PEC2 - Analítica descriptiva e inferencial

Gloria Manresa Santamaría

2022-12-02

1. Lectura del fichero

Cargamos el archivo de datos y lo guardamos bajo el nombre "gpa", mostramos las primeras y las últimas líneas para asegurarnos de que se han cargado correctamente todas las entradas. También comprobamos que el número de entradas en la base de datos original coincide con el número de líneas cargadas en R.

```
gpa <- read.csv('gpa_clean.csv')
```

```
kable(head(gpa,3))
```

sat	tothrs	hsize	hsrank	hsperc	colgpa	athlete	female	white	black	gpaletter
920	43	0.10	4	40.00000	2.04	TRUE	TRUE	FALSE	FALSE	C
1170	18	9.40	191	20.31915	4.00	FALSE	FALSE	TRUE	FALSE	A
810	14	1.19	42	35.29412	1.78	TRUE	FALSE	TRUE	FALSE	C

```
kable(tail(gpa,3))
```

	sat	tothrs	hsize	hsrank	hsperc	colgpa	athlete	female	white	black	gpaletter
4135	1340	62	0.45	1	2.222222	4.00	FALSE	FALSE	TRUE	FALSE	A
4136	980	12	0.35	23	65.714287	2.83	FALSE	TRUE	TRUE	FALSE	B
4137	1420	128	3.13	38	12.140580	3.94	FALSE	FALSE	TRUE	FALSE	A

Confirmamos que las 4137 entradas coinciden con las líneas del documento excel original.

Comprobamos el tipo de datos con el que R ha interpretado cada variable:

```
kable(sapply(gpa,class),col.names="Tipo variable")
```

	Tipo variable
sat	integer
tothrs	integer
hsize	numeric
hsrank	integer
hsperc	numeric
colgpa	numeric
athlete	logical
female	logical
white	logical
black	logical
gpaletter	character

Confirmamos que el tipo de datos de cada variable es correcto.

2. Estadística descriptiva y visualización

2.1 Análisis descriptivo

Empezaremos con mostrar el nombre de las variables y las dimensiones de la tabla (número de entradas y número de variables).

```
# Nombre de variables:
names(gpa)
```

```
##  [1] "sat"       "tothrs"    "hsize"     "hsrank"    "hsperc"    "colgpa"
##  [7] "athlete"   "female"    "white"     "black"     "gpaletter"
```

```
# Dimensiones tabla
dim(gpa)
```

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

Observamos que los datos están organizados en 11 variables y existen 4137 entradas.

A continuación mostramos una serie de estadísticos de las variables numéricas.

No se muestra el código ya que es el mismo utilizado en la entrega anterior (PEC1) y con la intención de no entregar un documento demasiado largo.

- **Nota de acceso "sat"**

Estadísticos de la variable "sat":

	SAT_HOMBRES	SAT_MUJERES
Media aritmética	1049.70	1006.62
Mediana	1050.00	1000.00
Media recortada	1050.15	1005.26
Desviación estándar	144.98	128.37
Rango intercuartílico	180.00	170.00
Desviación absoluta respecto de la mediana	133.43	133.43

Summary de la variable "sat":

```
summary(gpa$sat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     470     940    1030    1030    1120    1540
```

Observamos algunos de los estadísticos de la nota de acceso de los hombres y de las mujeres. Encontramos en el caso de los hombres los datos más dispersos (rango intercuartílico mayor). En el caso de las mujeres los datos se encuentran menos dispersos en torno a una mediana menor.

- **Horas totales cursadas en el semestre, "tothrs"**

Estadísticos de la variable “tothrs”:

	TOTHRS
Media aritmética	52.83
Mediana	47.00
Media recortada	51.27
Desviación estándar	35.33
Rango intercuartílico	63.00
Desviación absoluta respecto de la mediana	45.96

Summary de la variable “tothrs”:

```
summary(gpa$tothrs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   17.00   47.00   52.83   80.00  137.00
```

La desviación típica es bastante alta lo que implica unos datos muy dispersos en torno a la media.

- **Nota media del estudiante al final del primer semestre, “colgpa”**

Estadísticos de la variable “colgpa”:

	COLGPA
Media aritmética	2.65
Mediana	2.66
Media recortada	2.66
Desviación estándar	0.66
Rango intercuartílico	0.91
Desviación absoluta respecto de la mediana	0.67

Summary de la variable “colgpa”:

```
summary(gpa$colgpa)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.210   2.660   2.654   3.120   4.000
```

- **Número total de estudiantes, “hsize”**

Estadísticos de la variable “hsize”:

	HSIZE
Media aritmética	2.80
Mediana	2.51
Media recortada	2.71
Desviación estándar	1.74
Rango intercuartílico	2.03
Desviación absoluta respecto de la mediana	1.42

Summary de la variable “hsize”:

```
summary(gpa$hsize)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.03    1.65    2.51    2.80    3.68    9.40
```

- **Ranking del estudiante, "hsrank"**

Estadísticos de la variable "hsrank":

	HSRANK
Media aritmética	52.83
Mediana	30.00
Media recortada	43.99
Desviación estándar	64.68
Rango intercuartílico	59.00
Desviación absoluta respecto de la mediana	35.58

Summary de la variable "hsrank":

```
summary(gpa$hsrank)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   11.00   30.00   52.83   70.00  634.00
```

- **Ranking relativo del estudiante, "hsperc"**

Estadísticos de la variable "hsperc":

	HSPERC
Media aritmética	19.24
Mediana	14.58
Media recortada	17.73
Desviación estándar	16.57
Rango intercuartílico	21.28
Desviación absoluta respecto de la mediana	14.05

Summary de la variable "hsperc":

```
summary(gpa$hsperc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1667  6.4328 14.5833 19.2371 27.7108 92.0000
```

A continuación un resumen de las variables categóricas:

- **Indicador si el estudiante practica algún deporte en la universidad, "athlete"**

```
kable(table(gpa$athlete),col.names = c("Athlete","Frequency"))
```

Athlete	Frequency
FALSE	3943
TRUE	194

- **Indicador de si el estudiante es mujer, "female"**

```
kable(table(gpa$female),col.names = c("Female","Frequency"))
```

Female	Frequency
FALSE	2277
TRUE	1860

- **Indicador de si el estudiante es de raza blanca o no, "white"**

```
kable(table(gpa$white),col.names = c("White","Frequency"))
```

White	Frequency
FALSE	308
TRUE	3829

- **Indicador de si el estudiante es de raza negra o no, "black"**

```
kable(table(gpa$black),col.names = c("Black","Frequency"))
```

Black	Frequency
FALSE	3908
TRUE	229

- **Letra que indica el nivel de la nota colgpa, "gpaletter"**

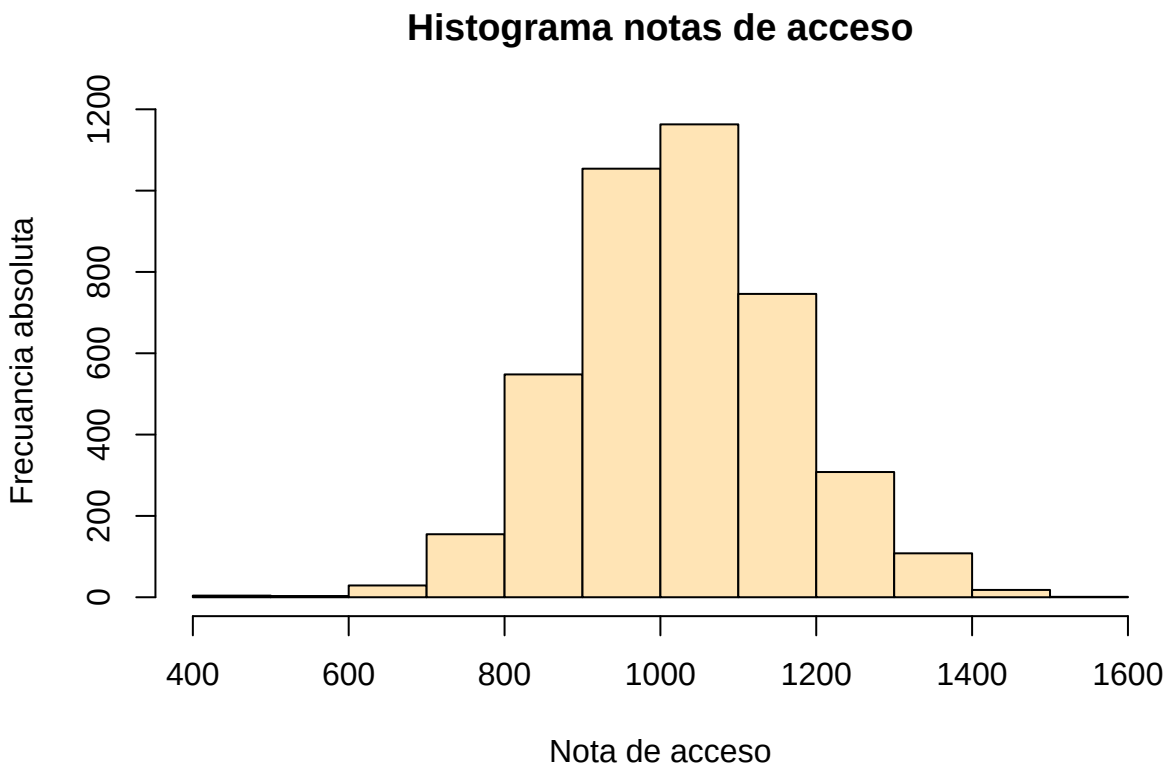
```
kable(table(gpa$gpaletter),col.names = c("GPA Letter","Frequency"))
```

