Estadística Descriptiva

Grado en Ingeniería Informática

2025/26

1. Introducción

El objetivo de este documento es presentar las técnicas más utilizadas de Estadística descriptiva para resumir la información de un conjunto de datos de una y dos variables. Los datos que vamos a analizar en este guión están en el archivo `AlumnosIndustriales.xlsx`. Estos datos corresponden a 95 estudiantes de Ingeniería Industrial, a quienes se les preguntó acerca de algunas variables como altura, peso, número de hermanos y otras siete variables. De esta manera, vamos a utilizar un conjunto de datos simple que nos ayudará a aprender la función descriptiva básica de R.

Primero leemos y vemos el archivo de datos. La figura muestra las primeras cinco observaciones de este archivo de datos.

```
library(readxl)
AlumnosIndustriales <- read_excel("AlumnosIndustriales.xlsx")
head(AlumnosIndustriales, 5)
```

```
## # A tibble: 5 x 11
##   nacimiento altura  peso zapato  sexo dinero tiempo locomocion residencia
##        <dbl>  <dbl> <dbl>  <dbl> <dbl>  <dbl>  <dbl>      <dbl>      <dbl>
## 1          1    180    72     44     1   1100     35          3          1
## 2          1    161    55     39     0    287     45          4          1
## 3          1    180    45     41     1   2000    100          4          3
## 4          1    180    99     44     1     25     40          3          2
## 5          1    178    68     41     1   3225     40          1          3
## # i 2 more variables: hermanos <dbl>, Variables <chr>
```

2. Descripción de variables categóricas.

La variable `residencia` corresponde a la ubicación del hogar de los estudiantes. Esta variable es categórica. El conjunto de valores que puede tener es

- Madrid Sur (1)
- Madrid Centro (2)
- Madrid-otros (3)
- Fuera de Madrid (4)

Para describir esta variable, primero obtenemos una tabla de frecuencias.

```r
ABStable <- table(AlumnosIndustriales$residencia)
lbls <- c("Madrid Sur", "Madrid Centro", "Madrid-otros", "Fuera de Madrid")
row.names(ABStable) <- lbls
ABStable
```

```
##
##      Madrid Sur   Madrid Centro    Madrid-otros Fuera de Madrid
##              46              36              12               1
```

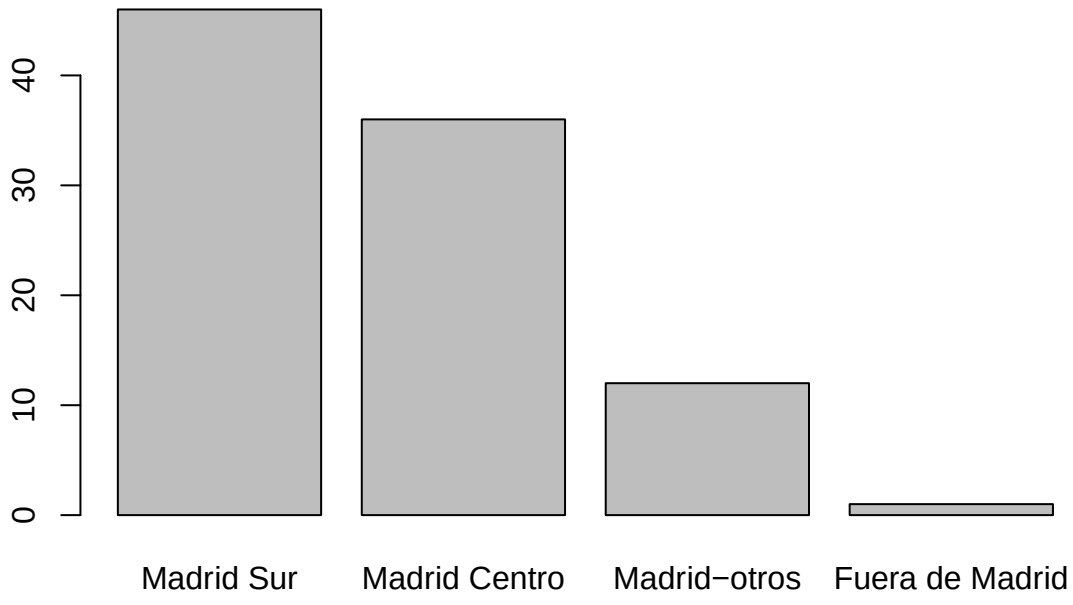
```r
RELtable <- prop.table(ABStable)
RELtable
```

```
##
##      Madrid Sur   Madrid Centro    Madrid-otros Fuera de Madrid
##      0.48421053      0.37894737      0.12631579      0.01052632
```

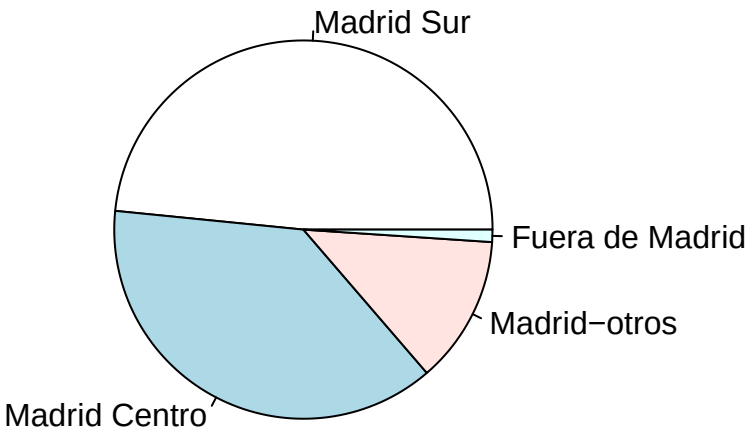
donde podemos ver que el grupo más grande de estudiantes son los que provienen de Madrid Sur, conformado por 46 estudiantes que representan el 48.4% de la muestra.

Para obtener un diagrama de barras o un diagrama de sectores o circular, utilizamos las siguientes intrucciones:

```r
barplot(ABStable)
```



```
pie(ABStable)
```



3. Descripción de variables cuantitativas

3.1 Análisis gráfico de una variable discreta con solo unos pocos valores

En el caso de variables discretas cuantitativas que tenga solo unos pocos valores, el análisis gráfico es el mismo que vimos para las variables categóricas. Ahora podemos obtener un gráfico de barras. La tabla de frecuencias sería similar a la que se genera al usar el análisis de datos categóricos.

