A2 - Análisis Estadístico I

Lukaz Martin Doehne

2022-12-02

Introducción

En esta actividad se introduce la inferencia estadística. Para ello usamos el conjunto de datos **gpa_clean.csv** que contiene atributos sobre universitarios. Las variables incluidas son:

· ***sat***: nota de acceso (medida en escala de 400 a 1600 puntos)

· ***tothrs***: horas totales

· ***hsize***: numero total de estudiantes en la cohorte de graduados del bachillerato (en cientos)

· ***hsrank***: ranking del estudiante, dado por la nota media del bachillerato, en su cohorte de graduados del bachillerato

· ***hsprec***: ranking relativo del estudiante (hsrank/hsize)

· ***colgpa***: nota media del estudiante al final del primer semestre (medida en escala de 0 a 4 puntos)

· ***athlete***: indicador de si el estudiante practica algún deporte en la universidad

· ***female***: indicador de si el estudiante es mujer

· ***white***: indicador de si el estudiante es de raza blanca o no

· ***black***: indicador de si el estudiante es de raza negra o no

· ***gpaletter***: letra que indica el nivel de la nota (A,B,C,D)

1. Lectura del fichero

A continuación leemos el fichero **gpa_clean.csv** y lo guardamos en un objeto *gpa*. Seguidamente, verificamos cada tipo de variable.

```
gpa<-read.csv("gpa_clean.csv",stringsAsFactors=TRUE)
head(gpa)
```

```
##    sat tothrs hsize hsrank   hsperc colgpa athlete female white black gpaletter
## 1  920     43  0.10      4 40.00000   2.04    TRUE   TRUE FALSE FALSE         C
## 2 1170     18  9.40    191 20.31915   4.00   FALSE  FALSE  TRUE FALSE         A
## 3  810     14  1.19     42 35.29412   1.78    TRUE  FALSE  TRUE FALSE         C
## 4  940     40  5.71    252 44.13310   2.42   FALSE  FALSE  TRUE FALSE         C
## 5 1180     18  2.14     86 40.18692   2.61   FALSE  FALSE  TRUE FALSE         B
## 6  980    114  2.68     41 15.29851   3.03   FALSE   TRUE  TRUE FALSE         B
```

```
dim=dim(gpa)
dim
```

```
## [1] 4137   11
```

```
str(gpa)
```

```
## 'data.frame':    4137 obs. of  11 variables:
##  $ sat      : int  920 1170 810 940 1180 980 880 980 1240 1230 ...
##  $ tothrs   : int  43 18 14 40 18 114 78 55 18 17 ...
##  $ hsize    : num  0.1 9.4 1.19 5.71 2.14 ...
##  $ hsrank   : int  4 191 42 252 86 41 161 101 161 3 ...
##  $ hsperc   : num  40 20.3 35.3 44.1 40.2 ...
##  $ colgpa   : num  2.04 4 1.78 2.42 2.61 ...
##  $ athlete  : logi  TRUE FALSE TRUE FALSE FALSE FALSE ...
##  $ female   : logi  TRUE FALSE FALSE FALSE FALSE TRUE ...
##  $ white    : logi  FALSE TRUE TRUE TRUE TRUE TRUE ...
##  $ black    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ gpaletter: Factor w/ 4 levels "A","B","C","D": 3 1 3 3 2 2 3 2 2 3 ...
```

Variables **cualitativas** (factor): ***athlete***, ***female***, ***white***, ***black***, ***gpaletter***

Variables **cuantitativas** (númericas): ***sat***, ***tothrs***, ***hsize***, ***hsperc***, ***colgpa***

2. Estadística descriptiva y visualización

2.1 Análisis descriptivo

A continuación se presenta el análisis descriptivo númerico de los datos. Vemos los valores centrales: *Median* y *Mean*. La dispersión de los datos: *Min*, *Max*, *1st Q* y *3rd Q*. Y los posibles valores para las variables categóricas.

```
summary(gpa)
```

```
##       sat           tothrs           hsize          hsrank
##  Min.   : 470   Min.   :  6.00   Min.   :0.03   Min.   :  1.00
##  1st Qu.: 940   1st Qu.: 17.00   1st Qu.:1.65   1st Qu.: 11.00
##  Median :1030   Median : 47.00   Median :2.51   Median : 30.00
##  Mean   :1030   Mean   : 52.83   Mean   :2.80   Mean   : 52.83
##  3rd Qu.:1120   3rd Qu.: 80.00   3rd Qu.:3.68   3rd Qu.: 70.00
##  Max.   :1540   Max.   :137.00   Max.   :9.40   Max.   :634.00
##      hsperc            colgpa       athlete          female
##  Min.   : 0.1667   Min.   :0.000   Mode :logical   Mode :logical
##  1st Qu.: 6.4328   1st Qu.:2.210   FALSE:3943      FALSE:2277
##  Median :14.5833   Median :2.660   TRUE :194       TRUE :1860
##  Mean   :19.2371   Mean   :2.654
##  3rd Qu.:27.7108   3rd Qu.:3.120
##  Max.   :92.0000   Max.   :4.000
##    white           black         gpaletter
##  Mode :logical   Mode :logical   A: 458
```

```
##  FALSE:308       FALSE:3908      B:1999
##  TRUE :3829      TRUE :229       C:1536
##                                  D: 144
##
##
```

```
cat("Las dimensiones del dataset son: (", dim, ") y la cantidad de valores nulos son: ", sum(is.na(gpa)))
```

```
## Las dimensiones del dataset son: ( 4137 11 ) y la cantidad de valores nulos son:  0
```

2.2 Visualización

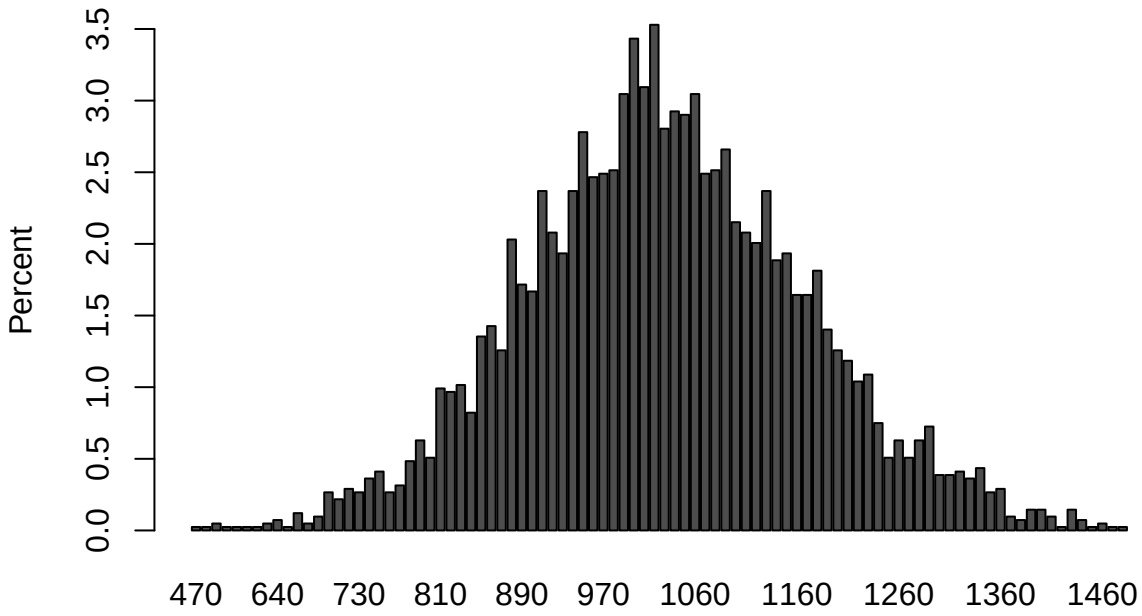
Estudiaremos de forma visual la distribución de las variables ***sat*** y ***colgpa***