GPA Letter	Frequency
A	458
B	1999
C	1536
D	144

2.2 Visualización

Distribución variables "sat" y "colgpa", por separado A continuación un histograma para observar la distribución de la variable "sat":

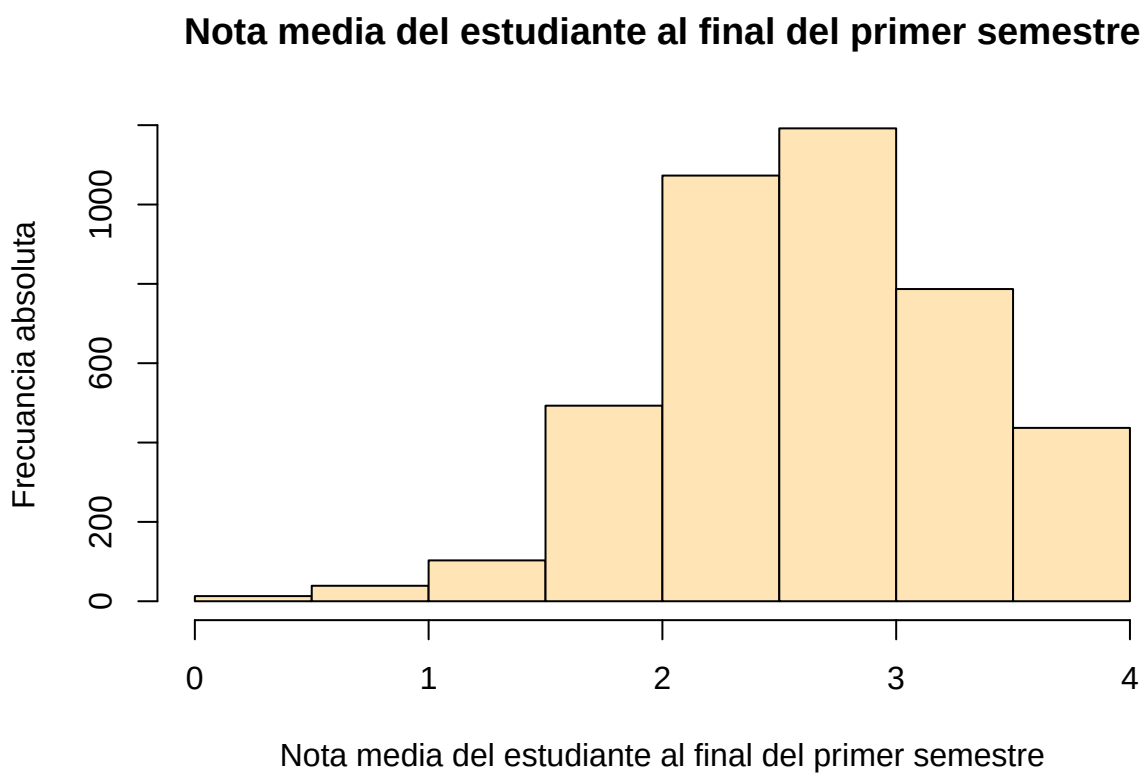
```
hist(gpa$sat,
    main = "Histograma notas de acceso",
    col="moccasin",
    xlab = "Nota de acceso",
    ylab = "Frecuancia absoluta"
)
```



Los datos de la variable "sat" que muestra las notas de acceso siguen una distribución normal según podemos observar en el gráfico.

Distribución de la variable "colgpa":

```
hist(gpa$colgpa,
    main = "Nota media del estudiante al final del primer semestre",
    col="moccasin",
    xlab = "Nota media del estudiante al final del primer semestre",
    ylab = "Frecuancia absoluta"
)
```

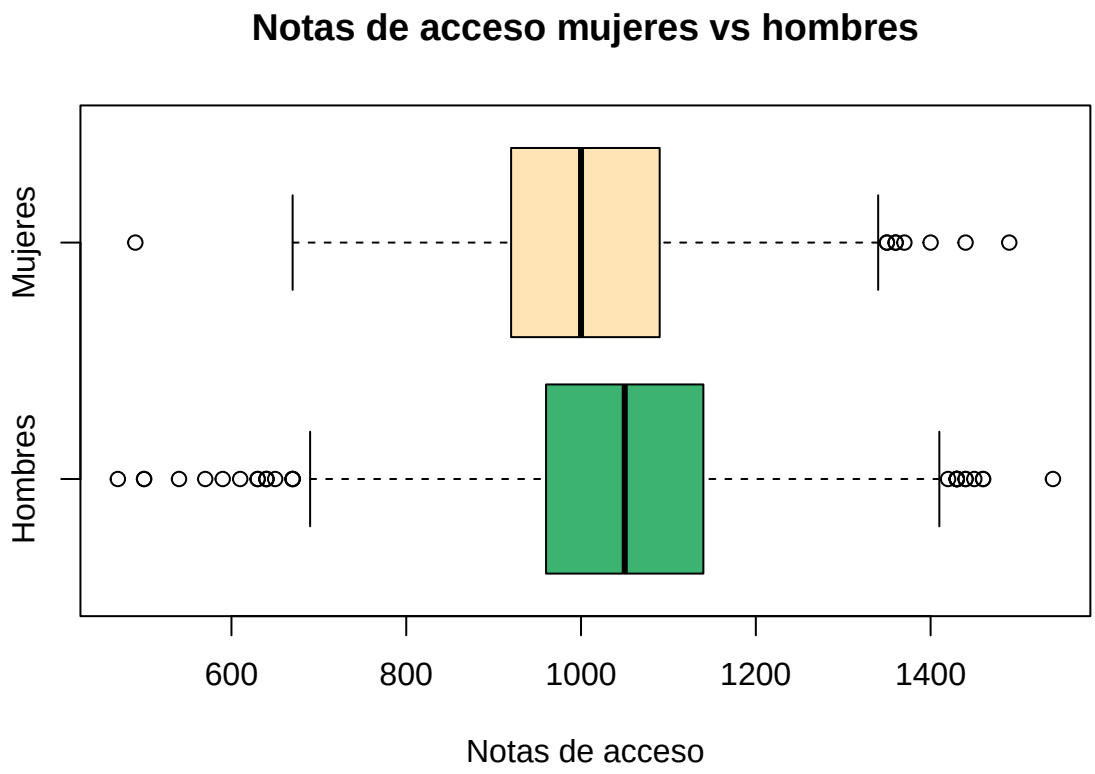


En este caso encontramos una cierta asimetría en los datos encontrando una cola a la izquieda.

Distribución de la variable "sat" con respecto a la variable género, atleta y raza.

- **1.Distribución de la variable "sat" con respecto a la variable género (female)**

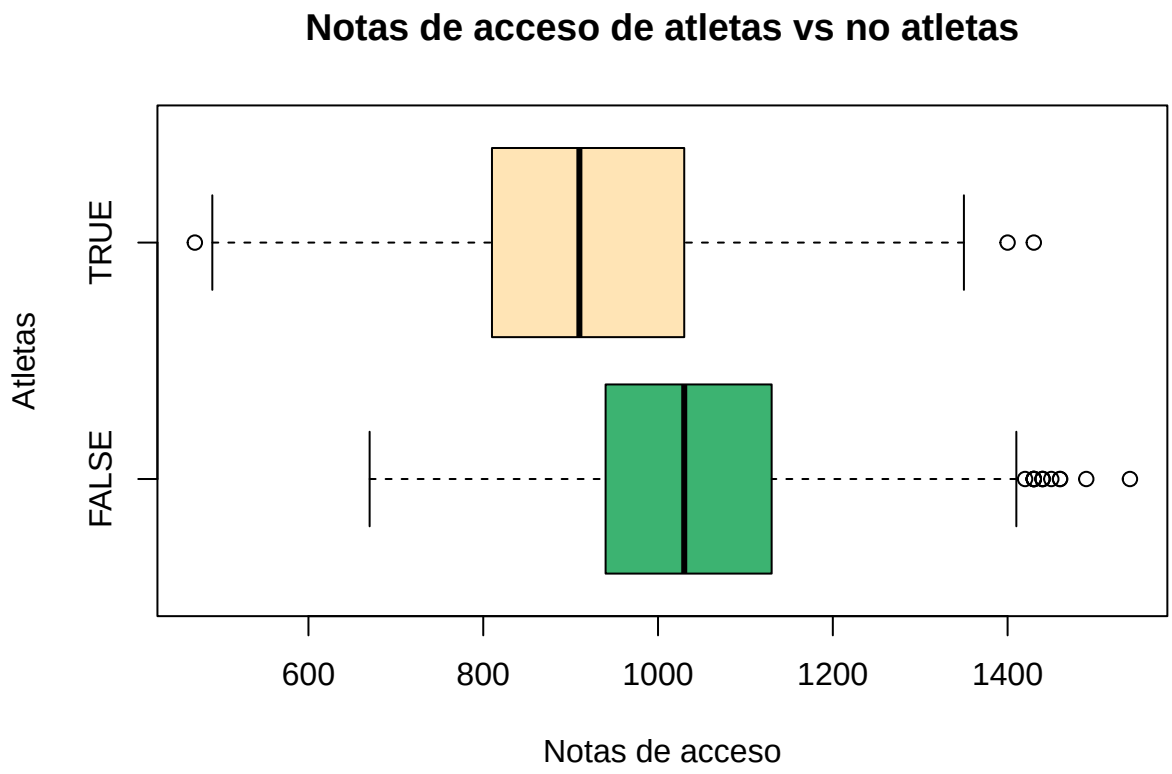
```
boxplot(gpa$sat~gpa$female,
  main = "Notas de acceso mujeres vs hombres",
  xlab = "Notas de acceso",
  ylab = NULL,
  col=c("mediumseagreen","moccasin"),
  horizontal = T,
  names = c("Hombres","Mujeres") )
```



En el diagrama de cajas observamos que la mediana de las mujeres es ligeramente inferior a la de los hombres. En lo que se refiere a la dispersión de los datos en el caso de los hombres encontramos los datos más dispersos ya que el rango intercuartílico es ligeramente mayor y los bigotes también son más largos. Por el contrario, en el caso de las mujeres los datos se encuentran menos dispersos en torno a una mediana menor. También cabe destacar que los hombres presentan más valores atípicos, sobre todo en el límite inferior.