Como ejemplo, utilizaremos la variable `hermanos` que tiene cada estudiante.

```
ABStable <- table(AlumnosIndustriales$hermanos)
ABStable
```

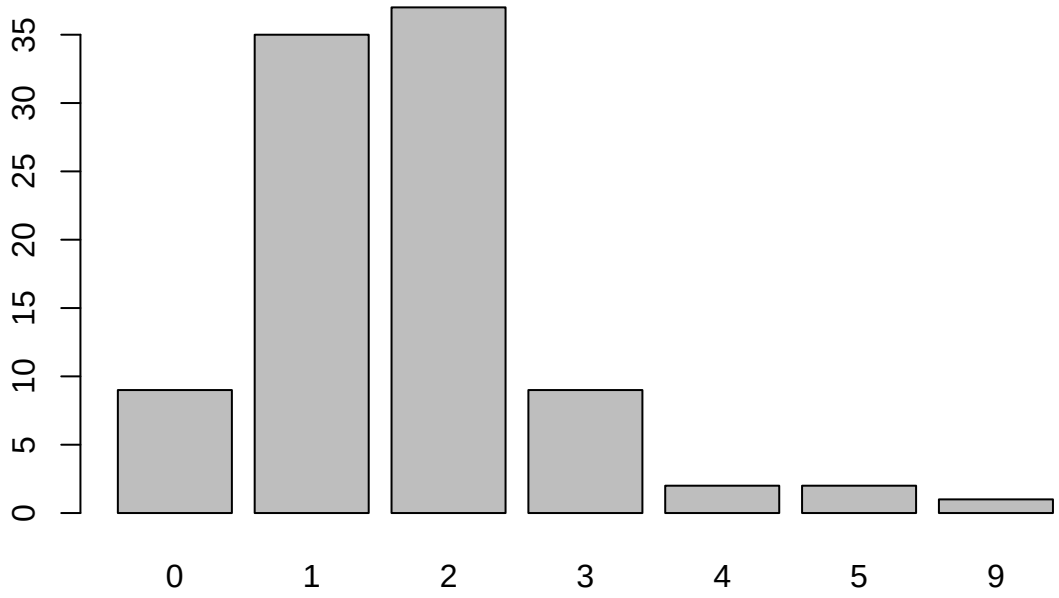
```
##
##  0  1  2  3  4  5  9
##  9 35 37  9  2  2  1
```

```
RELtable <- prop.table(ABStable)
RELtable
```

```
##
```

```
##          0          1          2          3          4          5          9
## 0.09473684 0.36842105 0.38947368 0.09473684 0.02105263 0.02105263 0.01052632
```

```r
barplot(ABStable)
```



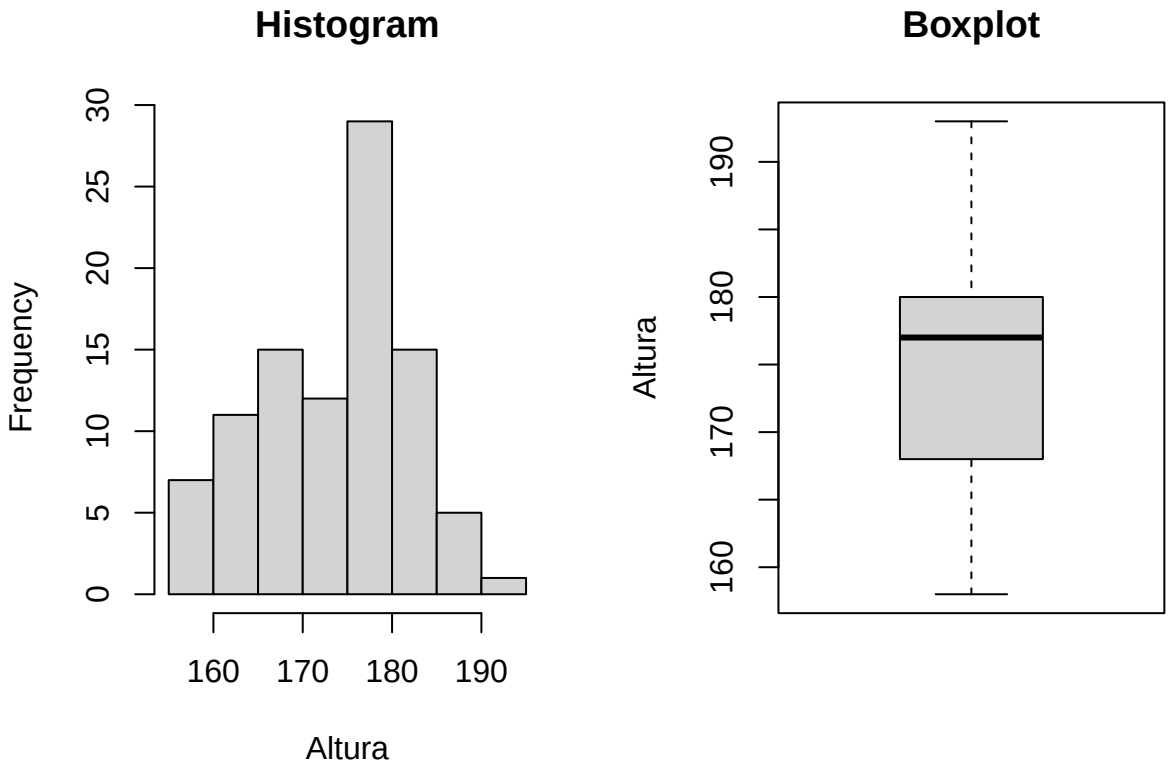
En la imagen podemos ver que las familias más frecuentes (entre las que estudian Ingeniería Industrial en UC3M) son las que tienen 2 ó 3 hijos (1 ó 2 hermanos).

3.2 Análisis gráfico de variables cuantitativas

El análisis gráfico de las variables cuantitativas generales puede realizarse utilizando las funciones `hist` y `boxplot`.

La variable `altura` contiene las alturas de los estudiantes. Aquí, trazamos su histograma y su diagrama de caja.

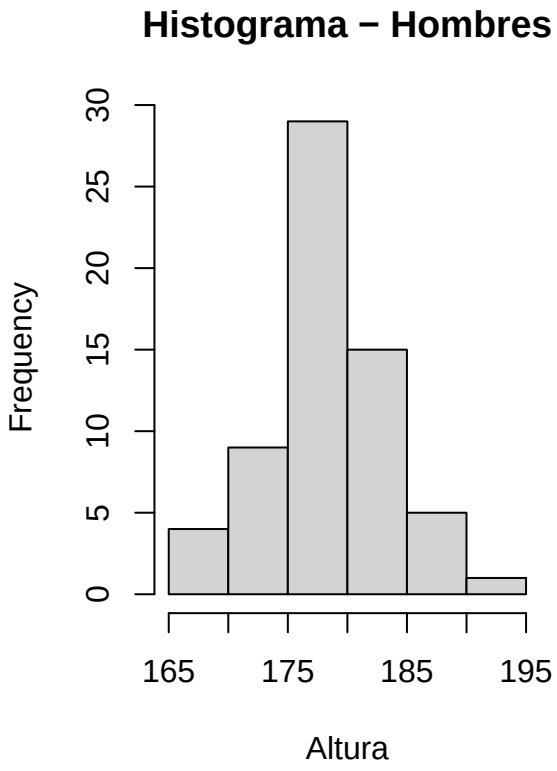
```r
par(mfrow=c(1,2))
hist(AlumnosIndustriales$altura, xlab = "Altura", main = "Histogram")
boxplot(AlumnosIndustriales$altura, ylab = "Altura", main = "Boxplot")
```

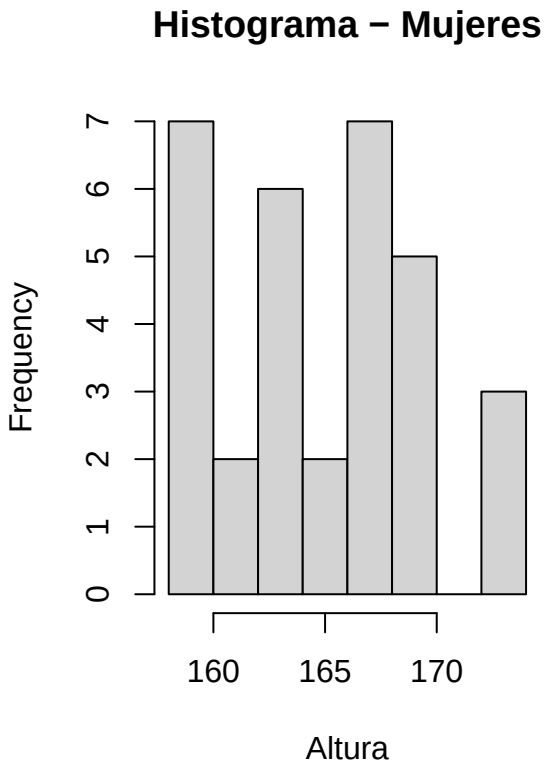


El diagrama de caja muestra que la distribución de alturas es asimétrica. La caja central muestra una asimetría a la izquierda, aunque la cola no es muy larga. Este efecto también es visible en el histograma. Además, podemos ver que existen dos modas que sugieren que la muestra no es homogénea. Probablemente esto se deba a la presencia conjunta de alturas masculinas y femeninas.

