Activitat 2: Anàlisi descriptiva i inferencial

Proposta de solució

Semestre 2019.2

Índex

En aquesta activitat usarem el conjunt de dades sobre el pes en néixer que s'ha prepocesado en l'activitat anterior. Una vegada el fitxer estigui preparat, es realitzarà una anàlisi descriptiva i inferencial sobre les dades. Us proporcionem el fitxer **BWprocessed.csv** perquè tots treballeu amb el mateix fitxer de dades, independentment del resultat obtingut en l'activitat 1.

Recordem que el fitxer de dades conté el pes i grandària dels nens i nenes nascuts a Espanya durant l'any 2019. L'arxiu conté 300 registres i 11 variables.

Aquestes variables són: ID, HP, City, Time, Day, BW, BPD, AD, Sex, Ge, Sm, donde:

- Id: identificador numèric.
- HP: nom de l'hospital.
- City: nom de la ciutat on s'ha produït el naixement.
- Time: hora del naixement (valor entre 0 i 24).
- Day: dia de la setmana (valor entre 1 i 7).
- Sex: sexe del nadó.
- BW: pes al naixement.
- BPD: diàmetre biparietal (en mm), determinat per ultrasons, mesurat abans de néixer.
- AD: diàmetre abdominal (en mm), determinat per ultrasons, mesurat abans de néixer.
- Ge: setmanes de gestació.
- Sm: si la mare és fumadora ('S' en cas afirmatiu, 'N' en cas negatiu).

L'anàlisi es desenvoluparà sobre la base de les preguntes de recerca següents:

1. Interval de confiança al 95% del pes en néixer
2. ¿El pes mitjà en néixer difereix significativament de 3.5kg?
3. ¿El pes en néixer de les nenes és diferent al dels nens?
4. ¿La proporció de naixements de nens és inferior als naixements de nenes?
5. ¿Existeix una relació entre baix pes en néixer i el fet que la mare sigui fumadora?

Notes importants a tenir en compte per lliurar l'activitat:

- És necessari 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 mateix (pas a pas).
- S'ha de respectar la mateixa numeració dels apartats que l'enunciat.
- No es poden realitzar llistats complets de les dades en la solució. El motiu és que es generen fitxers de sortida amb centenars de pàgines que són molt difícil de traçar i corregir. Per comprovar les funcionalitats del codi, podeu usar **head** i **tail** que només mostren unes línies del fitxer.
- El nivell de confiança per defecte és del 95%, tret que s'indiqui un altre valor.

---

1 Anàlisi descriptiva

1.1 Lectura del fitxer

Leer el fitxer **BWprocessed.csv**. Validar que els tipus de dades llegides són correctes. Si no és així, realitzar les conversions oportunes.

```
dim(ds)
```

```
## [1] 300  11
```

```
summary(ds)
```

```
##        ID                                    HP             City
##  Min.   :  1.00   H.U.de Bellvitge           : 40   Barcelona:127
##  1st Qu.: 75.75   H.U.Quirón Dexeus          : 36   Granada  : 30
##  Median :150.50   H.U.Politècnic La Fe       : 31   Madrid   : 71
##  Mean   :150.50   H.M.I. Virgen de las Nieves: 30   Pamplona : 27
##  3rd Qu.:225.25   C.U.Navarra                : 27   Sevilla  : 14
##  Max.   :300.00   H.G.U.Gregorio Marañón     : 27   Valencia : 31
##                   (Other)                    :109
##       Time            Day           BW            BPD               AD
##  00:00  : 12   Domingo  :26   Min.   :1025   Min.   : 64.00   Min.   : 71.00
##  09:21  :  4   Jueves   :41   1st Qu.:2124   1st Qu.: 86.75   1st Qu.: 93.75
##  06:50  :  3   Lunes    :33   Median :2680   Median : 91.00   Median :102.00
##  07:16  :  3   Martes   :43   Mean   :2661   Mean   : 88.94   Mean   :100.97
##  07:24  :  3   Miércoles:45   3rd Qu.:3100   3rd Qu.: 93.00   3rd Qu.:108.00
##  09:34  :  3   Sábado   :53   Max.   :4850   Max.   :100.00   Max.   :133.00
##  (Other):272   Viernes  :59
##  Sex           Ge        Sm
##  F:168   Min.   :33.00   N:238
##  M:132   1st Qu.:33.00   S: 62
##          Median :36.00
##          Mean   :36.24
##          3rd Qu.:38.00
##          Max.   :43.00
##
```

```
str(ds)
```

```
## 'data.frame':    300 obs. of  11 variables:
##  $ ID  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ HP  : Factor w/ 11 levels "C.U.Navarra",..: 3 6 6 2 5 3 4 1 7 1 ...
##  $ City: Factor w/ 6 levels "Barcelona","Granada",..: 3 1 1 1 3 3 2 4 3 4 ...
##  $ Time: Factor w/ 252 levels "00:00","01:52",..: 177 125 1 236 100 21 89 32 61 161 ...
##  $ Day : Factor w/ 7 levels "Domingo","Jueves",..: 7 3 4 2 5 6 5 7 5 7 ...
##  $ BW  : num  2200 4100 4200 1300 1150 3400 2500 3100 2550 2600 ...
##  $ BPD : int  82 97 97 74 68 94 87 92 86 93 ...
##  $ AD  : int  90 129 133 71 80 110 105 109 106 102 ...
##  $ Sex : Factor w/ 2 levels "F","M": 2 2 1 2 1 1 1 1 1 1 ...
##  $ Ge  : int  33 43 43 33 33 39 34 37 34 34 ...
##  $ Sm  : Factor w/ 2 levels "N","S": 1 1 1 2 2 1 1 1 1 1 ...
```

