Tarea HE Big Data-1

David Sánchez y Adrián Díaz

3 de enero de 2018

1. CÁLCULO DE NUEVAS VARIABLES, RECODIFICACIÓN Y FILTRADO

1.1. Importar el fichero JaenIndicadores.txt y denominar a la hoja de datos (data frame) Datos.Jaen

```
Datos.Jaen <- read.delim2("E:/datos david.sanchez/Dropbox/00 PERSONAL/56
MASTER BIG DATA/01 ASIGNATURAS/03 Herramientas Estadisticas Para/Tarea
sesion parte 1/JaenIndicadores.txt", stringsAsFactors=FALSE)
as.numeric(Datos.Jaen$Población)
```

```
##  [1]   1474  21523  11261    573  37903   5696   3951   1921  15276
17672
## [11]   2700   8061   3161   1964    565   2229   3063   2119   2117
683
## [21]  15029   3684   5036    826   8394   1317   1818    948    770
1858
## [31]   3340   1194    674   2029   2804   1893    691    493    702
6054
## [41]   2741   3063   1853   1283   2533 111406   3344   1546  11979
516
## [51]  57796   3993   1052   9270   7431  22702   8360   1974   5066
2231
## [61]   2204   5176   3085   7042   5483   1979   2654   6033   3753
4181
## [71]   1019    911   4826   2332   2159   2567   1301   2992  13327
13848
## [81]   7266   1773   1020  32764   4501   4989  11073   3308   8691
1246
## [91]   4772    446   1528   3253   4336   4338     NA
```

```
Datos.Jaen$Población<-as.integer(Datos.Jaen$Población)
Datos.Jaen$Consumo.de.energía.eléctrica<-
as.integer(Datos.Jaen$Consumo.de.energía.eléctrica)
Datos.Jaen$Consumo.de.agua..Invierno<-
as.integer(Datos.Jaen$Consumo.de.agua..Invierno)
Datos.Jaen$Consumo.de.agua..Verano<-
as.integer(Datos.Jaen$Consumo.de.agua..Verano)

Datos.Jaen$Residuos.sólidos.urbanos..Cantidad<-
gsub(",",".",Datos.Jaen$Residuos.sólidos.urbanos..Cantidad)
Datos.Jaen$Residuos.sólidos.urbanos..Cantidad<-
```

```r
as.numeric(Datos.Jaen$Residuos.sólidos.urbanos..Cantidad)

summary(Datos.Jaen)
```

```
##    CodigoINE      Municipio         Consumo.de.energía.eléctrica
##  Min.   :23001   Length:97          Min.   :   463
##  1st Qu.:23028   Class :character   1st Qu.:  3316
##  Median :23053   Mode  :character   Median :  6978
##  Mean   :23094                      Mean   : 22115
##  3rd Qu.:23079                      3rd Qu.: 14978
##  Max.   :23905                      Max.   :349561
##                                     NA's   :1
##  Consumo.de.agua..Invierno Consumo.de.agua..Verano
##  Min.   :  50.0            Min.   :   89
##  1st Qu.: 312.8            1st Qu.:  480
##  Median : 572.0            Median :  820
##  Mean   :1102.1            Mean   : 1488
##  3rd Qu.:1129.2            3rd Qu.: 1656
##  Max.   :8896.0            Max.   :10326
##  NA's   :3                 NA's   :3
##  Residuos.sólidos.urbanos..Destino Residuos.sólidos.urbanos..Cantidad
##  Length:97                         Min.   :  113.5
##  Class :character                  1st Qu.:  377.3
##  Mode  :character                  Median :  602.3
##                                    Mean   : 1872.7
##                                    3rd Qu.: 1329.9
##                                    Max.   :39197.5
##                                    NA's   :1
##    Población
##  Min.   :   446
##  1st Qu.:  1716
##  Median :  3028
##  Mean   :  6727
##  3rd Qu.:  5780
##  Max.   :111406
##  NA's   :1
```

```r
head(Datos.Jaen)
```

```
##   CodigoINE           Municipio Consumo.de.energía.eléctrica
## 1     23001 Albanchez de Mágina                         2165
## 2     23002      Alcalá la Real                        93991
## 3     23003           Alcaudete                        34985
## 4     23004        Aldeaquemada                          853
## 5     23005             Andújar                       139971
## 6     23006              Arjona                        12576
##   Consumo.de.agua..Invierno Consumo.de.agua..Verano
## 1                       298                    400
## 2                      4882                   6342
## 3                      1537                   2633
## 4                       123                    500
```

```
## 5                      8896                    10326                      
## 6                      1134                     2542                      
##   Residuos.sólidos.urbanos..Destino Residuos.sólidos.urbanos..Cantidad
## 1              Vertedero controlado                              370.49
## 2                        Compostaje                             6774.11
## 3                        Compostaje                             3680.95
## 4              Vertedero controlado                              113.53
## 5              Vertedero controlado                             11775.50
## 6              Vertedero controlado                              1222.79
##   Población
## 1      1474
## 2     21523
## 3     11261
## 4       573
## 5     37903
## 6      5696
```

