# Activitat 2: Anàlisi descriptiva i inferencial

Marc Cervera Rosell

03/04/2024

```
setRepositories(ind=2) # Per descarregar paquets de CRAN
```

## 1 Estadística descriptiva

En primer lloc, fem una anàlisi descriptiva d'algunes variables d'interès i la relació amb readingScore. Seguiu els passos que s'indiquen a continuació.

### 1.1 Distribució de variables

En primer lloc, mostreu visualment la distribució de gènere a la població, així com la proporció dels estudiants que parlen anglès a casa en relació als que no parlen anglès. Mostreu un gràfic per cada cas.

Abans de realitzar cap anàlisi cal llegir el fitxer. L'operació de lectura es realitzarà dins d'un bloc *tryCatch()*, així si no és possible realitzar la lectura es llançarà un error.

```
tryCatch({
  data <- read.csv("pisa_clean.csv", header = TRUE)
  print("Fitxer llegit correctament")
}, error = function(e){
  cat("ERROR en el moment de llegir el fitxer:",conditionMessage(e), "\n")
})
```

```
## [1] "Fitxer llegit correctament"
```

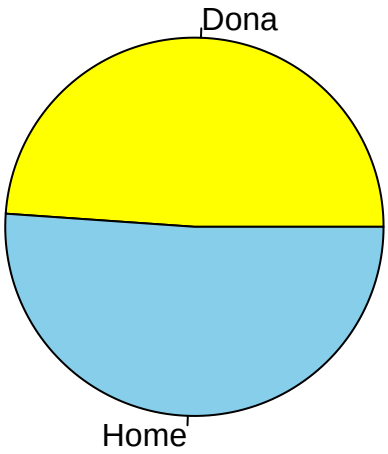
Per a mostrar la distribució de gènere de la població de l'informe PISA s'escull fer-ho mitjançant un diagrama de pastís. Abans de mostrar cap gràfic, però cal seleccionar les dades i calcular les freqüències de cada valor de la variable d'estudi. La funció *table()*, és una funció que precisament acompleix aquesta tasca, és a dir, crea una taula de freqüència de la variable. Bàsicament, compta el nombre d'ocurrències de cada valor únic i organitza la informació en una taula.

```
gender <- table(data$male)
cat("Hi ha un total de",gender["0"], "dones i un total de",gender["1"], "homes")
```

```
## Hi ha un total de 1791 dones i un total de 1872 homes
```

Com es pot observar, hi ha un nombre superior d'homes. Per tant, en el diagrama que mostrarà la distribució del gènere hi haurà una mica més de predominança de la part masculina.

```
pie(gender, main = "Distribució de gènere", col = c("yellow", "skyblue"),
    labels = c("Dona", "Home"))
```

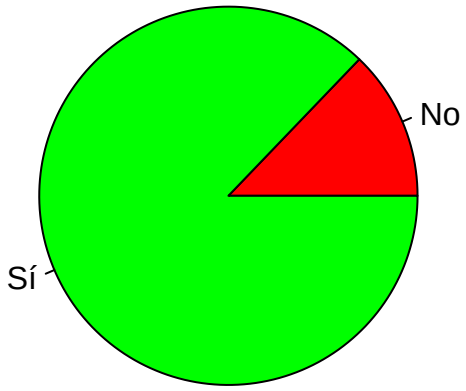



Per mostrar visualment la proporció d'alumnes que parlen anglès a casa i els que no, s'escull el mateix tipus de diagrama. Per mostrar aquesta proporció, es repeteix el procés anterior. En aquest cas, s'ha descartat l'ús d'un diagrama de densitat perquè en haver tant poca diferència entre les dades és considera molt més visual la interpretació d'un diagrama de pastís.

```
englishHome <- table(data$englishAtHome)
cat("Un total de",englishHome["0"], "no parlen anglès a casa i un total de"
    ,englishHome["1"], "alumnes parlen anglès a casa")
```

```
## Un total de 461 no parlen anglès a casa i un total de 3131 alumnes parlen anglès a casa
```

```
pie(englishHome, main = "Distribució d'alumnes que parlen anglès a casa i
    alumnes que no parlen anglès a casa", col = c("red", "green"), labels = c("No", "Sí"))
```




Com s'observa, hi ha una gran majoria d'alumnes que sí que parlen anglès a casa. En aquest segon cas es descarta un diagrama de distribució atès que es considera més senzill d'interpretar un "Sí" o un "No" que un 1 o un 0.