- **2. Distribución de la variable "sat" con respecto a la variable atleta (athlete)**

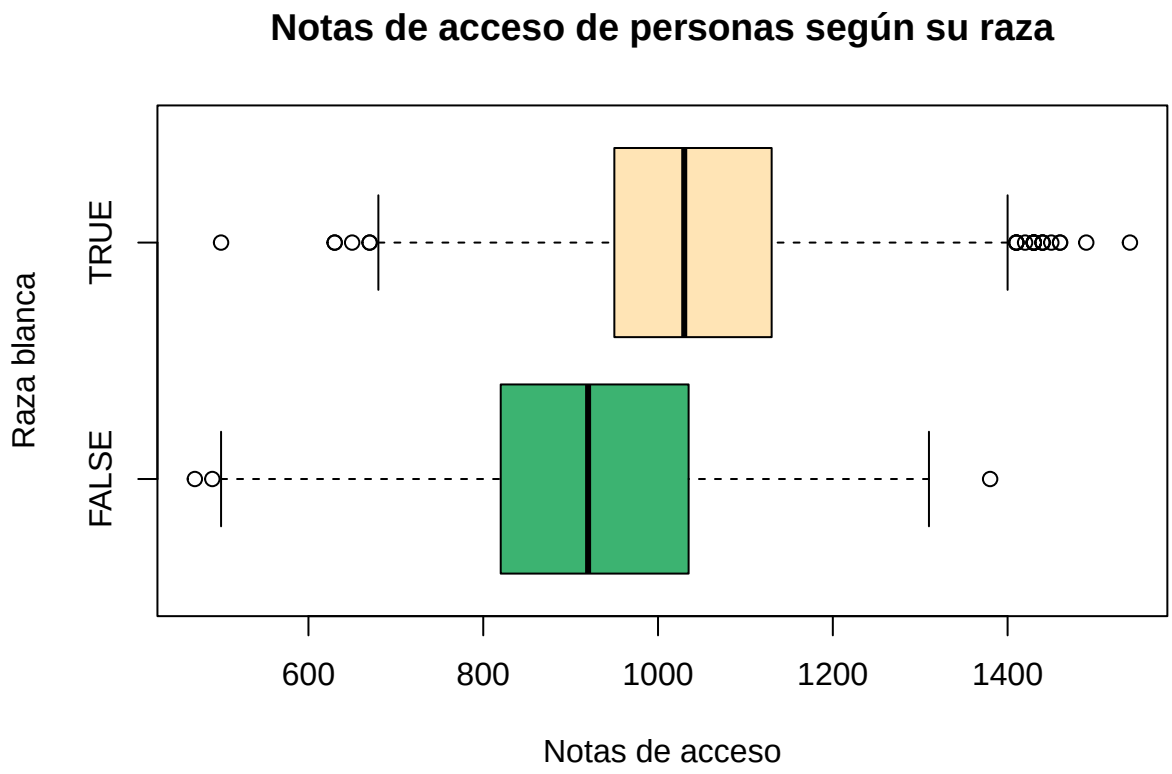
```
boxplot(gpa$sat~gpa$athlete,
  main = "Notas de acceso de atletas vs no atletas",
  xlab = "Notas de acceso",
  ylab = "Atletas",
  col=c("mediumseagreen","moccasin"),
  horizontal = T,
   )
```



Observamos que los atletas tienen una mediana en la variable de las notas de acceso menor a los estudiantes que no son atletas. También observamos que los no atletas presentan más valores atípicos en el límite superior de los datos.

- **3. Distribución de la variable "sat" con respecto a la raza white**

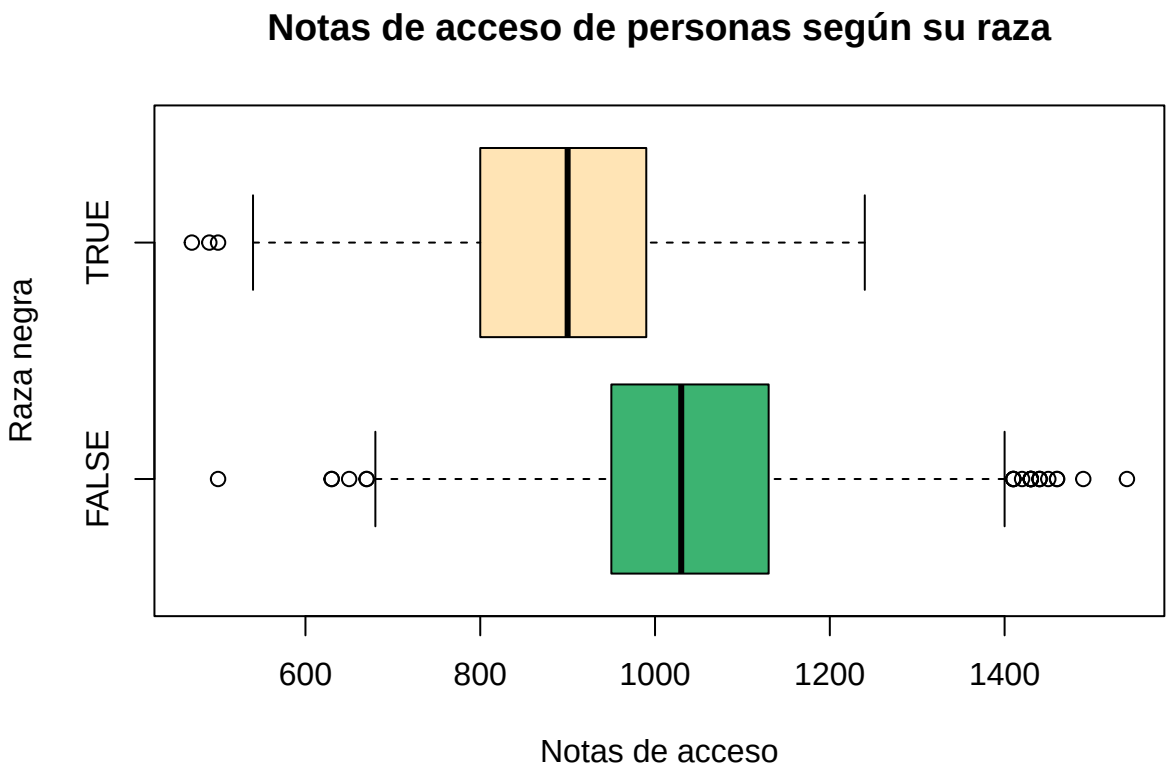
```
boxplot(gpa$sat~gpa$white,
  main = "Notas de acceso de personas según su raza",
  xlab = "Notas de acceso",
  ylab = "Raza blanca",
  col=c("mediumseagreen","moccasin"),
  horizontal = T,
   )
```



La mediana de las notas de acceso de las personas de raza blanca es superior a las personas de otras razas. También observamos que el valor mínimo (extremo inferior) de las personas de otras razas es bastante inferior al de las personas de raza blanca. De hecho, en las personas de raza blanca se consideran valores atípicos mientras que en las personas de otra raza todavía no ha alcanzado el valor mínimo.

- **4. Distribución de la variable "sat" con respecto a la raza black**

```
boxplot(gpa$sat~gpa$black,
  main = "Notas de acceso de personas según su raza",
  xlab = "Notas de acceso",
  ylab = "Raza negra",
  col=c("mediumseagreen","moccasin"),
  horizontal = T,
   )
```

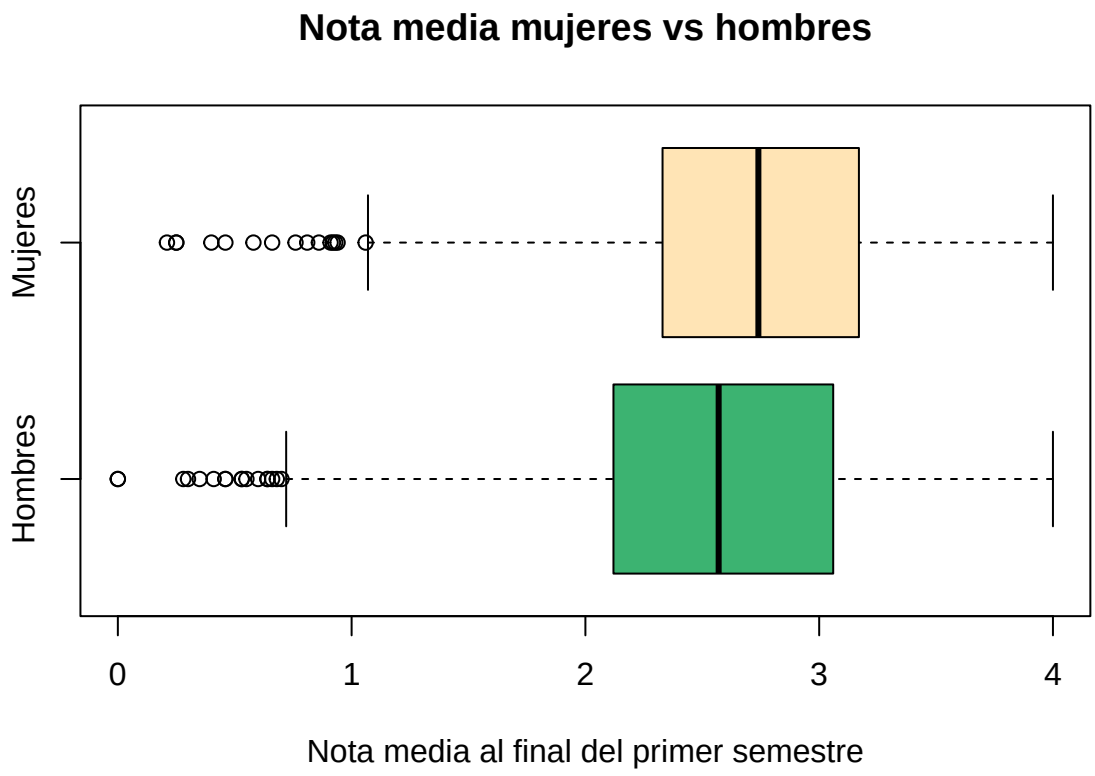


Las personas de raza negra presentan una mediana inferior a las personas de otras razas. También observamos que las personas de otras razas tienen más casos en el límite superior.

