A2 - Analítica descriptiva e inferencial

Solución

Semestre 2022.1

Índice

Introducción

En esta actividad nos introducimos en la inferencia estadística. Para ello, usaremos el conjunto de datos `gpa.csv` que se ha preprocesado en la actividad anterior. Este conjunto de datos contiene la nota media de estudiantes universitarios tras el primer semestre de clases (GPA: grade point average, en inglés), así como información sobre la nota de acceso, la cohorte de graduación en el instituto y algunas características de los estudiantes.

Este conjunto de datos surge de una encuesta realizada a una muestra representativa de estudiantes de una universidad de EEUU (por razones de confidencialidad el conjunto de datos no incluye el nombre de la universidad). Las variables incluidas en el conjunto de datos son:

- sat: nota de acceso (medida en escala de 400 a 1600 puntos)
- tothrs: horas totales cursadas en el semestre
- 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
- 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
- hsperc: ranking relativo del estudiante, en porcentaje (hsrank/hsize)
- 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 colgpa (A,B,C,D)

El objetivo de esta actividad es estudiar la nota de los estudiantes a partir de las variables de interés así como la proporción de atletas entre la población de estudiantes. Para ello, las preguntas que nos planteamos son:

- P1. ¿Cuál es el intervalo de confianza de la nota entre los estudiantes?
- P2. ¿Ser atleta influye en la nota?
- P3. ¿Las mujeres obtienen mejor nota que los hombres?
- P4. ¿Hay diferencias significativas en la nota según la raza?
- P5. ¿La proporción de atletas en la población es inferior al 5%?
- P6. ¿Hay más atletas entre los hombres que entre las mujeres?

Las variables de interés para este estudio son:

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

---

1 Lectura del fichero

```
gpa<-read.csv("gpa_clean.csv",stringsAsFactors=TRUE)
#gpa <- gpa[ complete.cases(gpa$colgpa), ]
```

---

2 Estadística descriptiva y visualización

2.1 Análisis descriptivo

Realizad un análisis descriptivo numérico de los datos (resumid los valores de las variables numéricas y categóricas). Mostrad el número de observaciones y el número de variables.

```
cat("Observaciones: ", nrow(gpa), " Variables: ", length(gpa), "\n")
```

```
## Observaciones:  4137  Variables:  11
```

```
quantitative=c("sat","colgpa" )
means<-sapply( gpa[, quantitative ], mean)
sds <- sapply( gpa[, quantitative ], sd)
df1 <- data.frame(means);
dfs <- cbind( df1, sds)
colnames(dfs)<-c("mean","sd")
dfs
```

```
##               mean          sd
## sat    1030.331158 139.4013922
## colgpa    2.654131   0.6578132
```

```
qualitative<-c("athlete", "female", "white", "black")
sapply( gpa[,qualitative], table)
```

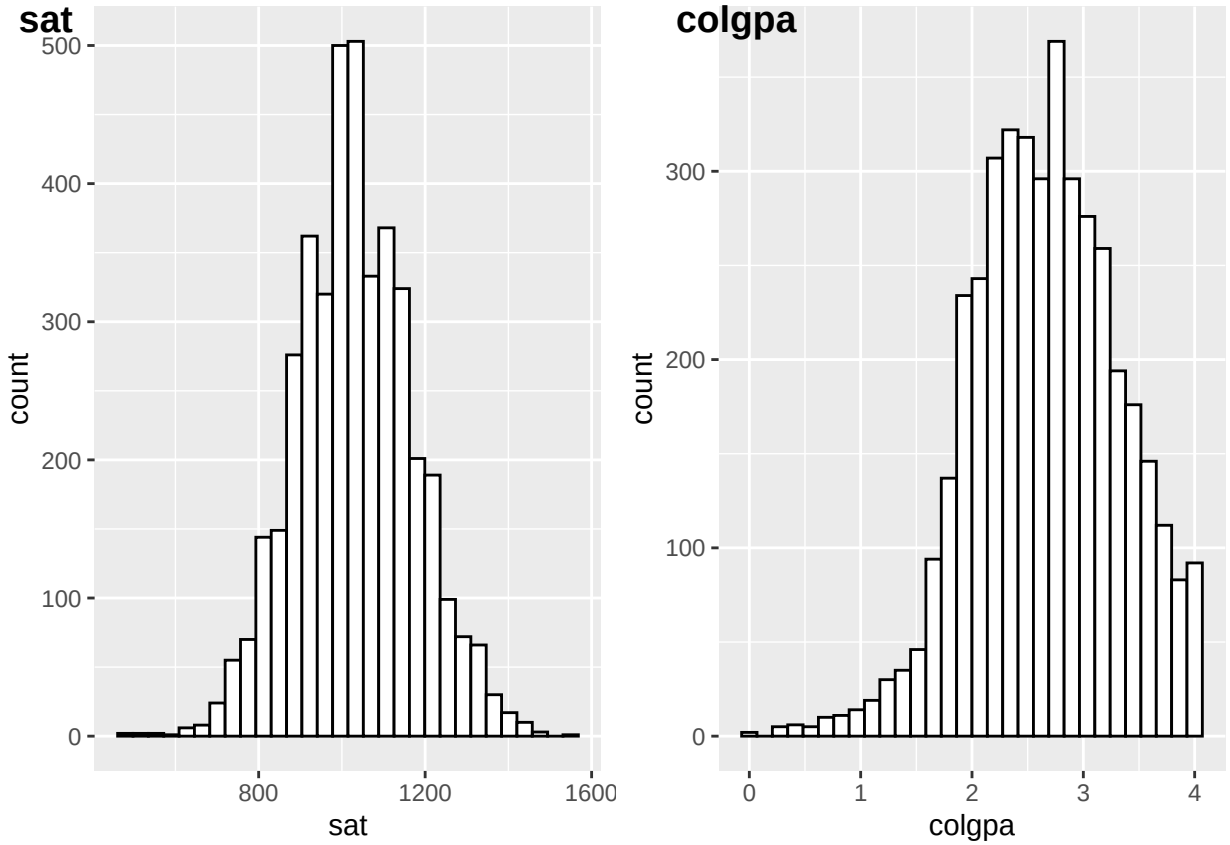
```
##       athlete female white black
## FALSE    3943   2277   308  3908
## TRUE      194   1860  3829   229
```