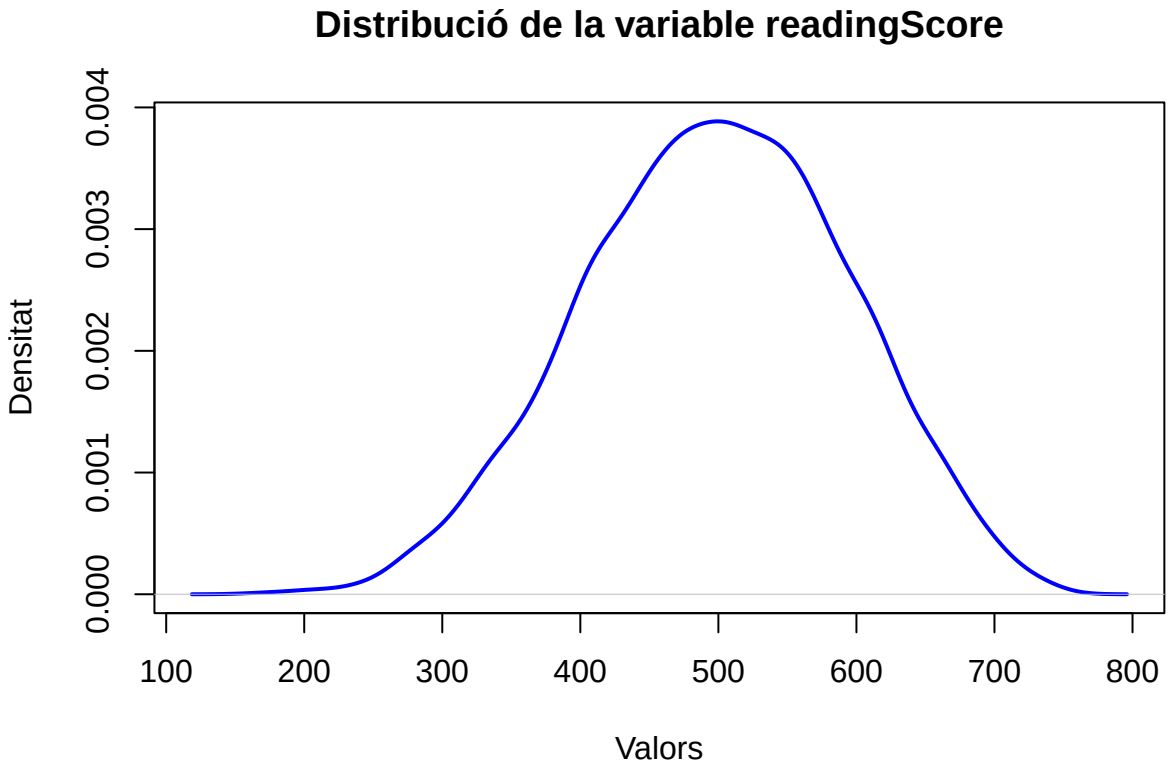
### 1.2 Anàlisi descriptiva de readingScore

Mostreu visualment la distribució de la variable readingScore. Interpreteu el gràfic.

```
scoring <- data$readingScore
```

Com en aquest cas hi ha molts valors diferents, per mostrar la distribució de la varibale s'utilitzarà un gràfic de densitat.

```
plot(density(scoring),
     main = "Distribució de la variable readingScore",
     xlab = "Valors",
     ylab = "Densitat",
     col = "blue",
     lwd = 2 # Gruix de la línia
     )
```




Si s'observa amb atenció la gràfica de distribució, es pot observar una tendència notòria en les puntuacions de lectura a adoptar valors propers a 500. Aquest fenomen es manifesta en la concentració màxima de la densitat (punt més àlgid de la gràfica) al voltant del valor 500. A mesura que els valors s'allunyen de 500, es pot veure una disminució gradual en la densitat de la corba (la corba s'aplana), indicant una menor freqüència d'observacions a mesura que augmenta la distància respecte a aquest màxim.

```
install.packages("ggplot2")
```

```
## Installing package into 'C:/Users/mcr99/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

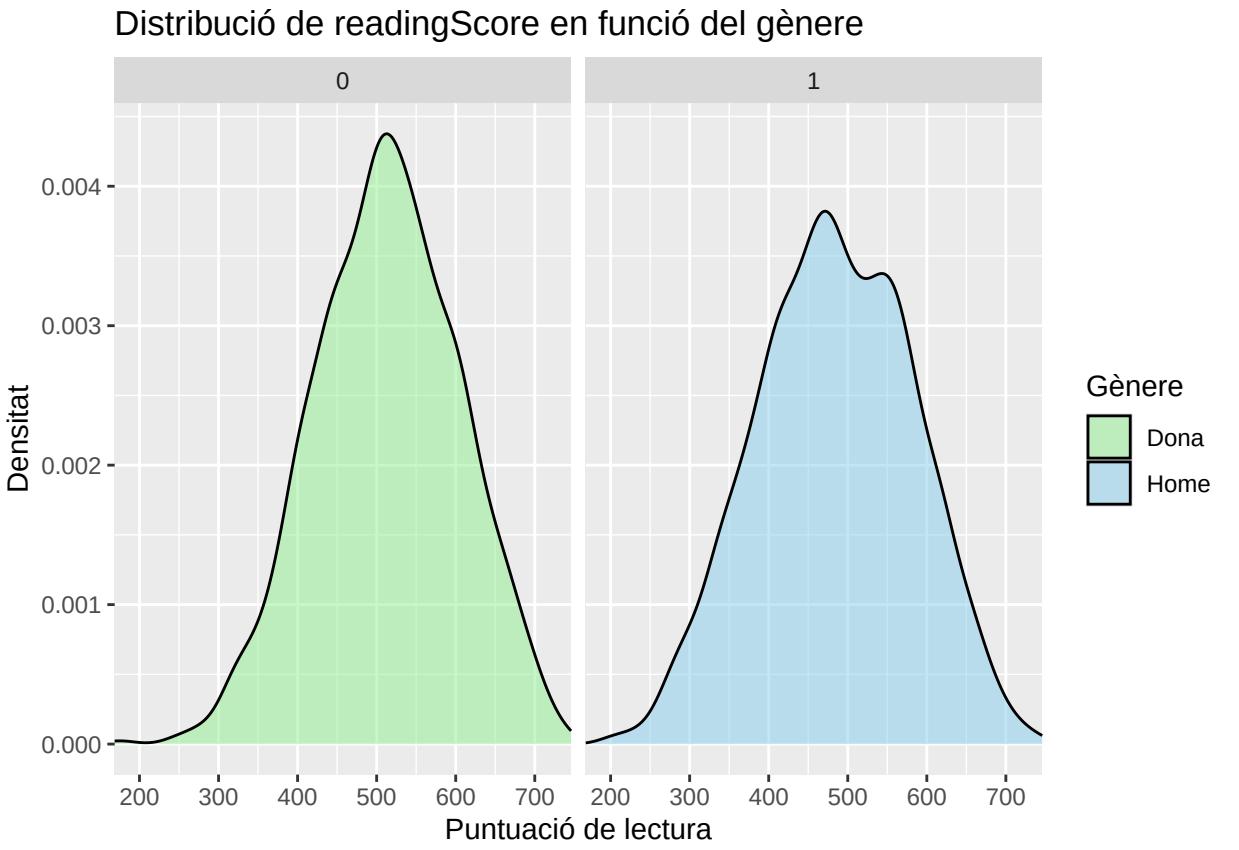
```
## Warning: package 'ggplot2' is not available for this version of R
##
## A version of this package for your version of R might be available elsewhere,
## see the ideas at
## https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages
```