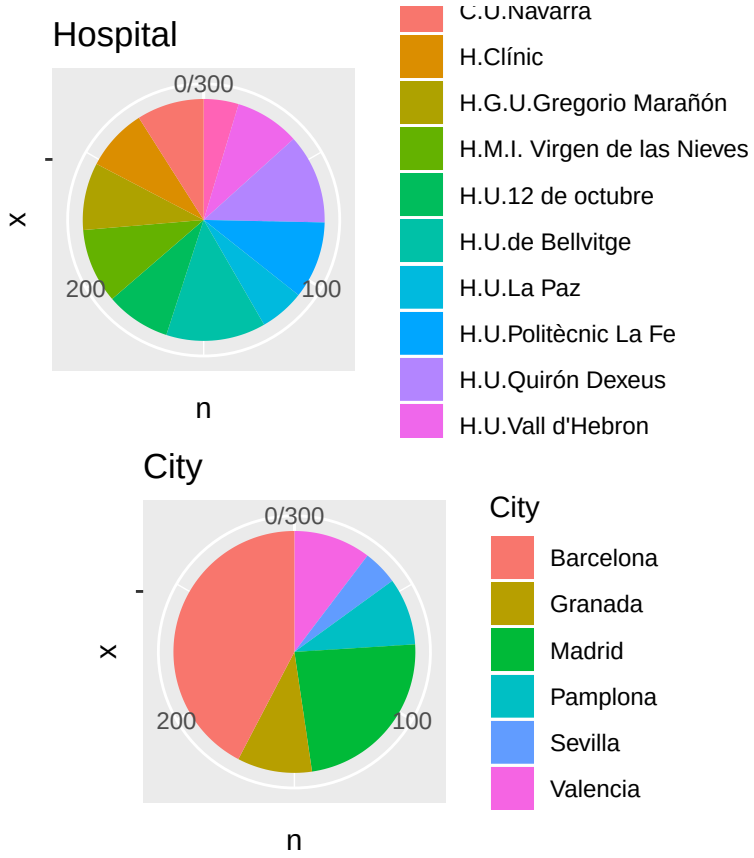
1.2 Anàlisi descriptiva visual

Representar de manera visual les variables del conjunt de dades i les distribucions dels seus valors. Trieu la representació més apropiada en cada cas.

```
hp <- summarize( group_by(ds, HP), n=length(HP))
g1 <- ggplot( hp, aes(x="", y=n, fill=HP)) +
  geom_bar(width = 1, stat = "identity") +
    coord_polar("y", start=0) + ggtitle("Hospital")

cit <- summarize( group_by(ds, City), n=length(City))
g2 <- ggplot( cit, aes(x="", y=n, fill=City)) +
  geom_bar(width = 1, stat = "identity") +
    coord_polar("y", start=0) + ggtitle("City")

sex <- summarize( group_by(ds,Sex), n=length(Sex))
g3 <- ggplot( sex, aes(x="", y=n, fill=Sex)) +
  geom_bar(width = 1, stat = "identity") +
    coord_polar("y", start=0) + ggtitle("Sex")

smm <- summarize( group_by(ds,Sm), n=length(Sm))
smm
```

```
## # A tibble: 2 x 2
##   Sm        n
##   <fct> <int>
## 1 N       238
## 2 S        62
```

```
g4 <- ggplot( smm, aes(x="", y=n, fill=Sm)) +
  geom_bar(width = 1, stat = "identity") +
    coord_polar("y", start=0) + ggtitle("Smoking")

grid.arrange(g1, g2,  ncol=1)
```




```
grid.arrange(g3, g4, ncol=2)
```

