A2 - Analítica descriptiva i inferencial

Solució

Semestre 2022.1

Índice

Introducció

En aquesta activitat ens introduïm a la inferència estadística. Per fer-ho, farem servir el conjunt de dades `gpa.csv` que s'ha preprocessat a l'activitat anterior. Aquest conjunt de dades conté la nota mitjana d'estudiants universitaris després del primer semestre de classes (GPA: grade point average, en anglès), així com informació sobre la nota d'accés, la cohort de graduació a l'institut i algunes característiques dels estudiants.

Aquest conjunt de dades sorgeix d'una enquesta realitzada a una mostra representativa d'estudiants d'una universitat dels EUA (per raons de confidencialitat el conjunt de dades no inclou el nom de la universitat). Les variables incloses al conjunt de dades són:

- sat: nota d'accés (mesura a escala de 400 a 1600 punts)
- tothrs: hores totals cursades al semestre
- colgpa: nota mitjana de l'estudiant al final del primer semestre (mesura en escala de 0 a 4 punts)
- athlete: indicador de si l'estudiant practica algun esport a la universitat
- hsize: nombre total d'estudiants a la cohort de graduats del batxillerat (en centenars)
- hsrank: rànquing de l'estudiant, donat per la nota mitjana del batxillerat, en la cohort de graduats del batxillerat
- hsperc: rànquing relatiu de l'estudiant, en percentatge (hsrank/hsize)
- female: indicador de si l'estudiant és dona
- white: indicador de si l'estudiant és de raça blanca o no
- black: indicador de si l'estudiant és de raça negra o no
- gpaletter: lletra que indica el nivell de la nota colgpa (A,B,C,D)

L'objectiu d'aquesta activitat és estudiar la nota dels estudiants a partir de les variables d'interès així com la proporció d'atletes entre la població d'estudiants. Per això, les preguntes que ens plantegem són:

- P1. Quin és l'interval de confiança de la nota en els estudiants?
- P2. Ser atleta influeix a la nota?
- P3. Les dones obtenen millor nota que els homes?
- P4. Hi ha diferències significatives a la nota segons la raça?
- P5. La proporció d'atletes a la població és inferior al 5%?
- P6. Hi ha més atletes entre els homes que entre les dones?

Al final de l'anàlisi, us demanem un resum executiu on cal resumir i explicar breument les conclusions de l'estudi.

Nota important a tenir en compte per lliurar l'activitat:

- Cal lliurar el fitxer Rmd i el fitxer de sortida (PDF o html). El fitxer de sortida ha d'incloure: el codi i el resultat de l'execució del codi (pas a pas).
- Cal respectar la mateixa numeració dels apartats que l'enunciat.
- No es poden realitzar llistats complets del conjunt de dades a la solució. Això generaria un document amb centenars de pàgines i dificulta la revisió del text. Per comprovar les funcionalitats del codi sobre les dades, es poden fer servir les funcions **head** i **tail** que només mostren unes línies del fitxer de dades.
- Es valora la precisió dels termes utilitzats (cal utilitzar de manera precisa la terminologia de l'estadística).

- Es valora també la concisió a la resposta. No es tracta de fer explicacions gaire llargues o documents molt extensos. Cal explicar-ne el resultat i argumentar la resposta a partir dels resultats obtinguts de manera clara i concisa.

---

1 Lectura del fitxer

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

```
## 'data.frame':    4096 obs. of  10 variables:
##  $ sat    : int  920 1170 810 940 1180 980 880 980 1240 1230 ...
##  $ tothrs : Factor w/ 125 levels "100h","101h",..: 67 46 42 64 46 16 103 79 46 45 ...
##  $ 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 ...
```

---

2 Estadística descriptiva i visualització

2.1 Anàlisi descriptiva

Realitzeu una anàlisi descriptiva numèrica de les dades (resumeu els valors de les variables numèriques i categòriques). Mostreu el nombre d'observacions i el nombre de variables.

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

```
## Observacions:  4096  Variables:  10
```

```
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    1030.905762 139.3353660
## colgpa    2.654546   0.6601642
```