La variable `sexo` contiene el género de los estudiantes (1 = masculino, 0 = femenino). Podemos usar esta variable para seleccionar las alturas de hombres o mujeres y de esta manera verificar si estos dos grupos tienen los datos de altura correspondientes concentrados alrededor de las dos modas.

```
par(mfrow=c(1,2))
hist(AlumnosIndustriales$altura[AlumnosIndustriales$sexo == 1], xlab = "Altura",
     main = "Histograma - Hombres")
hist(AlumnosIndustriales$altura[AlumnosIndustriales$sexo == 0], xlab = "Altura",
     main = "Histograma - Mujeres")
```





Vemos que al mostrar solo las alturas de los hombres, la distribución parece más simétrica, unimodal y con una moda en el intervalo [175, 180], con una alta concentración a su alrededor. Repitiendo lo mismo para las mujeres, obtenemos que su distribución no tiene una forma de campana unimodal como la de los hombres. Quizás esto se deba al hecho de que tenemos menos datos (solo 32 personas) o quizás porque son heterogéneos por sí mismos.

Tabla de frecuencia

La tabla de frecuencias nos da la misma información contenida en el histograma correspondiente, pero nos permite ver los valores numéricos de las frecuencias en cada intervalo. Para obtener la tabla de frecuencias podemos usar el siguiente código:

```
Altura <- AlumnosIndustriales$altura
range(Altura)
```

```
## [1] 158 193
```

```
breaks = seq(155, 195, by=5)
breaks
```

```
## [1] 155 160 165 170 175 180 185 190 195
```

```r
Altura.cut = cut(Altura, breaks, right=TRUE)
Altura.table = table(Altura.cut)
Altura.table
```

```
## Altura.cut
## (155,160] (160,165] (165,170] (170,175] (175,180] (180,185] (185,190] (190,195]
##         7        11        15        12        29        15         5         1
```

```r
prop.table(Altura.table)
```

```
## Altura.cut
##  (155,160]  (160,165]  (165,170]  (170,175]  (175,180]  (180,185]  (185,190]
## 0.07368421 0.11578947 0.15789474 0.12631579 0.30526316 0.15789474 0.05263158
##  (190,195]
## 0.01052632
```

Estas tablas muestran que el intervalo modal está alrededor de los valores (punto medio), 167.5 y 177.5, y que el intervalo con la frecuencia más alta, con el punto medio 177.5, contiene más del 30% de los estudiantes.

3.3 Medidas de resumen numérico de variables cuantitativas

Para calcular las medidas resumen de la variable `altura` podemos usar la función `summary`:

```r
summary(Altura)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   158.0   168.0   177.0   174.6   180.0   193.0
```

Proporciona medidas resumen de posición como la media, la mediana, el primer y tercer cuartiles, el mínimo y el máximo.

Se pueden obtener otras estadísticos de resumen utilizando la función `descr` del paquete `summarytools`:

```r
suppressWarnings(library(summarytools))
descr(Altura)
```

```
## Descriptive Statistics
## Altura
## N: 95
##
##                     Altura
## ------------------ --------
##              Mean   174.62
##           Std.Dev     8.23
##               Min   158.00
##                Q1   168.00
##            Median   177.00
##                Q3   180.00
##               Max   193.00
##               MAD     7.41
```

```
##                IQR    12.00
##                 CV     0.05
##           Skewness    -0.29
##        SE.Skewness     0.25
##           Kurtosis    -0.92
##            N.Valid    95.00
##          Pct.Valid   100.00
```

Incluye las estadísticas de posición anteriores y algunas medidas de dispersión y forma, como la desviación estándar, el coeficiente de variación, el MAD, los coeficientes de asimetría y curtosis.

Es necesario aclarar algunas cosas sobre estas medidas resumen:

- La varianza (y como consecuencia, la desviación estándar) que se usa se calcula usando la siguiente fórmula $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ en lugar de la fórmula de varianza $\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$. La conveniencia de dividir entre $n-1$ en lugar de n no es inmediata, y su justificación teórica se verá en temas más avanzados.
- El coeficiente de curtosis calculado es lo que se conoce como 'exceso de curtosis', definido como $\kappa = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{ns^4} - 3$. Por lo tanto, para una variable en forma de campana, la curtosis es igual a 0. Si κ es positivo, decimos que la variable tiene exceso de curtosis.

A continuación, se muestra una comparación entre hombres y mujeres mediante algunas medidas resumen:

```
suppressWarnings(library(summarytools))
descr(Altura[AlumnosIndustriales$sexo==1])
```

```
## Descriptive Statistics
## Altura
## N: 63
##
##                      Altura
## ----------------- --------
##              Mean   179.35
##           Std.Dev     5.04
##               Min   165.00
##                Q1   177.00
##            Median   180.00
##                Q3   182.00
##               Max   193.00
##               MAD     2.97
##               IQR     4.50
##                CV     0.03
##          Skewness    -0.28
##       SE.Skewness     0.30
##          Kurtosis     0.66
##           N.Valid    63.00
##         Pct.Valid   100.00
```

```
descr(Altura[AlumnosIndustriales$sexo==0])
```

```
## Descriptive Statistics
```

```
## Altura
## N: 32
##
##                      Altura
## ----------------- --------
##              Mean   165.31
##           Std.Dev     4.43
##               Min   158.00
##                Q1   161.50
##            Median   165.00
##                Q3   168.50
##               Max   174.00
##               MAD     5.19
##               IQR     6.50
##                CV     0.03
##          Skewness     0.21
##       SE.Skewness     0.41
##          Kurtosis    -1.14
##           N.Valid    32.00
##         Pct.Valid   100.00
```

Ahora podemos ver que los hombres de esta muestra son en promedio más altos que las mujeres. La altura promedio de los hombres es de 179 cm, mientras que para las mujeres es de 165 cm. En ambos sexos, la media es casi igual a la mediana. Esto es claramente visible mirando la simetría del gráfico de caja (ver sección 4) y el bajo valor absoluto de la asimetría. Esta concentración alrededor del valor mediano es visible también cuando se observa el rango intercuartílico. El 50% de los hombres (mujeres) ubicados en las posiciones centrales tienen una altura que difiere del valor mediano en menos de 3 cm.

3.4 Percentiles

Para obtener los percentiles, utilizamos la función `quantile`

```
quantile(Altura, probs = seq(0, 1, 0.1))
```

```
##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
## 158.0 163.0 167.0 169.2 173.0 177.0 179.0 180.0 181.0 184.6 193.0
```

Entonces, podemos concluir que el 20% del los estudiantes mide menos de 167 cm de altura y el 80% de ellos mide menos de 181 cm. El argumento `probs` puede tener cualquier valor en el intervalo $[0, 1]$.