Distribución de la variable "colgpa" con respecto a la variable género, atleta y raza.

- **1.Distribución de la variable "colgpa" con respecto a la variable género (female)**

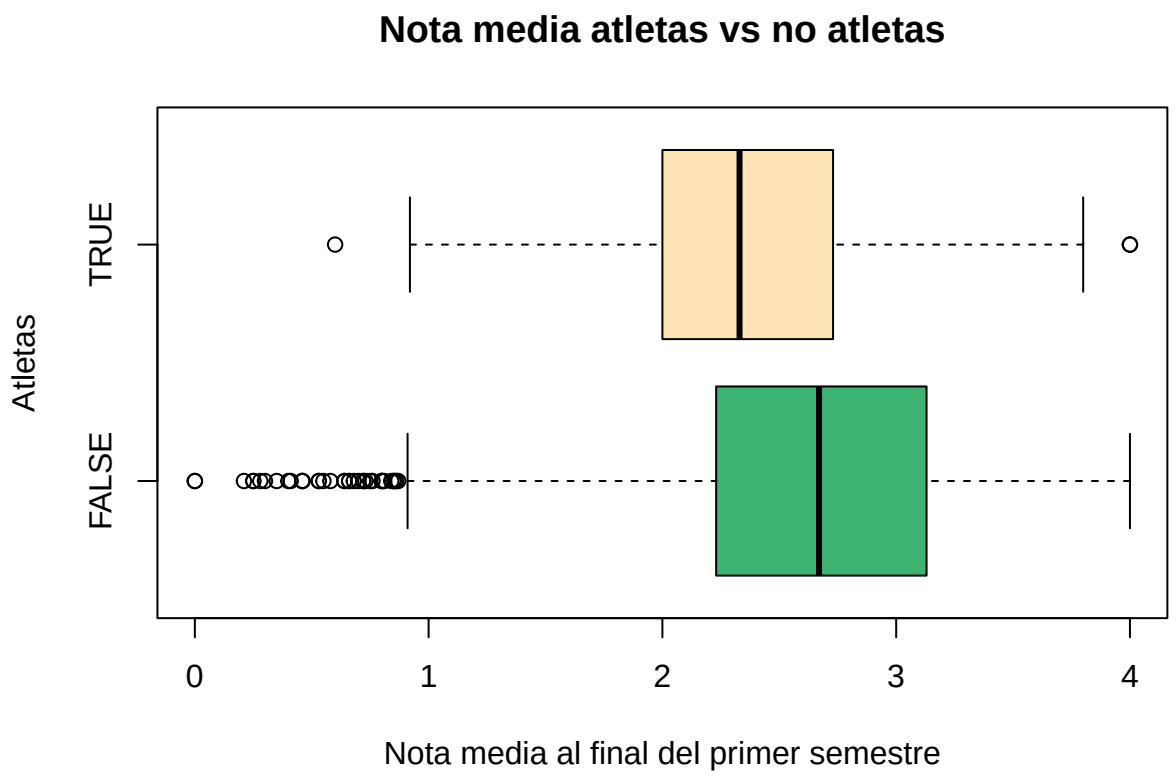
```
boxplot(gpa$colgpa~gpa$female,
  main = "Nota media mujeres vs hombres",
  xlab = "Nota media al final del primer semestre",
  ylab = NULL,
  col=c("mediumseagreen","moccasin"),
  horizontal = T,
  names = c("Hombres","Mujeres") )
```



Observamos que el extremo superior se asemeja bastante en los hombres y en las mujeres. En cambio, el extremo inferior de los hombres presenta una nota media menor a las mujeres, la mediana de éstos también es ligeramente menor. Tanto los hombres como las mujeres tienen valores atípicos en el extre inferior.

- **2. Distribución de la variable "colgpa" con respecto a la variable atleta (athlete)**

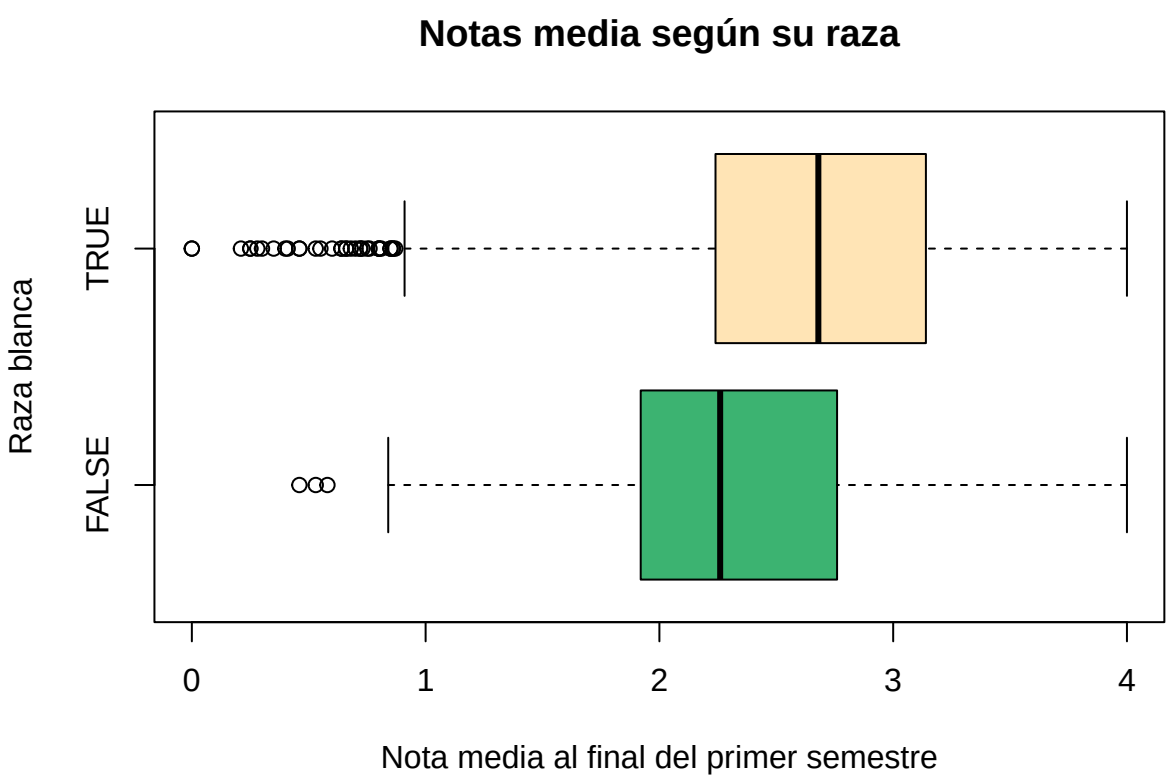
```
boxplot(gpa$colgpa~gpa$athlete,
  main = "Nota media atletas vs no atletas",
  xlab = "Nota media al final del primer semestre",
  ylab = "Atletas",
  col=c("mediumseagreen","moccasin"),
  horizontal = T,
   )
```



Los no atletas tienen una mediana superior que los atletas pero observamos que también presentan muchos más valores atípicos en el límite inferior de los datos.