```
tryCatch({
  library(ggplot2)
  print("Paquet importat correctament")
}, error = function(e){
  cat("ERROR en el moment d'imoprtar el paquet:",conditionMessage(e), "\n")
}
)
```

```
## [1] "Paquet importat correctament"
```

```
ggplot(data, aes(x = readingScore, fill = factor(male))) +
  geom_density(alpha = 0.5) +
```

```
  facet_wrap(~ data$male, nrow = 1) +
  scale_fill_manual(values = c("lightgreen", "skyblue"), name = "Gènere",
                    labels = c("Dona", "Home")) +
  scale_x_continuous(labels = scales::number_format(accuracy = 1),
                     expand = c(0, 0)) +
  # Augment de la precisió de l'escala de l'eix x
  labs(title = "Distribució de readingScore en funció del gènere",
       x = "Puntuació de lectura", y = "Densitat")
```

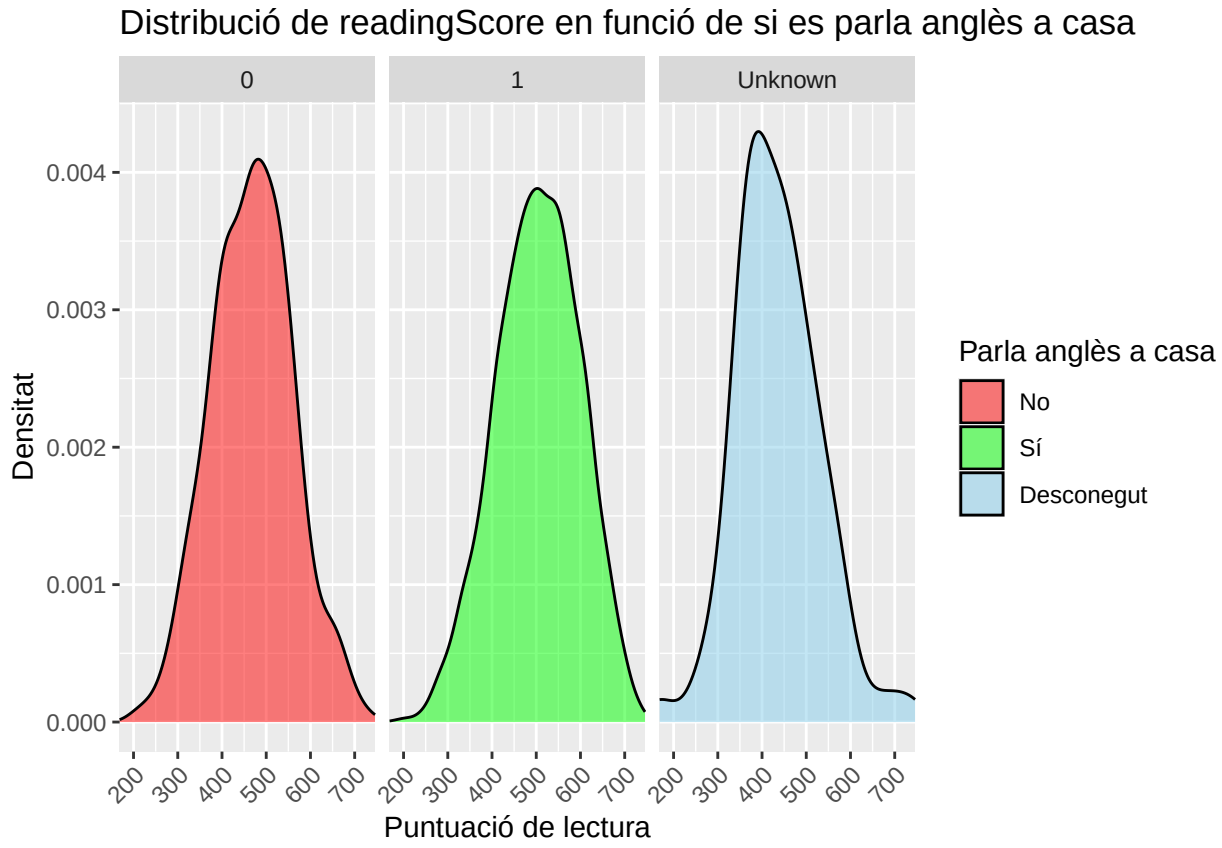



```
# facet_wrap(~ data$male, nrow = 1) -> Comanda per dividir la gràfica en dues pantalles
#(una per gènere)
```

En analitzar i comparar ambdues gràfiques, es pot observar que les puntuacions de lectura de les noies (gràfica esquerra) tendeixen a concentrar-se en valors d'entre 500 i 550. En el cas dels nois (gràfica dreta), la màxima densitat es concentra en valors entre 450 i 500. Per consegüent, es pot concloure que donades les gràfiques de distribució anteriors, les noies tenen una major puntuació de lectura que els nois atès que el punt més àlgid de la gràfica femenina és més elevat que el punt més àlgid de la gràfica masculina.