4 Descripción simultánea de más de una variable

En muchos casos estamos interesados en comparar varias variables, o en comparar los valores de una variable dividida en dos o más grupos de individuos como en el caso de la altura por género. En tales casos, es más interesante producir gráficos y resúmenes estadísticos en la misma ventana para facilitar esta comparación, en lugar de realizar el análisis univariado de cada variable por separado. Por ejemplo, nos gustaría generar los diagramas de caja de cada variable en el mismo gráfico.

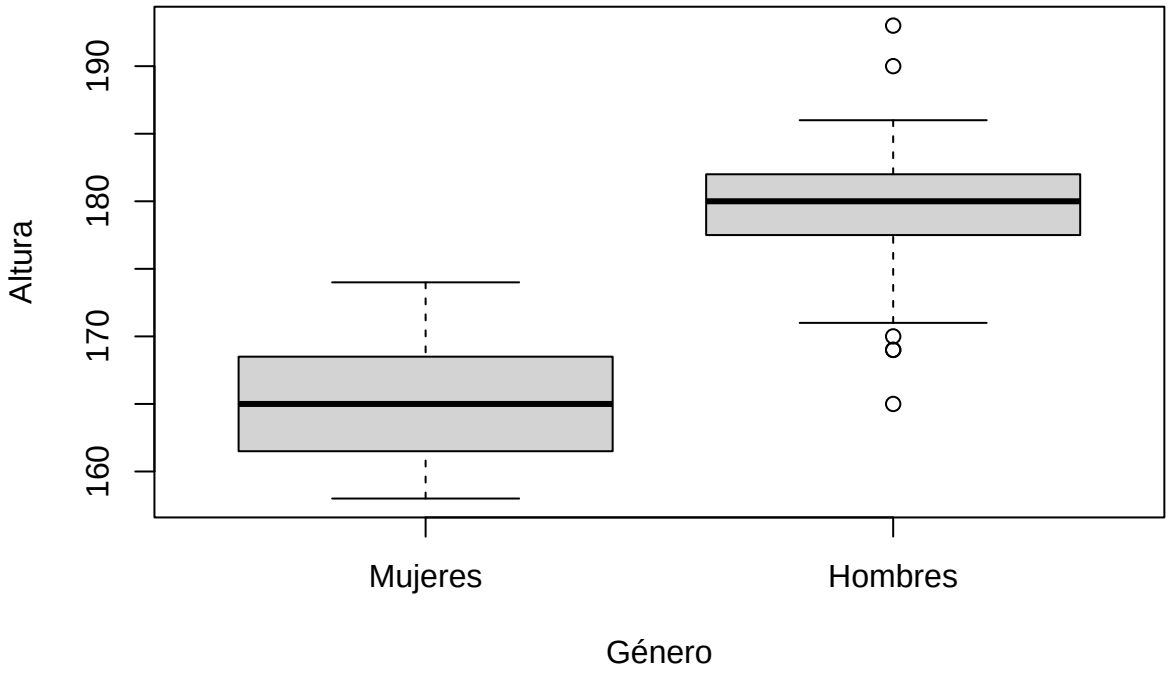
4.1 Múltiples diagramas de caja

Nuestro objetivo es crear un gráfico que tenga los diagramas de caja de varias variables o de la misma variable dividida en más de un subgrupo. Este gráfico permitirá una mejor comparación de estas variables.

4.1.1 Una variable cuantitativa por subgrupos

Estamos interesados en analizar cómo es la distribución de los valores de una variable cuando el conjunto de datos se subdivide en subgrupos de acuerdo con algún criterio. Por ejemplo, queremos estudiar las alturas de un grupo de estudiantes según su género. La variable `sexo` solo toma los valores 1 y 0. Estos valores son necesarios solo para distinguir a los miembros de cada grupo, por lo que el número que suponen es irrelevante. Podría ser -1 y 1, o incluso caracteres.

```
boxplot(AlumnosIndustriales$altura ~ AlumnosIndustriales$sexo, xlab = "Género",
        ylab = "Altura", names = c("Mujeres", "Hombres"))
```

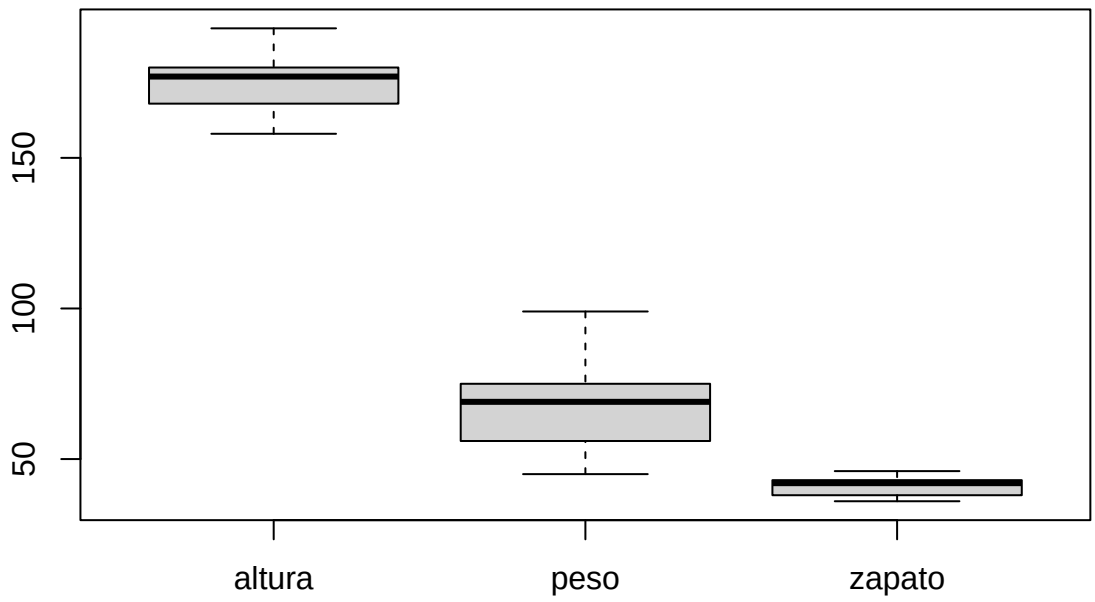


Puede verse que los hombres son generalmente más altos que las mujeres. Al mostrar simultáneamente los dos diagramas de caja, podemos interpretar que aproximadamente solo el 25% de las mujeres tienen alturas más altas comparables al 25% de los hombres más bajos. Además, son visibles algunos valores atípicos en el diagrama de caja de los hombres.