- **3. Distribución de la variable "colgpa" con respecto a la raza white**

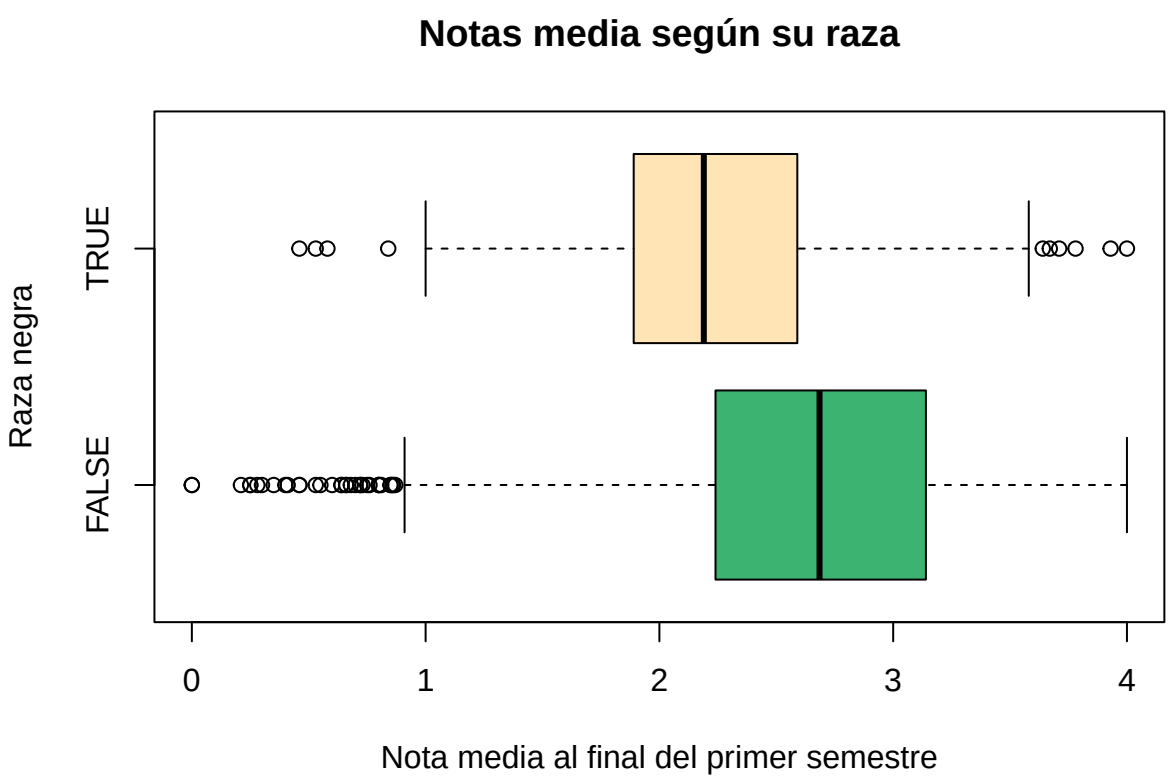
```
boxplot(gpa$colgpa~gpa$white,
  main = "Notas media según su raza",
  xlab = "Nota media al final del primer semestre",
  ylab = "Raza blanca",
  col=c("mediumseagreen","moccasin"),
  horizontal = T,
   )
```



Las personas de raza blanca tienen una mediana superior al resto pero también presentan muchos más valores atípicos en el extremo inferior.

- **4. Distribución de la variable "colgpa" con respecto a la raza black**

```
boxplot(gpa$colgpa~gpa$black,
  main = "Notas media según su raza",
  xlab = "Nota media al final del primer semestre",
  ylab = "Raza negra",
  col=c("mediumseagreen","moccasin"),
  horizontal = T,
   )
```



La mediana de las personas de raza negra es menor al resto, también encontramos sus notas menos dispersas en comparación con las de personas de otras razas que presentan bastantes valores atípicos en el límite inferior.

3. Intevalo de confianza de la media poblacional de la variable "sat" y "colgpa"

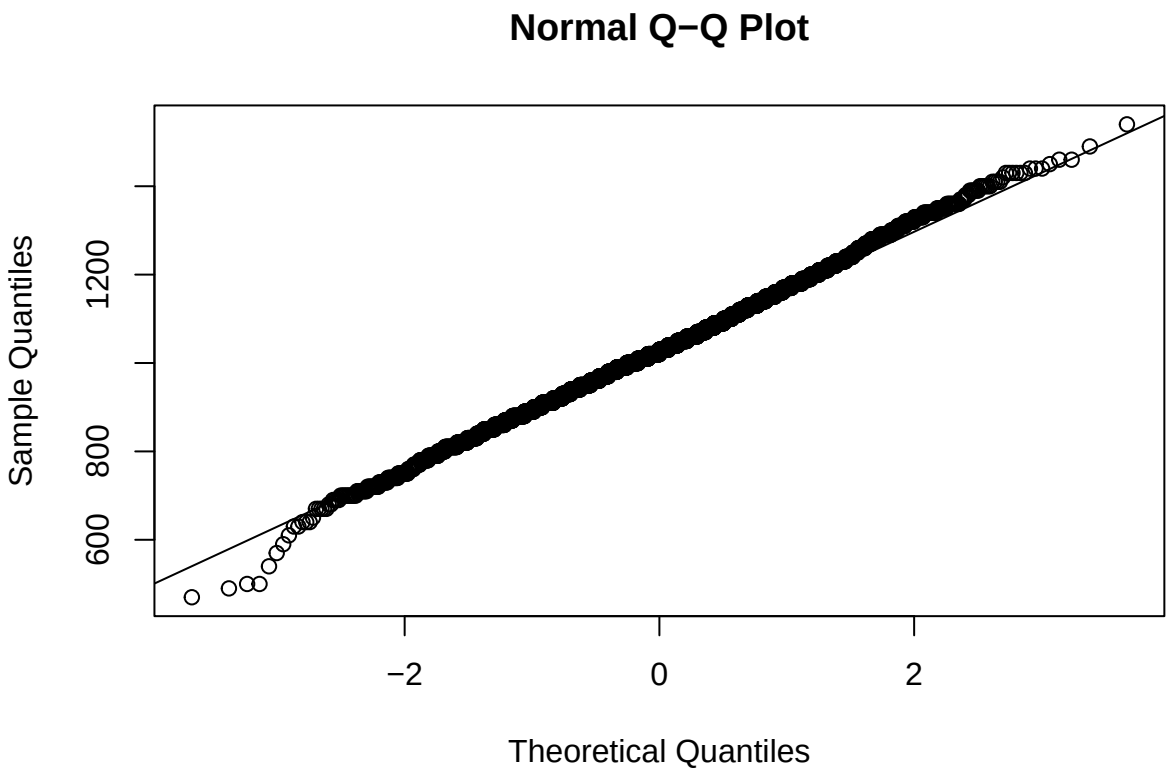
3.1 Supuestos

Comprobaremos si podemos asumir la normalidad de los datos a partir de las gráficas Q-Q de las variables "sat" y "colgpa.

En ambas gráficas mostradas a continuación podemos observar que en su mayoría se asemejan a una distribución normal por lo que la asumiremos en el estudio de éstas.

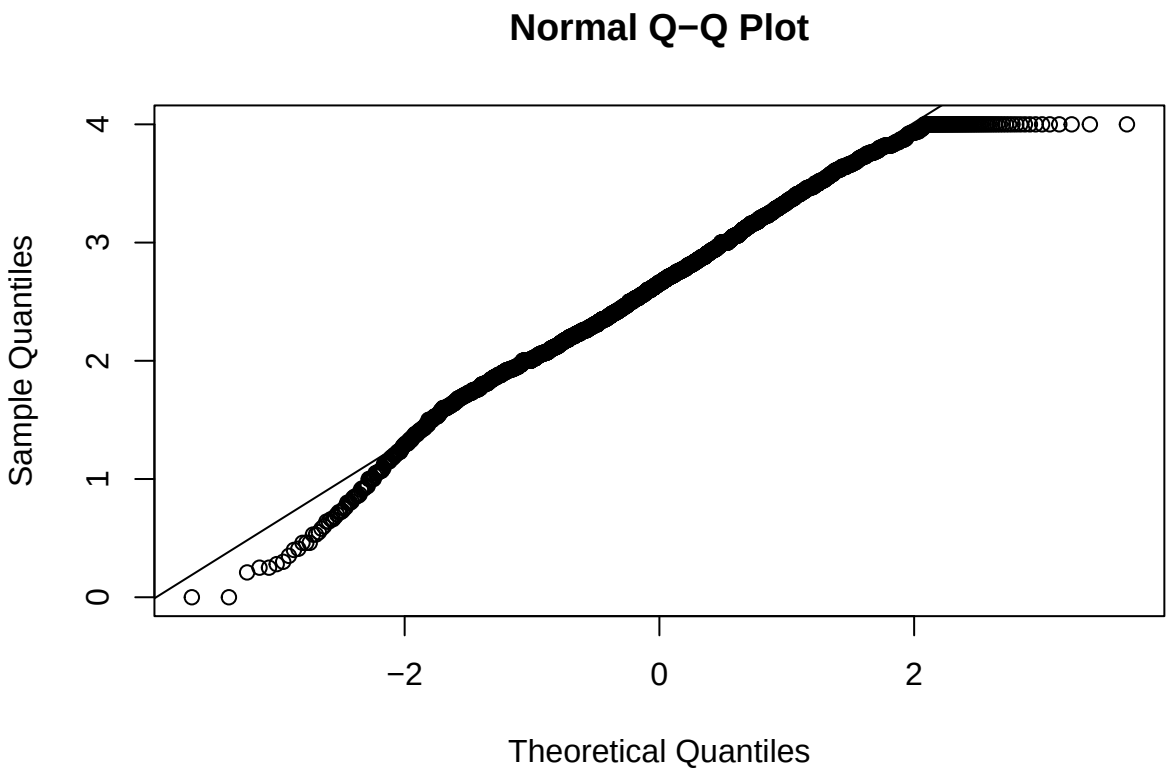
SAT

```
qqnorm(gpa$sat)
qqline(gpa$sat)
```



COLGPA

```
qqnorm(gpa$colgpa)
qqline(gpa$colgpa)
```



Con respecto a la varianza, la estiaremos a partir de la desviación de la muestra por lo que seguiremos una distribución t de Student con n-1 grados de libertad.

3.2 Función del cálculo del intervalo de confianza

A continuación escribimos la función para el cálculo del intervalo de confianza sobre la media con varianza desconocida.

```
IC <- function(x,NC){
alfa <- 1-NC
sd <- sd(x)
n <- length(gpa$colgpa)
SE <- sd/sqrt(n)
z <- qt(alfa/2,df=n-1,lower.tail=FALSE)
L <- mean(x)-z*SE
U <- mean(x)+z*SE
return(round(c(L,U),2))
}
```

3.3 Intervalo de confianza de la variable sat

A continuación se calcula el intervalo de confianza de la media poblacional de la variable sat de los estudiantes con una confianza del 90%:

```r
ICsat0.9 <- IC(gpa$sat,0.9)   #L, U
ICsat0.9
```

```
## [1] 1026.77 1033.90
```

El intervalo con una confianza del 95% es el siguiente:

```r
ICsat0.95 <- IC(gpa$sat,0.95)
ICsat0.95
```

```
## [1] 1026.08 1034.58
```

3.4 Intervalo de confianza de la variable colgpa

El intervalo de confianza al 90% de la media poblacional de la variable colgpa:

```r
ICcolgpa0.9 <- IC(gpa$colgpa,0.9)
ICcolgpa0.9
```

```
## [1] 2.64 2.67
```

El intervalo de confianza al 95% de la media poblacional de la variable colgpa:

```r
ICcolgpa0.95 <- IC(gpa$colgpa,0.95)
ICcolgpa0.95
```

```
## [1] 2.63 2.67
```

3.5 Interpretación

El intervalo de confianza al 90% de la nota media del estudiante al final del primer semestre es [2.64, 2.67].