```
data$englishAtHome[is.na(data$englishAtHome)] <- "Unknown" # Convertim els valors NA
ggplot(data, aes(x = readingScore, fill = englishAtHome)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ data$englishAtHome, nrow = 1) +
  scale_fill_manual(values = c("red", "green", "skyblue"), name = "Parla anglès a casa",
                    labels = c("No", "Sí", "Desconegut")) +
  scale_x_continuous(labels = scales::number_format(accuracy = 1),
                     expand = c(0, 0)) +
  labs(title = "Distribució de readingScore en funció de si es parla anglès a casa",
```

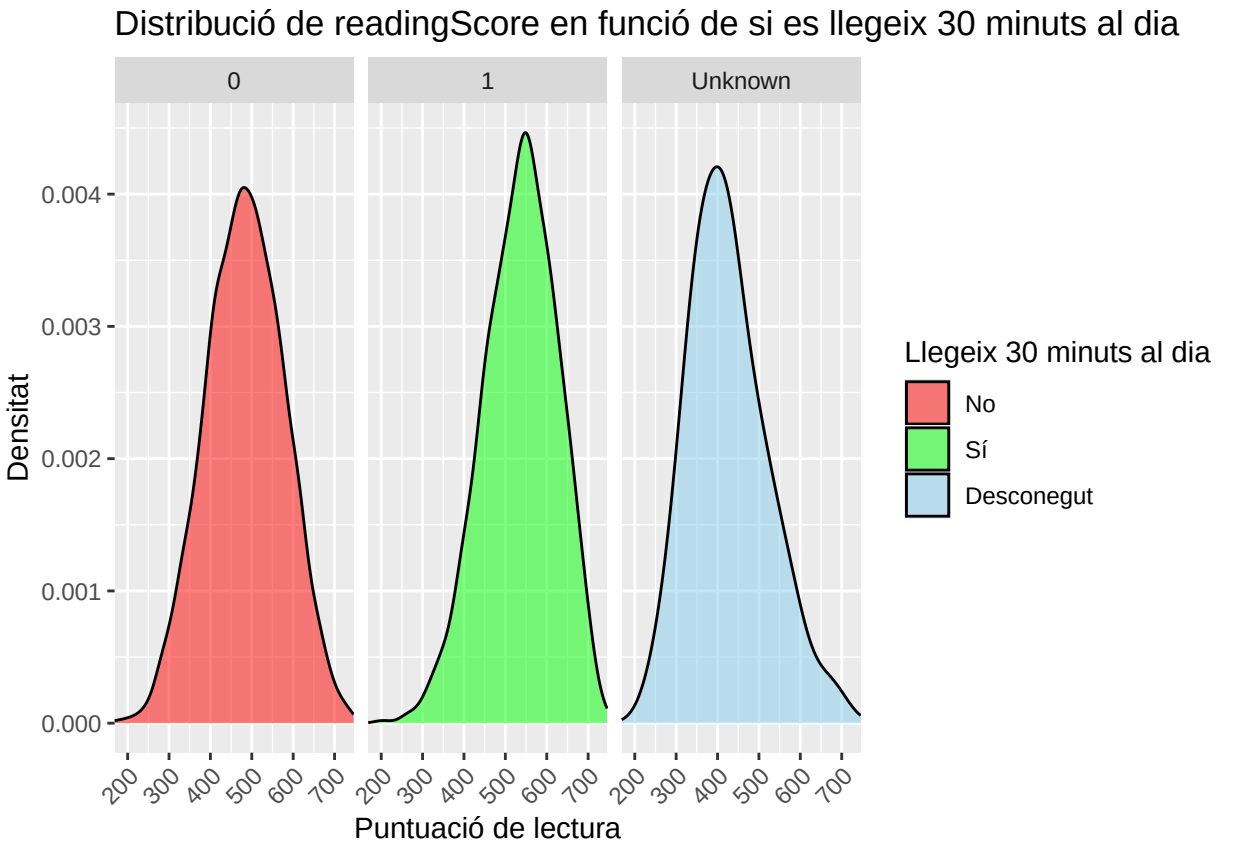
```
      x = "Puntuació de lectura", y = "Densitat") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




Observant les gràfiques, s'observa que aquelles persones que no parlen anglès a casa, tendeixen a treure puntuacions de lectura entre 450 i 500. Aquelles persones que sí que parlen anglès a casa, també concentren les seves puntuacions màximes entre 450 i 500, però en aquest cas el punt més àlgid de la gràfica no és tan agut com en el cas de les persones que no parlen anglès a casa, és a dir, es veu que les màximes puntuacions es troben concentrades entre 450 i 500 essent el punt més àlgid un valor molt proper a 500. El descens d'aquestes puntuacions no és tan pronunciat com en el cas de les persones que no parlen anglès a casa. El tercer cas, és el cas d'aquelles persones que no se sap si parlen o no l'anglès a casa. En aquest darrer cas, podem tornar a observar un descens molt pronunciat de les puntacions de lectura a mesura que aquestes s'allunyen de la puntuació que concentra la major densitat i un punt màxim molt agut. Les persones de les quals no coneixem l'idioma que parlen a casa, concentren les seves puntuacions de lectura entre 350 i 400.