4.1.2 Varias variables cuantitativas

Si tenemos varias variables de la misma magnitud o diferentes magnitudes, también podemos hacer un diagrama de caja múltiple de estas variables en un gráfico. Como ejemplo, queremos ver las distribuciones de tres variables físicas de las personas en la muestra: `altura`, `peso` y `zapato`.

```
boxplot(AlumnosIndustriales[,2:4])
```



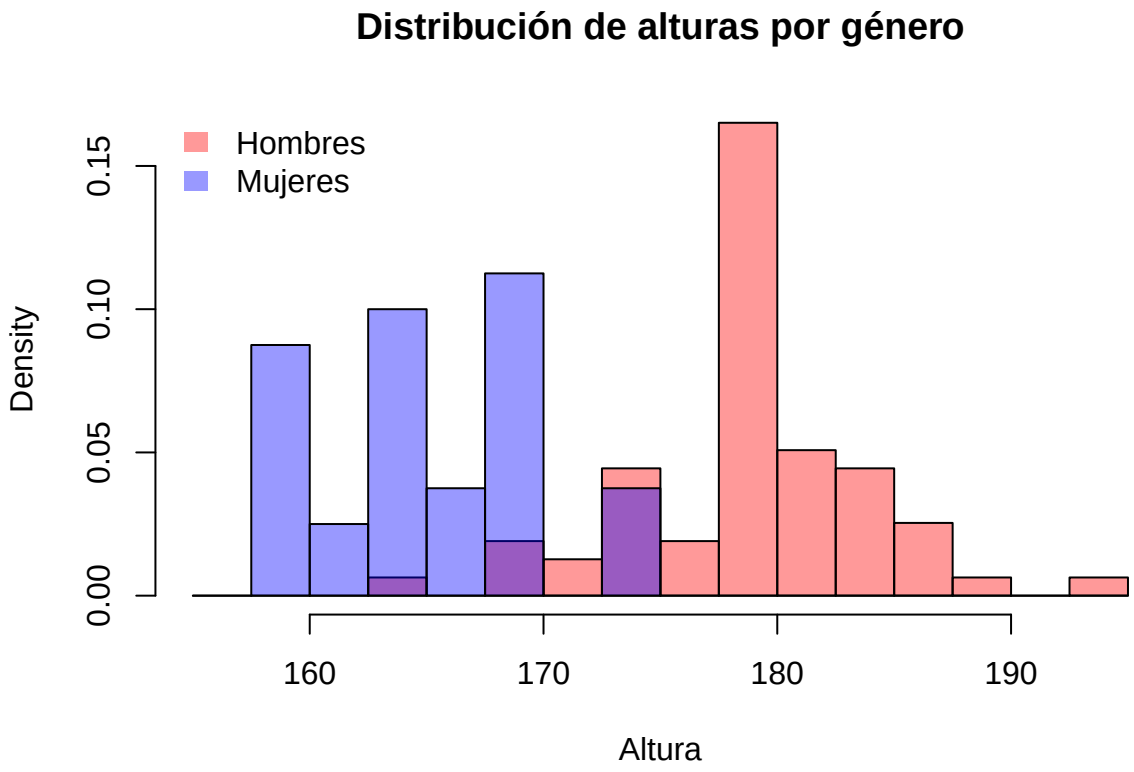
Aunque, en este caso, la representación es adecuada, intuimos que cuando hay grandes diferencias en escalas de las variables a representar es conveniente usar boxplot independientes.

4.2 Dos histogramas superpuestos

Si queremos comparar dos histogramas, es útil ponerlos en el mismo gráfico. Hay muchas formas de producir histogramas superpuestos, pero aquí mostramos cómo hacerlo utilizando los gráficos base R. Utilizaremos el ejemplo de alturas por género:

```
# notar que plot = FALSE
histHombres <- hist(Altura[AlumnosIndustriales$sexo == 1], breaks = seq(155,195,2.5),
                    plot = FALSE)
histMujeres <- hist(Altura[AlumnosIndustriales$sexo == 0], breaks = seq(155,195,2.5),
                      plot = FALSE)
# calcular el rango de los gráficos
xlim <- range(histHombres$breaks,histMujeres$breaks)
ylim <- range(0,histHombres$density,histMujeres$density)
# dibujar el primer histograma
plot(histHombres,xlim = xlim, ylim = ylim, col = rgb(1,0,0,0.4), xlab = 'Altura',
     freq = FALSE, ## frecuencia relative
     main = 'Distribución de alturas por género')
## dibujar el segundo histograma encima del anterior
opar <- par(new = FALSE)
plot(histMujeres,xlim = xlim, ylim = ylim,
     xaxt = 'n', yaxt = 'n', ## no agregar ejes
```

```
     col = rgb(0,0,1,0.4), add = TRUE,
     freq = FALSE)
## agregar una legenda
legend('topleft',c('Hombres','Mujeres'), fill = rgb(1:0,0,0:1,0.4), bty = 'n', border = NA)
```

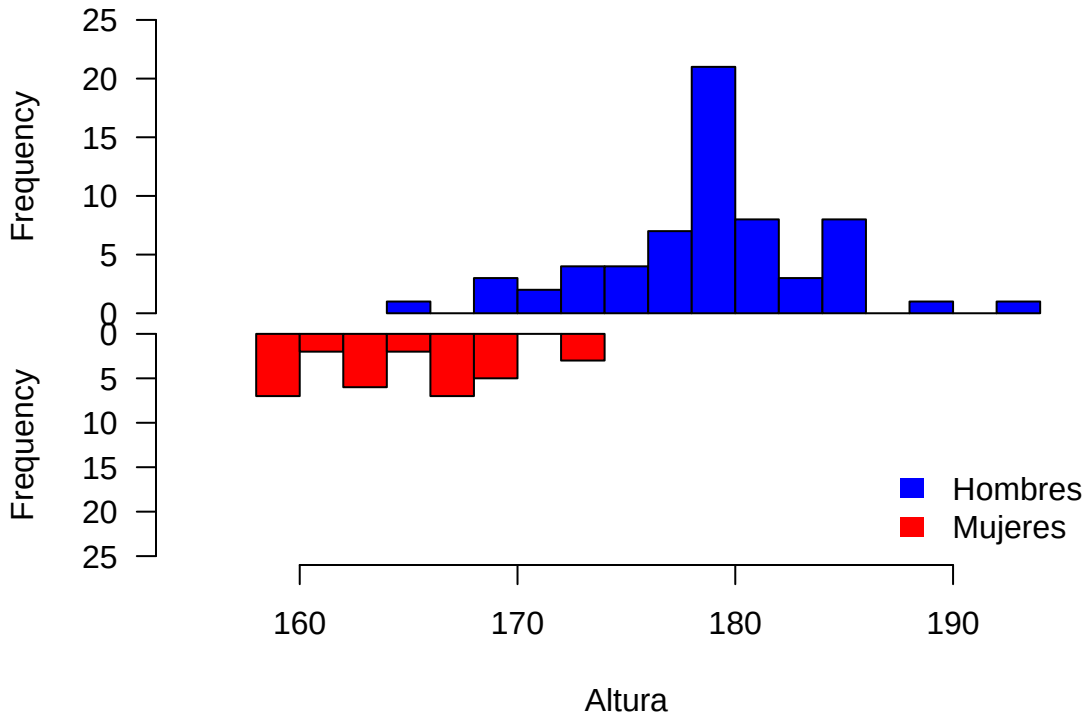



```
par(opar)
```

Otra posibilidad es obtener histogramas contrapuestos:

```
par(mfrow=c(2,1))

#Make the plot
par(mar=c(0,5,3,3))
hist(Altura[AlumnosIndustriales$sexo == 1], main="" , xlim=c(155,195),
     ylab="Frequency", xlab="", ylim=c(0,25) , xaxt="n", las=1 ,
     col="blue", breaks=10)
par(mar=c(5,5,0,3))
hist(Altura[AlumnosIndustriales$sexo == 0], main="" , xlim=c(155,195),
     ylab="Frequency", xlab="Altura", ylim=c(25,0), las=1,
     col="red"  , breaks=10)
legend('bottomright',c('Hombres','Mujeres'), fill = c("blue", "red"), bty = 'n', border = NA)
```

