PAC2 Anàlisi Estadística

Àlex Franco Granell

Novembre 2022

1. Lectura del fitxer

Carregue l'arxiu gpa_clean com es demana a l'enunciat i comprove que totes les variables tenen el tipus que els pertoca.

```
# Carregue les dades
gpa <- read.csv('./gpa_clean.csv', sep=',')

# Mostre el tipus
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: chr  "C" "A" "C" "C" ...
```

A través d'una inspecció visual del resultat, i d'una comparació amb la descripció oferida a l'enunciat, es pot dir que els tipus de les dades són correctes.

2. Estadística descriptiva i visualització

2.1 Anàlisi descriptiva

A continuació faré un anàlisi descriptiu de les dades a través dels seus valors numèrics. Primer els extrauré:

```
# Valors generals del dataframe:
cat("Hi ha ", nrow(gpa)," observacions i ", length(names(gpa))," atributs\n\n")
```

```
## Hi ha  4137  observacions i  11  atributs
```

```
# Paràmetres numèrics d'anàlisi:
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   Length:4137
##  FALSE:308       FALSE:3908      Class :character
##  TRUE :3829      TRUE :229       Mode  :character
##
##
##
```

```
# Contar lletres:
cat("Hi ha ", sum(gpa$gpaletter=="A")," valors d'A, \n
Hi ha ", sum(gpa$gpaletter=="B")," valors de B, \n
Hi ha ", sum(gpa$gpaletter=="C")," valors de C, \n
Hi ha ", sum(gpa$gpaletter=="D")," valors de D. \n")
```

```
## Hi ha  458  valors d'A,
##
## Hi ha  1999  valors de B,
##
## Hi ha  1536  valors de C,
##
## Hi ha  144  valors de D.
```

Com es pot observar en aquestes dades, s'hi poden extraure unes poques conclusions:

-sat: Podem veure que les dades tendeixen a agrupar-se entre el primer quartil i el tercer. Els mínims i màxims tenen una distància al quartil semblant, de 400 punts aproximadament.

-tothrs: En aquest podem veure que les dades tendeixen a agrupar-se entre la mediana i el tercer quartil. El valor del màxim pot haver desviat la mitjana, però sembla que hi ha tendènia cap a valors elevats.

-hsize: Al igual que l'anterior sembla que les dades s'agrupen cap als valors elevats del tercer quartil. Així i tot, hi ha un valor màxim que podria qualificar-se d'outliner perquè és molt extrem.

-hsrank: Per a aquest s'hi veu clarament una gran influència en la mitjana per part del màxim, que s'hauria de qualificar segur com un outliner dintre de la distribució. Aquesta, pel valor de la mediana, sembla més pròxima al primer quartil que al tercer.

-hsperc: En aquesta distribució s'hi veu també una gran influència del màxim en la mitjana, tot i que la majoria de les dades (50%) es troba entre 6.4 i 27.7. El mínim no té tant de pes com el màxim i la distribució tendeix al tercer quartil.

-colgpa: Per a aquesta distribució s'hi veu que la majoria de dades es troba pròxima al 2.5, tot i que pel que s'hi veu al valor del tercer quartil, la distribució tendeix a ser major. Q1 = 2, Q3 = 3, per tant tendeix cap a 4.

-athlete: Hi ha considerablement més gent que no és athleta que que sí ho és. Els athletes són una minoria marginal a la mostra.

-female: Hi ha més homens que dones, però semblen valors pròxims. Distància de 400 registres.

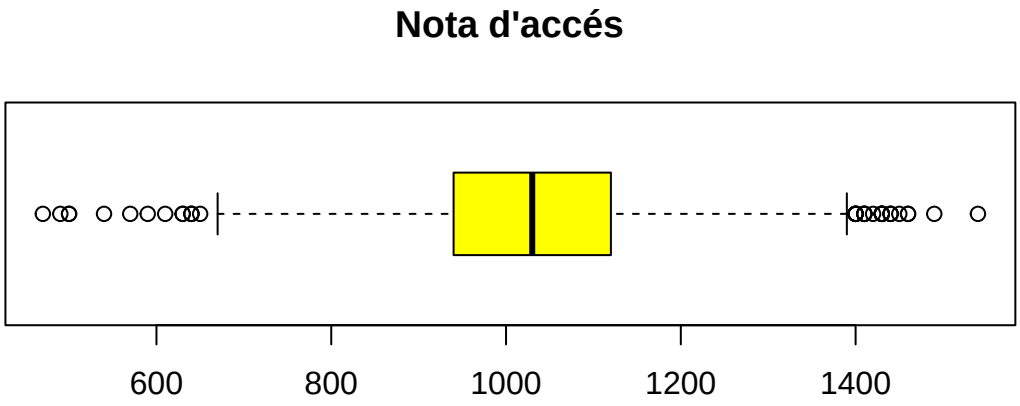
-white: Hi ha una clara majoria de persones blanques en aquesta mostra. La gent no blanca és una minoria quasi marginal.

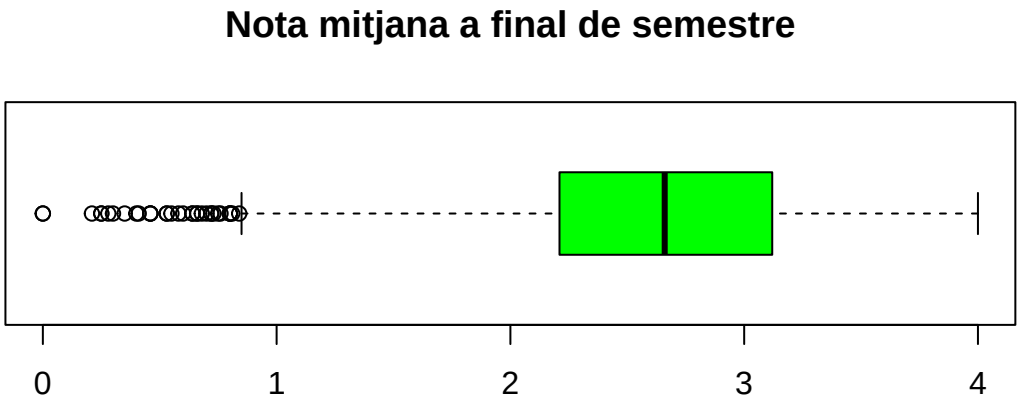
-black: La gent negra és clarament una minoria. Aquests valors, menys els de la gent que no és blanca, ens demostren que hi ha gent que no és blanca, però que no és negra. 308 - 229 = 79 persones no negres ni blanques. Una minoria absoluta.

-gpaletter: De les lletres s'hi pot veure que la majoria es troben entre la C i la B, amb tendència cap a la A. Ho podem saber al recompte de les variables. Sabem, a més, que aquestes dades corresponen a una discretització de colgpa. I efectivament concorden.