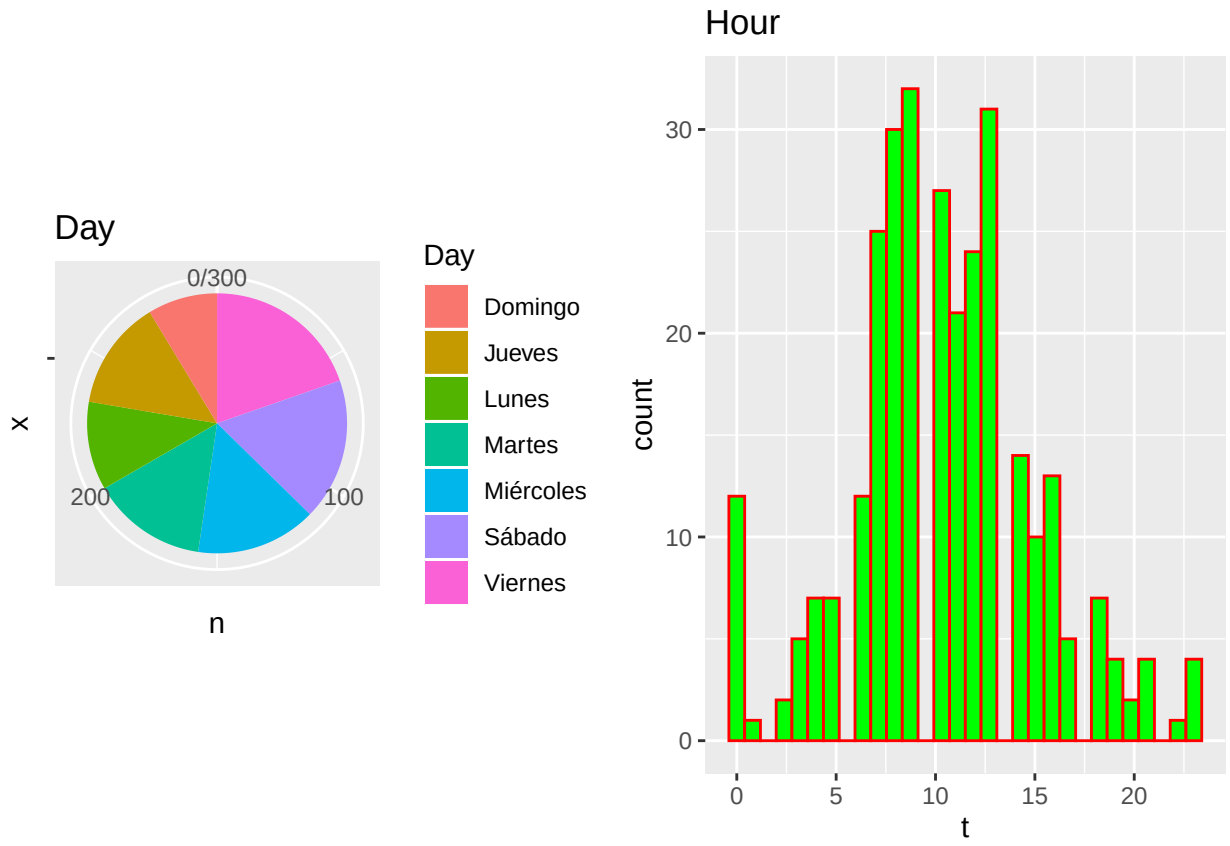

```
#Day and Time
day <- summarize( group_by(ds,Day), n=length(Day))
gDay <- ggplot( day, aes(x="", y=n, fill=Day)) +
  geom_bar(width = 1, stat = "identity") +
    coord_polar("y", start=0) + ggtitle("Day")

summary(ds$Time)
```

```
##   00:00   09:21   06:50   07:16   07:24   09:34   13:10   07:19   07:41   08:00
##      12       4       3       3       3       3       3       2       2       2
##   08:25   08:48   08:57   09:11   09:39   09:48   10:24   10:27   10:33   10:54
##       2       2       2       2       2       2       2       2       2       2
##   10:56   11:00   11:56   12:13   12:43   12:58   13:12   13:34   13:46   14:00
##       2       2       2       2       2       2       2       2       2       2
##   14:48   01:52   02:36   02:38   03:15   03:25   03:39   03:43   03:59   04:00
##       2       1       1       1       1       1       1       1       1       1
##   04:03   04:12   04:15   04:36   04:45   04:48   05:23   05:34   05:36   05:39
##       1       1       1       1       1       1       1       1       1       1
##   05:40   05:41   05:51   06:15   06:16   06:25   06:33   06:37   06:39   06:43
##       1       1       1       1       1       1       1       1       1       1
##   06:51   06:55   07:01   07:06   07:08   07:14   07:20   07:23   07:27   07:34
##       1       1       1       1       1       1       1       1       1       1
##   07:39   07:43   07:45   07:46   07:49   07:56   07:58   08:01   08:04   08:14
##       1       1       1       1       1       1       1       1       1       1
##   08:15   08:21   08:22   08:24   08:28   08:30   08:31   08:37   08:40   08:41
##       1       1       1       1       1       1       1       1       1       1
```

```
##    08:42    08:44    08:46    08:49    08:51    08:54    08:55    08:56    08:58 (Other)
##        1        1        1        1        1        1        1        1        1      153
```

```
ds$t<- as.numeric( str_extract( ds$Time, "\\d+")  )
gTime<-ggplot( ds, aes(t)) +
      geom_histogram(fill="green",col="red") +
      labs( title="Hour")

grid.arrange(gDay, gTime, ncol=2)
```

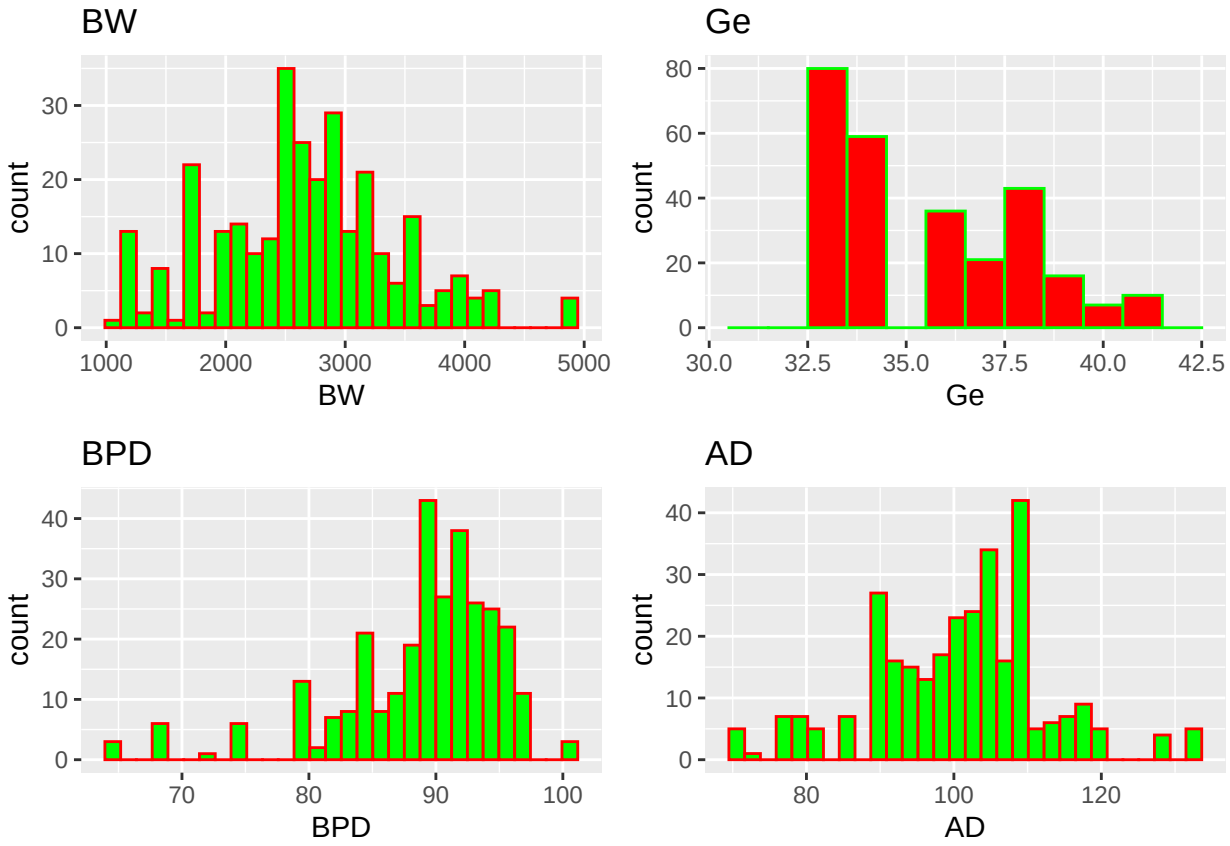
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Quantitative vars
g6<-ggplot( ds, aes(BW)) +
   geom_histogram(fill="green",col="red") +
      labs( title="BW")