En 90 de cada 100 alumnos el valor de su nota media al final del primer semestre está incluida entre 2.64 y 2.67. El procedimiento garantiza que el 90% de las muestras dan lugar a un intervalo que contiene el valor real del parametro. Si aumentamos el intervalo de confianza al 95% observamos que el intervalo aumenta ligeramente, desde 2.63 a 2.67.

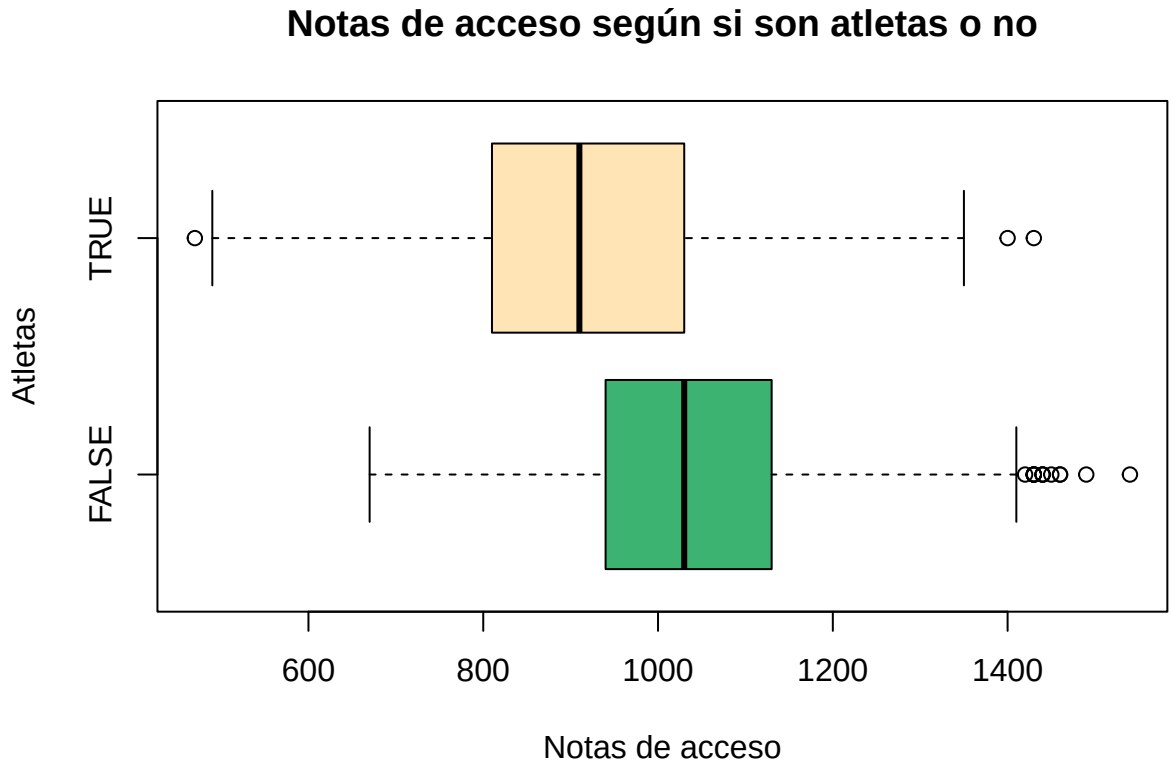
De esta misma manera se puede interpretar los resultados de la variable "sat". Donde el intervalo de confianza al 90% es [1026.77, 1033.90] y con el 95% de confianza aumenta a [1026.08, 1034.58].

4. ¿Ser atleta influye en la nota?

4.1 Análisis visual

Realizamos un gráfico de cajas para comparar las notas de acceso en función de si los estudiantes son atletas o no.

```
boxplot(gpa$sat~gpa$athlete,
    main = "Notas de acceso según si son atletas o no",
    xlab = "Notas de acceso",
    ylab = "Atletas",
    col=c("mediumseagreen","moccasin"),
    horizontal = T )
```



4.2 Función para el contraste de medias

Test sobre la media de dos poblaciones independientes con varianza desconocida igual:

```
# Test sobre la media de dos poblaciones independientes con varianza desconocida igual:
test4.2 <- function(mu1,mu2,alfa){
S <- sqrt(((length(mu1)-1)*(sd(mu1))^2+(length(mu2)-1)*(sd(mu2))^2)/(length(mu1)+length(mu2)-2))
tobs <- (mean(mu1)-mean(mu2))/(S*sqrt((1/length(mu1))+(1/length(mu2))))   #Estadístico
tcritL <- qt( alfa/2, (length(mu1))+(length(mu2))-2)
tcritU <- qt( 1-alfa/2, (length(mu1))+(length(mu2))-2)
pvalue <-pt( abs(tobs), df=(length(mu1))+(length(mu2))-2, lower.tail=FALSE)*2
return(c(tobs,tcritL,tcritU,pvalue))
}
```

4.3 Pregunta de investigación

¿Presentan los atletas notas significativamente diferentes a los no atletas?

4.4 Hipótesis nula y alternativa

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

4.5 Justificación del test a aplicar

Estamos ante dos poblaciones independientes (atletas y no atletas) donde asumimos distribución normal ya que según el teorema del límite central podemos asumir la normalidad de los datos en el caso de que todas las muestras sean superiores a 30, como es nuestro caso.

Puesto que la pregunta es si las notas son diferentes se trata de un test bilateral.

Nos quedaría realizar el test de igualdad de varianzas para determinar el test a aplicar.

```
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
```

Podemos asumir que las varianzas son iguales puesto que el valor observado cae dentro del intervalo de aceptación del test. El valor p es mayor que el valor de significancia por lo que confirmamos que las varianzas son iguales con un nivel de confianza del 95%.

Con esto confirmamos que el test a realizar es el test sobre la media de dos poblaciones independientes con varianza desconocida igual.

4.6 Cálculo

Cálculo con un nivel de confianza del 95%.

```
c4.6 <- test4.2(gpa$colgpa[gpa$athlete==TRUE],gpa$colgpa[gpa$athlete==FALSE],0.05)
c4.6
```

```
## [1] -5.910309e+00 -1.960538e+00  1.960538e+00  3.689891e-09
```

```
#tobs,tcritL,tcritU,pvalue
```

El valor observado se encuentra fuera de la zona de aceptación por lo que todo apunta a que deberíamos rechazar la hipótesis nula.

El valor p es inferior a alfa (significancia) por lo que confirmamos que podemos rechazar la hipótesis nula.

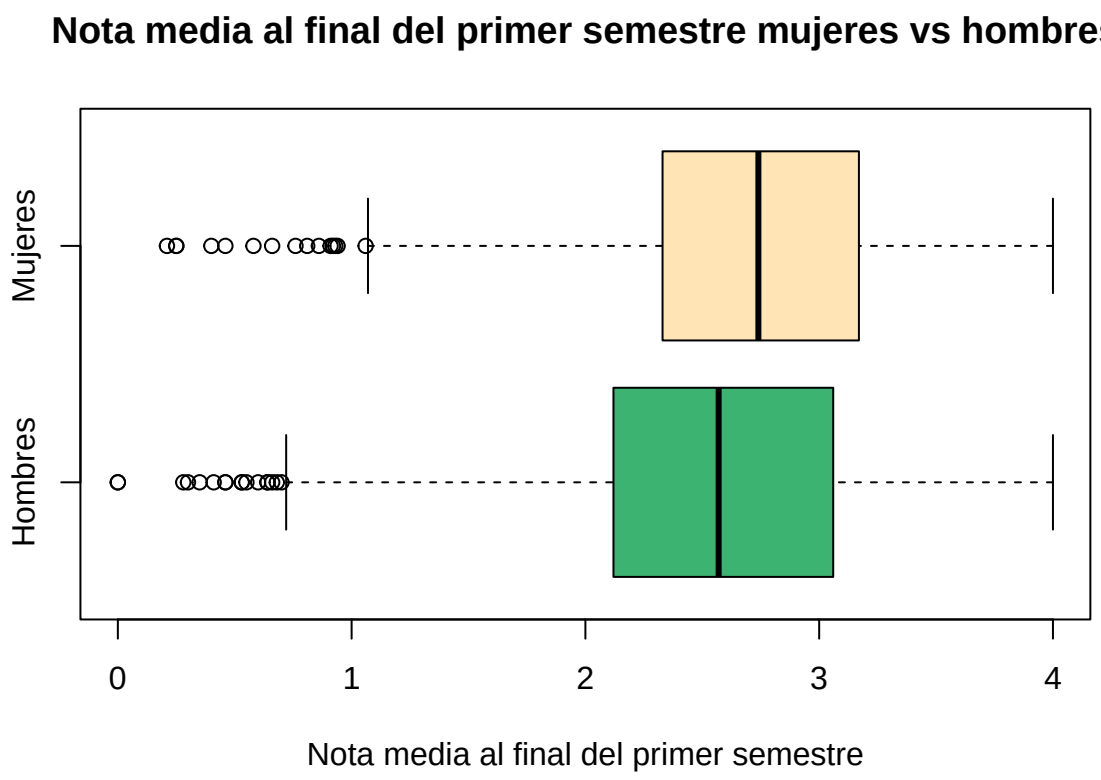
4.7 Interpretación del test

Según los resultados del test concluimos que la variable "colgpa" es decir la nota media de los estudiantes en el primer semestre es diferente para los estudiantes atletas que para los estudiantes no atletas con un nivel de confianza del 95%.