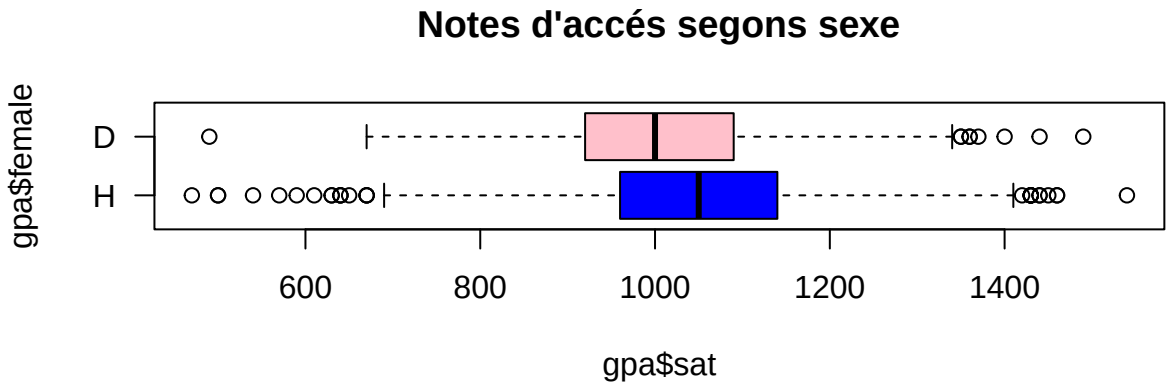
2.2 Visualització

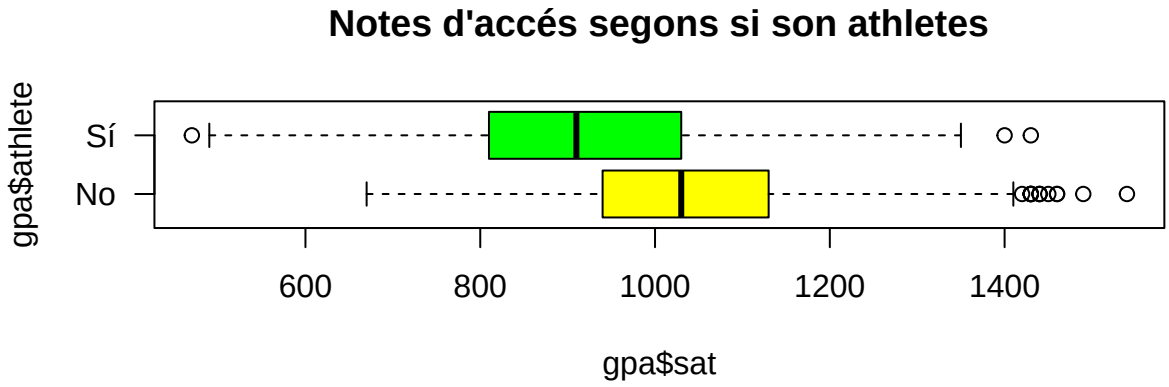
A continuació anem a fer una breu inspecció visual de les dades i extraure unes conclusions preliminars. Primerament anem a mostrar les distribucions de la nota d'accés i la nota mitjana a final del semestre, i després faré un comentari general.

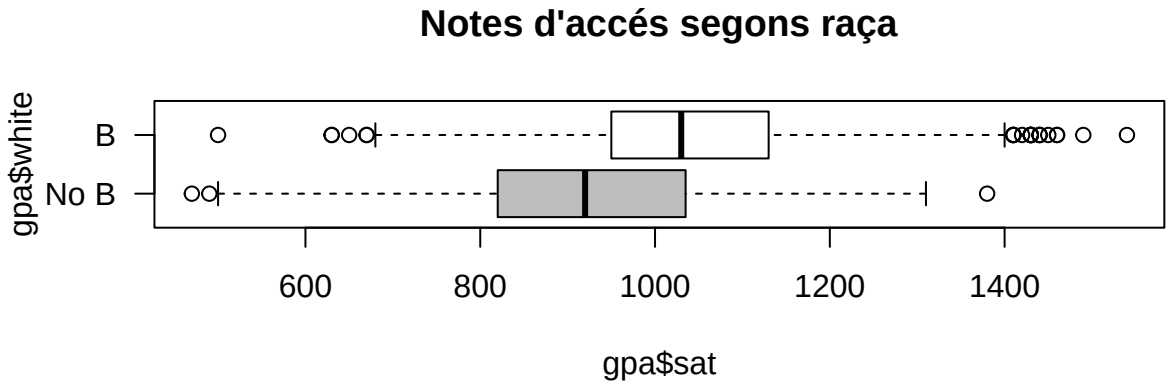




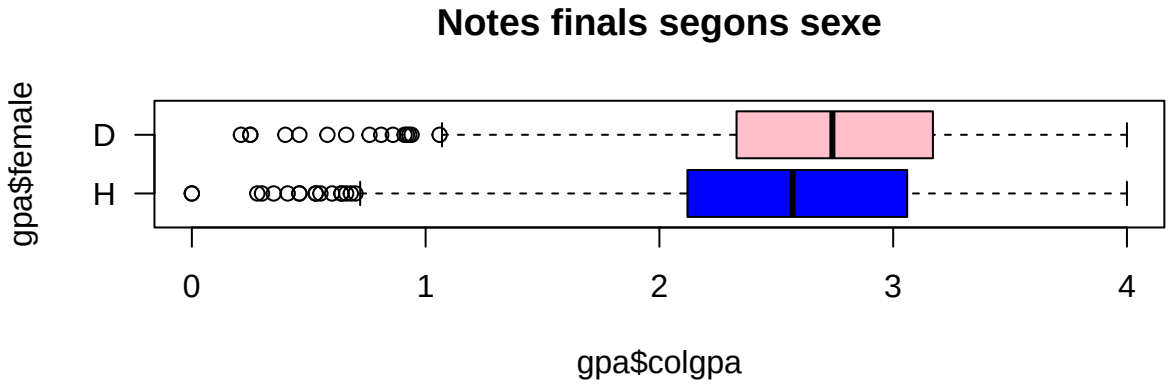
A continuació vaig a representar la variable sat, és a dir la nota d'accés, segons el gènere, si és athleta i la raça:

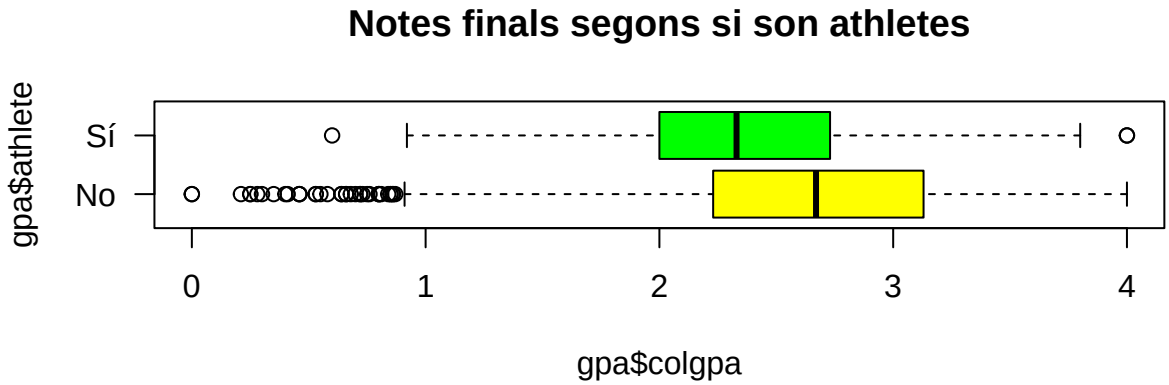


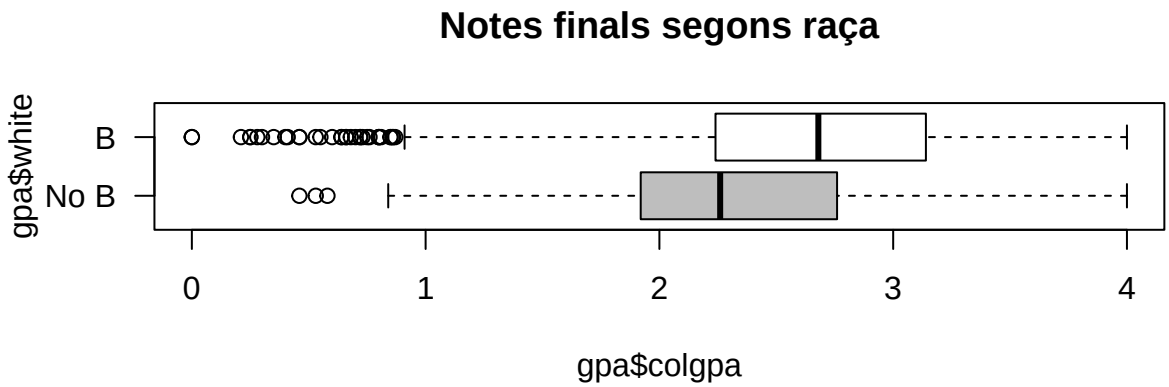




Ara vaig a fer el mateix, però per la variable colgpa, és a dir la nota del final de semestre:







Conclusions Notes d'accés: sat A través d'una inspecció visual s'hi veu que la distribució general de les notes d'accés tendeix cap al màxim. La majoria de notes es troben entre el 950 i el 1100 aprox.

Si ens fixem en les distribucions segons els sexes, veiem que els homens solen tenir notes més elevades en aquesta distribució. S'hi veu clarament en les diferències interquartíliques del Q3 i Q1 (o IQR). Fins i tot el màxim de les dones és menor, tot i que el mínim és equivalent.

Pel que fa a si són athletes, s'hi veu que els que sí ho són accedeixen amb una nota d'accés menor que la resta. Considerablement menor, de quasi 100 punts de desplaçament per al IQR. Cal destacar que el mínim dels athletes és molt inferior al mínim de la resta d'alumnes.

Finalment pel que fa als alumnes no blancs, s'hi veu que aquests solen tenir menys notes que els que són blancs. Amb un desplaçament del IQR també de quasi 100 punts. Destaca el mínim de les persones no blanques, i que el seu IQR concentra les notes quasi per baix del 1050 aprox.

Pel que fa als no blancs i als athletes, s'hi observa que tenen distribucions força semblants. S'ha de tenir en compte també que estem parlant de minories a la nostra mostra, i tal vegada no estan ben representades.

Conclusions Notes finals: colgpa

A través d'una inspecció visual s'hi veu que la distribució general de les notes de final ed semestre pot considerar-se positiva. El 50% dels alumnes (IQR) està aprovat. S'hi veu també que els casos d'abandonament de l'assignatura (notes pròximes al 0) són escasos i s'han delimitat com outliners.

Pel que fa a les distribucions per sexe, s'hi veu que tot i que les notes d'accés de les dones era menor, aquestes ténen notes majors al final de semestre. El IQR concentra les seues notes al voltant del 2.8 aprox, mentre que per als homens s'amplia el rang interquartílic per aproximar-lo al 2. Hi ha més homens que dones que passen justos el semestre. S'hi veu també en la diferència de mínims, les dones ténen un mínim més elevat.

Pel que respecta a si són athletes, s'hi veu que aquests tenen un IQR més pròxim al 2, amb un desplaçament cap al suspens considerable respecte a la resta d'alumnes. El IQR de la resta es situa quasi un 0.5 més elevat que el dels que són atheletes. Els mínims són equivalents, i de màxims sembla no haver suficients notes per als athletes. En general els athletes tenen notes menors

Finalment pel que fa a les persones no blanques, s'hi veu que el seu IQR està força desplaçat cap al suspens, amb el Q1 essent menor que l'aprovat. La mediana de les persones no blanques també ens indica aquesta tendència cap al suspens. Les persones blanques es diferèncien únicament amb les que no ho son per un major nombre de dades (fixar-se als outliners del mínim d'ambdues) i per un desplaçament de les notes cap al 3, amb un $Q3 > 3$ i una mediana més centrada.

3 Interval de confiança de la mitjana poblacional de la variable sat i colgpa

3.1 Supòsits

Primerament, per al càlcul dels intervals de confiança hem de suposar que la distribució poblacional de les notes d'accés, sat, i les notes de final de semestre, colgpa, segueixen una distribució normal.

Com volem calcular la mitjana poblacional, però no tenim la variança poblacional, concretament considerarem que aquestes distribucions seguiran una distribució de T. student amb n-1 graus de llibertat. Com tenim una mostra considerable, suposem que la distribució s'aproximarà prou a la distribució normal desitjada.

3.2 Funció de càlcul de l'interval de confiança

Com he comentat, per a aquest càlcul s'analitzarà la mitjana a través d'una T. Student. La funció a aplicar serà aquesta:

```
# x = variable a calcular
# NC = nivell de confiança
IC <- function( x, NC) {
  alfa <- 1-NC
  sd <- sd(x)
  n <- length(x)
  SE <- sd / sqrt(n)
  z <- qt(alfa/2, df=n-1, lower.tail = FALSE)
  L <- mean(x) - z*SE
  H <- mean(x) + z*SE
  round(c(L,H), 2)
}
```

3.3 Interval de confiança de la variable sat

Ara procediré a calcular l'interval de confiança de la variable sat. Per fer això faré servir la funció creada anteriorment. El nivell de confiança escollit serà primer del 90% i després del 95%:

```
# 90%
IC(gpa$sat,0.90)
```

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

```
# 95%
IC(gpa$sat,0.95)
```

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

3.4 Interval de confiança de la variable colgpa

Ara faré el mateix per calcular l'interval de confiança de la variable colgpa per a una confiança del 90% i del 95%:

```r
# 90%
IC(gpa$colgpa,0.90)
```

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

```r
# 95%
IC(gpa$colgpa,0.95)
```

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

3.5 Interpretació

A través de la funció generada, s'hi pot veure que la mitjana poblacional de la nota d'accés estaria en l'interval de 1026.77 fins 1033.90 per a una confiança del 90%. En canvi, per al segon test amb una confiança del 95% estaria entre 1026.08 i 1034.58. Com es pot veure, és segur que el mínim de l'interval ha d'estar sobre 1026, i el màxim és el que s'amplia per garantir més precisió. Així i tot, el màxim de l'interval estaria com a molt en els 1034. Açò ens ve a dir que la nota d'accés mitjana per a la població d'alumnes estaria sobre el 1030 aproximadament.