2.2 Visualización

Estudiaremos de forma visual la distribución de las variables `sat` y `colgpa`.

```
v.sat.hist<-ggplot(gpa, aes(x=sat)) +
  geom_histogram(color="black", fill="white")

v.gpa.hist<-ggplot(gpa, aes(x=colgpa)) +
  geom_histogram(color="black", fill="white")
ggarrange(v.sat.hist, v.gpa.hist,
          labels = c("sat", "colgpa"),
          ncol = 2, nrow = 1)
```

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



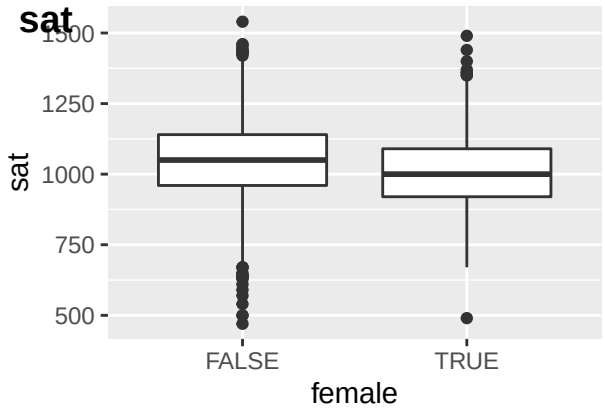

```
v.sat.fem<-ggplot(gpa, aes(x=female, y=sat)) +   geom_boxplot()
v.sat.at<-ggplot(gpa, aes(x=athlete, y=sat)) +   geom_boxplot()

gpa.f<-gpa[ gpa$white==TRUE | gpa$black==TRUE, ]
gpa.f$race <- ifelse( gpa.f$white==TRUE, "White", ifelse( gpa.f$black==TRUE, "Black", "NA"))
nrow(gpa.f)
```

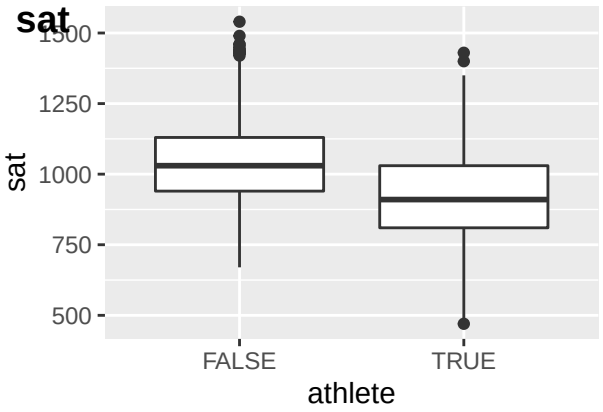
```
## [1] 4058
```

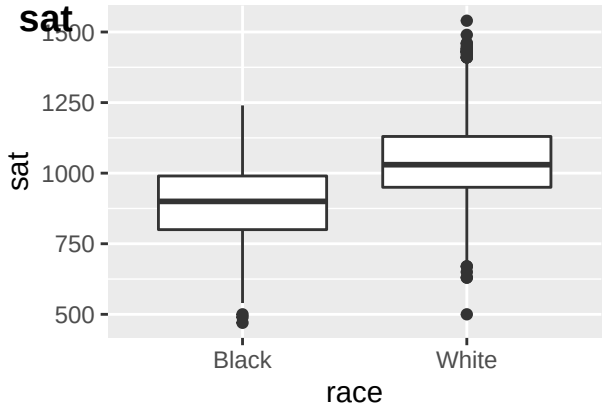
```
sum( complete.cases(gpa.f) )
```

```
## [1] 4058
```

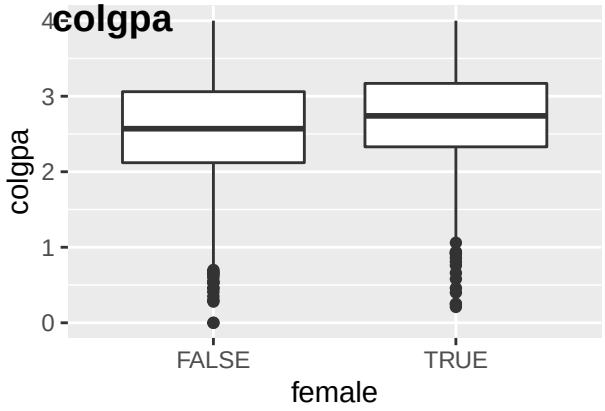
```
v.sat.white<-ggplot(gpa.f, aes(x=race, y=sat)) +   geom_boxplot()
ggarrange(v.sat.fem, v.sat.at, v.sat.white,
          labels = c("sat", "sat", "sat"),
          ncol = 2, nrow = 2)
```

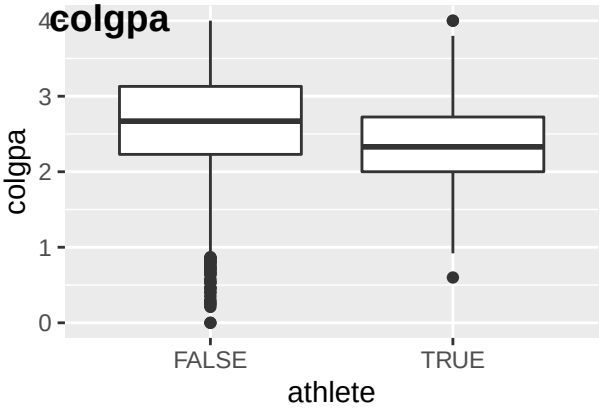


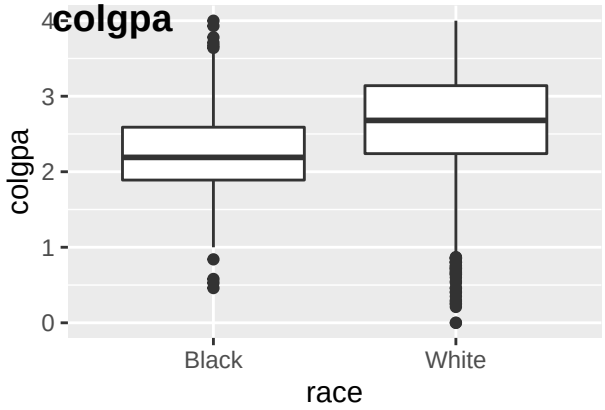





```
v.gpa.fem<-ggplot(gpa, aes(x=female, y=colgpa)) +   geom_boxplot()
v.gpa.at<-ggplot(gpa, aes(x=athlete, y=colgpa)) +   geom_boxplot()
v.gpa.white<-ggplot(gpa.f, aes(x=race, y=colgpa)) +   geom_boxplot()
ggarrange(v.gpa.fem, v.gpa.at, v.gpa.white,
          labels = c("colgpa", "colgpa", "colgpa"),
          ncol = 2, nrow = 2)
```