1.2. Recodificar la variable Poblacion en una variable cualitativa tipo factor llamada Tamaño con tres categorías:

Si la población es inferior a 2000 habitantes, Tamaño será "Pequeño".
Si la población está entre 2000 y 4500 habitantes, Tamaño será "Mediano".
Si la población es superior a 4500 habitantes, Tamaño será "Grande".

```
Datos.Jaen$Tamaño[ Datos.Jaen$Población < 2000 ] <- "Pequeño"
Datos.Jaen$Tamaño[ Datos.Jaen$Población > 2000 & Datos.Jaen$Población <=
4500 ] <- "Mediano"
Datos.Jaen$Tamaño[ Datos.Jaen$Población > 4500 ] <- "Grande"

Datos.Jaen$Tamaño<-as.factor(Datos.Jaen$Tamaño)
```

1.3. Calcular los siguientes promedios que se especifican a continuación y añadirlos como nuevas variables al fichero Datos.Jaen obtenidas a partir de las variables existentes:

Variable elec.hab que contendrá el consumo de energía eléctrica por habitante, obtenida como Consumo.de.energia.electrica/Poblacion

Variable agua.hab que contendrá el consumo medio de agua por habitante y día, obtenida como (Consumo.de.agua..Invierno + Consumo.de.agua..V erano)/Poblacion

Variable res.hab que contendrá los residuos sólidos urbanos por habitante, obtenida como Residuos.solidos.urbanos..Cantidad/Poblacion

```
Datos.Jaen$elec.hab<-
(Datos.Jaen$Consumo.de.energía.eléctrica/Datos.Jaen$Población)
Datos.Jaen$agua.hab<-((Datos.Jaen$Consumo.de.agua..Invierno +
Datos.Jaen$Consumo.de.agua..Verano)/
                        Datos.Jaen$Población)
Datos.Jaen$res.hab<-
(Datos.Jaen$Residuos.sólidos.urbanos..Cantidad/Datos.Jaen$Población)
```

1.4. Crear una nueva hoja de datos con todas las variables que contiene actualmente el data frame Datos.Jaen, pero referida sólo a los municipios de tamaño mediano y denominarla Datos.Jaen.Medianos

```
Datos.Jaen.Medianos<-subset(Datos.Jaen, (Tamaño=="Mediano"))
```

1.5. Guardar la hoja de datos Datos.Jaen con las nuevas variables creadas en los apartados anteriores y la hoja que contiene los datos de las poblaciones medianas (Datos.Jaen.Medianos) en un archivo de datos de R y llamadlo JaenIndicadores.RData

```
save.image("E:/datos david.sanchez/Dropbox/00 PERSONAL/56 MASTER BIG DATA/01 ASIGNATURAS/03 Herramientas Estadisticas Para/Tarea sesion parte 1/JaenIndicadores.RData")
```

---

2. ANÁLISIS ESTADÍSTICO DESCRIPTIVO DE DATOS

Instalar el paquete Hmisc si es preciso

```
if(!is.element('e1071', installed.packages())) install.packages('e1071', repos = 'https://cran.rediris.es/', dependencies = T)
```