2.2.1. Distribución de las variables *sat* y *colgpa* (por separado).

Distribución sat

Como es de esperar la distribución de sat sigue una distribución normal.

```
porc_sat<- t(prop.table(table(gpa$sat))) * 100
barplot(porc_sat, ylab = "Percent")
```

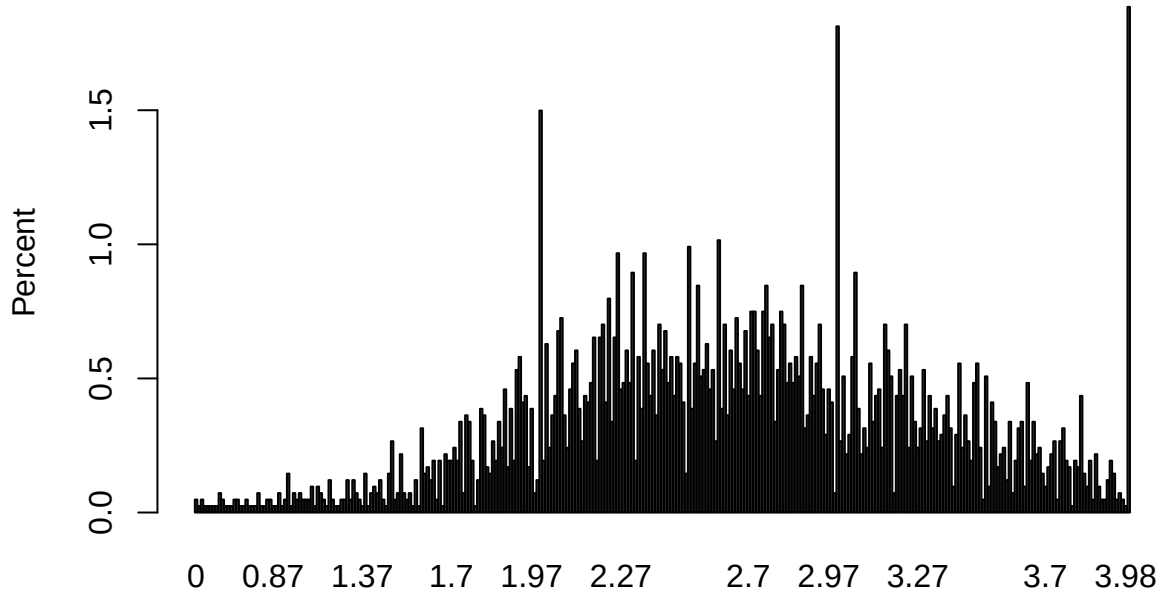


Pocos estudiantes sacan muy buenas o muy malas notas. Por norma general todos siguen una media. Durante los años las universidades han podido adaptar la dificultad del contenido de las asignaturas a los estudiantes.

Distribución colgpa

En la distribución de colgpa también se intuye una distribución normal.

```
porc_colgpa<- t(prop.table(table(gpa$colgpa))) * 100
barplot(porc_colgpa, ylab = "Percent")
```



Encontramos picos con los valores 2, 3 y 4.

```
names(porc_colgpa[1,which(porc_colgpa>1.1)])
```

```
## [1] "2" "3" "4"
```

En 1 también encontramos un pico si lo comparamos con sus vecinos

```
names(porc_colgpa[1,30:34])
```

```
## [1] "0.93000001" "0.94"       "1"          "1.03"       "1.05"
```

```
porc_colgpa[30:34]
```

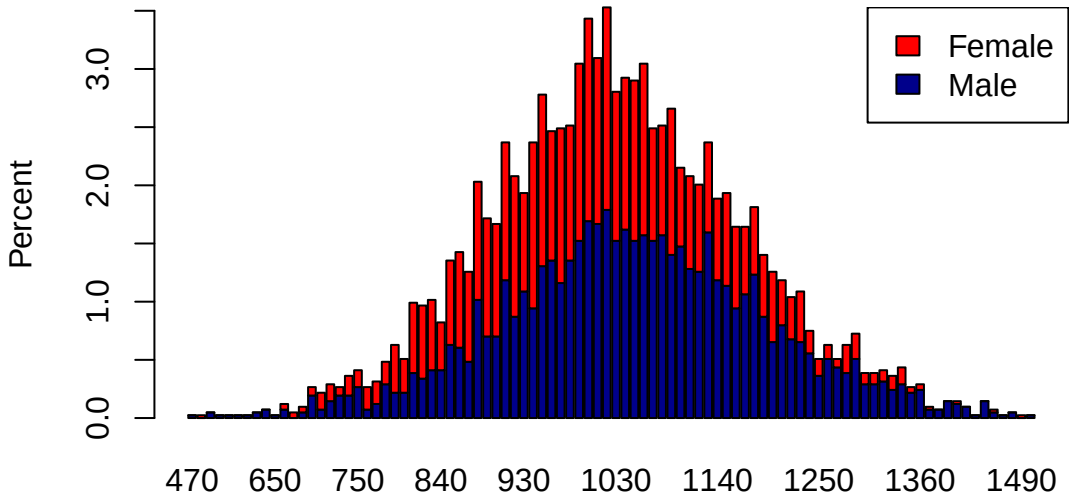
```
## [1] 0.02417211 0.04834421 0.14503263 0.02417211 0.07251632
```

Vemos que en *0.94* y *1.03* hay 0.048% y 0.024% de estudiantes respectivamente. En *1* hay 0.15%

2.2.2 Distribución de la variable *sat* con respecto a la variable *female*, *athlete*, *white* y *black*

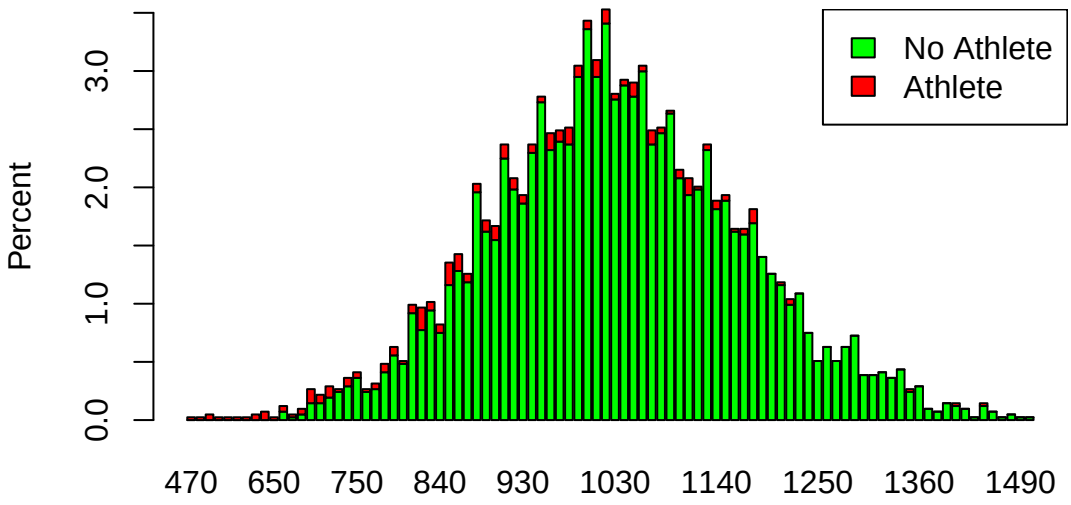
Distribución sat en función del sexo

```
porc_sat_gender<-t(prop.table(table(gpa$sat,gpa$female))) * 100
barplot(porc_sat_gender,ylab="Percent",col=c("darkblue","red"))
legend("topright",legend = c("Female", "Male"),fill = c("red", "darkblue"))
```