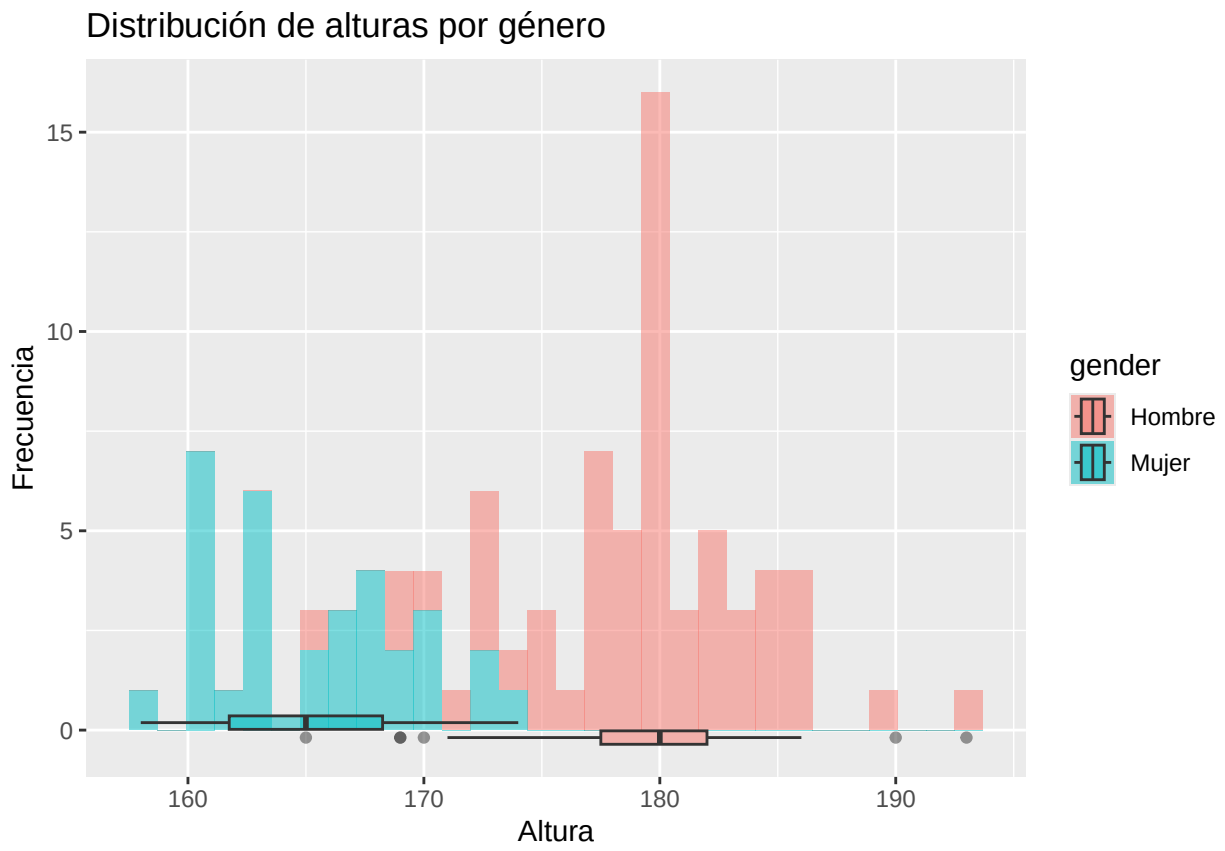


El paquete `ggplot2`, creado por Hadley Wickham, ofrece un lenguaje gráfico poderoso para crear gráficos elegantes y complejos, y es muy popular en la comunidad R. Dominar el idioma `ggplot2` puede ser un desafío, pero podemos usar una IA generativa para salir del apuro. *Recordad que el análisis de los resultados y las conclusiones son vuestra responsabilidad.* A modo de ejemplo, las siguientes líneas muestras como histogramas y diagramas de caja en el mismo gráfico:

```
suppressWarnings(library(ggplot2))
AlumnosIndustriales$gender <- "Hombre"
AlumnosIndustriales$gender[AlumnosIndustriales$sexo == 0] <- "Mujer"
qplot(altura, data=AlumnosIndustriales, geom=c("histogram", "boxplot"), fill=gender,
      alpha=I(.5), main="Distribución de alturas por género",
      xlab="Altura", ylab="Frecuencia")
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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



4.3 Medidas de resumen numérico de varias variables

En general, comenzamos nuestro análisis de datos mirando los gráficos. En un segundo momento buscamos medidas que puedan resumir de forma cuantitativa las características de mayor interés.

Utilizaremos el data.frame `AlumnosIndustriales`. Debe tenerse en cuenta que las variables `sexo`, `locomocion` y `residencia` están codificadas como numéricas y la variable `gender` es un carácter, por lo que podemos cambiar su clase para obtener estadísticas de resumen que sean válidas.

```
AlumnosIndustriales$sexo <- as.factor(AlumnosIndustriales$sexo)
AlumnosIndustriales$residencia <- as.factor(AlumnosIndustriales$residencia)
AlumnosIndustriales$locomocion <- as.factor(AlumnosIndustriales$locomocion)
AlumnosIndustriales$sexo <- as.factor(AlumnosIndustriales$sexo)
summary(AlumnosIndustriales)
```

```
##    nacimiento         altura           peso           zapato      sexo
##  Min.   : 1.000   Min.   :158.0   Min.   :45.00   Min.   :36.00   0:32
##  1st Qu.: 3.000   1st Qu.:168.0   1st Qu.:56.00   1st Qu.:38.00   1:63
##  Median : 5.000   Median :177.0   Median :69.00   Median :42.00
##  Mean   : 5.463   Mean   :174.6   Mean   :67.77   Mean   :40.98
##  3rd Qu.: 7.500   3rd Qu.:180.0   3rd Qu.:75.00   3rd Qu.:43.00
##  Max.   :12.000   Max.   :193.0   Max.   :99.00   Max.   :46.00
##      dinero           tiempo       locomocion residencia    hermanos
##  Min.   :   0.0   Min.   :  1.00   1:19       1:46       Min.   :0.000
##  1st Qu.: 217.5   1st Qu.: 20.00   2: 2       2:36       1st Qu.:1.000
```

```
##  Median : 655.0   Median : 40.00   3:29       3:12       Median :2.000
##  Mean   :1039.2   Mean   : 41.42   4:37       4: 1       Mean   :1.716
##  3rd Qu.:1300.0   3rd Qu.: 60.00   5: 8                  3rd Qu.:2.000
##  Max.   :5000.0   Max.   :120.00                         Max.   :9.000
##   Variables            gender
##  Length:95          Length:95
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