Recorde que el que volen dir els intervals de confiança és que és segur al 90/95% que la mitjana es troba en aquest interval. Per tant podem confiar prou en que la mitjana de la població per la nota d'accés es troba al voltant 1030.

Pel que fa a la nota al final de semestre, veiem que l'interval per a una confiança del 90% és de 2.64 i 2.67. Per a la confiança del 95% estaria entre 2.63 i 2.67. La diferència entre les dos és tan menuda que quasi podem assegurar que la mitjana poblacional és 2.6, que pel que fa a una nota mitjana d'un alumne ja seria suficient precisió.

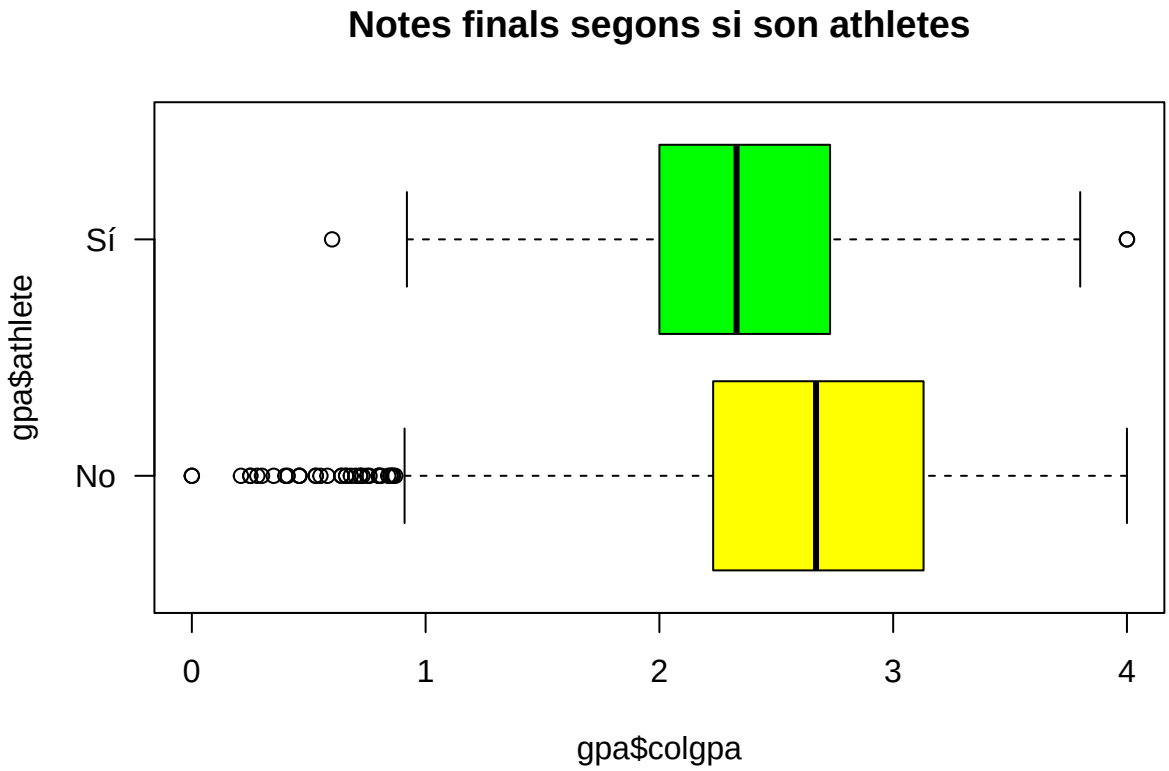
4 Ser atleta influex a la nota?

En aquest apartat ens preguntem si ser atleta influeix a la nota, amb una confiança del 95%. Les variables que utilitzaré seran colgpa per a les notes, i athlete. Per fer això crearé una funció de contrast entre mitjanes que retornarà l'estadístic de contrast, el valor crític i el valor p.

4.1 Anàlisi visual

Per saber com orientar la pregunta, observem una vegada més les distribucions que tenim a la mostra:

```
boxplot(gpa$colgpa ~ gpa$athlete,
        main = "Notes finals segons si son athletes",
        names = c("No", "Sí"),
        horizontal = TRUE,
        col = c("Yellow", "green"),
        las = 1)
```



Aparentment pel que s'observa a les distribucions sabriem que els atletes tenen una mitja inferior a la resta. Però com tenim poques dades d'athletes caldria comprovar-ho sobre la població.

4.2 Funció per al contrast de mitjanes

Cree la funció per al contrast de mitjanes sense la desviació poblacional desconeguda.

```r
test.mitjanes.nodp <- function(x, mu, alfa){
  mean <- mean(x)
  sd <- sd(x)
  n <- length(x)
  # Estadístic de contrast
  tobs <- (mean-mu)/(sd/sqrt(n))
  # Regió d'acceptació
  tcrit.L <- qt(alfa/2, df=n-1)
  tcrit.H <- qt(1-alfa/2, df=n-1)
  # Valor p
  pvalue <- pt(abs(tobs), lower.tail=FALSE, df=n-1)*2
  return (c(tcrit.L, tcrit.H, tobs, pvalue))
}
```

4.3 Pregunta de recerca

En aquest problema volem saber si la nota final dels atletes és diferent a la de la resta de persones. Concretament, a través de l'anàlisi visual podem intuïr que la mitjana seria menor.

4.4 Hipòtesi nul · la i l'alternativa

A través del que s'ha explicat abans, definisc les hipotesis com:

H0: La mitjana poblacional dels atletes és igual a la mitjana de la resta

H1: La mitjana poblacional dels atletes és diferent a la mitjana de la resta

o

H0: Mathlete = 2.6

H1: Mathlete != 2.6

4.5 Justificació del test a aplicar

Com comentava a l'exercici passat, asumim que les notes finals es comporten com una distribució normal, i per al cas de la desviació poblacional desconeguda es farà amb una distribució T. Student amb n-1 graus de llibertat. En el cas actual, volem comprovar que la mitjana dels atletes és igual o diferent a la de la resta. Faré un contràst d'hipòtesi sobre la mitjana per comprovar sí és certa aquesta assumpció. Com hem descobert a l'exercici anterior, la mitjana poblacional del total hauria d'estar al voltant de 2.6.

4.6 Càlcul

A continuació aplique la funció definida anteriorment per a la mitjana de 2.6 i una confiança del 95%:

```r
# Valors dels atletes:
athletes <- gpa[gpa$athlete == TRUE,]

# Test:
test.mitjanes.nodp(athletes$colgpa,2.6,0.05)
```

```
## [1] -1.972332e+00  1.972332e+00 -5.073683e+00  9.134190e-07
```

```
remove(athletes)
```

4.7 Interpretació del test

Com s'ha comprovat a través del calcul d'hipòtesis per a mostres amb desviacions poblacionals desconegudes, es pot concloure que cal rebutjar la hipòtesi nul · la. Això és per dos motius:

Primerament cal destacar que les regions d'acceptació se'ns han desviat un poc de la distribució normal. Això és perquè la distribució T. Student és més aplanada i ampla. Així mateix, l'interval s'ens ha quedat quasi com corresponia (-1.97, + 1.97). A partir d'aquest, s'hi veu que l'estadístic de contrast queda fora de la regió d'acceptació (EC = -5.074), per tant per aquest apartat podem descartar la hipòtesi nul · la.