Distribución sat en función de atleta

```
porc_sat_athlete<-t(prop.table(table(gpa$sat,gpa$athlete))) * 100
barplot(porc_sat_athlete,ylab="Percent",col=c("green","red"))
legend("topright",legend = c("No Athlete", "Athlete"),fill = c("green", "red"))
```



Distribución sat en función de la raza

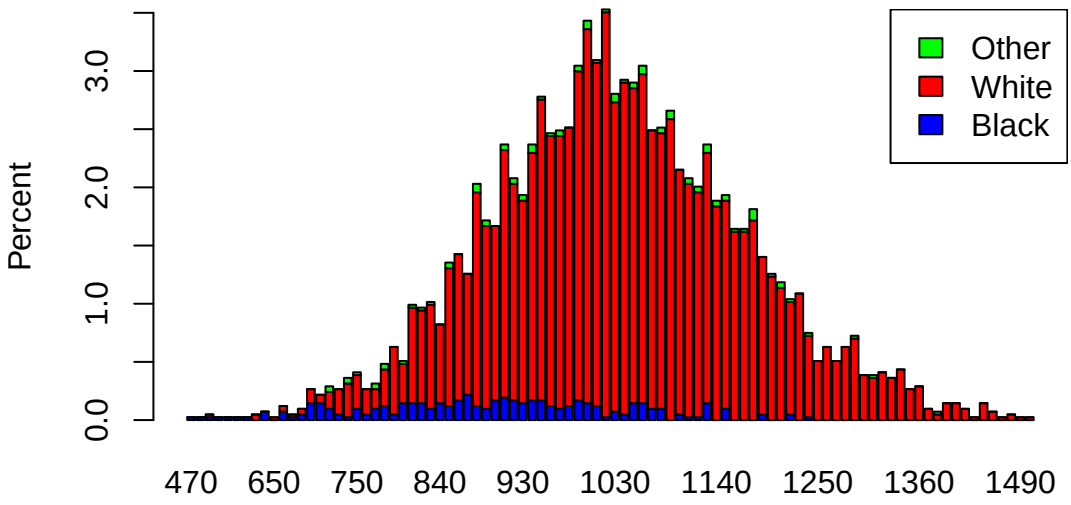
Dado que no hay ningun estudiante con las variables ***white*** y ***black*** a *TRUE*

```
sum(which(gpa$white==TRUE & gpa$black==TRUE))
```

```
## [1] 0
```

Podemos visualizar***sat*** con respecto a la raza en un plot

```
other<-which(gpa$white==FALSE & gpa$black==FALSE)
unification_race<-gpa$white
unification_race[other]=2
porc_sat_race<-t(prop.table(table(gpa$sat,unification_race))) * 100
barplot(porc_sat_race,ylab="Percent",col=c("blue","red", "green"))
legend("topright",legend = c("Other", "White", "Black"),fill = c("green", "red", "blue"))
```

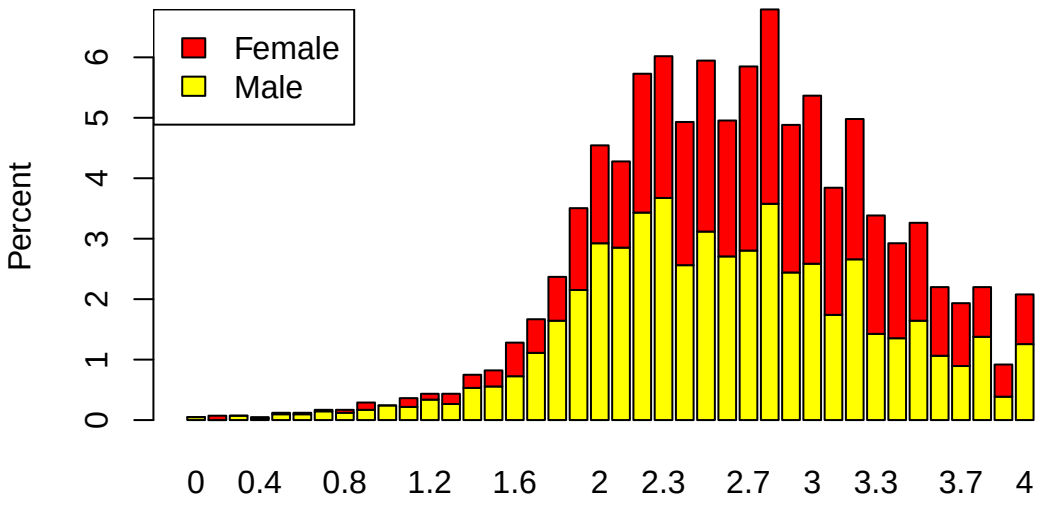


2.2.3. Mismo tipo de visualizaciones con la variable *colgpa* respecto a *female*, *athlete* y *white/black*

Para tener una visualización más sencilla de procesar he redondeado la variable ***colgpa*** a 1 decimal

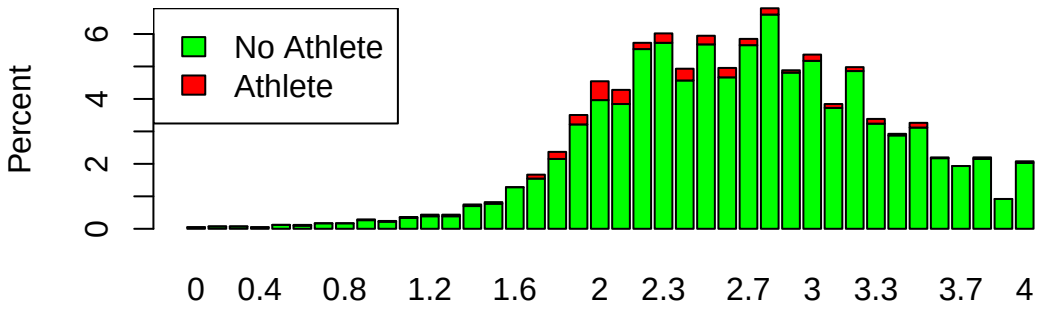
Distribución colgpa en función del sexo

```
porc_colgpa_gender<-t(prop.table(table(round(gpa$colgpa,1),gpa$female))) * 100
barplot(porc_colgpa_gender,ylab="Percent",col=c("yellow","red"))
legend("topleft",legend = c("Female", "Male"),fill = c("red", "yellow"))
```