La variable `sat` se distribuye de forma similar a una normal, con valores centrados alrededor de 1000, y con mínimo de 400 y máximo 1600. La variable `colgpa` tiene una distribución parecida a una normal pero bastante asimétrica. Concretamente la cola de la izquierda es más alargada que la cola de la derecha, aunque presenta menos frecuencia de valores en los valores por debajo de 2.

Se observan diferencias en la variable `sat` según si el estudiante es atleta o no y también entre razas. Las diferencias en `sat` entre sexos son menos apreciables. Un comportamiento parecido ocurre con la variable `colgpa`, aunque en el caso de `colgpa` parece que el sexo femenino presenta mejor nota que el sexo masculino (al contrario de lo que sucede con `sat`).

---

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

3.1 Supuestos

Asumimos distribución normal por el Teorema del Límite Central, puesto que el tamaño de la muestra es suficientemente grande (n=4137). El matiz es 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.

3.2 Función de cálculo del intervalo de confianza

Función que calcula el intervalo de confianza dada una variable de interés `x` y un nivel de confianza dado `NC`.

```r
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))
}
```

3.3 Intervalo de confianza de la variable sat

Calculamos el intervalo de confianza al 90% de la media poblacional de la variable `sat` y al 95%.

```r
  ic.sat.90<-IC(gpa$sat, 90)
  ic.sat.95<-IC(gpa$sat, 95)

  ic.sat.90; ic.sat.95
```

```
## [1] 1026.765 1033.897
```

```
## [1] 1026.082 1034.580
```

```r
  #Comprobación
  t.test( gpa$sat, conf.level=0.90 )$conf.int
```

```
## [1] 1026.765 1033.897
## attr(,"conf.level")
## [1] 0.9
```

```r
  t.test( gpa$sat, conf.level=0.95 )$conf.int
```

```
## [1] 1026.082 1034.580
## attr(,"conf.level")
## [1] 0.95
```

3.4 Intervalo de confianza de la variable colgpa.

Cálculo del intervalo de confianza de la variable `colgpa` con un nivel de confianza del 90% y del 95%.

```r
  ic.colgpa.90<-IC(gpa$colgpa, 90)
  ic.colgpa.95<-IC(gpa$colgpa, 95)

  ic.colgpa.90; ic.colgpa.95
```

```
## [1] 2.637305 2.670957
```

```
## [1] 2.634080 2.674182
```

```r
  #Comprobación
  t.test( gpa$colgpa, conf.level=0.90 )$conf.int
```

```
## [1] 2.637305 2.670957
## attr(,"conf.level")
## [1] 0.9
```

```r
t.test( gpa$colgpa, conf.level=0.95 )$conf.int
```

```
## [1] 2.634080 2.674182
## attr(,"conf.level")
## [1] 0.95
```

3.5 Interpretación

La interpretación del intervalo de confianza es que si realizamos un muestreo elevado de muestras de la población, aproximadamente el NC% (95% o 90%) de los intervalos de confianza obtenidos de estas muestras contienen el valor de la media poblacional de colgpa/sat.

---

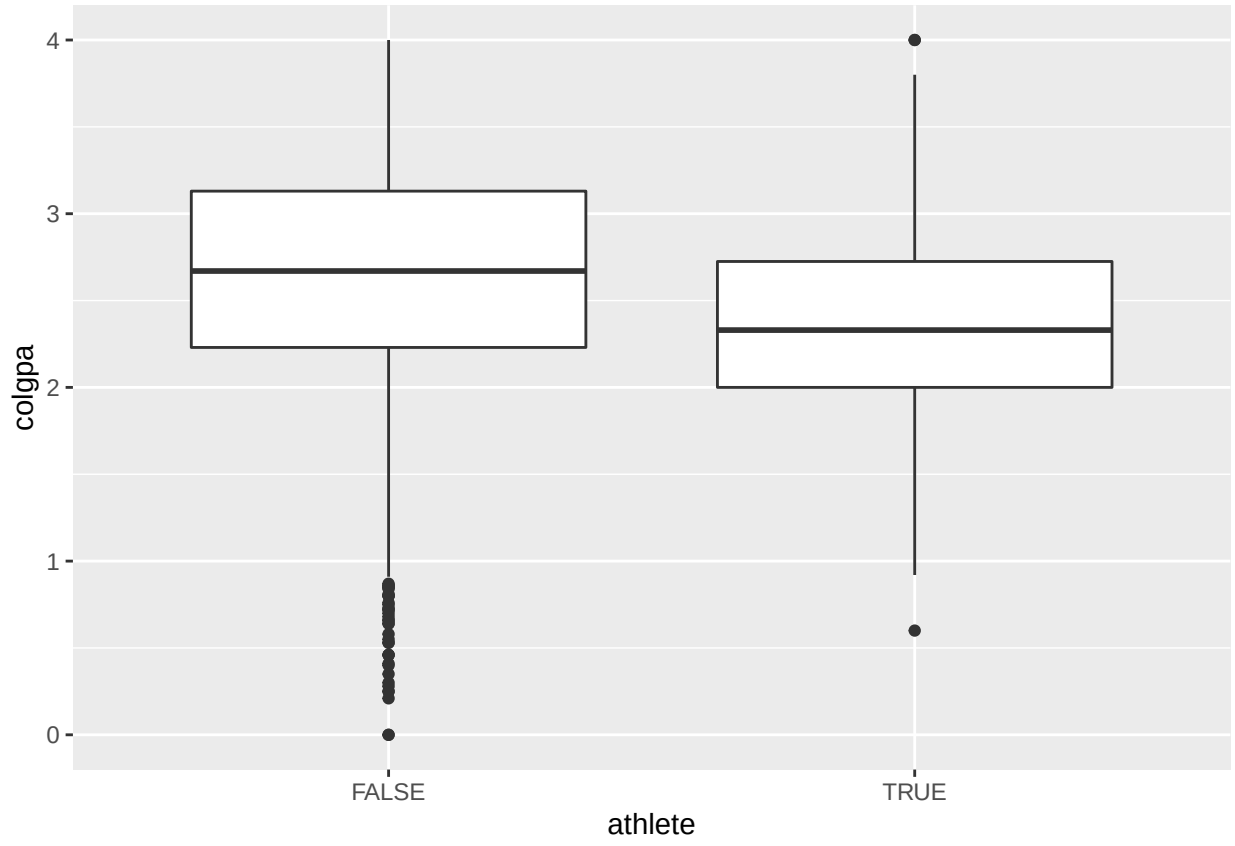
4 ¿Ser atleta influye en la nota?