Cargar paquete

```
library(e1071)
```

2.1. Importar el fichero Andalucia.txt y denominar a la hoja de datos (data frame) Datos.Andalucia. Comprobar si en el archivo .txt hay datos faltantes y cómo están codificados.

```
Datos.Andalucia <- read.delim2("E:/datos david.sanchez/Dropbox/00 PERSONAL/56 MASTER BIG DATA/01 ASIGNATURAS/03 Herramientas Estadisticas Para/Tarea sesion parte 1/Andalucia.txt", stringsAsFactors=FALSE)

Datos.Andalucia$Poblacion.2001 <-
as.integer(Datos.Andalucia$Poblacion.2001)

summary(Datos.Andalucia)
```

```
##    Codigo.INE     Municipio         Tasa.actividad.2001
Lineas.ADSL.2007
##  Min.   : 4001   Length:770         Min.   :26.92       Min.   :
0.0
##  1st Qu.:14047   Class :character   1st Qu.:46.73       1st Qu.:
26.0
##  Median :18910   Mode  :character   Median :51.88       Median :
181.0
##  Mean   :20945                      Mean   :51.44       Mean   :
1075.8
##  3rd Qu.:29013                      3rd Qu.:56.22       3rd Qu.:
723.5
##  Max.   :41903                      Max.   :74.21       Max.
```

```
:73274.0
##
##  Edad.media.2007 Renta.familiar.por.habitante.2003
##  Min.   :31.00   Length:770
##  1st Qu.:38.20   Class :character
##  Median :41.05   Mode  :character
##  Mean   :41.46
##  3rd Qu.:44.40
##  Max.   :60.80
##
##  Crecimiento.vegetativo.2006 Numero.parados.2007 Poblacion.2007
##  Min.   : -71.0              Min.   :    0.0     Min.   :    50
##  1st Qu.:  -6.0              1st Qu.:   36.0     1st Qu.:  1014
##  Median :   0.0              Median :  108.0     Median :  2761
##  Mean   :  42.7              Mean   :  639.4     Mean   : 10467
##  3rd Qu.:  23.0              3rd Qu.:  346.2     3rd Qu.:  7050
##  Max.   :2082.0              Max.   :45968.0     Max.   :699145
##
##  Poblacion.2006   Poblacion.2003   Poblacion.2001
##  Min.   :    44   Min.   :    50   Min.   :    61
##  1st Qu.:  1010   1st Qu.:  1007   1st Qu.:  1017
##  Median :  2720   Median :  2576   Median :  2638
##  Mean   : 10358   Mean   :  9879   Mean   :  9628
##  3rd Qu.:  6977   3rd Qu.:  6828   3rd Qu.:  6532
##  Max.   :704414   Max.   :709975   Max.   :702520
##                                    NA's   :1
```

```
head(Datos.Andalucia)
```

```
##   Codigo.INE Municipio Tasa.actividad.2001 Lineas.ADSL.2007
## 1       4001      Abla               47.05               44
## 2       4002  Abrucena               49.42               20
## 3       4003      Adra               62.10             2200
## 4       4004 Albánchez               43.66               38
## 5       4005 Alboloduy               51.50               15
## 6       4006     Albox               52.86             1128
##   Edad.media.2007 Renta.familiar.por.habitante.2003
## 1            44.3               Entre 8.300 y 9.300
## 2            44.4               Entre 8.300 y 9.300
## 3            36.0              Entre 9.300 y 10.200
## 4            50.1                                ..
## 5            48.3                                ..
## 6            40.0              Entre 9.300 y 10.200
##   Crecimiento.vegetativo.2006 Numero.parados.2007 Poblacion.2007
## 1                          -7                  49           1514
## 2                          -5                  39           1379
## 3                         131                1159          23742
## 4                          -7                   7            697
## 5                          -3                  26            728
## 6                          31                 361          11166
```

```
##   Poblacion.2006 Poblacion.2003 Poblacion.2001
## 1           1505           1480           1517
## 2           1339           1391           1437
## 3          23545          21704          21810
## 4            660            638            575
## 5            727            765            800
## 6          11000          10409           9661
```