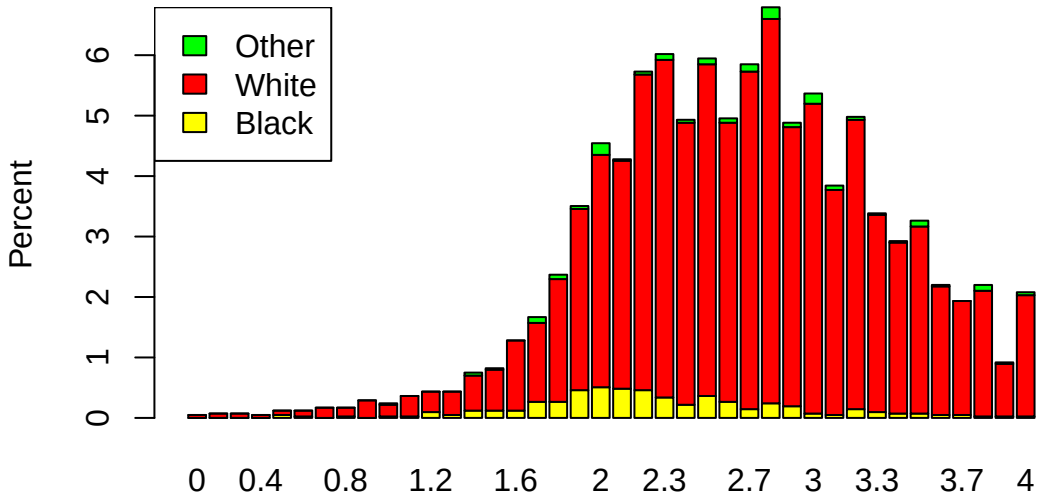
Distribución colgpa en función de atleta

```
porc_colgpa_athlete<-t(prop.table(table(round(gpa$colgpa,1),gpa$athlete))) * 100
barplot(porc_colgpa_athlete,ylab="Percent",col=c("green","red"))
legend("topleft",legend = c("No Athlete", "Athlete"),fill = c("green", "red"))
```



Distribución colgpa en función de la raza

```
porc_colgpa_race<-t(prop.table(table(round(gpa$colgpa,1),unification_race))) * 100
barplot(porc_colgpa_race,ylab="Percent",col=c("yellow","red", "green"))
legend("topleft",legend = c("Other", "White", "Black"),fill = c("green", "red", "yellow"))
```



2.2.4. Interpretación breve de los grafos

Para ***sat*** encontramos una distribución normal.

Para ***colgpa*** también hay una distribución normal, aunque no tan precisa como para ***sat***. Además, encontramos picos en los valores 1,2,3 y 4.

Al comparar estas variables con el sexo, atleta y la raza hemos observado:

female: la distribución hombre-mujer se distribuye relativamente equitativo. Como es de esperar ya que la población mundial tiende a 50% hombres y 50% mujeres.

athelete: Encontramos una gran descompensación. Solo el 4.7% de los estudiantes son atletas.

```
 t(prop.table(table(gpa$athlete))) * 100
```

```
##
##            FALSE      TRUE
##   [1,] 95.310612  4.689388
```

white/black: De misma manera encontramos una gran descompensación. 5.5% de raza negra, 92.6% de raza blanca y 1.9% de otra raza.

```
t(prop.table(table(unification_race))) * 100
```

```
##       unification_race
##                  0         1         2
##   [1,]  5.535412 92.554992  1.909596
```

3. Intervalo de confianza de la media poblacional de la variable sat y colgpa

3.1 Supuestos

La estadística inferencial estudia características de una población a partir de una muestra finita de datos.

Para que la muestra sea fiable ha de ser representativa. (Sin sesgos y de un tamaño adecuado)

Para averiguar si la muestra de la población es normal haremos uso de el **teorema del límite central** (TLC)

TLC

Este teorema establece que el contraste de hipótesis sobre la media de una muestra se aproxima a una distribución normal aunque la población original no siga una distribución normal, siempre que el tamaño de la muestra n sea suficientemente grande. Por suficientemente grande, se suele considerar superior a 30 elementos, $n > 30$.

```
dim[1]
```

```
## [1] 4137
```

```
mean(gpa$sat)
```

```
## [1] 1030.331
```

```
sd(gpa$sat)
```

```
## [1] 139.4014
```

Por el teorema del límite central, podemos asumir normalidad, puesto que tenemos una muestra de tamaño grande N=4137 y en el plot hemos visto seguimos una distribución normal

3.2 Función de cálculo del intervalo de confianza

La función IC calcula el intervalo de confianza de una variable a partir del nivel de confianza

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

  alpha<-1-NC
  sd<-sd(x)
  N<-length(x)
```

```
  mean<-mean(x)

  z<-qnorm(alpha/2, lower.tail=FALSE)

  L<-mean-(z*(sd/sqrt(N)))
  U<-mean+(z*(sd/sqrt(N)))

  return (c(L,U))
}
```

3.3 Intervalo de confianza de la variable sat

Intervalo de confianza al 90%

```
sat_ic_90<-IC(gpa$sat,0.9)
sat_ic_90
```

```
## [1] 1026.766 1033.896
```

Intervalo de confianza al 95%

```
sat_ic_95<-IC(gpa$sat,0.95)
sat_ic_95
```

```
## [1] 1026.083 1034.579
```

3.4 Intervalo de confianza de la variable colgpa

Intervalo de confianza al 90%