En este apartado queremos analizar si ser atleta influye en la nota `colgpa`. Es decir, si hay diferencias significativas entre atletas y no atletas en la nota `colgpa`, con un nivel de confianza del 95%.

4.1 Análisis visual

Se muestra un boxplot con la variable colgpa comparando atletas con no atletas.

```r
ggplot( gpa, aes(x=athlete,y=colgpa)) + geom_boxplot()
```



4.2 Función para el contraste de medias

Se implementa una función que calcula el contraste (paramétrico) de medias de dos muestras y que devuelve: el valor del estadístico de contraste, el valor crítico, y el valor p. El contraste es bilateral. La función asume normalidad en la variable de interés.

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

  #varianzas iguales
  if (var.equal){
    S <- sqrt( ( (n1-1)*sd1^2 + (n2-1)*sd2^2 ) / (n1+n2-2) )
    t <- (mean1-mean2) / (S * sqrt(1/n1+1/n2) )
    df <- n1+n2-2
  }
  else{
    #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
    t<- (mean1-mean2) / Sb   #valor observado
  }

  tcritical <- qt( alfa/2, df, lower.tail=FALSE )    #two sided
  pvalue<-pt( abs(t), df, lower.tail=FALSE )*2       #two sided

  #Guardamos el resultado en un named vector
  info<-c(mean1, mean2, t,tcritical,pvalue,df)
  names(info)<-c("mean1", "mean2", "t","tcritical", "pvalue", "df")
  return (info)
}
```

4.3 Pregunta de investigación

```
## [1] "¿Hay diferencias significativas en la nota `colgpa` entre atletas y no atletas?"
```

4.4 Hipótesis nula y la alternativa

$H_0 : colgpa_{at} = colgpa_{noat}$

$H_1 : colgpa_{at} != colgpa_{noat}$

4.5 Justificación del test a aplicar

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:

```
at <- gpa[gpa$athlete==TRUE,]
no.at <- gpa[gpa$athlete==FALSE,]
var.test( at$colgpa, no.at$colgpa )
```

```
##
##  F test to compare two variances
##
```

```
## data:  at$colgpa and no.at$colgpa
## 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 no muestra diferencias significativas entre varianzas. Por tanto, aplicaremos un test de dos muestras independientes sobre la media con varianzas desconocidas iguales. El test es bilateral.

4.6 Cálculo

```
dif.colgpa.at <- my.ttest.bilateral( at$colgpa, no.at$colgpa, CL=95, var.equal = TRUE)
dif.colgpa.at
```

```
##         mean1         mean2             t     tcritical        pvalue
##  2.382732e+00  2.667484e+00 -5.910309e+00  1.960538e+00  3.689891e-09
##            df
##  4.135000e+03
```

```
answer.P2 <- dif.colgpa.at
#Comprobación
t.test( at$colgpa, no.at$colgpa, alternative="two.sided", conf.level=0.95, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  at$colgpa and no.at$colgpa
## 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

Existen diferencias significativas en la nota `colgpa` entre los atletas y no atletas (p=3.6898912×10^{-9}). Se puede observar asimismo que el valor crítico es 1.9605379 y el valor observado es 5.9103092, notablemente superior al valor crítico y por tanto, fuera de la zona de aceptación de la hipótesis nula.

---

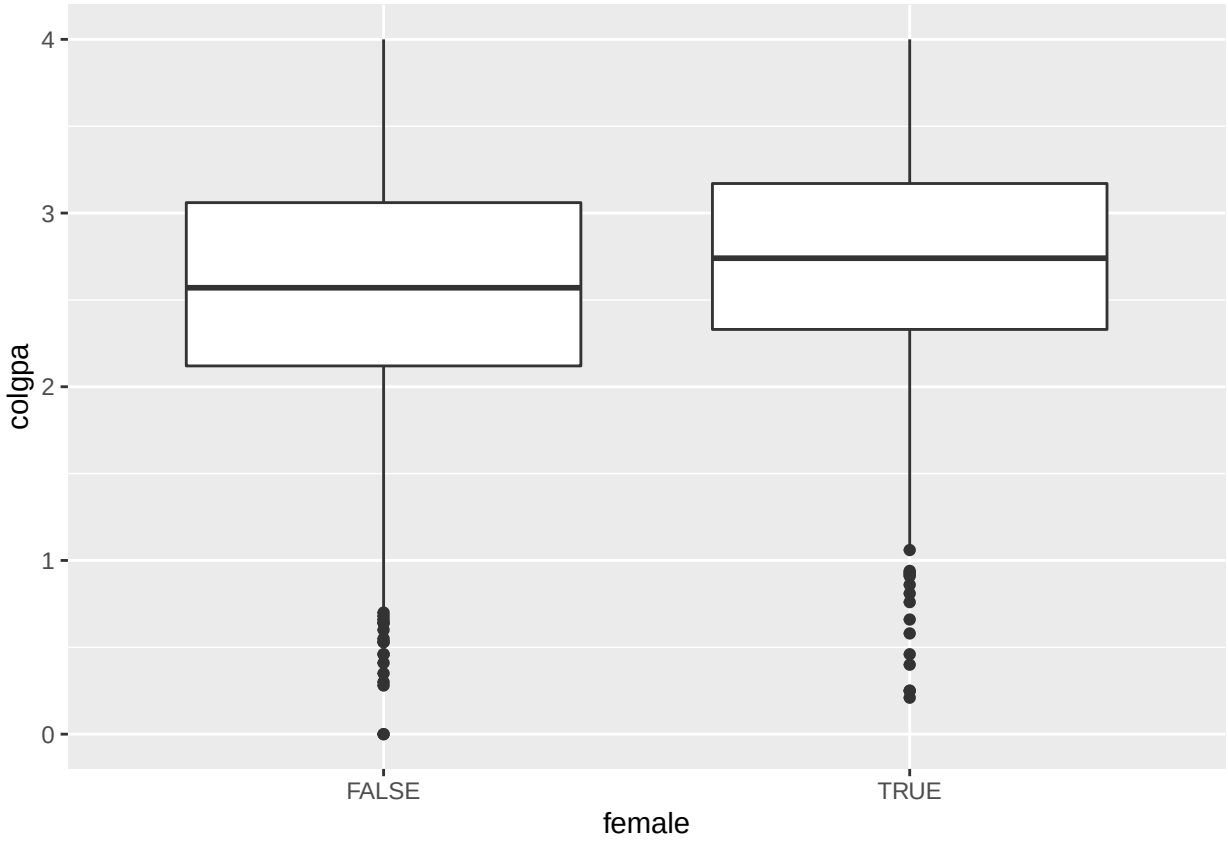
5 ¿Las mujeres tienen mejor nota que los hombres?

Se calcula para un nivel de confianza del 95% y del 90%.

5.1 Análisis visual

Se muestra un boxplot con la variable colgpa comparando mujeres con hombres.