5. ¿Las mujeres tienen mejor nota que los hombres?

5.1 Análisis visual

```
boxplot(gpa$colgpa~gpa$female,
  main = "Nota media al final del primer semestre mujeres vs hombres",
  xlab = "Nota media al final del primer semestre",
  ylab = NULL,
  col=c("mediumseagreen","moccasin"),
  horizontal = T,
  names = c("Hombres","Mujeres") )
```



5.2 Función

Utilizaremos el test sobre la media de dos poblaciones independientes con varianzas desconocidas diferentes.

```r
# Test sobre la media de dos poblaciones independientes con varianzas desconocidas diferentes:
test5.2 <- function(mu1,mu2,alfa){
s1 <- sd(mu1)
s2 <- sd(mu2)
n1 <- length(mu1)
n2 <- length(mu2)
x1 <- mean(mu1)
x2 <- mean(mu2)

v <- (((s1^2/n1)+(s2^2/n2))^2)/((((s1^2/n1)^2)/(n1-1))+((((s2^2/n2)^2)/(n2-1))))
tobs <- (x1-x2)/(sqrt(((s1^2)/n1)+((s2^2)/n2)))
tcrit <- qt(1-alfa,n1+n2-2)
pvalue <- pt(abs(tobs),df=n1+n2-2,lower.tail=FALSE)
return(c(tobs,tcrit,pvalue))

}
```

5.3 Pregunta de investigación

¿La nota media de las mujeres al final del primer semestre (colgpa) es significativamente mayor que la de los hombres?

5.4 Hipótesis nula y la alternativa

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

5.5 Justificación del test a aplicar

Estamos ante dos poblaciones independientes (mujeres y hombres) donde asumimos distribución normal ya que los tamaños de las muestras son lo suficientemente grande en ambos casos.

Se trata de un test unilateral por la derecha.

Nos quedaría realizar el test de igualdad de varianzas para determinar el test a aplicar.

```r
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 valor observado cae dentro del intervalo de aceptación del test. En cambio, el valor p es menor que el valor de significancia por lo que deberíamos rechazar la hipótesis nula y confirmar que las varianzas son diferentes.

Con esto confirmamos que el test a realizar es el test sobre la media de dos poblaciones independientes con varianza desconocidas diferentes.

5.6 Cálculo

Con un nivel de confianza del 95%

```
c5.2.95 <- test5.2(gpa$colgpa[gpa$female==TRUE],gpa$colgpa[gpa$female==FALSE],0.05)
c5.2.95  # tobs,tcrit,pvalue
```

```
## [1] 7.078735e+00 1.645222e+00 8.506574e-13
```

Con un nivel de confianza del 90%

```
c5.2.90 <- test5.2(gpa$colgpa[gpa$female==TRUE],gpa$colgpa[gpa$female==FALSE],0.1)
c5.2.90  # tobs,tcrit,pvalue
```

```
## [1] 7.078735e+00 1.281756e+00 8.506574e-13
```

El valor p obtenido en ambos casos es demasiado pequeño por lo que deberíamos rechazar la hipótesis nula y aceptar la alternativa.

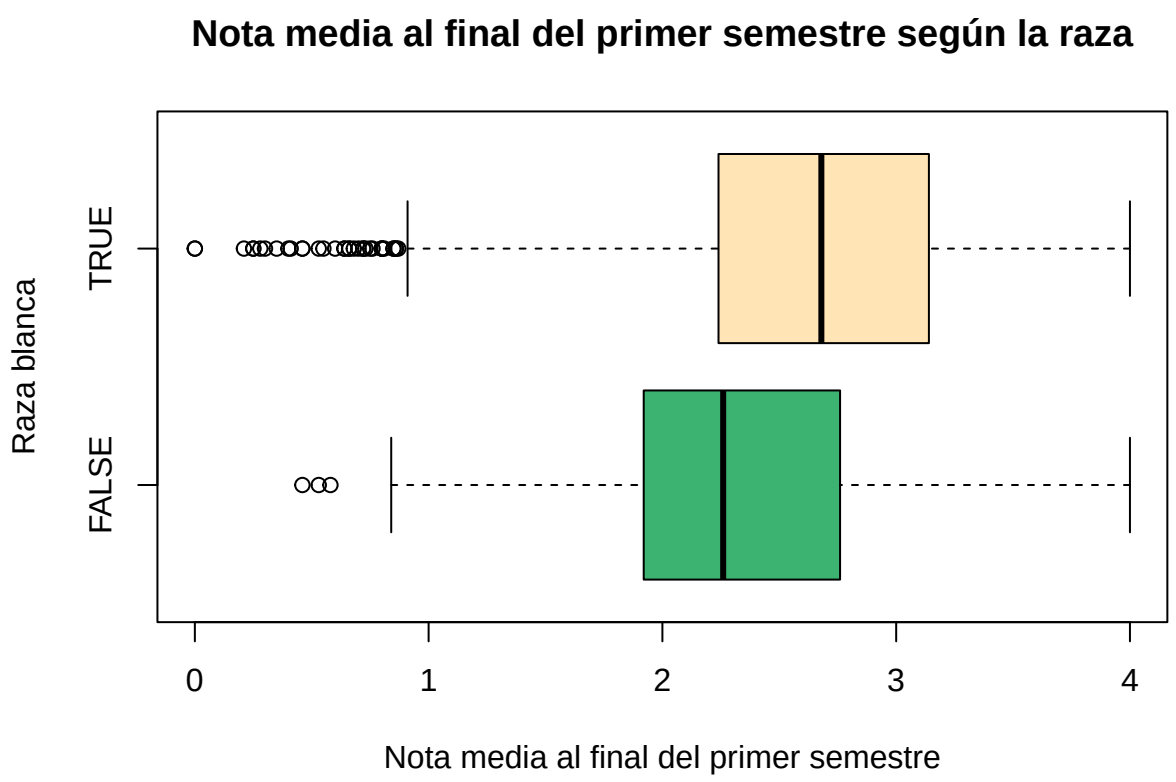
5.7 Interpretación del test

El test acepta la hipótesis alternativa, que es que la nota al final del primer semestre de las mujeres es mayor a la de los hombres.

6. ¿Hay diferencias según la raza?

6.1 Análisis visual

```
boxplot(gpa$colgpa~gpa$white,
  main = "Nota media al final del primer semestre según la raza",
  xlab = "Nota media al final del primer semestre",
  ylab = "Raza blanca",
  col=c("mediumseagreen","moccasin"),
  horizontal = T )
```



6.2 Función

Utilizaremos el test sobre la media de dos poblaciones independientes con varianzas desconocidas iguales esrita en el apartado 4.2.

6.3 Pregunta de investigación

¿La nota media de las personas de raza blanca al final del primer semestre (colgpa) es diferente a la nota media al final del primer semestre de las personas de raza negra?

6.4 Hipótesis nula y la alternativa

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

6.5 Justificación del test a aplicar

Estamos ante dos poblaciones independientes donde asumimos distribución normal ya que los tamaños de las muestras son lo suficientemente grande en ambos casos.

Se trata de un test bilateral.

Nos quedaría realizar el test de igualdad de varianzas para determinar el test a aplicar.

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

```
##
##  F test to compare two variances
##
## data:  gpa$colgpa[gpa$white == TRUE] and gpa$colgpa[gpa$white == FALSE]
## F = 0.99665, num df = 3828, denom df = 307, p-value = 0.9491
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8404046 1.1682008
## sample estimates:
## ratio of variances
##          0.9966458
```

El valor observado cae dentro del intervalo de aceptación del test y el valor p es mayor al valor de significancia por lo que podemos asumir que las varianzas son iguales.

Con esto confirmamos que el test a realizar es el test sobre la media de dos poblaciones independientes con varianzas desconocidas iguales.

6.6 Cálculo

Con un nivel de confianza del 95

```
c6.6.95 <- test4.2(gpa$colgpa[gpa$white==TRUE],gpa$colgpa[gpa$white==FALSE],0.05)
c6.6.95  #tobs,tcritL,tcritU,pvalue
```

```
## [1]  8.423306e+00 -1.960538e+00  1.960538e+00  4.987122e-17
```

Con un nivel de confianza del 90%

```
c6.6.90 <- test4.2(gpa$colgpa[gpa$white==TRUE],gpa$colgpa[gpa$white==FALSE],0.1)
c6.6.90 ## tobs,tcritL,tcritU,pvalue
```