```
colgpa_ic_90<-IC(gpa$colgpa,0.9)
colgpa_ic_90
```

```
## [1] 2.637309 2.670953
```

Intervalo de confianza al 95%

```
colgpa_ic_95<-IC(gpa$colgpa,0.95)
colgpa_ic_95
```

```
## [1] 2.634086 2.674176
```

3.5 Interpretación

El intervalo de confianza nos indica que en el intervalo definido se incluye la media de la población cerca del intervalo de confianza (e.g. 95%,90%).

Es decir, que si cogemos muestras aleatorias la probabilidad de que se encuentre dentro del intervalo será igual al intervalo de confianza escogido.

En concreto, con las variables ***sat*** y ***colgpa*** vemos que los valores del intervalo no cambian excesivamente al cambiar de 95% a 90% de confianza.

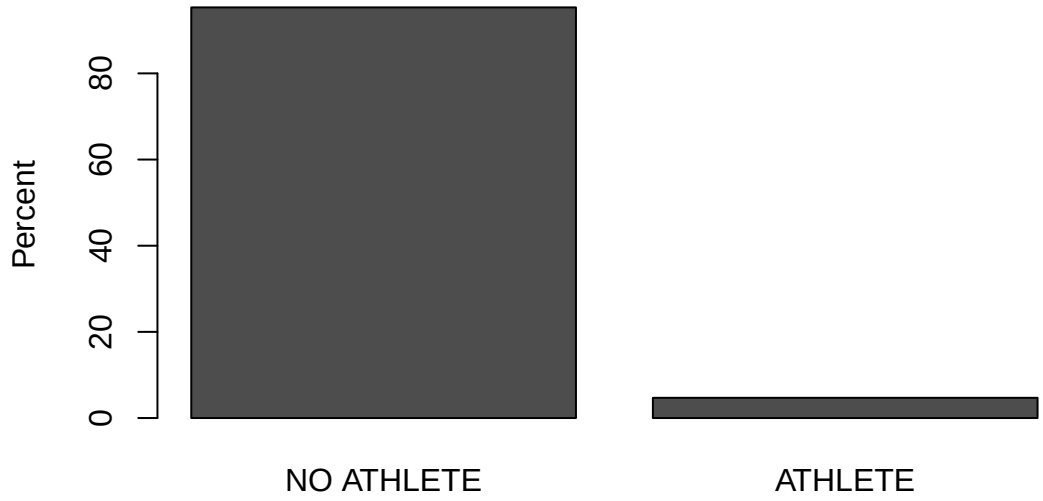
4. ¿Ser atleta influye en la nota?

En este apartado queremos analizar si ser atleta influye en la nota ***colgpa*** con un nivel de confianza del 95%.

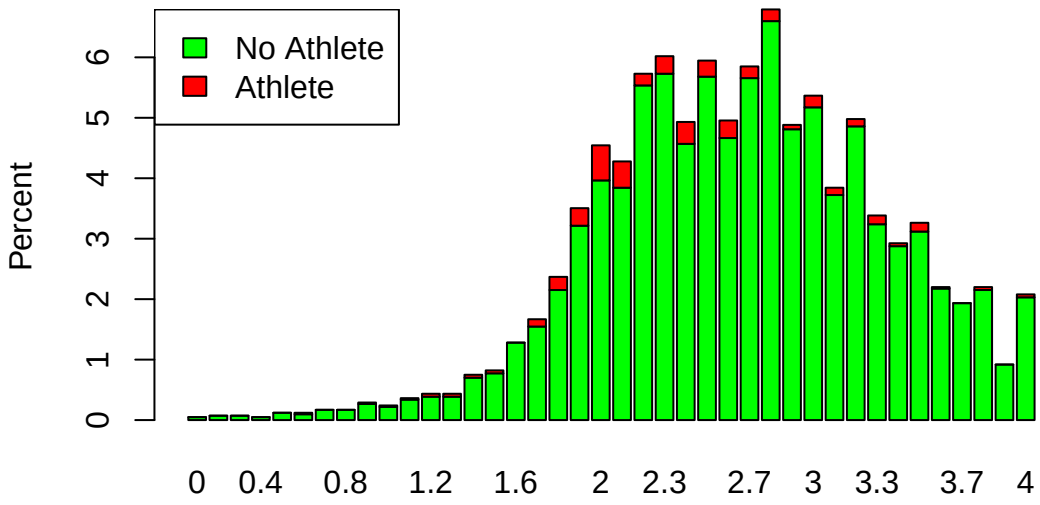
4.1 Análisis visual

Vemos que un gran porcentaje son No atletas. Por lo que no contaremos con una gran selección de atletas para averiguar de forma fiable la media de ***colpa***.

```
porc_athlete<- t(prop.table(table(gpa$athlete))) * 100
colnames(porc_athlete)<- c("NO ATHLETE", "ATHLETE")
barplot(porc_athlete, ylab = "Percent")
```




```
barplot(porc_colgpa_athlete,ylab="Percent",col=c("green","red"))
legend("topleft",legend = c("No Athlete", "Athlete"),fill = c("green", "red"))
```



4.2 Función para el contraste de medias

```
CM<-function(x, y, NC=0.95, var.equal=TRUE, alternative="two.sided"){
  sd.x <- sd(x)
  N.x<-length(x)
  mean.x<-mean(x)
  sd.y <- sd(y)
  N.y <- length(y)
  mean.y <- mean(y)
  alpha <- 1-NC

  if(var.equal){ # Varianzas desconocidas iguales (pg. 42 apuntes)
    S<-sqrt((((N.x-1)*(sd.x^2))+((N.y-1)*(sd.y^2))) / (N.x+N.y-2))
    denom<-S*sqrt((1/N.x)+(1/N.y))
    t<-(mean.x-mean.y)/denom
    gl<-N.x+N.y-2 # Grados de libertad
  }
  else{ # Varianzas desconocidas diferentes
    denom<-((((sd.x^2)/N.x)^2)/(N.x-1))+((((sd.y^2)/N.y)^2)/(N.y-1))
    gl<-((((sd.x^2)/N.x)+((sd.y^2)/N.y))^2)/denom
    t<-(mean.x-mean.y)/sqrt(((sd.x^2)/N.x)+((sd.y^2)/N.y))
  }

  if (alternative=="two.sided"){ # se reparte izquierda y derecha de la curva
    crit_value<-qt(1-(alpha/2),gl)
    pvalue<-pt(abs(t), gl, lower.tail=FALSE )*2 # por eso multiplicamos por 2 p
  }
```

```
  else if (alternative == "greater"){
    crit_value<-qt(1-alpha, gl)
    pvalue<-pt(t, gl, lower.tail=FALSE )
  }
  else if (alternative == "less"){
    crit_value <-qt(alpha, gl, lower.tail=TRUE )
    pvalue<-pt(t, gl, lower.tail=TRUE )
  }
  return (c(t,crit_value,pvalue,gl))
}
```

4.3 Pregunta de investigación

La pregunta de investigación será si la variable ***athlete*** influye significativamente a la variable ***colgpa***

Para ello definimos de manera más específica:

¿La media de ***colgpa*** de estudiantes atletas es igual a la media de ***colgpa*** de estudiantes que no son atletas?

Si la hipótesis se confirma entonces sabemos que la variable ***athlete*** no influye de manera significativa en la variable ***colgpa***

4.4 Hipótesis nula y la alternativa

H0:Mean_colgpa_ath=Mean_colgpa_non_ath

H1:Mean_colgpa_ath!=Mean_colgpa_non_ath

4.5 Justificación del test a aplicar

Aplicamos un test de hipótesis de dos muestras sobre la media.

Comprobamos la varianza de las poblaciones

```
var.test(gpa$colgpa[gpa$athlete==TRUE], gpa$colgpa[gpa$athlete==FALSE])
```

```
##
##  F test to compare two variances
##
## data:  gpa$colgpa[gpa$athlete == TRUE] and gpa$colgpa[gpa$athlete == FALSE]
## F = 0.82199, num df = 193, denom df = 3942, p-value = 0.07287
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6762059 1.0186147
## sample estimates:
## ratio of variances
##          0.8219902
```

El resultado del test es un valor p>0.05. Por tanto, asumimos igualdad de varianzas.

En consecuencia, el test es de dos muestras independientes sobre la media con varianzas desconocidas iguales. El test es bilateral.

4.6 Cálculo

```
info_athlete_colgpa<-CM(gpa$colgpa[gpa$athlete==TRUE], gpa$colgpa[gpa$athlete==FALSE], NC=0.95,
var.equal=TRUE, alternative="two.sided")
data.frame(T=c(info_athlete_colgpa[1]),CRIT_VALUE=c(info_athlete_colgpa[2]), PVALUE=c(info_athlete_colg
```

```
##           T CRIT_VALUE       PVALUE   GL
## 1 -5.910309   1.960538 3.689891e-09 4135
```

Confirmamos que los valores T, PVALUE y GL son iguales con la función original.

