	obs: 6.89; tcrit: -1.65; p: 1	El salario de Self-Employed no es inferior al resto NC=95 %
4	¿La proporción de SelfEmployed es superior a 0.1?	obs: 7.42; crit: 1.64; p: 0	La proporción de Self Employed es superior a 0.1 NC=95 %.
5	¿La proporción de Self Employed en mujeres es inferior a los hombres?	obs: -25.2 crit: -1.88 p: 0	Self Employed en mujeres es inferior a los hombres NC=97 %.
6	¿Las variables género y Self Employed están relacionadas?	chi: 635.14 p: 0	Las variables género y Self Employed están relacionadas NC=97 %.

8. Puntuación de la actividad

- Apartado 1 (10 %)
- Apartado 2 (10 %)
- Apartado 3 (15 %)
- Apartado 4 (10 %)
- Apartado 5 (20 %)
- Apartado 6 (15 %)
- Apartado 7 (10 %)
- Calidad del informe dinámico (10 %)