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

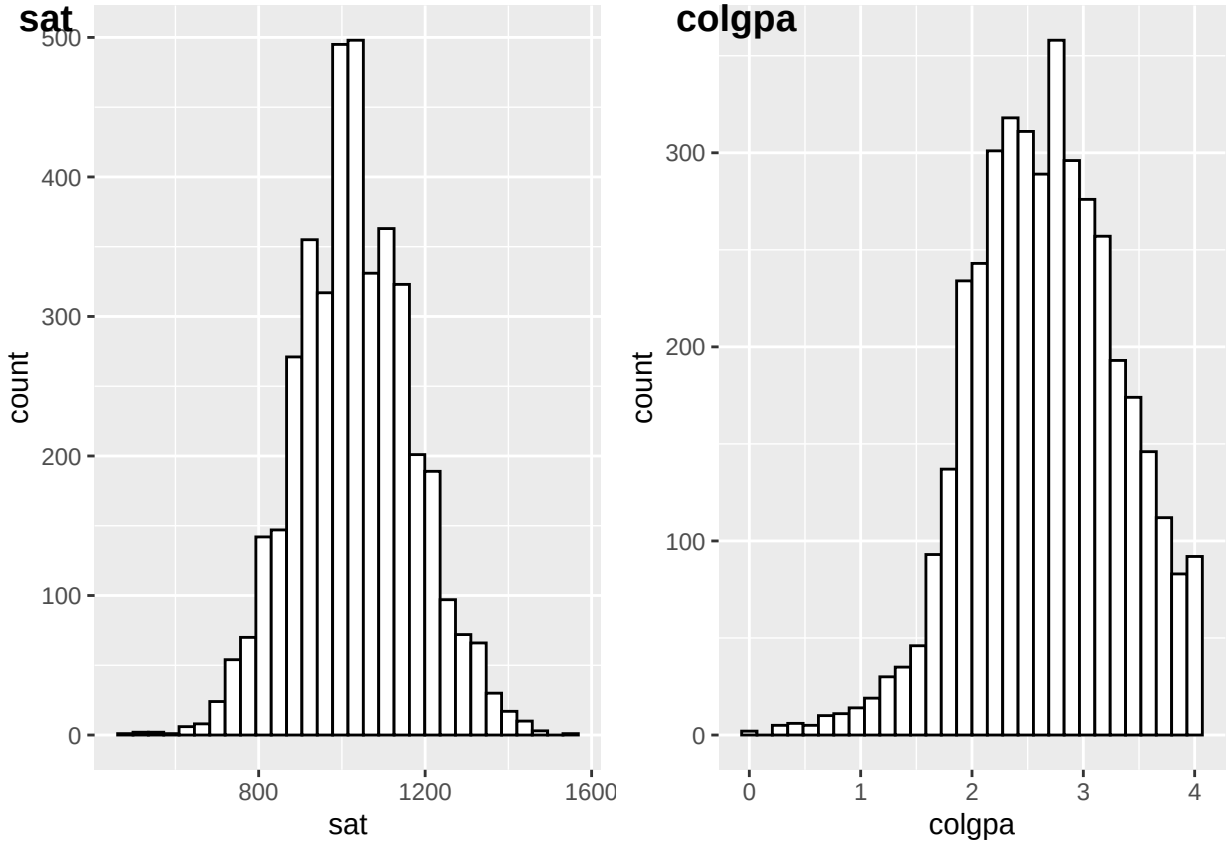
```
##       athlete female white black
## FALSE    3905   2253   304  3871
## TRUE      191   1843  3792   225
```

2.2 Visualització

Estudiem de forma visual la distribució de les variables `sat` i `colgpa`, així com la influència del sexe, raça i ser atleta en aquestes variables.

```
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`.
```



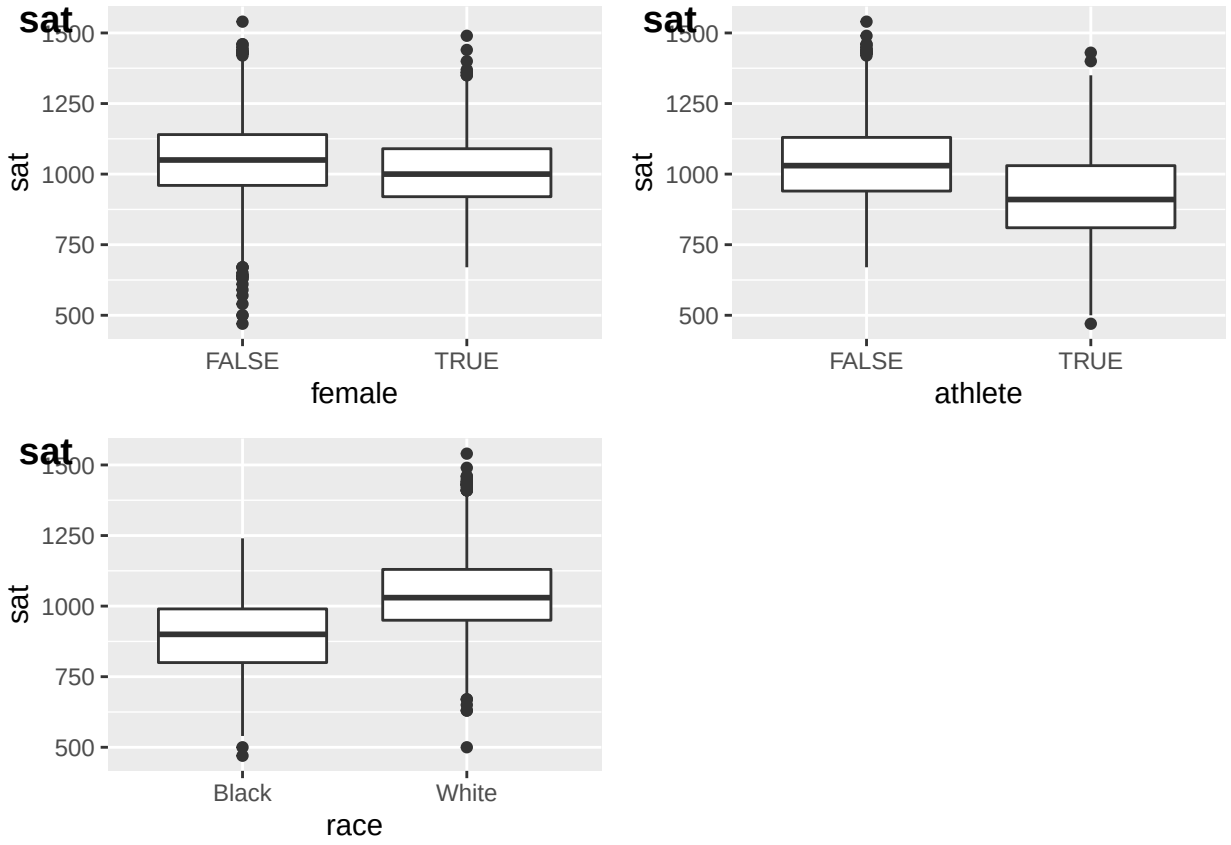

```
v.sat.fem<-ggplot(gpa, aes(x=female, y=sat)) +   geom_boxplot()
v.sat.at<-ggplot(gpa, aes(x=athlete, y=sat)) +   geom_boxplot()

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] 4017
```