Si solo nos interesan las variables cuantitativas, puede usarse la función `descr` del paquete `summaryTools`:

```
descr(AlumnosIndustriales)
```

```
## Non-numerical variable(s) ignored: sexo, locomocion, residencia, Variables, gender

## Descriptive Statistics
## AlumnosIndustriales
## N: 95
##
##                     altura    dinero   hermanos   nacimiento     peso   tiempo   zapato
## ----------------- -------- --------- ---------- ------------ -------- -------- --------
##              Mean   174.62   1039.23       1.72         5.46    67.77    41.42    40.98
##           Std.Dev     8.23   1200.14       1.25         3.26    11.80    24.74     2.73
##               Min   158.00      0.00       0.00         1.00    45.00     1.00    36.00
##                Q1   168.00    200.00       1.00         3.00    56.00    20.00    38.00
##            Median   177.00    655.00       2.00         5.00    69.00    40.00    42.00
##                Q3   180.00   1300.00       2.00         8.00    75.00    60.00    43.00
##               Max   193.00   5000.00       9.00        12.00    99.00   120.00    46.00
##               MAD     7.41    770.95       1.48         2.97    13.34    29.65     2.97
##               IQR    12.00   1082.50       1.00         4.50    19.00    40.00     5.00
##                CV     0.05      1.15       0.73         0.60     0.17     0.60     0.07
##          Skewness    -0.29      1.94       2.38         0.38     0.25     0.63    -0.33
##       SE.Skewness     0.25      0.25       0.25         0.25     0.25     0.25     0.25
##          Kurtosis    -0.92      3.42      10.77        -0.72    -0.59    -0.04    -1.11
##           N.Valid    95.00     95.00      95.00        95.00    95.00    95.00    95.00
##         Pct.Valid   100.00    100.00     100.00       100.00   100.00   100.00   100.00
```

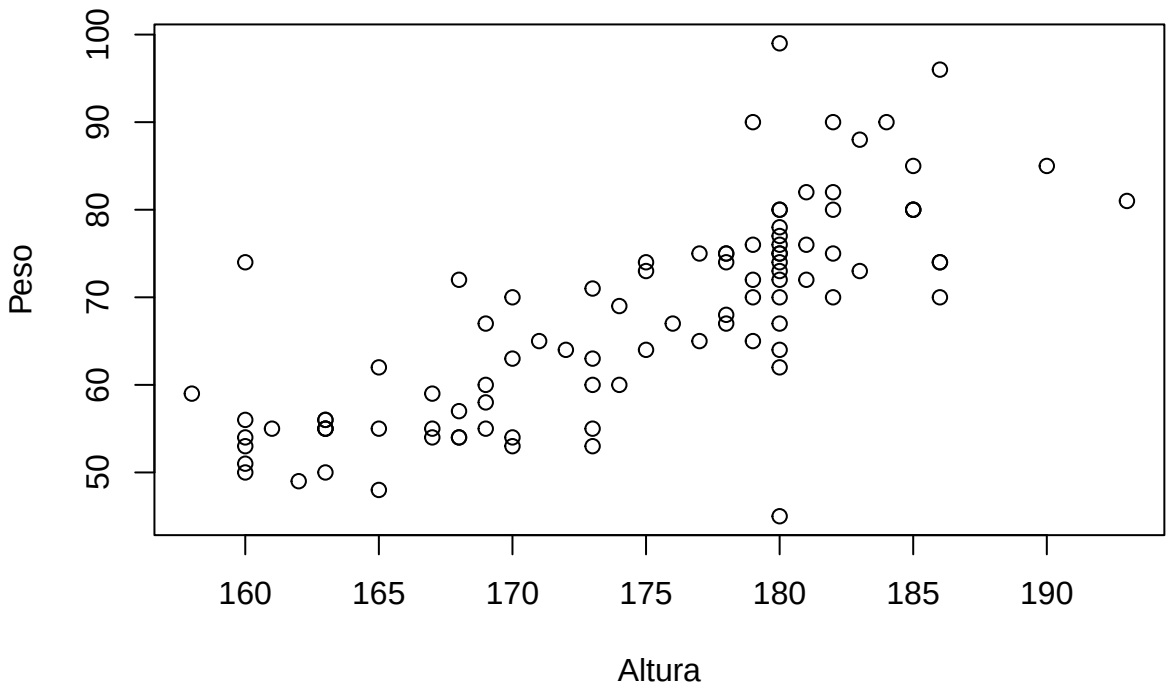
5. Análisis bivariante

En esta sección vamos a analizar dos variables observadas simultáneamente usando R. Vamos a estudiar su dependencia lineal y encontrar la recta de regresión que nos ayudará a predecir una variable en función de la otra. Usaremos como ejemplo las variables `altura` y `peso`.

5.1. Análisis gráfico

Obtendremos el diagrama de dispersión de estas dos variables usando la función `plot`. Dado que nuestro objetivo es usar `altura` como variable de entrada y `peso` como variable de salida, usaremos `altura` como x y `peso` como y. En general, si el propósito es generar el gráfico, esta distinción es arbitraria. El gráfico resultante es

```
plot(AlumnosIndustriales$altura, AlumnosIndustriales$peso, xlab = "Altura", ylab = "Peso")
```



donde podemos ver que la relación entre las dos variables es razonablemente lineal y fuerte. Por lo tanto, parece razonable usar la línea de regresión para predecir y como función de x. Es evidente que el peso de una persona depende de otros factores además de su altura. En el último tema de la asignatura, veremos como incorporar más variables en el modelo de regresión.

5.2. Medidas características bivariadas

Para calcular las medidas características que resumen esta relación lineal, podemos usar las siguientes instrucciones:

```
cov(AlumnosIndustriales$altura, AlumnosIndustriales$peso)
```

```
## [1] 74.49642
```

```
cov(AlumnosIndustriales[,2:3])
```

```
##          altura      peso
## altura 67.68466  74.49642
## peso   74.49642 139.24367
```

La primera línea proporciona la covarianza entre las dos variables y la segunda la matriz de covarianza. Usando la información contenida en esta matriz, podríamos obtener el coeficiente de correlación y los términos

que determinan la línea de regresión. Por ejemplo, el coeficiente de correlación de las dos variables viene dado por

$$corr(x,y) = \frac{cov(x,y)}{s_x s_y} = \frac{74.49642}{\sqrt{67.68466}\sqrt{139.24367}} = 0.767366.$$

Esta correlación coincide con el resultado calculado por R:

```
cor(AlumnosIndustriales$altura, AlumnosIndustriales$peso)
```

```
## [1] 0.767366
```

```
cor(AlumnosIndustriales[,2:3])
```

```
##          altura     peso
## altura 1.000000 0.767366
## peso   0.767366 1.000000
```

Un diagrama de dispersión razonablemente lineal y un coeficiente de correlación alto implican que la recta de regresión será adecuada para hacer predicciones (aunque mejorable si incluimos más variables).

5.3. Recta de regresión

Para calcular la recta de regresión, también llamada regresión simple (esto se debe al hecho de que solo tenemos una variable independiente), podemos usar

```
RegressionModel <- lm(peso ~ altura, data=AlumnosIndustriales)
print(RegressionModel)
```

```
##
## Call:
## lm(formula = peso ~ altura, data = AlumnosIndustriales)
##
## Coefficients:
## (Intercept)       altura
##    -124.426        1.101
```

La relación que queremos calcular es la recta de Mínimo Cuadrado

$$\widehat{y_i} = a + bx_i,$$

donde $b = \frac{cov(x,y)}{s_x^2}$ y $a = \bar{y} - b\bar{x}$.

Los valores de a y b que calculan las instrucciones anteriores son $a = -124.426$ y $b = 1.101$.