Però és que també veiem que el valor P és molt més menut que la significació demanada. És a dir, VP = 0.0000009 és menor a 0.05. Per tant, per aquest apartat també es podria rebutjar l'hipòtesi nul · la.

Per tant es pot concloure que la mitjana poblacional dels atletes és diferent a 2.6 per la nota del final del semestre i que ser atleta influeix a la nota final. Caldria contrastar com, tot i que suposem, per la distribució de la mostra, que serà a menys nota que la població general.

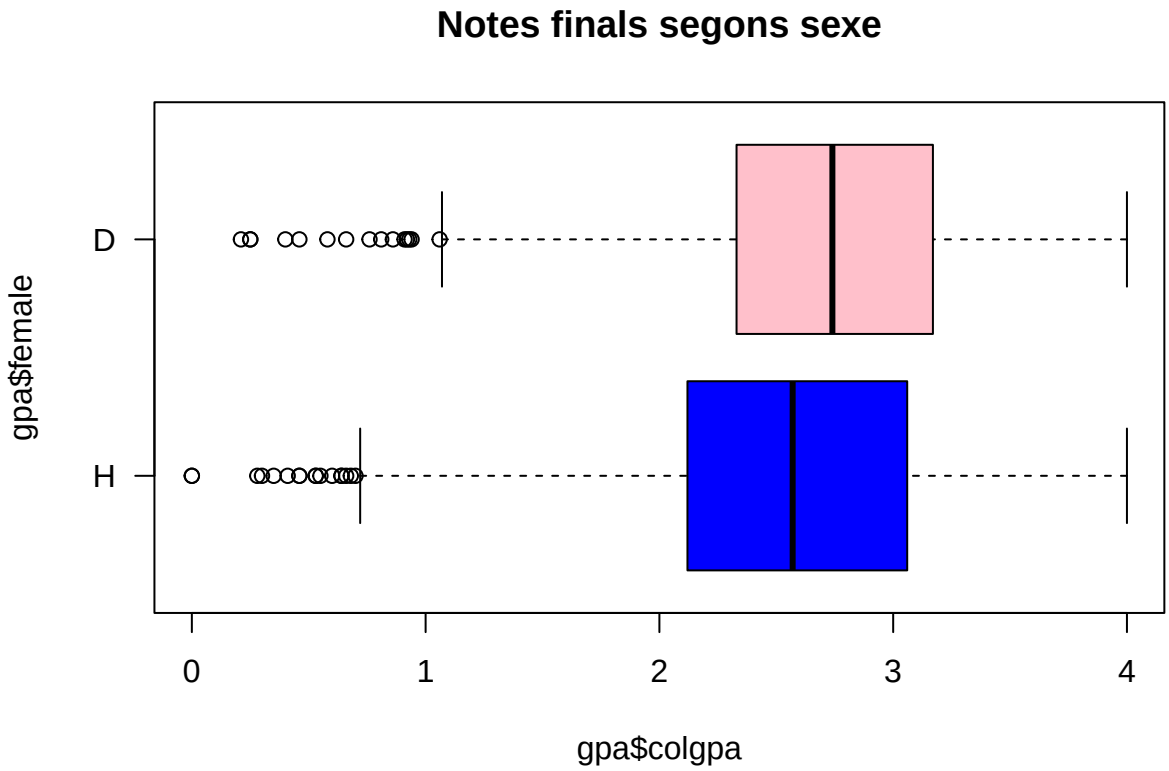
5 Les dones tenen millor nota que els homens?

Per a aquest exercici volem comprovar si les dones tenen més nota que el homens, específicament. Ho volem per una confiança del 90% i del 95%.

5.1 Anàlisi visual

Aquesta pregunta vindria donada per l'anàlisi visual fet anteriorment, on s'hi veia que les dones tenien un IQR més desplaçat cap al 4 que els hòmens:

```
boxplot(gpa$colgpa ~ gpa$female,
        main = "Notes finals segons sexe",
        names = c("H", "D"),
        horizontal = TRUE,
        col = c("blue", "pink"),
        las = 1)
```



5.2 Funció

Com no s'explicita el contrari, la funció que s'utilitzarà per al càlcul del test de dues mostres independents sobre la mitjana amb variàncies desconegudes serà la donada per t.test, d'R (ja que segueix una distribució de T. Student). Per tant no cal definir aqui cap funció.

5.3 Pregunta de recerca

En aquest problema ens interessa saber si la mitjana de la població de dones, pel que fa a la nota final, és major a la mitjana dels hòmens.

Per saber-ho també hauré de calcular si les variàncies són iguals o no.

5.4 Hipòtesi nul · la i l'alternativa

A través del que s'ha explicat abans, definisc les hipotesis com:

H0: La mitjana poblacional de les dones és igual a la mitjana dels hòmens

H1: La mitjana poblacional de les dones és major a la mitjana dels hòmens

o

H0: MfemaleTRUE = MfemaleFalse

H1: MfemaleTRUE > MfemaleFalse

5.5 Justificació del test a aplicar

El test que aplicaré per extraure una conclusió serà el test de dues mostres independents sobre la mitjana amb variàncies desconegudes. Açò és perquè volem comparar dues mostres, que podem suposar que seran independents. A més, tenim una mostra suficientment gran n > 30.