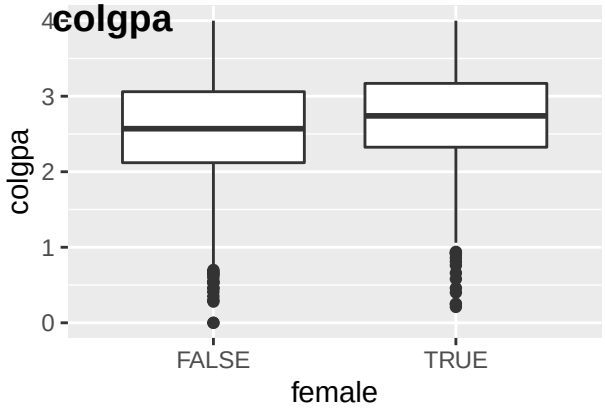
```
sum( complete.cases(gpa.f) )
```

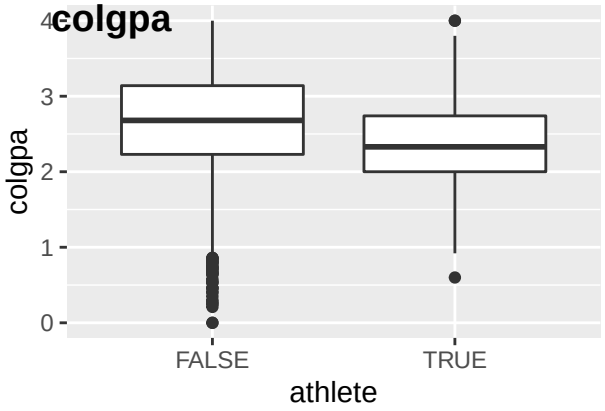
```
## [1] 4017
```

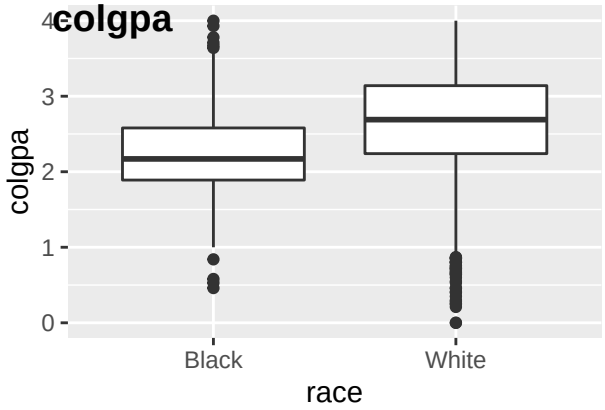
```
v.sat.white<-ggplot(gpa.f, aes(x=race, y=sat)) +   geom_boxplot()
ggarrange(v.sat.fem, v.sat.at, v.sat.white,
          labels = c("sat", "sat", "sat"),
          ncol = 2, nrow = 2)
```




```
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` es distribueix de forma similar a una normal, amb valors centrats al voltant de 1000, i amb mínim de 400 i màxim 1600. La variable `colgpa` té una distribució semblant a una normal però força asimètrica. Concretament, la cua de l'esquerra és més allargada que la cua de la dreta, encara que presenta menys freqüència de valors en els valors per sota de 2.

S'observen diferències a la variable `sat` segons si l'estudiant és atleta o no i també entre races. Les diferències en `sat` entre sexes són menys apreciables. Un comportament semblant passa amb la variable `colgpa`, encara que en el cas de `colgpa` sembla que el sexe femení presenta millor nota que el sexe masculí (al contrari del que passa amb `sat`).

---

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

3.1 Supòsits

Assumim distribució normal pel Teorema del Límit Central, ja que la mida de la mostra és prou gran (n=4096). El matís és que no coneixem la variància de la població i per tant usem la variància mostral per aproximar la variància de la població. En aquest cas, hem d'aplicar la distribució t de Student amb n-1 graus de llibertat. A la pràctica, com que la mida de mostra és prou gran, la distribució t de Student és molt similar a la distribució normal. Per tant, s'acceptaria també que es faci servir qnorm en lloc de qt.

3.2 Funció de càlcul de l'interval de confiança

Funció que calcula l'interval de confiança atesa una variable d'interès `x` i un nivell de confiança donat `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 Interval de confiança de la variable sat

Calculem l'interval de confiança al 90% de la mitjana poblacional de la variable `sat` i al 95%.

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

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

```
## [1] 1027.324 1034.488
```

```
## [1] 1026.637 1035.174
```

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

```
## [1] 1027.324 1034.488
## attr(,"conf.level")
## [1] 0.9
```

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

```
## [1] 1026.637 1035.174
## attr(,"conf.level")
## [1] 0.95
```

3.4 Interval de confiança de la variable colgpa.

Càlcul de l'interval de confiança de la variable `colgpa` amb un nivell de confiança del 90% i del 95%.

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

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

```
## [1] 2.637575 2.671517
```

```
## [1] 2.634323 2.674769
```

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

```
## [1] 2.637575 2.671517
## attr(,"conf.level")
## [1] 0.9
```

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