g7<-ggplot( ds, aes(Ge)) +
   geom_histogram(fill="red",col="green", breaks=seq(30.5,42.5,by=1)) +
   labs( title="Ge")

g8<-ggplot( ds, aes(BPD)) +
   geom_histogram(fill="green",col="red") +
      labs( title="BPD")

g9<-ggplot( ds, aes(AD)) +
   geom_histogram(fill="green",col="red") +
```

```
      labs( title="AD")

grid.arrange(g6, g7, g8, g9, ncol=2)
```

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



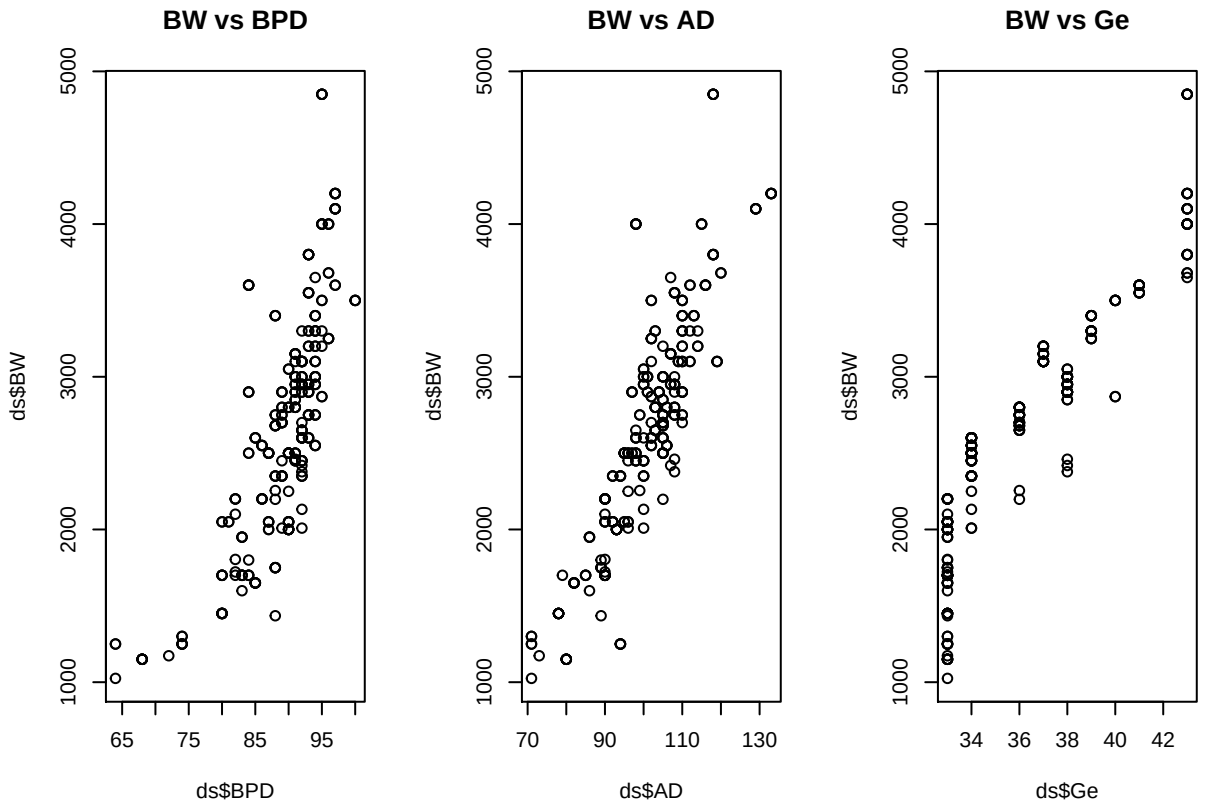
1.3 Correlació

Estudiar visualment les possibles correlacions entre:

- les variables pes en néixer i el diàmetre biparietal
- les variables pes i diàmetre abdominal
- les variables pes i setmanes de gestació.

Interpreteu els gràfics. Per avaluar la correlació numèricament, podeu usar la funció **cor**.