```
ggplot( gpa, aes(x=female,y=colgpa)) + geom_boxplot()
```



5.2 Función

Se implementa una función para el contraste de medias unilateral.

```
my.ttest.unilateral <- function( x1, x2, CL=95, alternative="less", var.equal=FALSE ){
  mean1<-mean(x1); n1<-length(x1); sd1<-sd(x1)
  mean2<-mean(x2); n2<-length(x2); sd2<-sd(x2)
  alfa <- (1-CL/100)
  #varianzas iguales
  if (var.equal){
    S <- sqrt( ( (n1-1)*sd1^2 + (n2-1)*sd2^2 ) / (n1+n2-2) )
    t <- (mean1-mean2) / (S * sqrt(1/n1+1/n2) )
    df <- n1+n2-2
  }
  else{
    #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
    t<- (mean1-mean2) / Sb   #valor observado
  }
  #less
  if (alternative=="less"){
    tcritical <- qt( alfa, df, lower.tail=TRUE )
    pvalue<-pt( t, df, lower.tail=TRUE )
  }
```

```
  else{ #greater
    tcritical <- qt( alfa, df, lower.tail=FALSE )
    pvalue<-pt( t, df, lower.tail=FALSE )
  }

  #Guardamos el resultado en un named vector
  info<-c(mean1, mean2, t,tcritical,pvalue,df)
  names(info)<-c("mean1", "mean2", "t","tcritical", "pvalue", "df")
  return (info)
}
```

5.3 Pregunta de investigación

```
## [1] "¿Las mujeres tienen mejor nota `colgpa` que los hombres?"
```

5.4 Hipótesis nula y la alternativa

Escribid las hipótesis nula y alternativa.

$H_0 : colgpa_{Female} = colgpa_{Male}$

$H_1 : colgpa_{Female} > colgpa_{Male}$

5.5 Justificación del test a aplicar

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:

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

```
##
##  F test to compare two variances
##
## data:  fem$colgpa and male$colgpa
## 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
```

```
pvalue<-var.test( fem$colgpa, male$colgpa )$p.value
```

El resultado del test muestra diferencias significativas entre varianzas (p=2.0238439×10^{-5}). Por tanto, aplicaremos un test de dos muestras independientes sobre la media con varianzas desconocidas diferentes. El test es unilateral por la derecha.

5.6 Cálculo

```
dif.colgpa.fem <- my.ttest.unilateral( fem$colgpa, male$colgpa, CL=95,
                                       var.equal = FALSE, alternative="greater")
dif.colgpa.fem
```

```
##        mean1        mean2            t    tcritical       pvalue           df
## 2.733016e+00 2.589693e+00 7.078735e+00 1.645227e+00 8.521971e-13 4.087407e+03
```