```
t.test(gpa$colgpa[gpa$athlete==TRUE], gpa$colgpa[gpa$athlete==FALSE],
var.equal=TRUE, alternative ="two.sided")
```

```
##
##  Two Sample t-test
##
## data:  gpa$colgpa[gpa$athlete == TRUE] and gpa$colgpa[gpa$athlete == FALSE]
## t = -5.9103, df = 4135, p-value = 3.69e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3792088 -0.1902956
## sample estimates:
## mean of x mean of y
##  2.382732  2.667484
```

4.7 Interpretación del test

El valor crítico para un nivel de confianza del 95% es 1.960538.

El valor T es -5.910309.

Por lo tanto **rechazamos la hipótesis nula**, ya que no se encuentra en el rango de valores aceptados (-1.96,1.96).

Queda confirmado por el pvalue de 3.69eˆ-09 siendo mucho menor al valor alpha=0.05

Al rechazar la hipótesis asumimos que la variable ***athlete*** tiene influencia en la nota de colgpa y no son independientes.

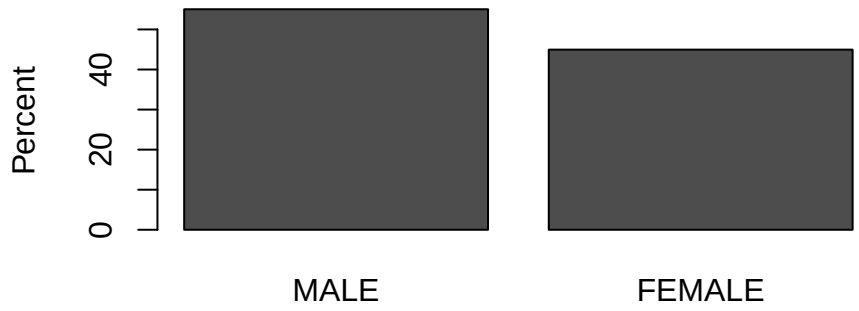
5 ¿Las mujeres tienen mejor nota que los hombres?

En este apartado queremos analizar si el género influye en la nota ***colgpa*** con un nivel de confianza de 95% y 90%.

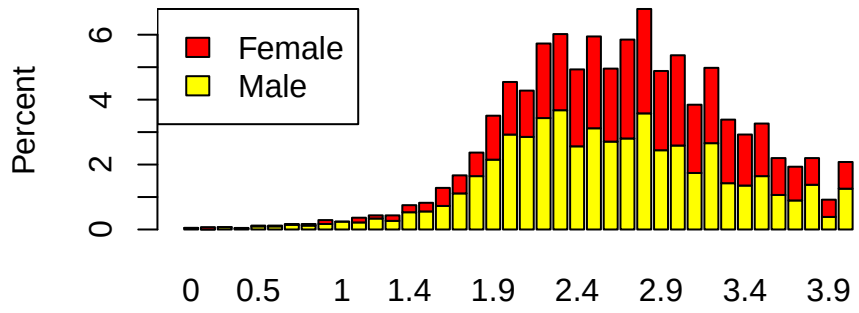
5.1 Análisis visual

Positivamente, podemos observar que la variable género está relativamente equilibrada. Lo que es bueno para evitar un dataset sesgado.

```
porc_gender<- t(prop.table(table(gpa$female))) * 100
colnames(porc_gender)<- c("MALE", "FEMALE")
barplot(porc_gender, ylab = "Percent")
```




```
barplot(porc_colgpa_gender,ylab="Percent",col=c("yellow","red"))
legend("topleft",legend = c("Female", "Male"),fill = c("red", "yellow"))
```



5.2 Función

Añadimos una condición *if* a la función anterior.

5.3 Pregunta de investigación

¿La media de notas de las mujeres es superior a la media de notas de los hombres?

5.4 Hipótesis nula y la alternativa

H0:Mean_colgpa_female=Mean_colgpa_male

H1:Mean_colgpa_female>Mean_colgpa_male

5.5 Justificación del test a aplicar

Haremos un test sobre la media con varianzas desconocidas. Asumimos normalidad (TLC).

```
var.test(gpa$colgpa[gpa$female==TRUE],gpa$colgpa[gpa$female==FALSE])
```

```
##
##  F test to compare two variances
##
## data:  gpa$colgpa[gpa$female == TRUE] and gpa$colgpa[gpa$female == FALSE]
## F = 0.82757, num df = 1859, denom df = 2276, p-value = 2.024e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7590051 0.9026724
## sample estimates:
## ratio of variances
##          0.8275687
```

El resultado es un valor p<0.05, por lo que haremos un test unilateral por la derecha con varianzas desconocidas diferentes.

5.6 Cálculo

95% Nivel de confianza

```
info_female_colgpa<-CM(gpa$colgpa[gpa$female==TRUE], gpa$colgpa[gpa$female==FALSE], NC=0.95,
var.equal=FALSE, alternative="greater")
data.frame(T=c(info_female_colgpa[1]),CRIT_VALUE=c(info_female_colgpa[2]), PVALUE=c(info_female_colgpa[
```

```
##          T CRIT_VALUE       PVALUE       GL
## 1 7.078735   1.645227 8.521971e-13 4087.407
```

90% Nivel de confianza

```
info_female_colgpa<-CM(gpa$colgpa[gpa$female==TRUE], gpa$colgpa[gpa$female==FALSE], NC=0.9,
var.equal=FALSE, alternative="greater")
data.frame(T=c(info_female_colgpa[1]),CRIT_VALUE=c(info_female_colgpa[2]), PVALUE=c(info_female_colgpa[
```

```
##          T CRIT_VALUE       PVALUE       GL
## 1 7.078735   1.281759 8.521971e-13 4087.407
```

5.7 Interpretación del test

El valor crítico para un nivel de confianza del 95% es 1.645227.

El valor T es 7.078735.

Por lo tanto **rechazamos la hipótesis nula**, ya que se encuentra fuera del rango de valores aceptados.

Queda confirmado por el pvalue de 8.52e-13 siendo mucho menor al valor alpha=0.05

Lo mismo para un nivel de confianza de 90%

Por lo tanto rechazamos la hipótesis nula y concluimos que la media de las notas de las mujeres es superior a la de los hombres

6 ¿Hay diferencias en la nota según la raza?

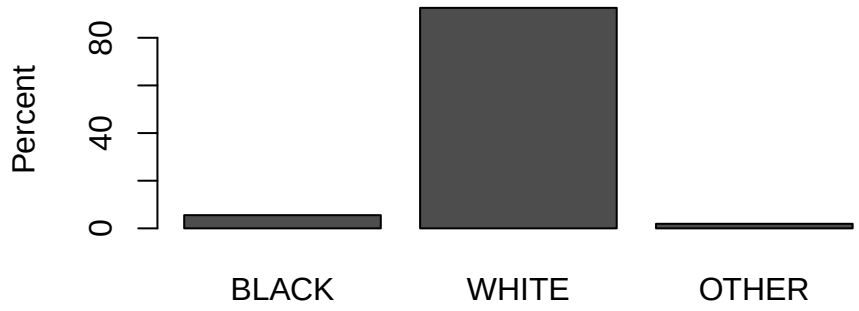
Antes del análisis matizar que **correlación no implica causalidad**.

Si la interpretación del test resulta que el hecho de que un estudiante sea de una raza determine una dependencia con su nota, no ha de ser por motivo de la raza sino por ejemplo por el entorno socioeconómico en el que estan viviendo.

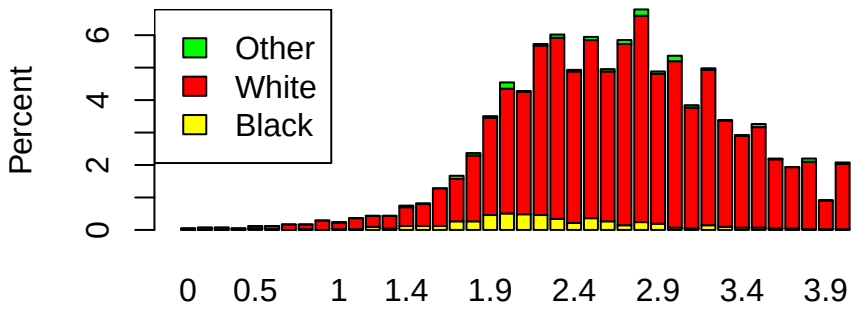
6.1 Análisis visual

Encontramos un desequilibrio en la cantidad de estudiantes por raza.