```
par(mfrow=c(1,3))
plot( ds$BW ~ ds$BPD, main="BW vs BPD" )
plot( ds$BW ~ ds$AD, main="BW vs AD"  )
plot( ds$BW ~ ds$Ge, main="BW vs Ge"  )
```




```
par(mfrow=c(1,1))
cor( ds[,c("BW","BPD","AD","Ge")])
```

```
##            BW       BPD        AD        Ge
## BW  1.0000000 0.7749223 0.8752187 0.9138659
## BPD 0.7749223 1.0000000 0.7357070 0.6181404
## AD  0.8752187 0.7357070 1.0000000 0.7823264
## Ge  0.9138659 0.6181404 0.7823264 1.0000000
```

Interpretació: Visualment, s'observa com existeix una relació, gairebé lineal, entre les variables pes, diàmetre biparietal i diàmetre abdominal. A major pes, major diàmetre biparietal i també major diàmetre abdominal. Existeix també relació entre les setmanes de gestació i el pes en néixer, encara que s'observa que la relació no és lineal i que, per una setmana de gestació determinada, existeix una certa variabilitat en el pes en néixer. Per tant, les setmanes de gestació no determinen exactament el pes en néixer del nen, la qual cosa és indicatiu que existeixen altres variables que poden influir.

Pel que fa a la relació entre diàmetre biparietal i abdominal amb el pes, la relació observada és útil per predir el pes en néixer (ja que com es va indicar en A1, diàmetre biparietal i abdominal es van mesurar abans de néixer).

---

2 Pes mitjà en néixer

2.1 Interval de confiança

Es desitja investigar l'interval de confiança del 95% del pes mitjà en néixer a partir de la mostra de dades. S'aconsella definir una funció IC que calculi l'interval de confiança d'una variable donada. Aquesta funció es podrà aprofitar més endavant.

2.1.1 Càlculs

Nota: heu de realitzar els càlculs manualment. No es poden usar funcions com a **t.test** o similars que ja calculin directament l'interval de confiança. En canvi, sí podeu fer servir funcions com **qnorm**, **pnorm**, **qt**, **pt** per conèixer els valors de la distribució normal o t-Student.

```
IC <- function( x, alfa=0.05 ){
  #Error típico
  n <- length(x)
  errorTipico <- sd(x) / sqrt( n )
  errorTipico

  t<-qnorm( 1-alfa/2 )
  t

  error<- t * errorTipico
  error

  return ( c( mean(x) - error, mean(x) + error ))
}

my.ic <- IC( ds$BW, alfa=0.05 )

#Comprovació
t.test(ds$BW)
```

```
##
##  One Sample t-test
##
## data:  ds$BW
## t = 60.567, df = 299, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2574.339 2747.245
## sample estimates:
## mean of x
##  2660.792
```

2.1.2 Interpretació

Interpreteu el resultat obtingut i el significat d'interval de confiança.

Interpretació: L'interval de confiança del 95% del pes en néixer és (2574.6884758, 2746.8955242). El significat d'aquest interval és el següent: si es prenen diferents mostres de la població, el 95% dels intervals calculats contenen el valor del paràmetre peso mitjà.

2.2 Contrast del valor mitjà amb 3.5kg

Independentment del resultat de l'apartat anterior, passem a desenvolupar el procediment per donar resposta a la pregunta següent:

¿El pes mitjà en néixer dels nens i nenes nascuts a Espanya és de 3.5kg?

Responeu a la pregunta amb un 95% de nivell de confiança. Seguiu els passos que s'indiquen a continuació. **Nota:** S'ha de calcular a partir de tota la mostra. No es demana el càlcul per separat de nens i nenes, sinó de totes les dades en global.

2.2.1 Escriviu la hipòtesi nul · la i alternativa.

$H0 : \mu = \mu_0$

$H1 : \mu \neq \mu_0$

donde $\mu_0 = 3500$.

2.2.2 Mètode

Detalleu quin mètode usareu per respondre aquesta pregunta.

Resposta: Apliquem el contrast d'una mostra sobre la mitjana (bilateral). Atès que el contrast és sobre la mitjana i que la grandària de la mostra és prou gran, es compleix el teorema del límit central i per tant, assumim normalitat.

2.2.3 Càlculs

Desenvolupeu tots els càlculs. No podeu usar funcions del tipus *t.test*. Heu de fer els càlculs manualment.

```
n <- nrow( ds )
alfa <- 0.05

tobs <- (mean(ds$BW)-3500) / (sd(ds$BW)/sqrt(n))

tcritical <- qnorm( alfa/2 )
pvalue <- pnorm( tobs ) * 2

data.frame(tcritical, tobs, pvalue) %>% kable()
```

tcritical	tobs	pvalue
-1.959964	-19.10279	0

```
#Comprovació
t.test( ds$BW, alternative="two.sided", mu=3500)
```

```
##
##  One Sample t-test
##
## data:  ds$BW
## t = -19.103, df = 299, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 3500
## 95 percent confidence interval:
##  2574.339 2747.245
## sample estimates:
## mean of x
##  2660.792
```

2.2.4 Conclusió

A partir dels valors obtinguts sobre el valor crític, el valor observat i el valor p, responeu la pregunta plantejada.

Resposta: el valor crític per a α=0.05 és -1.959964 i el valor observat és -19.102789. La regió d'acceptació de H_0 és (-1.959964, 1.959964). Per tant, el valor observat es troba fora de la regió d'acceptació, amb la qual cosa es rebutja la hipòtesi nul · la, concloent que el pes medio no és igual a 3.5kg. S'arriba a la mateixa conclusió amb el valor p que és igual a $2.3935344 \times 10^{-81}$, molt inferior α.

2.2.5 Relació amb l'interval de confiança.

A partir del resultat sobre l'interval de confiança obtingut en l'apartat anterior, com habriáis pogut respondre a aquesta mateixa pregunta?

Resposta: l'interval de confiança calculat anteriorment (2574.6884758, 2746.8955242) no conté el valor medio 3.5kg. Per tant, s'arriba a la mateixa conclusió.

---

3 Contrast de pes mitjà entre nens i nenes

Es desitja investigar si el pes en néixer de les nenes és diferent al dels nens. Per això, heu de realitzar un contrast sobre la mitjana del pes, considerant dues mostres independents. Seguiu els passos que s'especifiquen a continuació.

3.1 Hipòtesi

Escriviu la hipòtesi nul · la i la hipòtesi alternativa

$H0 : \mu_1 = \mu_2$

$H1 : \mu_1 \neq \mu_2$

on μ_1 és el valor mitjà del pes dels nens i μ_2 el valor mitjà del pes de les nenes.

3.2 Assumpció de normalitat

Justificar si es pot assumir normalitat en aquest cas.

Resposta: Es pot aplicar el teorema del límit central pel qual la distribució de la mitjana mostral d'una mostra prou gran és aproximadament normal. A més, la mitjana serà la mateixa que la variable d'interès i la desviació típica de la mitjana mostral serà aproximadament l'error estàndard.

3.3 Mètode

Explicar quin mètode usareu per al contrast a partir de l'anàlisi prèvia sobre l'assumpció de normalitat. Indiqueu també si es tracta d'un contrast unilateral o bilateral.

Resposta: Contrast de dues mostres independents sobre la mitjana. El contrast és bilateral.

3.4 Càlculs

Realitzeu tots els càlculs del contrast. No es pot usar funcions del tipus *t.test*. Heu de realitzar els càlculs manualment, seguint les fórmules que s'especifiquen en els materials. Sí que podeu usar funcions del tipus **qnorm**, **pnorm**, **pt**, **qt**, les quals codifiquen els valors de les distribucions de dades normals i t-Student.