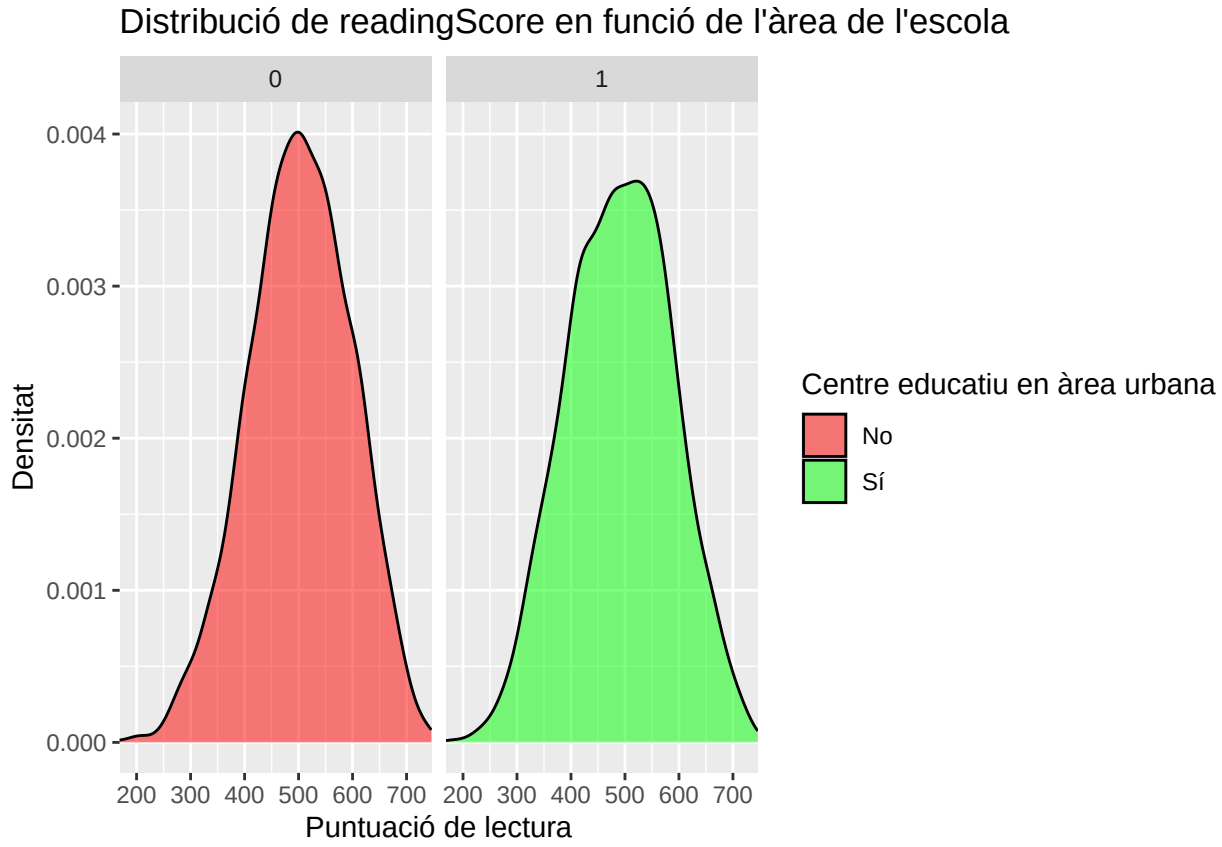
```
data$read30MinsADay[is.na(data$read30MinsADay)] <- "Unknown" # Convertim els valors NA
ggplot(data, aes(x = readingScore, fill = read30MinsADay)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ data$read30MinsADay, nrow = 1) +
  scale_fill_manual(values = c("red", "green", "skyblue"),
                    name = "Llegeix 30 minuts al dia",
                    labels = c("No", "Sí", "Desconegut")) +
  scale_x_continuous(labels = scales::number_format(accuracy = 1),
                     expand = c(0, 0)) +
  labs(title =
         "Distribució de readingScore en funció de si es llegeix 30 minuts al dia",
       x = "Puntuació de lectura", y = "Densitat") +
```

```
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



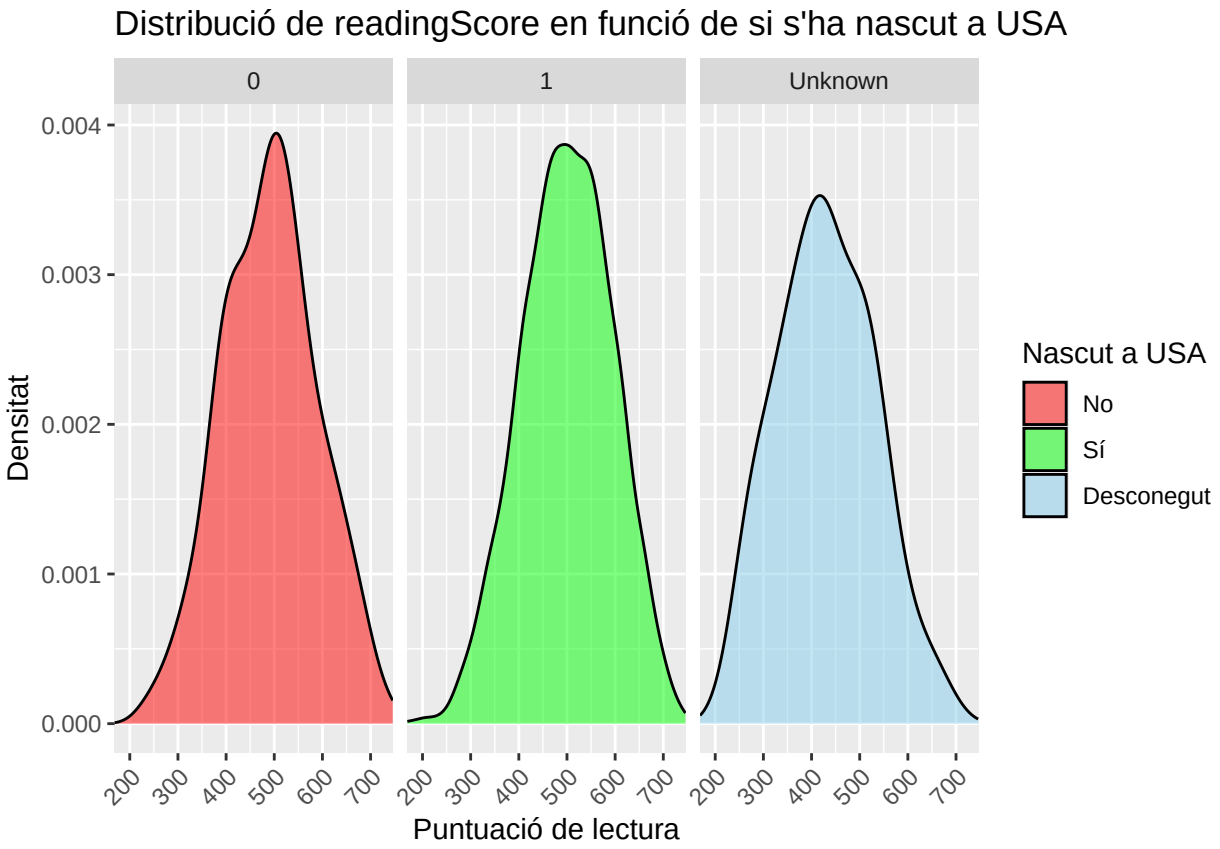

En fer l'anàlisi de les gràfiques d'aquest cas d'estudi, es pot observar que els alumnes que no llegeixen 30 minuts al dia tendeixen a treure unes puntuacions de lectura d'entre 450 i 500 essent el valor més elevat un nombre intermedi entre 450 i 500, mentre que en els alumnes que sí que llegeixen 30 minuts al dia, es pot observar una clara tendència a l'alça de les seves notes de lectura. Els alumnes lectors concentren la seva densitat de notes de lectura en valors d'entre 500 i 550 essent el valor més alt un nombre molt proper a 550. El tercer cas és el d'aquells alumnes dels quals no se sap si llegeixen o no. En el cas d'aquests alumnes, la densitat de notes es concentra en valors d'entre 350 i 400 essent el valor més àlgid un valor proper a 400.