```
## [1]  8.423306e+00 -1.645222e+00  1.645222e+00  4.987122e-17
```

El valor observado no cae en el intervalo de aceptación, además el valor p obtenido en ambos casos es demasiado pequeño por lo que deberíamos rechazar la hipótesis nula y aceptar la alternativa.

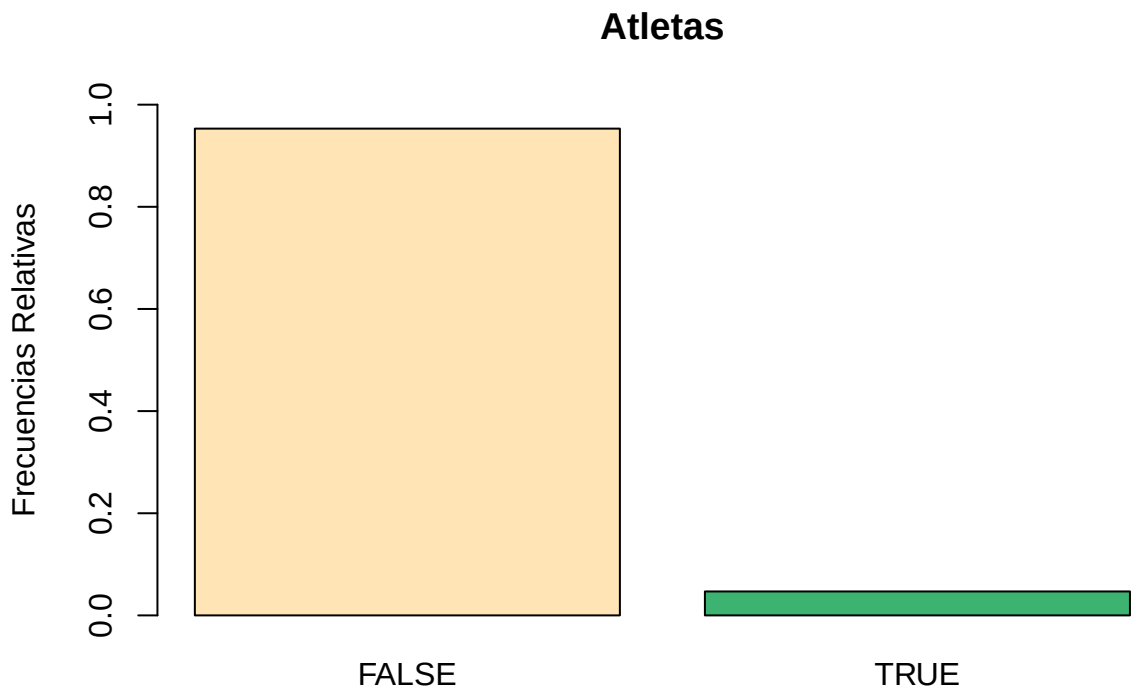
6.7 Interpretación del test

El test acepta la hipótesis alternativa, que confirma que la nota al final del primer semestre entre personas de raza blanca y raza negra son diferentes.

7. Proporción de atletas

7.1. Análisis visual

```r
barplot(prop.table(table(gpa$athlete)),
col=c("moccasin","mediumseagreen"),
ylim=c(0,1),
main="Atletas",
ylab ="Frecuencias Relativas")
```



7.2 Pregunta de investigación

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

7.3 Hipótesis nula y alternativa

$H_0 : p = 0.05$

$H_1 : p < 0.05$

7.4 Justificación del test a aplicar.

Se elige el estadístico de contraste para una distribución normal estándar. Se asume la distribución normal por encontrarnos con muestras grandes en ambos casos (atletas y no atletas).

Se trata de un test unilateral por la izquierda.

7.5 Cálculos del test

Cálculos para el contraste de hipótesis de una muestra sobre la proporción, test unilateral.

```
# Proporción observada en la muestra:
pmuestra <- 1-length(gpa$athlete[gpa$athlete==FALSE])/(length(gpa$athlete))
n <- length(gpa$athlete)
```

```
# Cálculos del test
test7.4 <- function(pmuestra,n,p0,alfa){
z<- (pmuestra-p0)/(sqrt((p0*(1-p0))/n))
q <- qnorm(alfa)
pvalue <- pnorm(z)
return(c(z,q,pvalue))
}
```

```
c7.5 <- test7.4(pmuestra,n,0.05,0.05) #z,q,pvalue
c7.5
```

```
## [1] -0.9166711 -1.6448536  0.1796575
```

Según los resultamos obtenidos averiguamos que el intervalo aceptación va desde -1.64 hasta infinito. Por lo que el valor observado cae dentro de esta aceptando la hipótesis nula. Por otro lado, comprobamos también que el valor p es mayor al valor de significancia por lo que no podemos rechazar la hipótesis nula con un valor de confianza del 95%.

7.6 Interpretación del test

Al aceptar la hipótesis nula concluimos que la proporción de atletas no 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

A continuación un gráfico con la frecuencia relativa de los atletas entre los hombres y entre las mujeres para poder compararlas.

```
par(mfrow=c(1,2))
barplot(prop.table(table(gpa$athlete[gpa$female==FALSE])),
  main = "Hombres",
  ylab="Frecuencia relativa",
  xlab = "Atletas",
  ylim = c(0, 1),
  col=c("mediumseagreen"),
  beside=TRUE)
  barplot(prop.table(table(gpa$athlete[gpa$female==TRUE])),
  main = "Mujeres",
  xlab = "Atletas",
  ylim = c(0, 1),
  col=c("moccasin"),
  beside=TRUE
)
```

8.2 Pregunta de investigación

¿La proporción de atletas es superior en los hombres que en las mujeres?

8.3 Hipótesis nula y alternativa

$H_0 : p_1 = p_2$

$H_1 : p_1 > p_2$

8.4 Justificación del test a aplicar

Estamos ante un caso de dos poblaciones independientes (hombres y mujeres) en el que queremos comparar la proporción de atletas en ambas poblaciones.

Puesto que podemos asumir la normalidad de los datos según el teorema del límite central por trabajar con grandes tamaños de muestras utilizaremos el contraste de hipótesis de dos muestras independientes sobre la proporción.

8.5 Cálculos del test

Contraste de hipótesis de dos muestras independientes sobre la proporción.

```r
alfa = 0.05

n1<- length(gpa$athlete[gpa$female==FALSE])
n2 <- length(gpa$athlete[gpa$female==TRUE])
x1 <- gpa$athlete[gpa$female==FALSE] #Hombres
x2 <- gpa$athlete[gpa$female==TRUE]  #Mujeres
p1 <- sum(x1==TRUE)/length( x1 )
p2 <- sum(x2==TRUE)/length( x2)
p<-(n1*p1 + n2*p2) / (n1+n2)
zobs <- (p1-p2)/( sqrt(p*(1-p)*(1/n1+1/n2)) )
zcrit <- qnorm(alfa, lower.tail=FALSE)
pvalue<- pnorm(zobs, lower.tail=FALSE)

c8.5 <- c(zobs, zcrit, pvalue)
c8.5
```

```
## [1] 6.241964e+00 1.644854e+00 2.160550e-10
```

Los resultados apuntan a que deberíamos rechazar la hipótesis nula ya que el valor p es más pequeño que el nivel de significancia.

8.6 Interpretación del test

La proporción de atletas en los hombres es mayor que en las mujeres con un nivel de confianza del 95%.

9 Resumen y conclusiones

N	Pregunta	Resultado	Conclusión
P3.0:	Intervalo de confianza de la media de “sat” al 90%	[1026.77, 1033.90]	El intervalo de confianza al 90% de la variable “sat” es entre los valores 1026.77 y 1033.90.
P3.1:	Intervalo de confianza de la media de “sat” al 95%	[1026.08, 1034.58]	El intervalo de confianza al 95% de la variable “sat” es entre los valores 1026.08 y 1034.58.
P3.2:	Intervalo de confianza de la media de “colgpa” al 90%	[2.64, 2.67]	El intervalo de confianza al 90% de la variable “colgpa” es entre los valores 2.64 y 2.67.
P3.3:	Intervalo de confianza de la media de “colgpa” al 95%	[2.63, 2.67]	El intervalo de confianza al 95% de la variable “colgpa” es entre los valores 2.63 y 2.67.
P4:	¿Presentan los atletas notas significativamente diferentes a los no atletas?	tobs=-5.910309, tcritL=-1.960538, tcritU=1.960538, pvalue=3.689891e-09	La variable “colgpa” es diferente para los estudiantes atletas que para los estudiantes no atletas con un nivel de confianza del 95%.

N	Pregunta	Resultado	Conclusión
P5:	¿La nota media de las mujeres al final del primer semestre (colgpa) es significativamente mayor que la de los hombres?	tobs=7.078735, tcrit=1.645222, pvalue=8.506574e-13	La nota al final del primer semestre de las mujeres es mayor a la nota de los hombres con un nivel de confianza del 95%.
P6:	¿La nota media de las personas de raza blanca al final del primer semestre es diferente a la de las personas de raza negra?	tobs=8.423306, tcritL=-1.960538, tcritU=1.960538 pvalue=4.987122e-17	La nota al final del primer semestre entre personas de raza blanca y raza negra son diferentes con un nivel de confianza del 95%.
P7:	¿La proporción de los atletas en la población es inferior al 5%?	z=-0.9166711, q=-1.6448536, pvalue=0.1796575	La proporción de atletas no es inferior al 5% con un nivel de confianza del 95%.
P8:	¿La proporción de atletas es superior en los hombres que en las mujeres?	zobs=6.241964, zcrit=1.644854, pvalue=2.160550e-10	La proporción de atletas en los hombres es mayor que en las mujeres con un nivel de confianza del 95%.

10. Resumen ejecutivo

Los estudios referentes a la nota de acceso nos indican que con una alta probabilidad, del 95%, la media poblacional se encuentra entre los valores siguientes [1026.08, 1034.58]. Con respecto a la nota media del estudiante al final del primer semestre la media poblacional se encuentra, también con un 95% de probabilidad, entre 2.63 y 2.67.

La nota media de los estudiantes atletas y los no atletas al final del primer semestre apunta a que será diferente.En referencia a los atletas todo apunta a que la proporción de éstos no es inferior al 5%. También hemos observado que la proporción de atletas en los hombres parece ser mayor que en las mujeres.

Por otro lado la nota media al final del primer semestre parece ser que es mayor en el caso de las mujeres que en el caso de los hombres. Si tenemos en cuenta la raza esta nota indica que será diferente para personas de diferentes razas.