5.6 Càlcul

```
# Comprovació de les variàncies
var.test(gpa$colgpa[gpa$female==TRUE],
         gpa$colgpa[gpa$female==FALSE],
         conf.level = 0.90)
```

```
##
##  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
## 90 percent confidence interval:
##  0.7696299 0.8901446
## sample estimates:
## ratio of variances
##          0.8275687
```

```
var.test(gpa$colgpa[gpa$female==TRUE],
         gpa$colgpa[gpa$female==FALSE],
         conf.level = 0.95)
```

```
##
##  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
```

```
# Les variàncies no són iguals
t.test(gpa$colgpa[gpa$female==TRUE],
       gpa$colgpa[gpa$female==FALSE],
       alternative="greater",
       var.equal=FALSE,
       conf.level = 0.90)
```

```
##
##  Welch Two Sample t-test
##
## data:  gpa$colgpa[gpa$female == TRUE] and gpa$colgpa[gpa$female == FALSE]
## t = 7.0787, df = 4087.4, p-value = 8.522e-13
## alternative hypothesis: true difference in means is greater than 0
## 90 percent confidence interval:
##  0.1173717       Inf
## sample estimates:
## mean of x mean of y
##  2.733016  2.589693
```

```
t.test(gpa$colgpa[gpa$female==TRUE],
       gpa$colgpa[gpa$female==FALSE],
       alternative="greater",
       var.equal=FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  gpa$colgpa[gpa$female == TRUE] and gpa$colgpa[gpa$female == FALSE]
## 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 Interpretació del test

Primerament he fet el test de les variàncies per veure si són equivalents entre homens i dones. El resultat per al 95% i per al 90% m'ha donat els mateixos resultats: el valor observat cau en la zona d'acceptació de la hipòtesi nul · la. Així mateix, el p valor era més menut que el nivell de significació per als dos casos.

Per tant, prenent la desició a partir del p valor, accepte l'hipòtesi alternativa i per tant les variàncies són diferents entre hòmens i dones.

Una vegada sabem això, he aplicat el t.test per a les variàncies diferents, a un nivell de confiança del 90% i del 95%. Per a ambdós casos el valor observat cau dintre de la zona d'acceptació de la hipòtesi nul · la, però com abans, el p valor és menor al nivell de significància. Per tant, triant a partir del p valor, hem de rebutjar la hipotèsi nul · la i hem d'acceptar que les dones ténen una mitjana major que la dels homes. Per tant acceptem l'hipòtesi alternativa en ambdós casos.

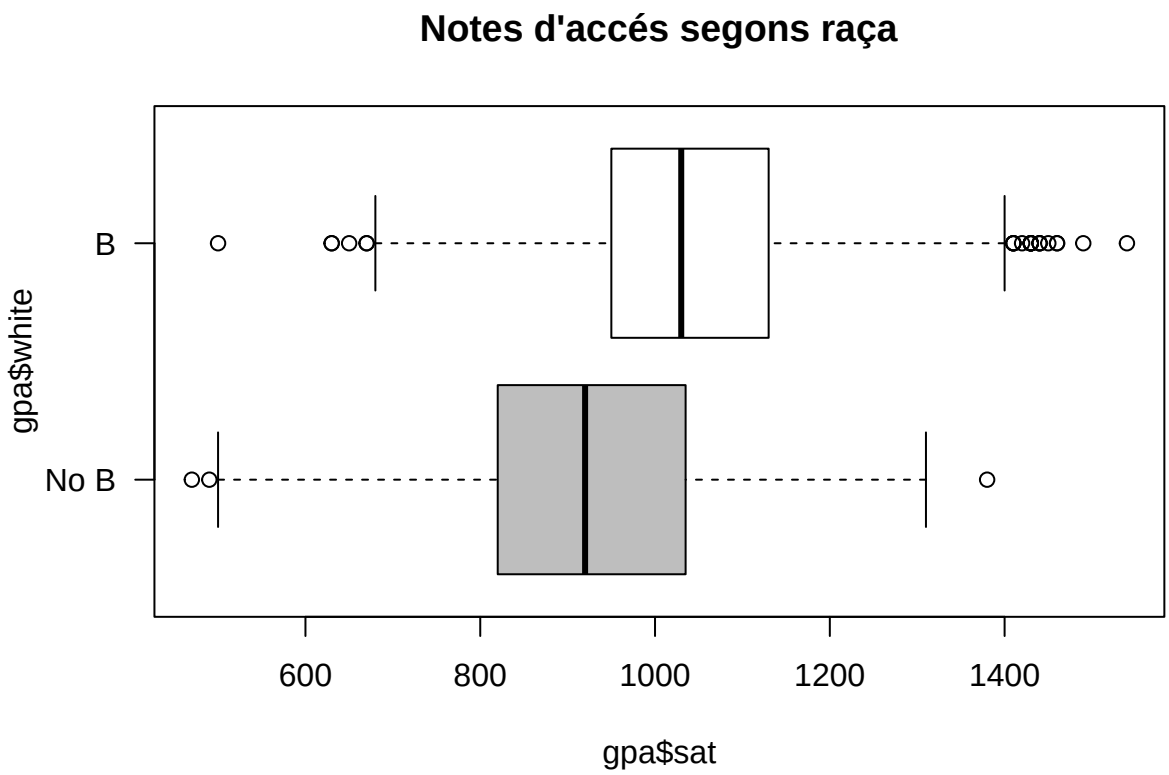
6 Hi ha diferències a la nota segons la raça?

En aquest exercici volem comprovar si la raça influeix a la nota per a una confiança del 90% i del 95%.

6.1 Anàlisi visual

Aquesta pregunta vindria donada per l'anàlisi visual fet anteriorment, on s'hi veia que les persones blanques tindrien un IQR més desplaçat cap al 4 que les que no son blanques:

```
boxplot(gpa$sat ~ gpa$white,
        main = "Notes d'accés segons raça",
        names = c("No B", "B"),
        horizontal = TRUE,
        col = c("grey", "white"),
        las = 1)
```



6.2 Funció

Com a l'exercici anterior, s'utilitzaran les funcions de t.test de R per fer el càlcul. No cal definir cap funció.

6.3 Pregunta de recerca

En aquest problema ens interessa saber si la mitjana de la població de les persones blanques és diferent a la de les persones que no ho son.

Per saber-ho també hauré de calcular si les variàncies són iguals o no.

6.4 Hipòtesi nul · la i l'alternativa

A través del que s'ha explicat abans, definisc les hipotesis com:

H0: La mitjana poblacional de les persones blanques és igual a la mitjana de les que no ho son

H1: La mitjana poblacional de les persones blanques és diferent a la mitjana de les que no ho son

o

H0: MwhiteTRUE = MwhiteFALSE

H1: MwhiteTRUE != MwhiteFALSE

6.5 Justificació del test a aplicar

Una vegada més el test que aplicaré és la comparació de dues mostres independents sobre la mitjana de variàncies desconegudes. Això és perquè podem suposar que seran independents i tenim una mostra major a 30.

6.6 Càlcul

```
# Comprovació de les variàncies
var.test(gpa$colgpa[gpa$white==TRUE],
         gpa$colgpa[gpa$white==FALSE],
         conf.level = 0.90)
```

```
##
##  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
## 90 percent confidence interval:
##  0.8639577 1.1389414
## sample estimates:
## ratio of variances
##          0.9966458
```

```
var.test(gpa$colgpa[gpa$white==TRUE],
         gpa$colgpa[gpa$white==FALSE],
         conf.level = 0.95)
```

```
##
##  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
```

```
# Les variàncies no són iguals
t.test(gpa$colgpa[gpa$white==TRUE],
       gpa$colgpa[gpa$white==FALSE],
       alternative="two.sided",
       var.equal=TRUE,
       conf.level = 0.90)
```

```
##
##  Two Sample t-test
##
## data:  gpa$colgpa[gpa$white == TRUE] and gpa$colgpa[gpa$white == FALSE]
## t = 8.4233, df = 4135, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  0.2618740 0.3890016
## sample estimates:
## mean of x mean of y
##  2.678360  2.352922
```