```
ggplot(data, aes(x = readingScore, fill = factor(urban))) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ data$urban, nrow = 1) +
  scale_fill_manual(values = c("red", "green"), name = "Centre educatiu en àrea urbana",
                    labels = c("No", "Sí")) +
  scale_x_continuous(labels = scales::number_format(accuracy = 1),
                     expand = c(0, 0)) +
  # Augment de la precisió de l'escala de l'eix x
  labs(title = "Distribució de readingScore en funció de l'àrea de l'escola",
       x = "Puntuació de lectura", y = "Densitat")
```




Observant amb atenció les gràfiques anteriors es pot observar una clara tendència a obtenir puntuacions més baixes si el centre educatiu no es troba en una àrea urbana. Si s'analitza l'escala de les puntuacions de lectura, es pot veure que la màxima densitat de puntuacions per les escoles d'àrees rural tot just arriba a 500 i les escoles d'àrees urbanes superen aquest 500 i es troben entre 500 i 550.

```
data$selfBornUS[is.na(data$selfBornUS)] <- "Unknown" # Convertim els valors NA
ggplot(data, aes(x = readingScore, fill = selfBornUS)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ data$selfBornUS, nrow = 1) +
  scale_fill_manual(values = c("red", "green", "skyblue"), name = "Nascut a USA",
                    labels = c("No", "Sí", "Desconegut")) +
  scale_x_continuous(labels = scales::number_format(accuracy = 1),
                     expand = c(0, 0)) +
  labs(title = "Distribució de readingScore en funció de si s'ha nascut a USA",
       x = "Puntuació de lectura", y = "Densitat") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




Després d'observar les tres gràfiques, es pot concloure que aquells alumnes que no llegeixen 30 minuts al dia tendeixen a obtenir unes notes de lectura al voltant dels 500 punts, sent el punt amb més densitat el mateix valor 500. Els alumnes que sí que llegeixen 30 minuts al dia, tendeixen a treure unes notes entre 450 i 500, però a diferència d'aquells alumnes no lectors, es pot veure com la corba en el seu punt més àlgid no té un màxim tan agut, és a dir en la part més alta de la corba es reparteixen una mica més les notes. Aquells alumnes que no se sap si llegeixen o no tendeixen a treure unes notes entre 400 i 450 i, un altre cop, es percep un màxim bastant agut, cosa que indica que les notes no estan tan repartides (en la part alta de la corba de densitat) com en la corba dels alumnes lectors.

```
any(is.na(data$minutesPerWeekEnglish))
```

```
## [1] TRUE
```