```
answer.P3 <- dif.colgpa.fem
#Comprobación
t.test( fem$colgpa, male$colgpa, alternative="greater", conf.level=0.95, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  fem$colgpa and male$colgpa
## t = 7.0787, df = 4087.4, p-value = 8.522e-13
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1100126       Inf
## sample estimates:
## mean of x mean of y
##  2.733016  2.589693
```

5.7 Interpretación del test

Se puede afirmar que las estudiantes de sexo femenino tienen una mejor nota que los estudiantes de sexo masculino (p=8.521971×10^{-13}). Se puede observar asimismo que el valor crítico es 1.6452265 y el valor observado es 7.0787353, notablemente superior al valor crítico y por tanto, fuera de la zona de aceptación de la hipótesis nula. No es necesario calcular para nivel de confianza del 90% puesto que se han encontrado diferencias significativas con el nivel de confianza del 95%.

---

6 ¿Hay diferencias en la nota según la raza?

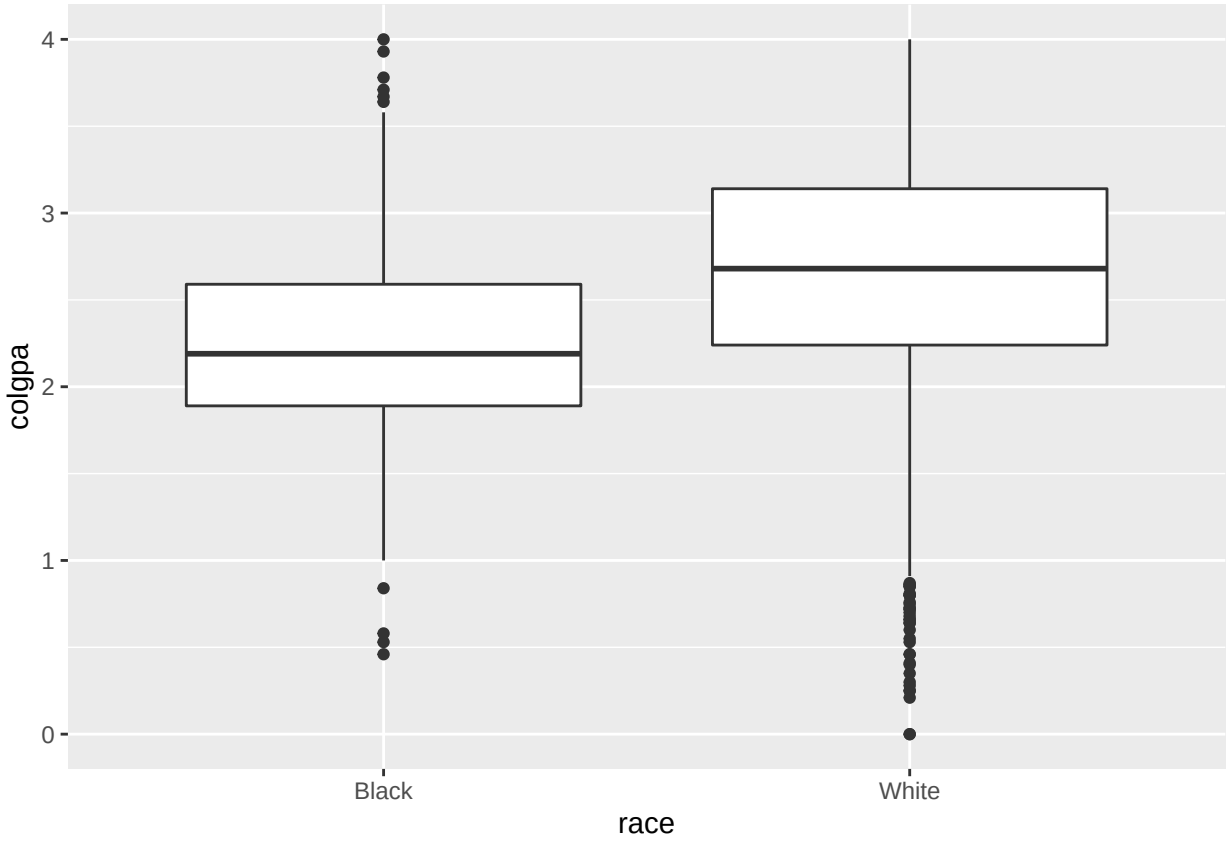
Antes de desarrollar este apartado, queremos hacer énfasis que el hecho que existan (si existen) diferencias entre razas (blancos y negros), no necesariamente nos lleva a concluir que la raza influye en la nota (existen factores socioeconómicos que pueden afectar). Por tanto, se debe ser muy cauteloso a la hora de realizar conclusiones en este sentido. Un tema similar puede suceder con las diferencias en cuanto a género. A pesar de estas puntualizaciones, el estudio es interesante puesto que en caso de detectar diferencias en uno u otro sentido, se pueden analizar los causantes de estas diferencias y realizar intervenciones educativas apropiadas.

Después de este matiz, nos gustaría estudiar si las personas de raza blanca tienen una nota diferente en `colgpa` que las personas de raza negra. Seguid los mismos apartados que anteriormente.

6.1 Análisis visual

Se muestra un boxplot con la variable colgpa comparando mujeres con hombres.

```
ggplot( gpa.f, aes(x=race,y=colgpa)) + geom_boxplot()
```



6.2 Función

Usaremos una de las funciones anteriores.

6.3 Pregunta de investigación

```
## [1] "¿Existen diferencias en `colgpa` entre los estudiantes de raza blanca y los de raza negra?"
```

6.4 Hipótesis nula y la alternativa

$H_0 : colgpa_{white} = colgpa_{black}$

$H_1 : colgpa_{white} != colgpa_{black}$

6.5 Justificación del test a aplicar

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:

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

```
##
##  F test to compare two variances
##
## data:  white$colgpa and black$colgpa
```

```
## 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 no muestra diferencias significativas entre varianzas. Por tanto, aplicaremos un test de dos muestras independientes sobre la media con varianzas desconocidas iguales. El test es bilateral.

6.6 Cálculo

```
dif.gpa.white <- my.ttest.bilateral( white$colgpa, black$colgpa, CL=95, var.equal = TRUE)
dif.gpa.white
```

```
##        mean1        mean2            t    tcritical       pvalue           df
## 2.678360e+00 2.255546e+00 9.559319e+00 1.960549e+00 1.990140e-21 4.056000e+03
```

```
answer.P4 <- dif.gpa.white
#Comprobación
t.test( white$colgpa, black$colgpa, alternative="two.sided", conf.level=0.95, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  white$colgpa and black$colgpa
## t = 9.5593, df = 4056, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3360979 0.5095302
## sample estimates:
## mean of x mean of y
##  2.678360  2.255546
```

6.7 Interpretación del test

Existen diferencias significativas en la nota `colgpa` entre los estudiantes de raza blanca y los de raza negra (p=$1.9901397 \times 10^{-21}$). Se puede observar asimismo que el valor crítico es 1.960549 y el valor observado es 9.5593191, el cual se encuentra fuera de la zona de aceptación de la hipótesis nula.