```
porc_race<-t(prop.table(table(unification_race))) * 100
colnames(porc_race)<- c("BLACK", "WHITE", "OTHER")
barplot(porc_race, ylab = "Percent")
```




```
barplot(porc_colgpa_race,ylab="Percent",col=c("yellow","red", "green"))
legend("topleft",legend = c("Other", "White", "Black"),fill = c("green", "red", "yellow"))
```



6.2 Función

Utilizamos la función ya definida

6.3 Pregunta de investigación

¿La media de notas de estudiantes de raza blanca varia significamente de la media de notas de estudiantes de raza negra?

6.4 Hipótesis nula y la alternativa

H0: Mean_golgpa_white = Mean_colgpa_black

H1: Mean_colgpa_white != Mean_colgpa_black

6.5 Justificación del test a aplicar

Haremos un test sobre la media con varianzas desconocidas. Asumimos normalidad (TLC).

```
var.test(gpa$colgpa[gpa$white==TRUE], gpa$colgpa[gpa$black==TRUE])
```

```
##
##  F test to compare two variances
##
## data:  gpa$colgpa[gpa$white == TRUE] and gpa$colgpa[gpa$black == TRUE]
## F = 1.127, num df = 3828, denom df = 228, p-value = 0.2343
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9250878 1.3509805
## sample estimates:
## ratio of variances
##           1.126968
```

El resultado del test es un valor p>0.05. Por tanto, asumimos igualdad de varianzas.

En consecuencia, el test es de dos muestras con varianzas desconocidas iguales. El test es bilateral.ç

6.6 Cálculo

95% Nivel de confianza

```
info_race_colgpa<-CM(gpa$colgpa[gpa$white==TRUE], gpa$colgpa[gpa$black==TRUE], NC=0.95,
var.equal=TRUE, alternative="two.sided")
data.frame(T=c(info_race_colgpa[1]),CRIT_VALUE=c(info_race_colgpa[2]), PVALUE=c(info_race_colgpa[3]),GL=
```

```
##          T CRIT_VALUE      PVALUE   GL
## 1 9.559319   1.960549 1.99014e-21 4056
```

90% Nivel de confianza

```
info_race_colgpa<-CM(gpa$colgpa[gpa$white==TRUE], gpa$colgpa[gpa$black==TRUE], NC=0.95,
var.equal=TRUE, alternative="two.sided")
data.frame(T=c(info_race_colgpa[1]),CRIT_VALUE=c(info_race_colgpa[2]), PVALUE=c(info_race_colgpa[3]),GL=
```

```
##          T CRIT_VALUE      PVALUE   GL
## 1 9.559319   1.960549 1.99014e-21 4056
```

6.7 Interpretación del test

El valor crítico para un nivel de confianza del 95% es 1.960549.

El valor T es 9.559319.

Por lo tanto **rechazamos la hipótesis nula**, ya que se encuentra fuera del rango de valores aceptados.

Queda confirmado por el pvalue de 1.99014e-21 siendo mucho menor al valor alpha=0.05

Lo mismo para un nivel de confianza de 90%

Por lo tanto rechazamos la hipótesis nula y concluimos que la media de las notas estudiantes de raza blanca difiere de las notas de estudiantes de raza negra.

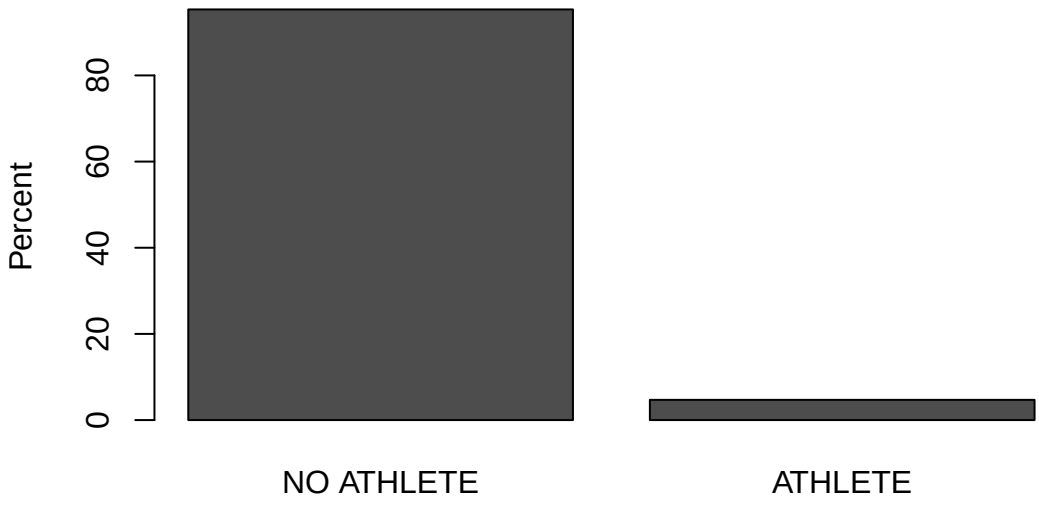
7. Proporción de atletas

Nos preguntamos si la proporción de atletas en la población es inferior al 5% con un nivel de confianza del 95%.

7.1 Análisis visual

Vemos que un gran porcentaje son No atletas en la muestra.

```
barplot(porc_athlete, ylab = "Percent")
```



7.2 Pregunta de investigación

¿La proporción de estudiantes atletas en la población es inferior a 0.05?

7.3 Hipótesis nula y la alternativa

H0: Prop_Athlete = 0.05

H1: Prop_Athlete < 0.05

7.4 Justificación del test a aplicar

Se trata de un contraste sobre la proporción sobre una muestra.

El test es unilateral por la izquierda.

7.5 Realizad los cálculos del test

```
proporcion<-function(x,y,n,NC=0.95,alternative="less"){
  prop_x<-sum(x)/n
  prop_y<-y
  SE<-sqrt((prop_x*(1-prop_x))/n) # pg. 32 apuntes
  z<-(prop_x-prop_y)/SE
  alpha<-1-NC
  if (alternative == "less"){
```

```
    crit_value <-qnorm(alpha, lower.tail=TRUE )
    pvalue<-pnorm(z, lower.tail=TRUE )
  }
  else if (alternative == "greater"){
    crit_value <-qnorm(1-alpha)
    pvalue<-pnorm(z, lower.tail=FALSE )
  }
  pL<-prop_x - crit_value*SE
  pU<-prop_x + crit_value*SE
  return (c(z, crit_value, pvalue, pL, pU))
}
```

```
info_prop_athlete<-proporcion(gpa$athlete[gpa$athlete==TRUE], 0.05,length(gpa$athlete),NC=0.95,alternati
data.frame(Z=c(info_prop_athlete[1]),CRIT_VALUE=c(info_prop_athlete[2]), PVALUE=c(info_prop_athlete[3])
```