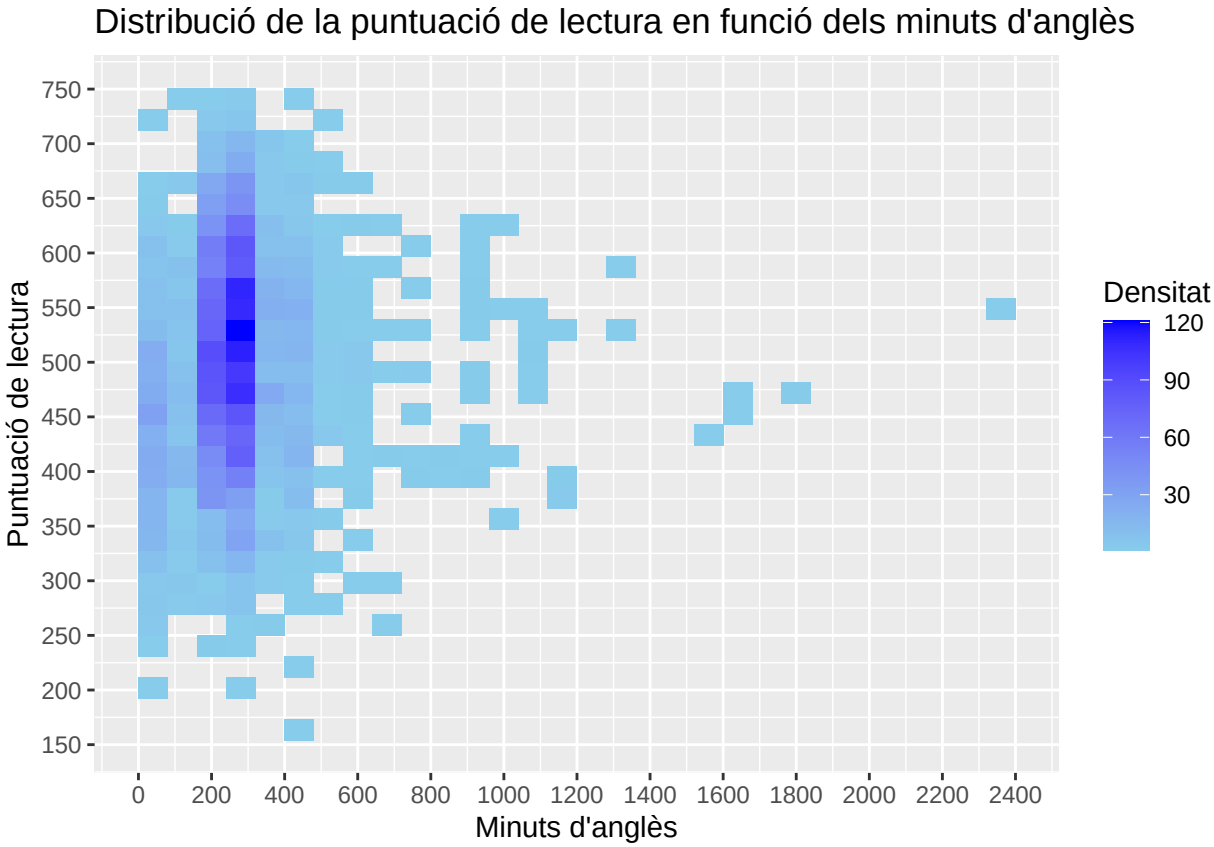
```
dataClean <- data[complete.cases(data$readingScore, data$minutesPerWeekEnglish), ]
any(is.na(dataClean$readingScore))
```

```
## [1] FALSE
```

```
any(is.na(dataClean$minutesPerWeekEnglish))
```

```
## [1] FALSE
```

```
ggplot(dataClean, aes(x = minutesPerWeekEnglish, y = readingScore)) +
  geom_bin2d() +
  scale_fill_gradient2(low = "grey", mid = "skyblue", high = "blue", name = "Densitat") +
  labs(x = "Minuts d'anglès", y = "Puntuació de lectura") +
  scale_y_continuous(breaks = seq(0, 1000, by = 50)) +
  scale_x_continuous(breaks = seq(0, 2500, by = 200)) +
  ggtitle("Distribució de la puntuació de lectura en funció dels minuts d'anglès")
```