```
t.test(gpa$colgpa[gpa$white==TRUE],
       gpa$colgpa[gpa$white==FALSE],
       alternative="two.sided",
       var.equal=TRUE,
       conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  gpa$colgpa[gpa$white == TRUE] and gpa$colgpa[gpa$white == FALSE]
## t = 8.4233, df = 4135, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2496916 0.4011840
## sample estimates:
## mean of x mean of y
##  2.678360  2.352922
```

6.7 Interpretació del test

Primerament he fet el test de les variàncies per veure si són equivalents entre blancs i no blancs. El resultat per al 90% i per al 95% m'ha donat els mateixos resultats: el valor observat cau en la zona d'acceptació de la hipòtesi nul · la i el p valor és més gran que el nivell de significació. Per tant no podem descartar la hipòtesi nul · la i asumim que les variànceis són iguals per ambdós casos.

Una vegada sabem això, he aplicat el t.test per a les variàncies iguals, a un nivell de confiança del 90% i del 95%. Per a ambdós casos el valor observat cau fora de la zona d'acceptació de la hipòtesi nul · la, i així mateix, el p valor és menor al nivell de significància. Per tant, podem rebutjar la hipotèsi nul · la i hem d'acceptar que existeix una diferència entre persones blanques i no blanques pel que fa la nota final. Per tant acceptem l'hipòtesi alternativa en ambdós casos.

7 Proporció d'atletes

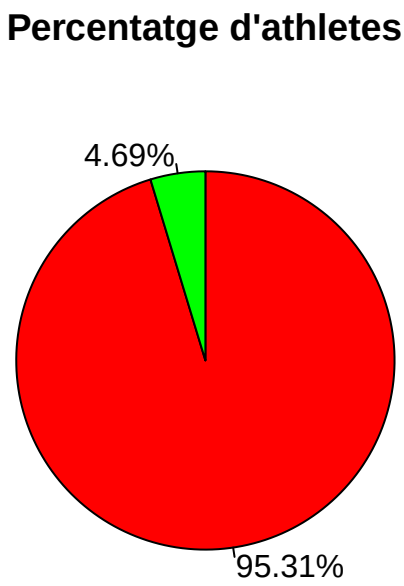
Per a aquest exercici treballaré amb proporcions sobre la mateixa categoria de athletes. Per tant aquest sí és més diferent que la resta.

7.1 Anàlisi visual

Com hem fet prèviament a la pac anterior, sabem que la proporció d'athletes és molt menuda i no aplega al 5% de la mostra. És per això que volem confirmar si açò es pot extrapolar a la població sancera.

```
# Cree la taula de freqüencies:
athletes <- table(gpa$athlete)
PropAth <- (athletes / sum(athletes)) * 100

# Cree un diagrama
pie(PropAth,
    labels = paste0(round(PropAth, digits = 2), "%"),
    main = "Percentatge d'athletes",
    clockwise = TRUE,
    col = c("red", "green")
    )
```




```
remove(athletes,PropAth)
```

7.2 Pregunta de recerca

Per a aquest exercici ens interessa saber si la proporció d'atletes és inferior al 5% de la població per a un nivell de confiança del 95%.

7.3 Hipòtesi nul · la i l'alternativa

Per el que s'ha comentat anteriorment, les hipòtesis són:

P0: La proporció d'atletes a la població és igual al 5%

P1: La proporció d'atletes a la població és menor al 5%

o

P0: $p = 0.05$

P1: $p < 0.05$

7.4 Justificació del test a aplicar

S'ha d'aplicar el test d'hipotesis d'una mostra sobre la proporció perquè sols estudiem el nombre d'atletes a la poblacio a partir de la mostra que tenim.

7.5 Realitzeu els càlculs del test

```
# Sabem que la proporció a la mostra és igual a 4.69% (s'hi veu al gràfic)

# Calcule l'estadístic de contrast:
z <- round((0.0469 - 0.05)/sqrt((0.05*(1-0.05))/length(gpa$athlete)), 2)

# Calcule el valor crític:
VC <- round(qt(0.05,(length(gpa$athlete)-1)), 2)

# Calcule el p-valor:
p <- 2*pt(-abs(z),df=(length(gpa$athlete)-1))

# Resultats:
cat("L'interval és: [",VC,", INF]\n")
```

```
## L'interval és: [ -1.65 , INF]
```

```
cat("L'estadístic de contrast és: ",z,"\n")
```

```
## L'estadístic de contrast és:  -0.91
```

```
cat("El p-valor és: ",p)
```

```
## El p-valor és:  0.3628755
```

7.6 Interpretació del test

L'estadístic de contrast cau dintre de l'interval d'acceptació de la hipòtesi nul · la, per tant no podem rebutjar-la a priori. Així mateix, com el p valor no és inferior a la significància fixada, no podem rebutjar la hipòtesi nul · la. Per tant podem concloure en que la població d'athletes no és inferior a 5%.

8 Hi ha més atletes entre els homes que entre les dones?

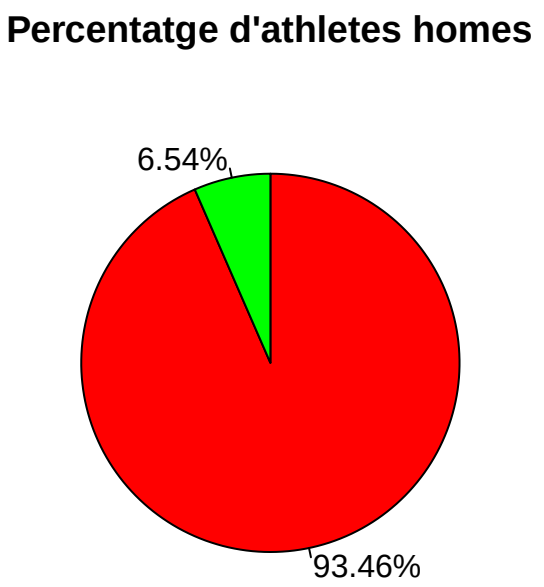
En aquest exercici treballarem sobre les proporcions de dues mostres diferents, per als homes i per a les dones.

8.1 Anàlisi visual

Tal i com hem fet a l'exercici anterior, podem construïr la proporció d'atletes per als homens com per les dones. S'hi veu que el nombre d'atletes homes és major al nombre de dones atletes.

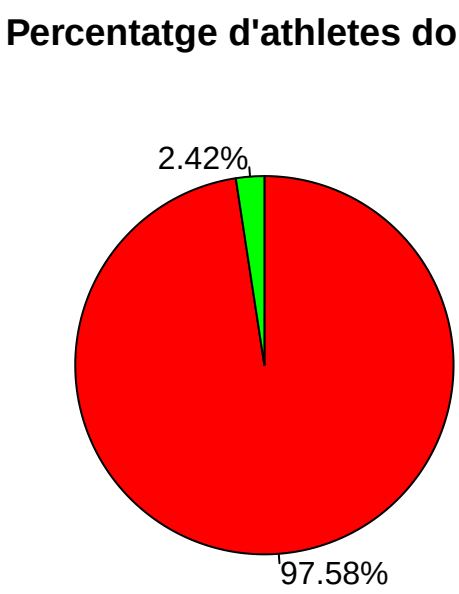
```
# Cree la taula de freqüencies d'homes:
homes <- gpa[gpa$female == FALSE,]
athletes <- table(homes$athlete)
PropAth <- (athletes / sum(athletes)) * 100

# Cree un diagrama
pie(PropAth,
    labels = paste0(round(PropAth, digits = 2), "%"),
    main = "Percentatge d'athletes homes",
    clockwise = TRUE,
    col = c("red", "green")
    )
```