---

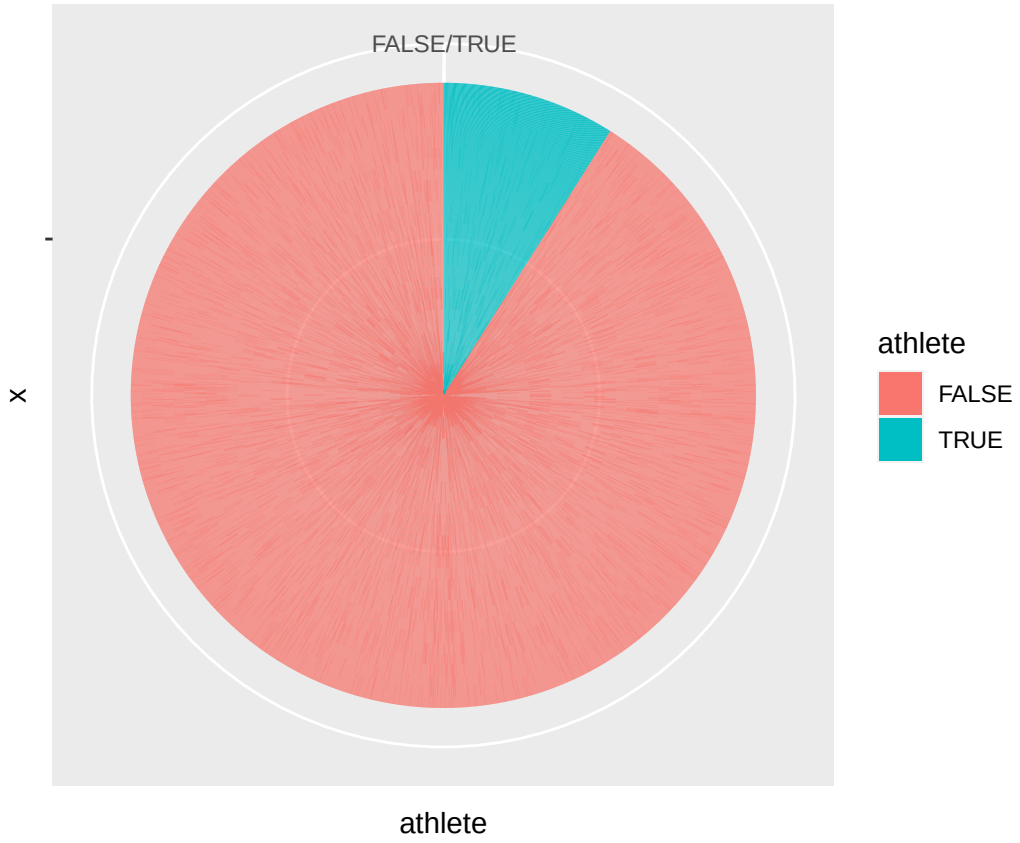
7 Proporción de atletas

Nos preguntamos si la proporción de atletas en la población es inferior al 5% con un nivel de confianza del 95%. Para ello, seguid los mismos pasos que en los casos anteriores.

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

7.1 Análisis visual

```
ggplot( gpa, aes(x="", y=athlete, fill=athlete)) +
  geom_bar(stat="identity", width = 1) +
  coord_polar("y",start=0)
```



7.2 Pregunta de investigación

```
## [1] "¿La proporción de atletas en la población es inferior a 0.05?"
```

7.3 Hipótesis nula y la alternativa

$H_0 : p_{at} = 0.05$

$H_1 : p_{at} < 0.05$

7.4 Justificación del test a aplicar

Es un contraste sobre la proporción de una muestra. Es un test unilateral por la izquierda. Asumimos muestras grandes.

7.5 Cálculos

```
#Test de proporción de una muestra unilateral por la izquierda.
my.proptest.left <-function( p, p0, n, CL=95 ){
  z <- (p-p0)/sqrt( (p0*(1-p0)/n))
  alfa <- 1 - CL/100

  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)
}
```

7.6 Cálculo

```
answer.P5 <- my.proptest.left( nrow(at)/nrow(gpa), p0=0.05, nrow(gpa)); answer.P5
```

```
##           p          p0           z   zcritical      pvalue
##  0.04689388  0.05000000 -0.91667115 -1.64485363  0.17965749
```

7.7 Interpretación del test

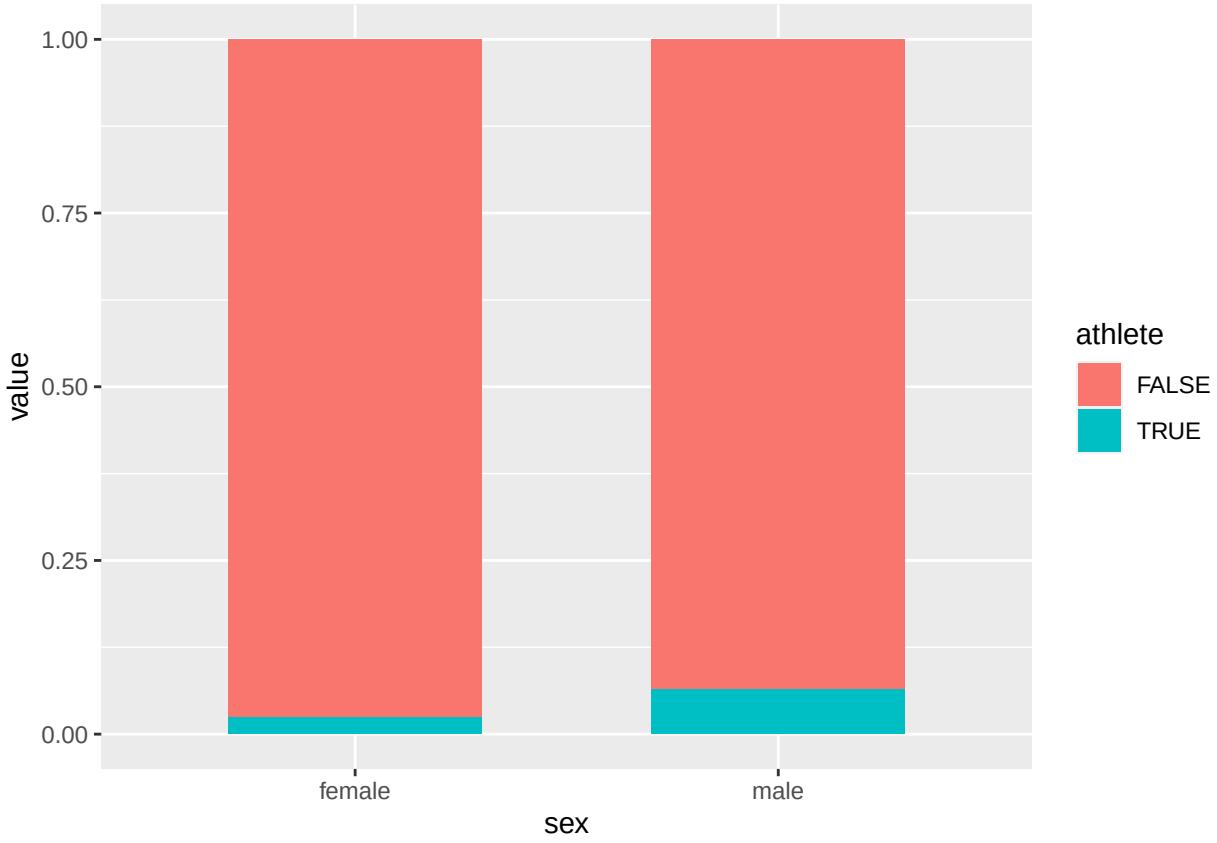
No se puede afirmar que la proporción de atletas en la población es inferior al 5% con un nivel de confianza del 95%.