```
## [1] 2.634323 2.674769
## attr(,"conf.level")
## [1] 0.95
```

3.5 Interpretació

La interpretació de l'interval de confiança és que si fem un mostreig elevat de mostres de la població, aproximadament el NC% (95% o 90%) dels intervals de confiança obtinguts d'aquestes mostres contenen el valor de la mitjana poblacional de colgpa/sat.

---

4 Ser atleta influeix a la nota?

En aquest apartat volem analitzar si ser atleta influeix a la nota `colgpa`. És a dir, si hi ha diferències significatives entre atletes i no atletes en aquesta nota, amb un nivell de confiança del 95%.

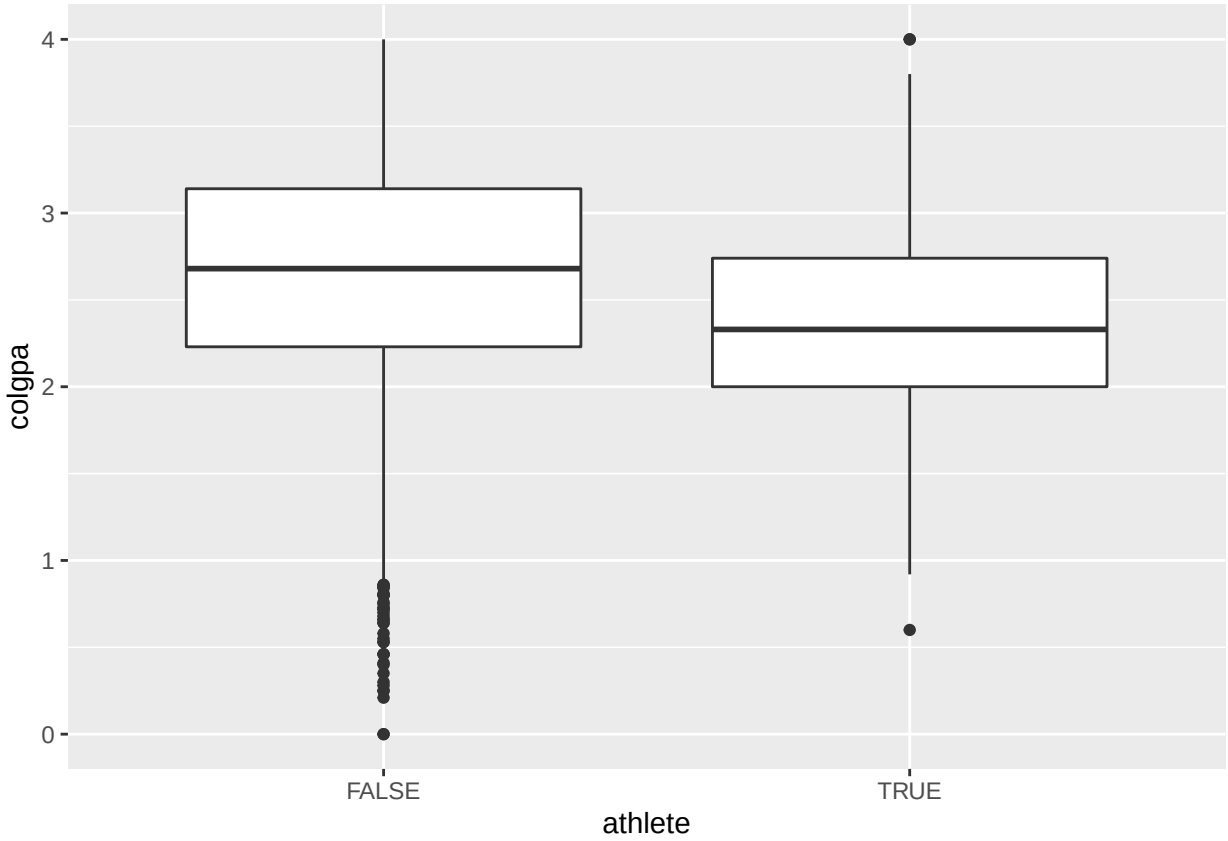
Com que farem preguntes similars en els apartats següents, es recomana implementar una funció que permeti realitzar tests d'hipòtesis per a la diferència de mitjanes. La funció ha de rebre com a paràmetre les dues mostres, el nivell de confiança i altres paràmetres que es puguin requerir.

Seguiu els passos que es detallen a continuació.

4.1 Anàlisi visual

Es mostra un boxplot de la variable colpga en funció de si l'estudiant és atleta o no.

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



4.2 Funció per al contrast de mitjanes

S'implementa una funció que calcula el contrast (paramètric) de mitjanes de dues mostres i que torna: el valor de l'estadística de contrast, el valor crític i el valor p. El contrast és bilateral. La funció assumeix normalitat a la variable d'interès.

```
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)

  #variances iguals
  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{
    #variances diferents
    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

  #Guardem el resultat 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 recerca

```
## [1] "Hi ha diferències significatives en la nota `colgpa` entre atletes i no atletes?"
```

4.4 Hipòtesis nul · la i alternativa

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

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

4.5 Justificació del test a aplicar

És un test de dues mostres sobre la mitjana amb variàncies desconegudes. Pel teorema del límit central, podem assumir normalitat. Comprovem igualtat de variàncies:

```
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 = 0.82843, num df = 190, denom df = 3904, p-value = 0.0874
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6805288 1.0283923
## sample estimates:
## ratio of variances
##           0.828433
```

El resultat del test no mostra diferències significatives entre variàncies. Per tant, aplicarem un test de dues mostres independents sobre la mitjana amb variàncies desconegudes iguals. El test és bilateral.

4.6 Càlcul

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

```
##         mean1         mean2             t     tcritical        pvalue
##  2.381728e+00  2.667890e+00 -5.873221e+00  1.960544e+00  4.613421e-09
##            df
##  4.094000e+03
```

```
answer.P2 <- dif.colgpa.at
#Comprovació
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 = -5.8732, df = 4094, p-value = 4.613e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3816861 -0.1906382
## sample estimates:
## mean of x mean of y
##  2.381728  2.667890
```

4.7 Interpretació del test

Hi ha diferències significatives a la nota `colgpa` entre els atletes i no atletes (p=4.6134206×10^{-9}). Es pot observar així mateix que el valor crític és 1.9605436 i el valor observat és 5.873221, notablement superior al valor crític i, per tant, fora de la zona d'acceptació de la hipòtesi nul · la.