```
myttest <- function( x1, x2, CL=0.95,equalvar=TRUE, alternative="bilateral" ){ #z test
  mean1<-mean(x1)
  n1<-length(x1)
  sd1<-sd(x1)
  mean2<-mean(x2)
```

```
  n2<-length(x2)
  sd2<-sd(x2)
  if (equalvar==TRUE){
    s <-sqrt( ((n1-1)*sd1^2 + (n2-1)*sd2^2 )/(n1+n2-2) )
    Sb <- s*sqrt(1/n1 + 1/n2)
    df<-n1+n2-2
  }
  else{ #equalvar==FALSE
    Sb <- sqrt( sd1^2/n1 + sd2^2/n2 )
    denom <- ( (sd1^2/n1)^2/(n1-1) + (sd2^2/n2)^2/(n2-2))
    df <- ((sd1^2/n1 + sd2^2/n2)^2) / denom
  }
  alfa <- (1-CL)
  t<- (mean1-mean2) / Sb
  if (alternative=="bilateral"){
     tcritical <- qnorm( alfa/2, lower.tail=FALSE ) #two sided
     pvalue<-pnorm( abs(t), lower.tail=FALSE )*2 #two sided
   }
   else if (alternative=="less"){
      tcritical <- qnorm( alfa, df, lower.tail=TRUE )
      pvalue<-pnorm( t, df, lower.tail=TRUE )
   }
   else{ #(alternative=="greater")
      tcritical <- qnorm( alfa, lower.tail=FALSE )
      pvalue<-pnorm( t, lower.tail=FALSE )
   }
  #Guardem en resultat en un data frame
      info<-data.frame(t,tcritical,pvalue)
      info %>% kable() %>% kable_styling()
      return (info)
}

info<-myttest( ds$BW[ds$Sex=="M"], ds$BW[ds$Sex=="F"], alternative="bilateral" )
info
```

```
##            t tcritical   pvalue
## 1 -0.8837318  1.959964 0.376841
```

```
info %>% kable()  # %>% kable_styling()
```

t	tcritical	pvalue
-0.8837318	1.959964	0.376841

```
#Comprovació
t.test( ds$BW[ds$Sex=="M"], ds$BW[ds$Sex=="F"], alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  ds$BW[ds$Sex == "M"] and ds$BW[ds$Sex == "F"]
## t = -0.89029, df = 288.53, p-value = 0.3741
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -251.21208   94.73091
## sample estimates:
## mean of x mean of y
```

```
##  2616.977  2695.218
```

3.5 Interpreteu els resultats

Sobre la base del valor p, valor observat i valor crític obtinguts en l'apartat anterior, donar resposta a la pregunta plantejada.

Resposta: el valor crític per a α=0.05 és 1.959964 i el valor observat és -0.8837318.La regió d'acceptació de H_0 és (-1.959964, 1.959964). Per tant, el valor observat es troba dins de la regió d'acceptació, amb la qual cosa no es rebutja la hipòtesi nul · la, concloent que el pes mitjà dels nens no és diferent significativament al de les nenes. S'arriba a la mateixa conclusió amb el valor p que és igual a 0.376841, sent no inferior a α.

3.6 Interval de confiança

Una altra manera d'avaluar si existeixen diferències del pes mitjà entre nens i nenes és a partir dels respectius intervals de confiança. Calcular aquests intervals de confiança i extreure les conclusions i com aquestes corroboren els resultats del contrast realitzat.

```
  #S'usa la funció IC definida anteriorment.
  IC( ds$BW[ds$Sex=="M"], alfa=0.05 )
```

```
## [1] 2491.638 2742.317
```

```
  IC( ds$BW[ds$Sex=="F"], alfa=0.05 )
```

```
## [1] 2577.071 2813.364
```

Interpretació: Els intervals de confiança al 95% del pes mitjà en néixer de nens i nenes se solapen. Per tant, no podem dir que el pes mitjà en néixer sigui significativament diferent entre nens i nenes.

---

4 Proporció de nens i nenes

A continuació, es planteja la pregunta següent:

> ¿La proporció de naixements de nens és inferior a la dels naixements de nenes?

Per respondre a aquesta pregunta, seguiu els passos que s'especifiquen a continuació.

4.1 Hipòtesi

Escriviu la hipòtesi nul · la i la hipòtesi alternativa.

$H0 : p = p_0$

$H1 : p < p0$

on p_0 és igual a 0.5.

4.2 Mètode

Especificar quin mètode usareu per a aquest contrast. També indiqueu si el contrast és unilateral o bilateral.

Resposta: Es tracta d'un contrast sobre la proporció per mostres grans (veure mòdul "Contrasts d'hipòtesis").

4.3 Càlculs

Realitzeu tots els càlculs manualment. Igual que en els apartats anteriors, només podeu usar les funcions **qnorm**, **pnorm**, **qt**, **pt**.