```
##            Z CRIT_VALUE    PVALUE         pL       pL.1
## 1 -0.9449996  -1.644854 0.1723295 0.05230035 0.04148742
```

7.6 Interpretación del test

Valor observado: -0.9449996

La zona de aceptación de la hipótesis nula es [-1.644854, infinito)

Como el valor observado está dentro del rango de aceptación de la hipótesis nula, no podemos rechazar H0.

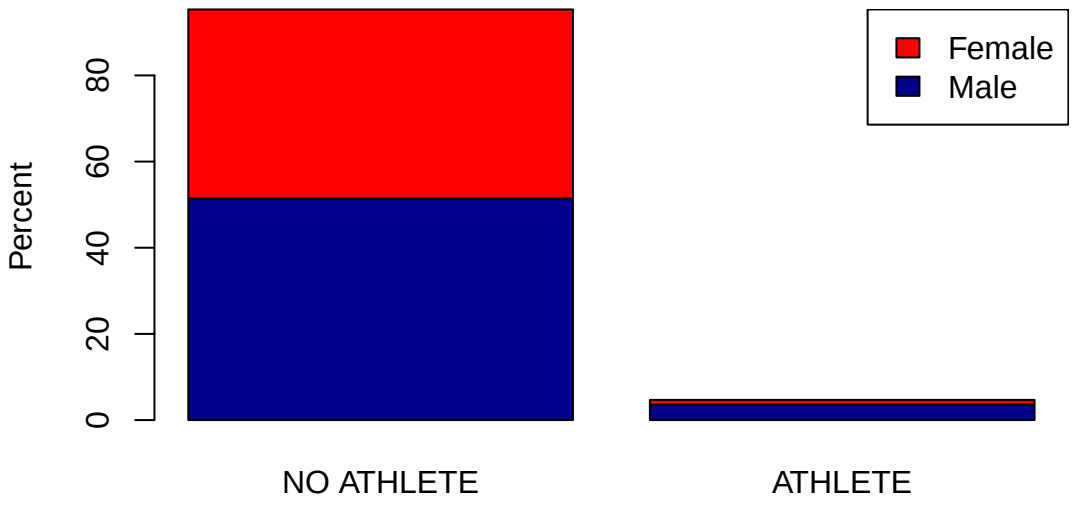
El pvalue es 0.172329 y, por lo tanto, no inferior a alpha. No podemos rechazar la hipótesis nula.

Por lo tanto, se concluye que la proporción de atletas no es inferior a 0.05 con un nivel de confianza del 95 %.

8. ¿Hay más atletas entre los hombres que entre las mujeres?

8.1 Análisis visual

```
porc_athlete_gender<- t(prop.table(table(gpa$athlete,gpa$female))) * 100
colnames(porc_athlete_gender)<- c("NO ATHLETE", "ATHLETE")
barplot(porc_athlete_gender,ylab="Percent",col=c("darkblue","red"))
legend("topright",legend = c("Female", "Male"),fill = c("red", "darkblue"))
```



8.2 Pregunta de investigación

¿La proporción de atletas que son hombres es superior a 0.5?

Si es mayor indica que hay más hombres que mujeres atletas

8.3 Hipótesis nula y la alternativa

H0: Prop_Athlete_male = 0.5

H1: Prop_Athlete_male > 0.5

8.4 Justificación del test a aplicar

Se trata de un contraste sobre la proporción sobre una muestra.

El test es unilateral por la derecha

8.5 Realizad los cálculos del test

```
info_prop_ath_fem<-proporcion(gpa$athlete[gpa$athlete==TRUE&gpa$female==FALSE], 0.5,sum(gpa$athlete),NC
data.frame(Z=c(info_prop_ath_fem[1]),CRIT_VALUE=c(info_prop_ath_fem[2]), PVALUE=c(info_prop_ath_fem[3])
```

```
##          Z CRIT_VALUE       PVALUE       pL      pL.1
## 1 8.845142   1.644854 4.570593e-19 0.718196 0.8178865
```

8.6 Interpretación del test

El valor observado: 8.845142

La zona de aceptación de la hipótesis nula es [-infinito, 1.644854)

Como el valor observado no está dentro del rango de aceptación de la hipótesis nula, podemos rechazar H0.

El pvalue es 4.570593e-19 y, por lo tanto, inferior a alpha. Se confirma que podemos rechazar la hipótesis nula.

Por lo tanto, se concluye que la proporción de atletas que son hombres es superior a la de mujeres, ya que es superior a 0.5.

9. Resumen y conclusiones

```
p0<-c("IC sat 90%", paste0("(",round(sat_ic_90[1],2),", ",round(sat_ic_90[2],2),")"), "Intervalo de con
p1<-c("IC sat 95%", paste0("(",round(sat_ic_95[1],2),", ",round(sat_ic_95[2],2),")"), "Intervalo de con
p2<-c("IC colgpa 90%", paste0("(",round(colgpa_ic_90[1],2),", ",round(colgpa_ic_90[2],2),")"), "Interval
p3<-c("IC colgpa 95%", paste0("(",round(colgpa_ic_90[1],2),", ",round(colgpa_ic_90[2],2),")"), "Interval
p4<-c("¿Colgpa de atletas es igual a la de no atletas?", paste0("obs:",round(info_athlete_colgpa[1],2),
p5<-c("¿Colgpa de mujeres es superior a la de hombres?", paste0("obs:",round(info_female_colgpa[1],2),"
p6<-c("¿Colgpa de raza blanca varia de raza negra?", paste0("obs:",round(info_race_colgpa[1],2)," crit:
p7<-c("¿Proporción de atletas es inferior a 0.05?", paste0("obs:",round(info_prop_athlete[1],2)," crit:
p8<-c("¿Proporción de atletas hombres es superior a 0.5 (superior a mujeres)?", paste0("obs:",round(inf
table<-data.frame(p1,p2,p3,p4,p5,p6,p7,p8)
knitr::kable(t(table))
```

p1	IC sat 95%	(1026.08, 1034.58)	Intervalo de confianza
p2	IC colgpa 90%	(2.64, 2.67)	Intervalo de confianza
p3	IC colgpa 95%	(2.64, 2.67)	Intervalo de confianza
p4	¿Colgpa de atletas es igual a la de no atletas?	obs:-5.91 crit:1.96 p:0	No son iguales con un NC de 95%
p5	¿Colgpa de mujeres es superior a la de hombres?	obs:7.08 crit:1.28 p:0	Si es superior con un NC de 95%
p6	¿Colgpa de raza blanca varia de raza negra?	obs:9.56 crit:1.96 p:0	Si varia con un NC de 95%
p7	¿Proporción de atletas es inferior a 0.05?	obs:-0.94 crit:-1.64 p:0.17	Si es inferior a 0.05 con un NC de 95%
p8	¿Proporción de atletas hombres es superior a 0.5 (superior a mujeres)?	obs:8.85 crit:1.64 p:0	Si es superior con un NC de 95%

10. Resumen ejecutivo

En este documento respondemos las siguientes preguntas:

1. ¿Cuál es el intervalo de confianza de la nota entre los estudiantes?

El intervalo de confianza al 95% se situa entre 2.63 y 2.67

2. ¿Ser atleta influye en la nota?

Podemos afirmar que sí influye con un nivel de confianza de 95%

3. ¿Las mujeres obtienen mejor nota que los hombres?

Podemos afirmar que sí sacan mejores notas con un nivel de confianza de 95%

4. ¿Hay diferencias significativas en la nota según la raza?

Podemos afirmar que sí influye con un nivel de confianza de 95% (No necesariamente por el hecho de ser de una raza sino por posibles motivos socioeconómicos)

5. ¿La proporción de atletas en la población es inferior al 5%?

Podemos afirmar que sí son inferior al 5% con un nivel de confianza de 95%

6. ¿Hay más atletas entre los hombres que entre las mujeres?

Podemos afirmar que sí son superiores en proporción respecto a las mujeres con un nivel de confianza de 95%