---

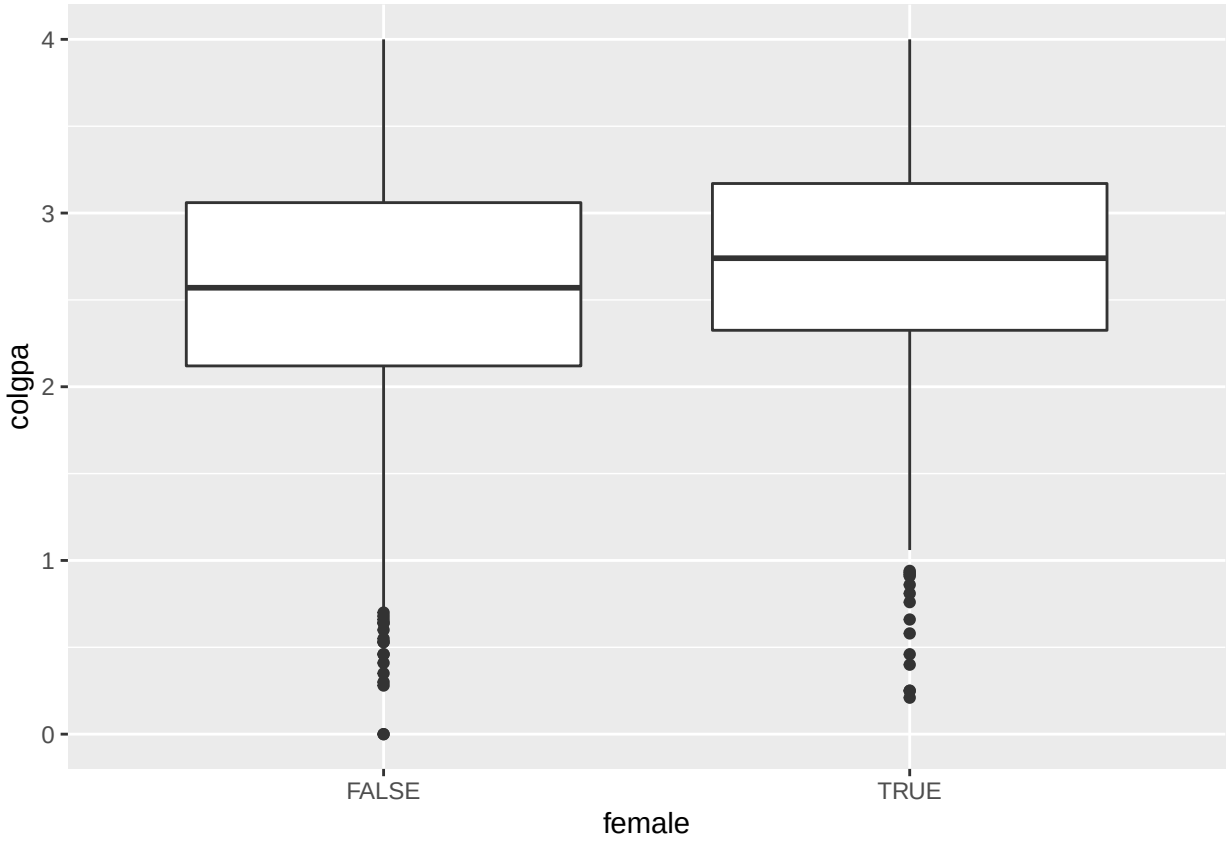
5 Les dones tenen millor nota que els homes?

Es calcula per a un nivell de confiança del 95% i del 90%.

5.1 Análisi visual

Es mostra un boxplot amb la variable colgpa comparant dones amb homes.

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



5.2 Funció

S'implementa una funció per al contrast de mitjanes 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)
  #variances iguals
  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{
    #variances diferents
    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 observat
  }
  #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 )
  }

  #Guardem el resultat 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 recerca

```
## [1] "Les dones tenen millor nota a "colgpa" que els homes?"
```

5.4 Hipòtesi nul · la i alternativa

Escriviu les hipòtesis nul · la i alternativa.

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

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

5.5 Justificació del test a aplicar

És un test de dues mostres sobre la mitjana amb variàncies desconegudes. Pel teorema del límit central, podem assumir normalitat. Comprovem igualtat de variàncies:

```
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 = 0.82788, num df = 1842, denom df = 2252, p-value = 2.305e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7589643 0.9033950
## sample estimates:
## ratio of variances
##          0.8278771
```

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

El resultat del test mostra diferències significatives entre variàncies (p=2.3046097×10^{-5}). Per tant, aplicarem un test de dues mostres independents sobre la mitjana amb variàncies desconegudes diferents. El test és unilateral per la dreta.

5.6 Càlcul

```
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
## 2.733511e+00 2.589951e+00 7.029779e+00 1.645230e+00 1.208469e-12 4.047939e+03
```

```
answer.P3 <- dif.colgpa.fem
#Comprovació
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 = 7.0298, df = 4047.9, p-value = 1.208e-12
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1099612       Inf
## sample estimates:
## mean of x mean of y
##  2.733511  2.589951
```

5.7 Interpretació del test

Es pot afirmar que les estudiants de sexe femení tenen una millor nota que els estudiants de sexe masculí (p=$1.2084693 \times 10^{-12}$). Es pot observar així mateix que el valor crític és 1.6452301 i el valor observat és 7.0297786, notablement superior al valor crític i, per tant, fora de la zona d'acceptació de la hipòtesi nul · la. No cal calcular per al nivell de confiança del 90%, ja que s'han trobat diferències significatives amb el nivell de confiança del 95%.