```r
# Cree la taula de freqüencies de dones:
dones <- gpa[gpa$female == TRUE,]
athletes <- table(dones$athlete)
PropAth <- (athletes / sum(athletes)) * 100

# Cree un diagrama
pie(PropAth,
    labels = paste0(round(PropAth, digits = 2), "%"),
    main = "Percentatge d'athletes dones",
    clockwise = TRUE,
    col = c("red", "green")
    )
```




```r
remove(athletes,PropAth)
```

8.2 Pregunta de recerca

A causa de l'anàlisi visual podem intuir que hi ha més atletes entre els homes que entre les dones. Per tant volem saber per a una confiança del 95% si per la població el nombre d'atletes entre els homes és superior al de les dones.

8.3 Hipòtesi nul · la i l'alternativa

Per el que s'ha comentat anteriorment, les hipòtesis són:

P0: La proporció d'atletes a la població masculina és igual a la proporció femenina

P1: La proporció d'atletes a la població masculina és major a la proporció femeninaal

o

P0: pH = pD

P1: pH > pD

8.4 Justificació del test a aplicar

Com hi ha prou dades per a homes i per a dones, podem dividir la mostra en dos suposant que siguen dades preses de manera independent. Com son mostres prou grans, podem fer un test de dues mostres sobre la proporció d'atletes per veure si les hipotesis es compleixen.

8.5 Càlculs del test

```
# homes guarda female == FALSE
# dones guarda female == TRUE

# Calculem les proporcions:
p1 <- sum(homes$athlete)/length(homes$athlete)
p2 <- sum(dones$athlete)/length(dones$athlete)

p <- (length(homes$athlete)*p1 + length(dones$athletet)*p2)/
  (length(homes$athlete)+length(dones$athlete))

# Estadístic de contrast:
z <- round((p1-p2)/(sqrt(p*(1-p)*(1/length(homes$athlete)+1/length(dones$athlete)))), 2)

# Valor crític al 95%:
VC <- round(qnorm(0.05, lower.tail = FALSE), 2)

# p-valor:
p <- pnorm(z, lower.tail = FALSE)

# Resultats:
cat("L'interval és: [INF,",VC,"]\n")
```

```
## L'interval és: [INF, 1.64 ]
```

```
cat("L'estadístic de contrast és: ",z,"\n")
```

```
## L'estadístic de contrast és:  7.08
```

```
cat("El p-valor és: ",p)
```

```
## El p-valor és:  7.207723e-13
```

8.6 Interpretació del test

L'estadístic de contrast en aquest cas cau fora de la zona d'acceptació de la hipòtesi nul · la. Així mateix, el p-valor és inferior al nivell de significació. Per tant podem afirmar que per a una confiança del 95% la proporció entre la població d'atletes és major entre els homes que en les dones. Per tant acceptem la hipotèsi alternativa.

9 Resum i conclusions

```
## Loading required package: knitr
```

```
## Warning: package 'knitr' was built under R version 4.1.2
```

Table 1: Taula resum del preprocesament

N	Pregunta	Resultat	Conclusió
3.3	Interval de confiança mitjana poblacional sat al 90%	(1026.77, 1033.90)	El resultat al 90% està entre 1026.77 i 1033.90
3.3	Interval de confiança mitjana poblacional sat al 95%	(1026.08, 1034.58)	El resultat al 95% està entre 1026.08 i 1034.58
3.4	Interval de confiança mitjana poblacional colgpa al 90%	(2.64, 2.67)	El resultat al 95% està entre 2.64 i 2.67
3.4	Interval de confiança mitjana poblacional colgpa al 95%	(2.63, 2.67)	El resultat al 95% està entre 2.63 i 2.67
4	Ser athleta influeix a la nota?	((-1.97, 1.97), -5.073683e+00, 9.134190e-07)	El resultat al 95% ens permet afirmar que ser athleta efectivament influeix a la nota final
5	Les dones tenen millor nota que els homes al 90%?	(t=7.0787, (0.1173717, Inf), p-value=8.522e-13)	El resultat al 90% ens permet afirmar que les dones tenen més nota final que els homes
5	Les dones tenen millor nota que els homes al 95%?	(t=7.0787, (0.1100126, Inf), p-value=8.522e-13)	El resultat al 95% ens permet afirmar que les dones tenen més nota final que els homes
6	Hi ha diferències a la nota segons la raça al 90%?	(t=8.4233, (0.2618740, 0.3890016), p-value<2.2e-16)	El resultat al 90% ens permet afirmar sí existeixen diferències entre races
6	Hi ha diferències a la nota segons la raça al 95%?	(t=8.4233, (0.2496916, 0.4011840), p-value<2.2e-16)	El resultat al 95% ens permet afirmar sí existeixen diferències entre races
7	És la proporció d'atletes a la població menor al 5% al 95% confiança?	((-1.65, INF), z = -0.91, p-value = 0.3628755)	El resultat al 95% ens permet descartar que els atletes a la població siguen menys del 5%
8	La proporció d'atletes entre els homes és major que entre les dones al 95%?	((INF, 1.64), z = 7.08, p-value = 0.0000000000007)	El resultat al 95% ens permet afirmar que entre els homes hi ha més atletes que entre les dones

10 Resum executiu

En aquesta PAC hem estudiat les notes d'accés dels estudiants, les notes mitjanes, i les distribucions relacionades amb el sexe, l'esport i la raça. Ho hem fer per poder extraure informació útil del dataset proporcionat. D'aquest s'ha extret:

Nota d'accés: la meitat dels alumnes ténen una nota d'accés d'entre 900 i 1100. Els que més nota d'accés tenen són els homes, la gent que no practica esport i les persones blanques. S'han estudiat les distribucions individualment, no creuant paràmetres (Ex: no s'han cercat les dones esportistes blanques). La nota mitja de la nota d'accés per a la població estaria entre 1026.08 i 1034.58.

Nota final: Per a aquesta s'ha realitzat un estudi més exhaustiu. Sabem que la mitjana poblacional serà 2.6 aproximadament. A més s'han fet diversos estudis per comprovar les hipotesis sorgides a l'inspecció visual de les distribucions:

- S'ha confirmat la hipòtesis que diu que ser atleta influeix a la nota final. Concretament, es presuposa que de manera negativa segons l'anàlisi visual.
- S'ha confirmat la hipòtesis que diu que les dones tenen millor mitja que els homes.
- S'ha confirmat la hipòtesis que diu que hi ha una diferència en la mitja segons les races. És a dir, els blancs suposadament tindrien una mitja major a les persones no blanques segons l'anàlisi visual.

Finalment s'ha de comentar que la meitat dels alumnes aprova.

Proporció d'atletes: S'ha descartat la hipòtesis que diu que hi ha menys d'un 5% d'atletes a la població d'alumnes. Aquesta venia donada per una inspecció visual de les dades. És probable que la mostra estiga esbiaixada per mostrar-ho així.

Proporció d'atletes segons sexe: S'ha confirmat la hipòtesis que diu que hi ha més atletes entre els homes que d'entre les dones.