---

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

8.1 Análisis visual

```
  sex<- c(rep("female",2), rep("male",2))
  athlete <-c( TRUE, FALSE, TRUE, FALSE)
  value <- c( sum(gpa$female==TRUE & gpa$athlete==TRUE),
              sum(gpa$female==TRUE & gpa$athlete==FALSE),
              sum(gpa$female==FALSE & gpa$athlete==TRUE),
              sum(gpa$female==FALSE & gpa$athlete==FALSE)
              )
  df <- data.frame(sex,athlete, value)
  ggplot( df, aes(x=sex, y=value, fill=athlete)) +
    geom_bar(position="fill", stat="identity", width=0.6)
```



8.2 Pregunta de investigación

```
## [1] "¿La proporción de atletas entre los hombres es mayor que entre las mujeres?"
```

8.3 Hipótesis nula y alternativa

$H_0 : p_{atM} = p_{atF}$

$H_1 : p_{atM} > p_{atF}$

8.4 Justificación del test a aplicar

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 por la derecha.

8.5 Cálculos

```
my.proptest2.D <-function ( x1,x2,n1,n2, CL=95){
 p1 <- x1/n1
 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)) )
 pvalue <- pnorm( zobs, lower.tail=FALSE)
 zcrit <- qnorm( alfa, lower.tail=FALSE )
 result <-  c(p1,p2,zobs, zcrit, pvalue)
```

```
 names(result) <- c("p1", "p2", "zobs","zcrit", "pvalue")
 return (result)
}

n.fem <- nrow( fem )
n.male <- nrow( male )
n.at.fem <- nrow( fem[fem$athlete==TRUE,] )
n.at.male <- nrow( male[male$athlete==TRUE,] )


answer.P6 <- my.proptest2.D( n.at.male, n.at.fem,n.male,n.fem,95)
answer.P6
```

```
##           p1           p2         zobs        zcrit       pvalue
## 6.543698e-02 2.419355e-02 6.241964e+00 1.644854e+00 2.160550e-10
```

```
#Validación con prop.test
success<-c( n.at.fem, n.at.male)
nn<-c(n.fem,n.male)
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 = 38.962, df = 1, p-value = 2.161e-10
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.03089911
## sample estimates:
##     prop 1     prop 2
## 0.02419355 0.06543698
```

8.6 Interpretación del test

Podemos afirmar que la proporción de atletas entre los estudiantes de sexo masculino es mayor a la proporción de atletas entre el sexo femenino, con un nivel de confianza del 95%. El valor p es $2.1605497 \times 10^{-10}$. Asimismo, se comprueba que el valor crítico es 1.6448536 y el valor observado es 6.2419642, estando éste último en de la región de rechazo de la hipótesis nula.

---

9 Resumen y conclusiones

En la tabla siguiente, se presentan los resultado de cada pregunta de investigación resumidos:

N	Pregunta	Resultado (valor observado, crítico, valor p...)	Conclusión
1a	Intervalo de confianza 95% y 90% sat	ic90=1026.765, 1033.897; ic95=1026.082, 1034.58	-
1b	Intervalo de confianza 95% y 90% colgpa	ic90=2.637, 2.671; ic95=2.634, 2.674	-
2	¿Hay diferencias significativas en la nota 'colgpa' entre atletas y no atletas?	obs=5.91; crit=1.961; pvalue=3.6898912×10^{-9}	Hay diferencias significativas en la nota 'colgpa' entre atletas y no atletas con un NC del 95%.
3	¿Las mujeres tienen mejor nota 'colgpa' que los hombres?	obs=7.079; crit=1.645; pvalue=8.521971×10^{-13}	Las mujeres tienen mejor nota en 'colgpa' que los hombres con NC 95%.
4	¿Existen diferencias en 'colgpa' entre los estudiantes de raza blanca y los de raza negra?	obs=9.559; crit=1.961; pvalue=$1.9901397 \times 10^{-21}$	Existen diferencias significativas al 95% en la nota colgpa entre estudiantes de raza blanca y los de raza negra
5	¿La proporción de atletas en la población es inferior a 0.05?	obs=-0.9166711; crit=-1.6448536; pvalue=0.1796575	No se puede afirmar que la proporción de atletas en la población es inferior al 5% con un nivel de confianza del 95%.
6	¿La proporción de atletas entre los hombres es mayor que entre las mujeres?	obs=6.2419642; crit=1.6448536; pvalue=$2.1605497 \times 10^{-10}$	La proporción de atletas en la población de hombres es mayor que en las mujeres con NC 95%.

10 Resumen ejecutivo

Se ha realizado un análisis descriptivo e inferencial del conjunto de datos gpa, el cual contiene los datos de una muestra de estudiantes de una universidad de EEUU. Entre otras variables, el conjunto de datos contiene la nota de acceso (sat) y la nota media de cada estudiante al finalizar el primer semestre (golgpa). Los datos contienen el sexo del estudiante, la raza y si el estudiante es atleta.

En base a estos datos, se han obtenido un conjunto de conclusiones que se resumen a continuación. Todas las conclusiones se extraen con un nivel de confianza del 95% y se pueden generalizar a la población de estudiantes de Estados Unidos, asumiendo que la muestra analizada sea suficientemente representativa de la población.

- La variable de nota de acceso toma el valor medio entre 1026.63 y 1035.17 puntos.
- La nota media al finalizar el primer semestre está entre 2.63 y 2.67.
- Se observan diferencias significativas en la nota al finalizar el primer semestre entre estudiantes de raza blanca y estudiantes de raza negra.
- Se puede afirmar que las estudiantes de sexo femenino obtienen mejor nota al finalizar el semestre que los estudiantes de sexo masculino.
- Asimismo, se observan diferencias significativas en la nota al finalizar el primer semestre entre los estudiantes atletas y los que no son atletas.
- La proporción de atletas en la población no es inferior al 5%.
- Se puede afirmar que la proporción de atletas entre los estudiantes de sexo masculino es mayor que entre las estudiantes de sexo femenino.