---

6 Hi ha diferències a la nota segons la raça?

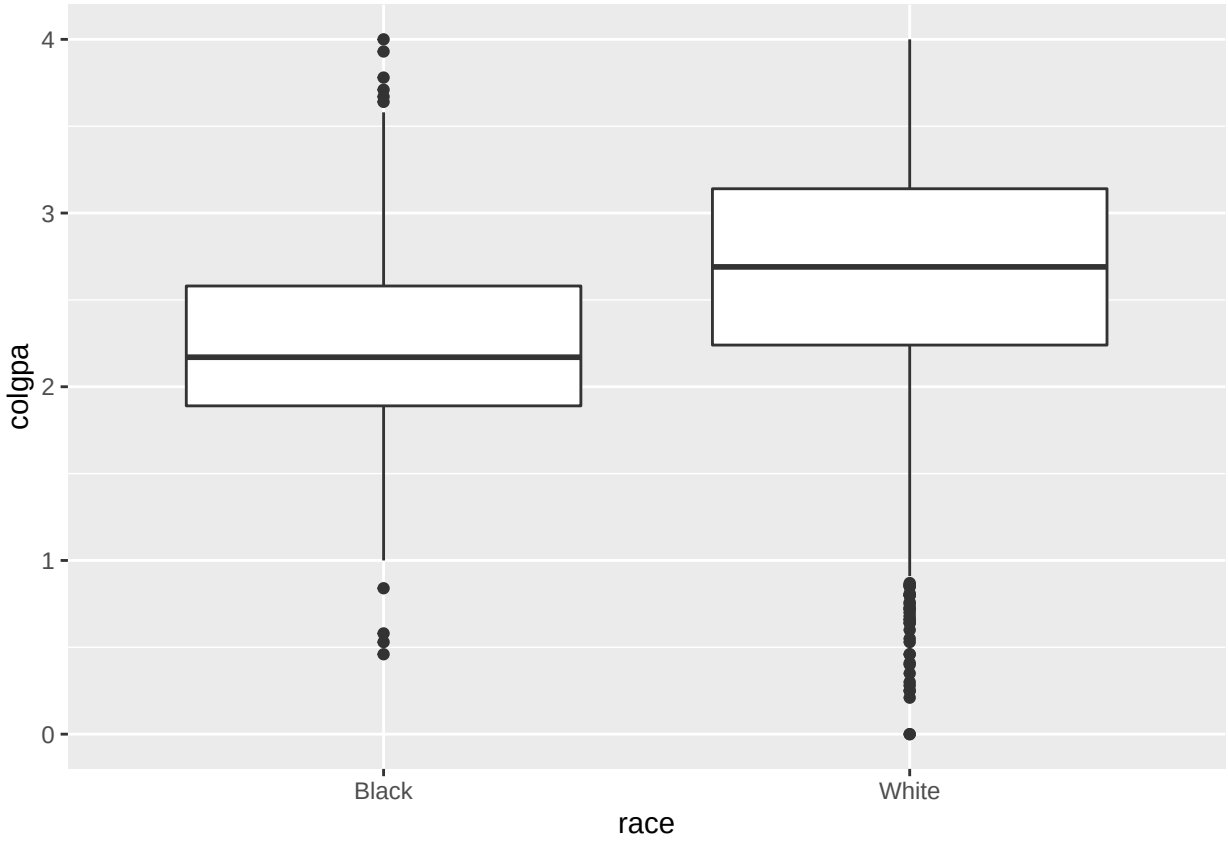
Abans de desenvolupar aquest apartat, volem fer èmfasi que el fet que hi hagi (si existeixen) diferències entre races (blancs i negres), no necessàriament ens porta a concloure que la raça influeix en la nota (hi ha factors socioeconòmics que poden afectar). Per tant, cal ser molt cautelós a l'hora de fer conclusions en aquest sentit. Un tema similar pot passar amb les diferències quant a gènere. Tot i aquestes puntualitzacions, l'estudi és interessant ja que en cas de detectar diferències en un sentit o l'altre, es poden analitzar els causants d'aquestes diferències i realitzar intervencions educatives apropiades.

Després d'aquest matís, ens agradaria estudiar si les persones de raça blanca tenen una nota diferent a "colgpa" que les persones de raça negra. Seguiu els mateixos apartats que anteriorment.

6.1 Anàlisi visual

Es mostra un boxplot amb la variable colgpa comparant dones amb homes.

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



6.2 Funció

Usem una de les funcions anteriors.

6.3 Pregunta de recerca

```
## [1] "Hi ha diferències en colgpa entre els estudiants de raça blanca i els de raça negra?"
```

6.4 Hipòtesis nul · la i alternativa

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

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

6.5 Justificació del test a aplicar

És un test de dues mostres sobre la mitjana amb variàncies desconegudes. Pel teorema del límit central, podem assumir normalitat. Comprovem igualtat de variàncies:

```
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 = 1.1236, num df = 3791, denom df = 224, p-value = 0.2505
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9206914 1.3489979
## sample estimates:
## ratio of variances
##           1.123629
```

El resultat del test no mostra diferències significatives entre variàncies. Per tant, aplicarem un test de dues mostres independents sobre la mitjana amb variàncies desconegudes iguals. El test és bilateral.

6.6 Càlcul

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

```
##        mean1        mean2            t    tcritical       pvalue           df
## 2.679045e+00 2.248444e+00 9.619516e+00 1.960555e+00 1.130982e-21 4.015000e+03
```

```
answer.P4 <- dif.gpa.white
#Comprovació
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 = 9.6195, df = 4015, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3428401 0.5183618
## sample estimates:
## mean of x mean of y
##  2.679045  2.248444
```

6.7 Interpretació del test

Hi ha diferències significatives a la nota `colgpa` entre els estudiants de raça blanca i els de raça negra (p=$1.1309822 \times 10^{-21}$). Es pot observar així mateix que el valor crític és 1.960555 i el valor observat és 9.6195156, el qual es troba fora de la zona d'acceptació de la hipòtesi nul · la.