```
  #Freqüència observada
  n <- nrow( ds )
  p = (sum( ds$Sex=="M") )/ n
  p
```

```
## [1] 0.44
```

```
  p0 <- 0.5

  #Estadístic de contrast
  pobs <- (p - p0)/ sqrt( p0*(1-p0)/n)

  #L'estadístic segueix una distribució normal estàndard baix H0
  z <- qnorm( 0.05, lower.tail=TRUE )
  pvalue <- pnorm( pobs, lower.tail=TRUE )
  cat("p=", p, "pobs=", pobs, "zcritical=", z, "pvalue=", pvalue)
```

```
## p= 0.44 pobs= -2.078461 zcritical= -1.644854 pvalue= 0.01883346
```

4.4 Interpretació

A partir del resultat de l'estadístic de contrast observat, el valor crític i el valor p, concloeu sobre si es pot afirmar que la proporció de nens que neixen és inferior a la de les nenes.

Resposta: el valor crític per a α=0.05 és -1.6448536 i el valor observat és -2.078461. Per tant, el valor observat es troba fora de la regió d'acceptació, amb la qual es rebutja la hipòtesi nul · la, concloent que la proporció de nens nascuts és inferior a la de les nenes. S'arriba a la mateixa conclusió amb el valor p que és igual a 0.0188335, sent aquest inferior α.

---

5 Relació entre baix pes i mare fumadora

A continuació, se'ns planteja la següent pregunta:

> ¿Existeix relació entre baix pes en néixer i el fet que la mare sigui fumadora?

Per investigar la relació entre baix pes en néixer i ser fill de mare fumadora, es realitzen les anàlisis següents.

5.1 Anàlisi sobre els casos de baix pes

En primer lloc, s'etiqueten els casos de baix pes en néixer com aquells que són inferiors a 2.5kg. Es construeix una taula de contingència on s'avalua la relació entre baix pes en néixer i ser fill de mare fumadora. Una vegada construïda la taula, s'ha d'aplicar el test d'independència de dues variables Chi quadrat (test ChiSquare). Realitzeu tots els càlculs d'aquest test i després, interpreteu si podem observar una relació entre baix pes i ser fill de mare fumadora amb un nivell de confiança del 95%.

Nota: per calcular el test d'independència de dues variables, no podeu usar la funció **chisq.test** per resoldre l'exercici. Heu de realitzar els càlculs manualment. Només podeu usar **chisq.test** per validar els vostres resultats. Per accedir als valors de la funció de distribució Chi Quadrat, podeu usar les funcions **dchisq**, **pchisq**, **qchisq** de la llibreria **stats**.

Podeu consultar el test d'independència de dues variables ChiSquare en textos d'estadística. Adjuntem alguns enllaços que poden ser útils:

- https://stattrek.com/chi-square-test/independence.aspx
- https://www.spss-tutorials.com/chi-square-independence-test/

Seguiu els passos que s'especifiquen a continuació.

5.1.1 Hipòtesi

Escriviu la hipòtesi nul · la i alternativa.

H_0: les variables sota pes en néixer i mare fumadora són independents.

H_1: existeix una relació de dependència entre les variables.

5.1.2 Càlculs

Realitzeu els càlculs manualment del test. Per això, es recomana construir una funció **my.chisq** que realitzi aquests càlculs i que pugueu reusar més endavant.

```
my.chisq <- function( x, y ){
  n00 <- sum( (x==0) & (y==0) )
  n11 <- sum( (x==1) & (y==1) )
  n01 <- sum( (x==0) & (y==1) )
  n10 <- sum( (x==1) & (y==0) )

  cat("Table values:", n00, " ", n01, " ", n10, " ", n11, "\n")

  nx0 <- sum( x==0)
  nx1 <- sum( x==1)
  ny0 <- sum( y==0)
  ny1 <- sum( y==1)

  cat(nx0, " ", nx1, " ", ny0, " ", ny1, "\n")
  total = sum(nx0, nx1)

  e00 <- (nx0*ny0) / total
  e01 <- (nx0*ny1) / total
  e10 <- (nx1*ny0) / total
  e11 <- (nx1*ny1) / total

  cat("Expected values:",  e00, " ", e01, " ", e10, " ", e11, "\n")

  chisq <- (n00-e00)^2 / e00 + (n01-e01)^2/e01 + (n10-e10)^2/e10 + (n11-e11)^2/e11
  df <- 1
  cat("chisq: ", chisq)


  return (chisq)

}

#Preparació de les dades: selecció dels casos amb baix pes en néixer
alfa <- 0.05
ds$lower.BW <- (ds$BW < 2500)
```

```r
  #Es crea un data frame amb dues columnes: baix pes i mare fumadora
  data <- ds[, c("lower.BW" , "Sm")]
  #Es codifiquen els valors com 0 i 1.
  data$lower.BW <- ifelse( data$lower.BW==TRUE, 1, 0)
  data$Sm       <- ifelse( data$Sm=='S', 1, 0)

  #La funció chisq retorna el valor observat
  chisq <- my.chisq( data$lower.BW, data$Sm )
```

```
## Table values: 191   2   47   60
## 193   107   238   62
## Expected values: 153.1133   39.88667   84.88667   22.11333
## chisq:  127.1823
```

```r
  #Càlcul de valor crític i valor p
  criticalv <- qchisq( alfa, df=1, lower.tail=FALSE )
  pvalue <- pchisq( chisq, df=1,lower.tail=FALSE )
  data.frame(chisq, criticalv, pvalue) %>% kable()
```

chisq	criticalv	pvalue
127.1823	3.841459	0