La siguiente instrucción proporciona un resumen detallado del modelo de regresión obtenido.

```
summary(RegressionModel)
```

```
##
## Call:
## lm(formula = peso ~ altura, data = AlumnosIndustriales)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.6887  -4.6823   0.0222   3.8113  25.3113
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.42643   16.67148  -7.463 4.41e-11 ***
## altura         1.10064    0.09537  11.541  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.607 on 93 degrees of freedom
## Multiple R-squared:  0.5889, Adjusted R-squared:  0.5844
## F-statistic: 133.2 on 1 and 93 DF,  p-value: < 2.2e-16
```

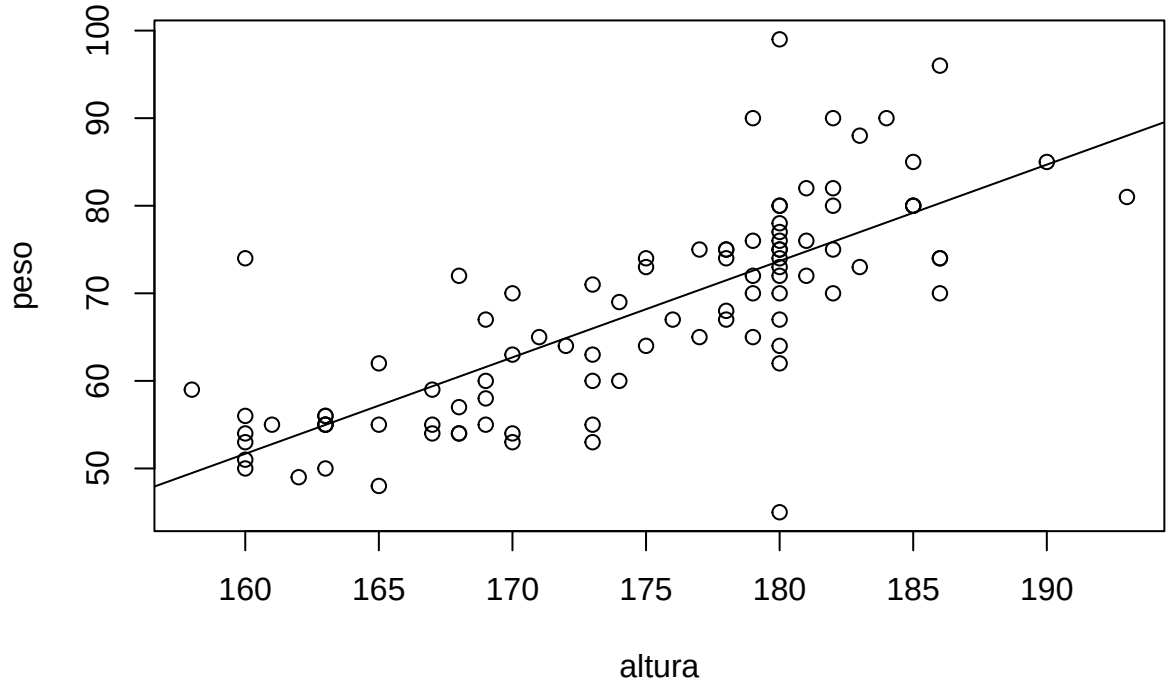
Sin embargo, en este momento, solo nos interesan los valores correspondientes a la columna `Estimate` (Estimaciones). El parámetro b corresponden al coeficiente asociado a la variable `altura`, que es la pendiente de la recta de regresión. El parámetro a es el coeficiente `Intercept` (Intercepto), que es el punto de intersección de la recta con el eje y).

Nuestra recta de regresión está finalmente dada por

$$\texttt{peso} = -124.42643 + 1.10064 \times \texttt{altura}.$$

La recta de regresión se puede trazar usando

```
plot(AlumnosIndustriales[,2:3])
abline(RegressionModel)
```

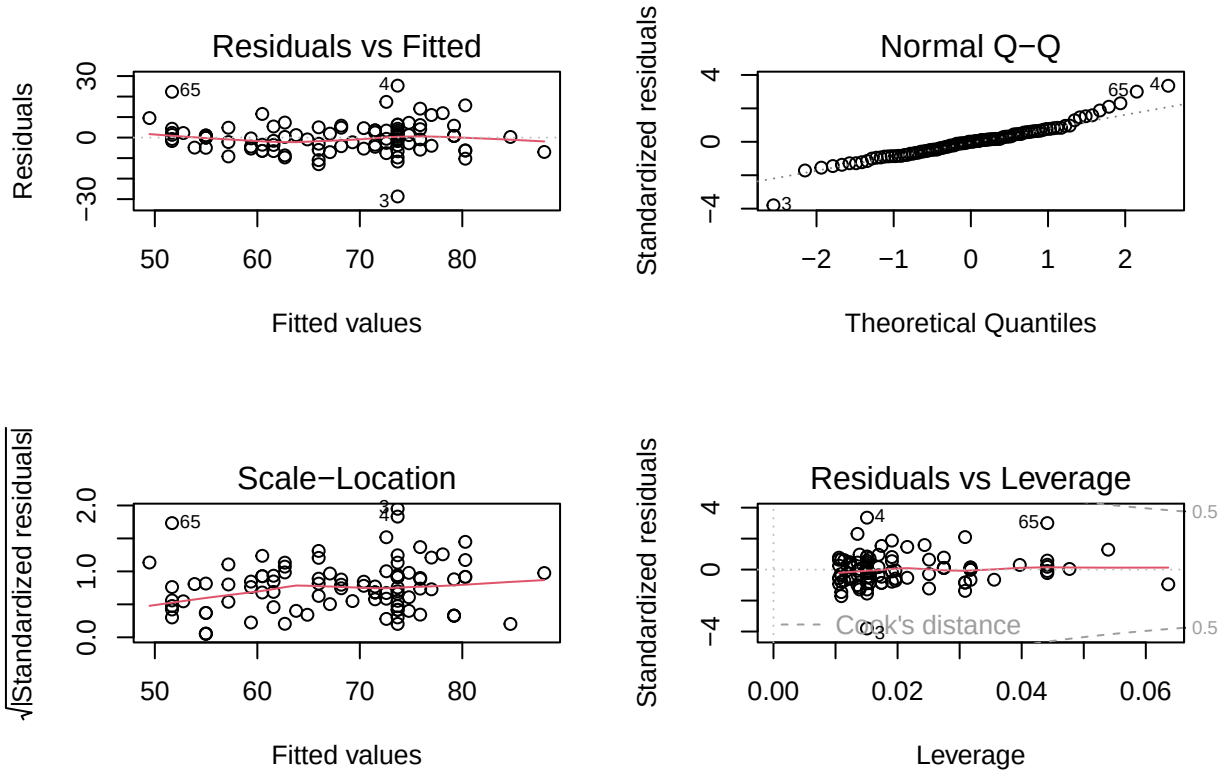


La recta de regresión junto con la nube de puntos muestra que la precisión de la predicción es razonablemente buena para casi todos los valores de las alturas.

El coeficiente de determinación, R^2, que se muestra en la salida `summary` es $R^2 = 0.5889$. Por lo tanto, si suponemos que la relación lineal entre las dos variables es aceptable, podemos decir que la variabilidad del peso se puede explicar en aproximadamente un 59% por la altura de la persona. Es decir, tenemos un predictor aceptable (que mejoraremos si incluimos otras variables=.

Para hacer un diagnóstico de linealidad para este modelo, mostramos el gráfico de residuos frente a predicciones, que está disponible en

```
par(mfrow=c(2,2))
plot(RegressionModel)
```



El gráfico de residuos frente a valores ajustados que obtenemos es razonable, no se ve ningún patrón que sugiera no linealidad.