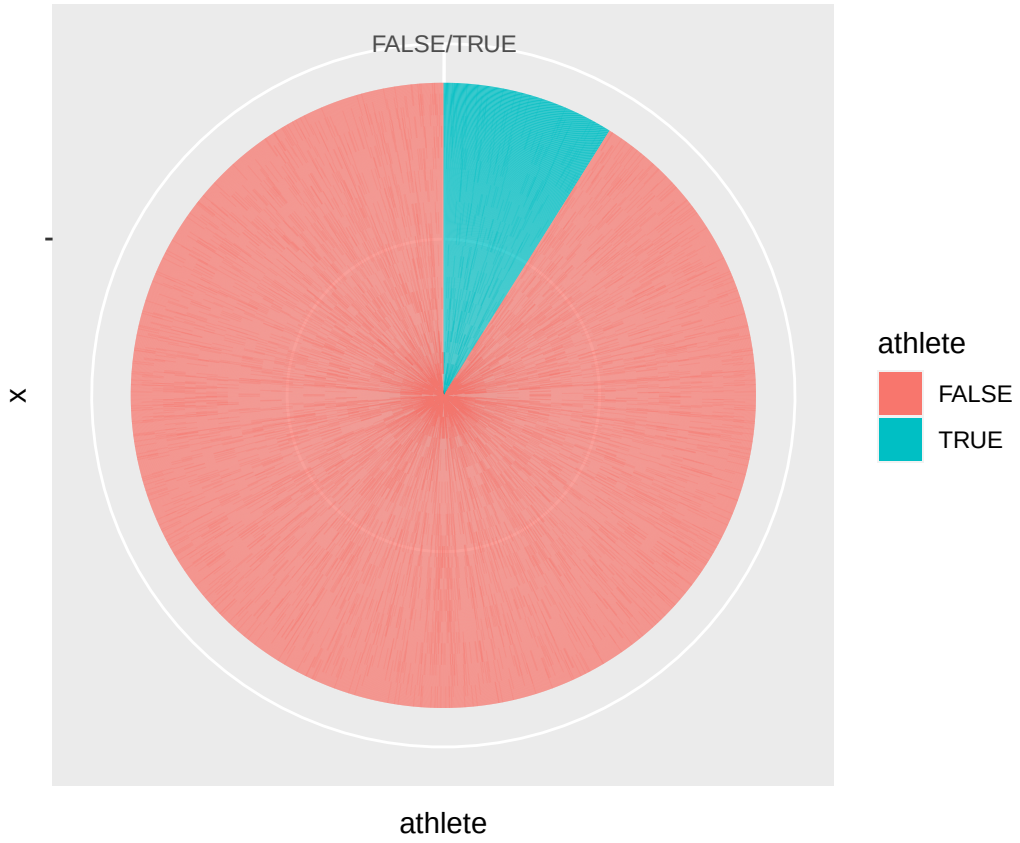
---

7 Proporció d'atletes

Ens preguntem si la proporció d'atletes a la població és inferior al 5% amb un nivell de confiança del 95%. Per fer-ho, seguiu els mateixos passos que en els casos anteriors.

7.1 Análisi visual

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



7.2 Pregunta de recerca

```
## [1] "La proporció d'atletes a la població és inferior a 0.05?"
```

7.3 Hipòtesi nul · la i alternativa

$H_0 : p_{at} = 0.05$

$H_1 : p_{at} < 0.05$

7.4 Justificació del test a aplicar

És un contrast sobre la proporció d'una mostra. És un test unilateral per l'esquerra. Assumim mostres grans.

7.5 Càlculs

```r
#Test de proporció d'una mostra unilateral per l'esquerra.
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 Càlcul

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

```
##           p          p0           z   zcritical      pvalue
##  0.04663086  0.05000000 -0.98935535 -1.64485363  0.16124466
```

7.7 Interpretació del test

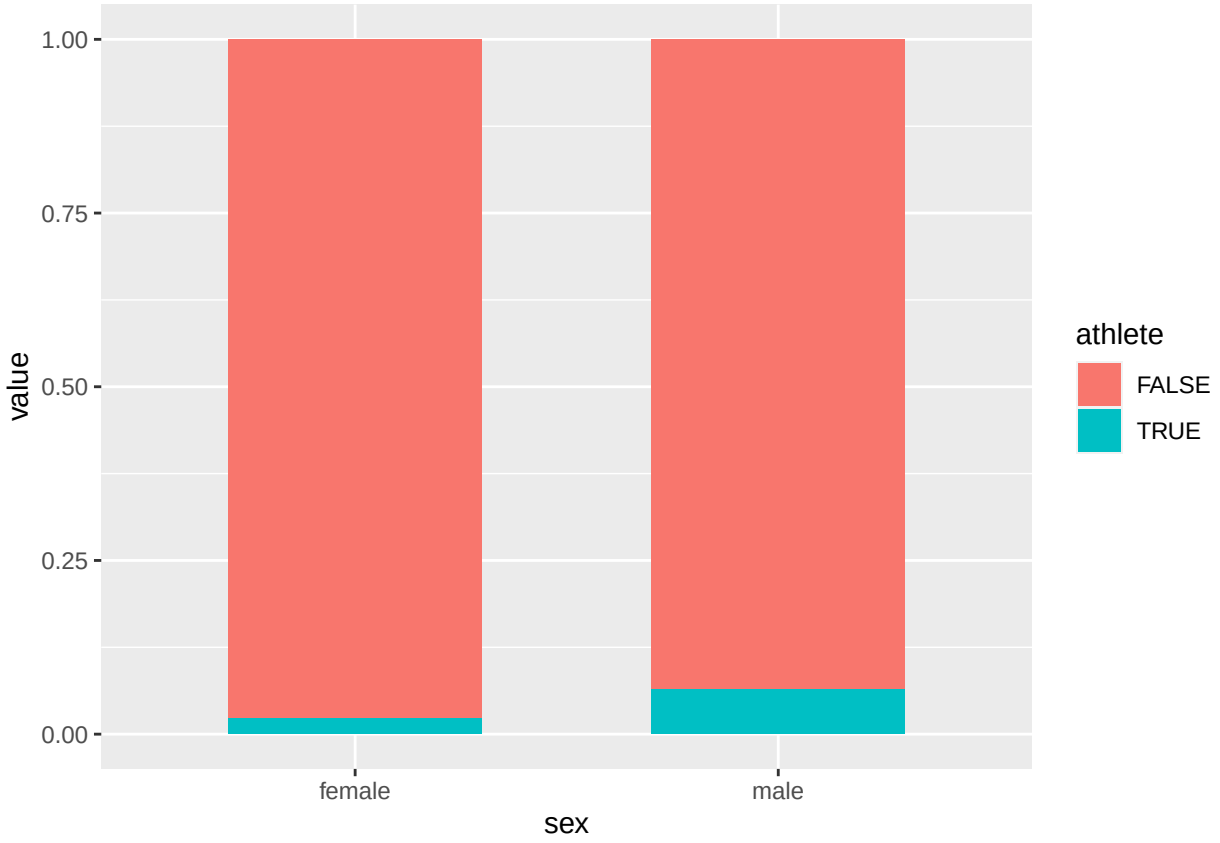
No es pot afirmar que la proporció d'atletes a la població és inferior al 5% amb un nivell de confiança del 95%.

---

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

8.1 Anàlisi 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 recerca