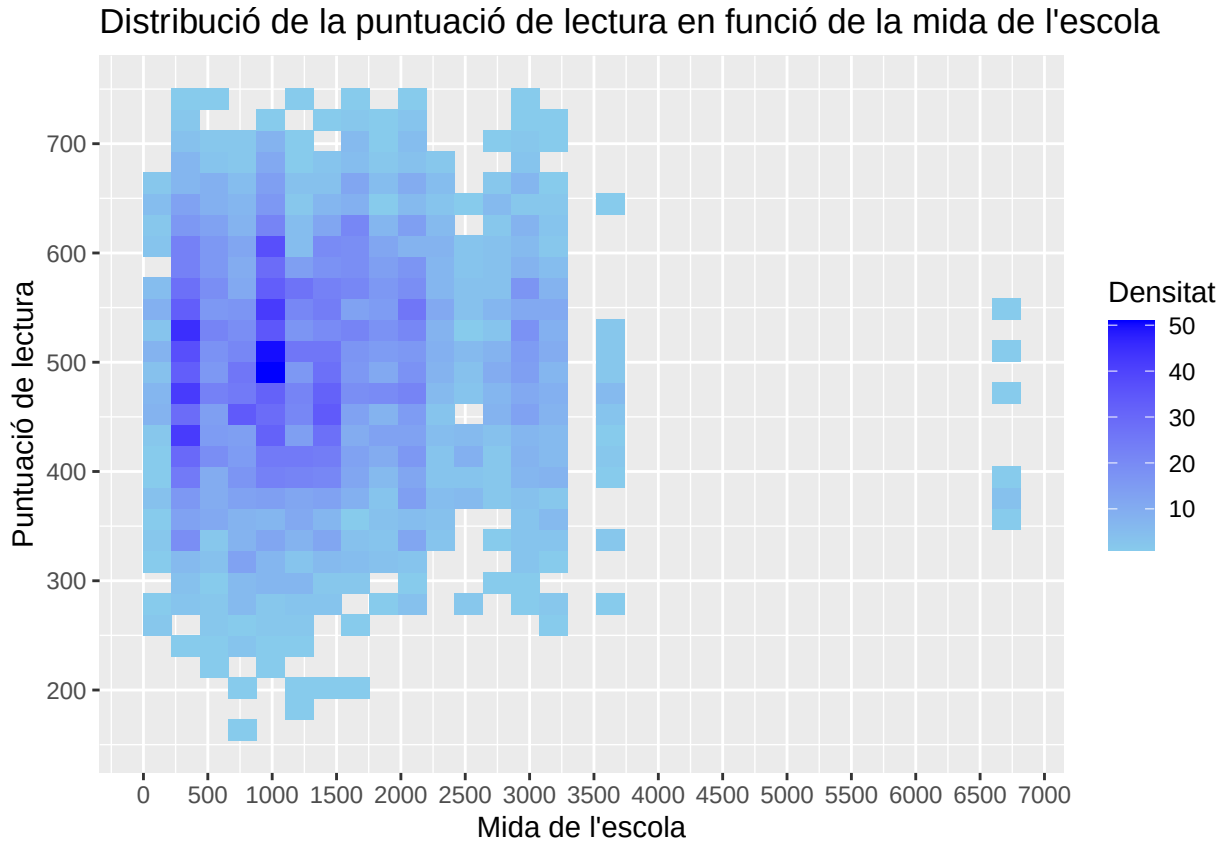



Com s'observa a la gràfica, la densitat més gran es concentra al voltant dels 300 minuts d'anglès a la setmana amb unes notes al voltant de 525.

```
any(is.na(data$schoolSize))
```

```
## [1] FALSE
```

```
ggplot(data, aes(x = schoolSize, y = readingScore)) +
  geom_bin2d() +
  scale_fill_gradient2(low = "grey", mid = "skyblue", high = "blue", name = "Densitat") +
  labs(x = "Mida de l'escola", y = "Puntuació de lectura") +
  scale_y_continuous(breaks = seq(0, 1000, by = 100)) +
  scale_x_continuous(breaks = seq(0, 8000, by = 500)) +
  ggtitle("Distribució de la puntuació de lectura en funció de la mida de l'escola")
```




Atès que ambdues variables tenen un gran nombre de valors diferents, en aquesta ocasió s'opta per representar la distribució mitjançant un mapa de calor, on la coloració més fosca implica una major densitat. S'observa que la llegenda del mapa de calor està numerada de 0 a 50. La determinació d'aquesta escala es fa de manera automàtica segons els valors de les dades, és a dir, *ggplot* busca el valor màxim i el mínim i a partir d'ells crea una gradació de color. En aquest cas, i segons el codi de coloració, es pot constatar que la densitat més gran es troba en les escoles que tenen al voltant de 1000 alumnes i unes puntuacions de lectura al voltant dels 500 punts.

## 2 Interval de confiança de reading score

Calculeu l'interval de confiança del valor mitjà de readingScore al 95% i al 97%. Interpreteu el resultat.

**Requisits:**

• Implementeu una funció que calculi l'interval de confiança i que pugueu utilitzar per obtenir l'IC per al nivell de confiança de 95% i 97% respectivament.

• No podeu fer servir funcions d'R que calculin l'interval de confiança. Sí podeu utilitzar funcions per calcular els valors de la distribució corresponent, com *qt*, *qnorm*, *pt*, *pnorm*.

```
calcul_interval <- function(nc){
  alpha <- 1-nc
  scorings <- as.numeric(data$readingScore)
  n <- length(scorings) # Nombre de dades
  desviation <- sd(scorings) # sd = Standard desviation
  error <- desviation / sqrt(as.numeric(n))
  t <- qt(alpha/2, df=n-1, lower.tail = FALSE)
```

```
  L <- mean(scorings) - t * error
  U <- mean(scorings) + t * error
  c((round(L, 2)), round(U, 2))
}
```

```
nc95 <- calcul_interval(0.95)
cat("L'interval de confiança per a un nivell de cinfiança del 95% és:", nc95)
```

```
## L'interval de confiança per a un nivell de cinfiança del 95% és: 494.82 501.01
```

```
nc97 <- calcul_interval(0.97)
cat("L'interval de confiança per a un nivell de cinfiança del 97% és:", nc97)
```

```
## L'interval de confiança per a un nivell de cinfiança del 97% és: 494.49 501.34
```

Els intervals de confiança donen un rang de valors en el qual pot estar la mitjana poblacional de la variable d'estudi. En aquest cas, amb un 95%/97% de confiança, la mitjana real de la varibale *readingScore* es trobarà dins d'aquest rang de valors. És a dir, s'està afirmant que hi ha una probabilitat del 95%/97% que la puntuació mitjana de lectura de tots els alumnes que han participat a la prova pisa estigui dins del rang de valors. En el cas del 95% de nivell de confiança, s'està afirmant que amb un 95% de probabilitat la puntuació mitjana de lectura es troba entre els 494.82 punts i els 501.34 punts. En el cas del 97% de nivell de confiança el rang s'estableix en 494.49 i 501.34.