```r
  #Comprovació usant la funció R chisq.test
  table <- table( data$lower.BW, data$Sm)
  table
```

```
##
##       0   1
##   0 191   2
##   1  47  60
```

```r
  chisq.test(table,correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 127.18, df = 1, p-value < 2.2e-16
```

5.1.3 Conclusió del test

Sobre la base dels resultats del test, es pot afirmar que existeix relació entre baix pes i mare fumadora? Expliqueu sobre la base del valor p, valor observat i valor crític.

Resposta: El valor p és igual a $1.6946557 \times 10^{-29}$ amb la qual cosa rebutgem la hipòtesi nul · la i concloem que existeix relació entre mare fumadora i el baix pes. S'arriba a la mateixa conclusió a partir del valor crític i el valor observat.

5.1.4 Interpretació

Analitzeu si existeixen altres variables que puguin influir en l'anàlisi i com afectarien les conclusions extretes del test.

Resposta: Tal com es va avaluar al principi d'aquesta activitat, les setmanes de gestació influeixen en el pes en néixer. Es va veure que existeix una correlació entre setmanes de gestació i pes. La relació té una certa variabilitat i una part d'aquesta variabilitat pot ser explicada pel fet que la mare sigui fumadora. La variable sexe del bebè sembla no ser un factor influent en el pes.

5.2 Anàlisi dels casos de baix pes (amb gestació >36)

Realitzeu la mateixa anàlisi per als naixements a partir de la setmana 36. Per realitzar aquesta anàlisi, es considerarà baix pes en néixer els casos que es trobin en el primer quartil del pes, considerant només els casos amb setmanes de gestació superiors a 36.

Podem concloure que es presenta el mateix patró de relació (d'independència o dependència entre les variables) que en el cas anterior? Interpreteu els resultats.

5.2.1 Hipòtesi

Escriviu la hipòtesi nul · la i alternativa.

H_0: les variables sota pes en néixer i mare fumadora (en gestacions superiors a 36 setmanes) són independents.

H_1: existeix una relació de dependència entre les variables

5.2.2 Càlculs

Realitzeu els càlculs. Podeu aprofitar la funció desenvolupada anteriorment. Mostreu els resultats dels càlculs.

```
  #Selecció de les dades: gestació superior a 36 i pes en el primer quartil
  ds.sel <- ds[ds$Ge>36,]
  fivenum( ds.sel$BW )
```

```
## [1] 2378 2950 3200 3600 4850
```

```
  threshold <-   fivenum( ds.sel$BW )[2]
  ds.sel$lower.BW <- (ds.sel$BW < threshold)
  data <- ds.sel[, c("lower.BW" , "Sm")]
  data$lower.BW <- ifelse( data$lower.BW==TRUE, 1, 0)
  data$Sm       <- ifelse( data$Sm=='S', 1, 0)

  #Càlcul test chisq : valor observat
  alfa <- 0.05
  chisq <- my.chisq( data$lower.BW, data$Sm )
```

```
## Table values: 104   0   16   5
## 104   21   120   5
## Expected values: 99.84   4.16   20.16   0.84
## chisq:  25.79365
```

```
  #Càlcul de valor crític i valor p
  criticalv <- qchisq( alfa, df=1, lower.tail=FALSE )
  pvalue <- pchisq( chisq, df=1,lower.tail=FALSE )

  data.frame(chisq, criticalv, pvalue) %>% kable()
```

chisq	criticalv	pvalue
25.79365	3.841459	4e-07

```
  #
  table <- table( data$lower.BW, data$Sm)
  table
```

```
##
##       0   1
##   0 104   0
##   1  16   5
```

```
  chisq.test(table, correct=FALSE)
```

```
## Warning in chisq.test(table, correct = FALSE): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 25.794, df = 1, p-value = 3.799e-07
```

```
  chisq.test(table, correct=TRUE)
```

```
## Warning in chisq.test(table, correct = TRUE): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table
## X-squared = 19.966, df = 1, p-value = 7.884e-06
```

5.2.3 Conclusió

Interpreteu els resultats obtinguts i responeu a la pregunta plantejada. Justifiqueu les vostres conclusions a partir del valor observat, valor crític i valor p.

Resposta: En els casos en què les setmanes de gestació són superiors a 36, és a dir, amb la gestació completa, el baix pes en néixer també està relacionat amb mares fumadores. No obstant això, es produeix en pocs casos i l'aproximació ChiSquare pot ser poc precisa.

6 Conclusions de l'anàlisi

Finalment, a partir de tot l'estudi realitzat, escriviu les conclusions d'aquest estudi, donant resposta ben fonamentada a les preguntes plantejades a l'inici de l'activitat.

Resposta: En aquest estudi s'ha investigat el pes en néixer dels nens i nenes nascuts a Espanya en 2019, a partir d'una mostra de 300 individus. Les conclusions obtingudes són:

- El pes en néixer és diferent de 3.5kg.
- No existeixen diferències significatives entre el pes de nens i nenes.
- Existeix una relació lineal positiva entre les dimensions abans de néixer (diàmetre biparietal i diàmetre abdominal) i el pes en néixer. Es podrien usar aquestes variables prenatals per estimar el pes en néixer.
- Existeix una correlació positiva entre les setmanes de gestació i el pes en néixer.
- Ser mare fumadora és un factor influent que explica el baix pes en néixer, tant si es tenen en compte tots els naixements com aquells que es produeixen amb la gestació completa.
- La proporció de nens en néixer és inferior al de les nenes.

7 Puntuació dels apartats

- Apartat 1 (15%)
- Apartat 2 (20%)
- Apartat 3 (15%)

- Apartat 4 (10%)
- Apartat 5 (20%)
- Apartat 6 (10%)
- Qualitat de l'informe dinàmic i del codi R (10%)