```
## [1] "La proporció d'atletes entre els homes és més gran que entre les dones?"
```

8.3 Hipòtesi nul · la i alternativa

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

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

8.4 Justificació del test a aplicar

Apliquem un contrast sobre la diferència de proporcions, assumint l'aproximació de la distribució binomial a una de normal per a mostres grans. El contrast és unilateral per la dreta.

8.5 Càlculs

```
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
## 6.524634e-02 2.387412e-02 6.247480e+00 1.644854e+00 2.085641e-10
```

```
#Validació 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 = 39.031, df = 1, p-value = 2.086e-10
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.03100639
## sample estimates:
##     prop 1     prop 2
## 0.02387412 0.06524634
```

8.6 Interpretació del test

Podem afirmar que la proporció d'atletes entre els estudiants de sexe masculí és més gran que la proporció d'atletes entre el sexe femení, amb un nivell de confiança del 95%. El valor p és $2.0856413 \times 10^{-10}$. Així mateix, es comprova que el valor crític és 1.6448536 i el valor observat és 6.2474798, estant aquest últim en de la regió de rebuig de la hipòtesi nul · la.

---

9 Resum i conclusions

A la taula següent, es presenten els resultats de cada pregunta de recerca resumits:

N	Pregunta	Resultat (valor observat, crítico, valor p...)	Conclusió
1a	Interval de confiança 95% i 90% sat	ic90=1027.324, 1034.488; ic95=1026.637, 1035.174	-
1b	Interval de confiança 95% i 90% colgpa	ic90=2.638, 2.672; ic95=2.634, 2.675	-
2	Hi ha diferències significatives en la nota 'colgpa' entre atletes i no atletes?	obs=5.873; crit=1.961; pvalue=4.6134206×10^{-9}	Hi ha diferències significatives en la nota 'colgpa' entre atletes i no atletes amb un NC del 95%.
3	Les dones tenen millor nota a "colgpa" que els homes?	obs=7.03; crit=1.645; pvalue=$1.2084693 \times 10^{-12}$	Les dones tenen millor nota a 'colgpa' que els homes amb NC 95%.
4	Hi ha diferències en colgpa entre els estudiants de raça blanca i els de raça negra?	obs=9.62; crit=1.961; pvalue=$1.1309822 \times 10^{-21}$	Hi ha diferències significatives al 95% a la nota colgpa entre estudiants de raça blanca i els de raça negra
5	La proporció d'atletes a la població és inferior a 0.05?	obs=-0.9893554; crit=-1.6448536; pvalue=0.1612447	No es pot afirmar que la proporció d'atletes a la població és inferior al 5% amb un nivell de confiança del 95%.
6	La proporció d'atletes entre els homes és més gran que entre les dones?	obs=6.2474798; crit=1.6448536; pvalue=$2.0856413 \times 10^{-10}$	La proporció d'atletes a la població d'homes és més gran que a les dones amb NC 95%.

10 Resum executiu

S'ha realitzat una anàlisi descriptiva i inferencial del conjunt de dades `gpa`, que conté les dades d'una mostra d'estudiants d'una universitat dels Estats Units. Entre altres variables, el conjunt de dades conté la nota d'accés (sat) i la nota mitjana de cada estudiant en finalitzar primer semestre (golgpa). Les dades contenen el sexe de l'estudiant, la raça i si l'estudiant és atleta.

Segons aquestes dades, s'han obtingut un conjunt de conclusions que es resumeixen a continuació. Totes les conclusions s'extreuen amb un nivell de confiança del 95% i es poden generalitzar a la població d'estudiants dels Estats Units, assumint que la mostra analitzada sigui prou representativa de la població.

- La variable de nota d'accés pren el valor mitjà entre 1026.63 i 1035.17 punts.
- La nota mitjana en finalitzar el primer semestre està entre 2.63 i 2.67.
- S'observen diferències significatives a la nota en finalitzar el primer semestre entre estudiants de raça blanca i estudiants de raça negra.
- Es pot afirmar que les estudiants de sexe femení obtenen millor nota en finalitzar el semestre que els estudiants de sexe masculí.
- Així mateix, s'observen diferències significatives a la nota en finalitzar el primer semestre entre els estudiants atletes i els que no són atletes.
- La proporció d'atletes a la població no és inferior al 5%.
- Es pot afirmar que la proporció d'atletes entre els estudiants de sexe masculí és més gran que entre les estudiants de sexe femení.