2.2.1. A partir de la variable código INE, construir una variable tipo factor que distinga la provincia de pertenencia de cada municipio, denominarla "Provincia" y añadirla al data frame.

```
Datos.Andalucia$Provincia [Datos.Andalucia$Codigo.INE<= 4999 ] <-
"Almeria"
Datos.Andalucia$Provincia [Datos.Andalucia$Codigo.INE >= 11000 &
Datos.Andalucia$Codigo.INE <= 11999  ] <-"Cádiz"
Datos.Andalucia$Provincia [Datos.Andalucia$Codigo.INE >= 14000 &
Datos.Andalucia$Codigo.INE <= 14999  ] <-"Córdoba"
Datos.Andalucia$Provincia [Datos.Andalucia$Codigo.INE >= 18000 &
Datos.Andalucia$Codigo.INE <= 18999  ] <-"Granada"
Datos.Andalucia$Provincia [Datos.Andalucia$Codigo.INE >= 21000 &
Datos.Andalucia$Codigo.INE <= 21999  ] <-"Huelva"
Datos.Andalucia$Provincia [Datos.Andalucia$Codigo.INE >= 23000 &
Datos.Andalucia$Codigo.INE <= 23999  ] <-"Jaén"
Datos.Andalucia$Provincia [Datos.Andalucia$Codigo.INE >= 29000 &
Datos.Andalucia$Codigo.INE <= 29999  ] <-"Málaga"
Datos.Andalucia$Provincia [Datos.Andalucia$Codigo.INE >= 41000 &
Datos.Andalucia$Codigo.INE <= 41999  ] <-"Sevilla"

Datos.Andalucia$Provincia <-as.factor(Datos.Andalucia$Provincia)

View(Datos.Andalucia)
summary(Datos.Andalucia$Provincia)
```

```
## Almeria   Cádiz Córdoba Granada  Huelva    Jaén  Málaga Sevilla
##     102      44      75     168      79      97     100     105
```

```
names(Datos.Andalucia)
```

```
##  [1] "Codigo.INE"
##  [2] "Municipio"
##  [3] "Tasa.actividad.2001"
##  [4] "Lineas.ADSL.2007"
##  [5] "Edad.media.2007"
##  [6] "Renta.familiar.por.habitante.2003"
##  [7] "Crecimiento.vegetativo.2006"
##  [8] "Numero.parados.2007"
##  [9] "Poblacion.2007"
## [10] "Poblacion.2006"
## [11] "Poblacion.2003"
```

```
## [12] "Poblacion.2001"
## [13] "Provincia"
```

```
str(Datos.Andalucia)
```

```
## 'data.frame':    770 obs. of  13 variables:
##  $ Codigo.INE                       : int  4001 4002 4003 4004 4005
4006 4007 4008 4009 4010 ...
##  $ Municipio                        : chr  "Abla" "Abrucena" "Adra"
"Albánchez" ...
##  $ Tasa.actividad.2001              : num  47 49.4 62.1 43.7 51.5 ...
##  $ Lineas.ADSL.2007                 : int  44 20 2200 38 15 1128 18 0
0 6 ...
##  $ Edad.media.2007                  : num  44.3 44.4 36 50.1 48.3 40
44.8 49.3 50.9 41.3 ...
##  $ Renta.familiar.por.habitante.2003: chr  "Entre 8.300 y 9.300"
"Entre 8.300 y 9.300" "Entre 9.300 y 10.200" ".." ...
##  $ Crecimiento.vegetativo.2006      : int  -7 -5 131 -7 -3 31 -5 -4 0
3 ...
##  $ Numero.parados.2007              : int  49 39 1159 7 26 361 38 12 0
31 ...
##  $ Poblacion.2007                   : int  1514 1379 23742 697 728
11166 957 611 154 710 ...
##  $ Poblacion.2006                   : int  1505 1339 23545 660 727
11000 967 617 142 703 ...
##  $ Poblacion.2003                   : int  1480 1391 21704 638 765
10409 1018 627 152 676 ...
##  $ Poblacion.2001                   : int  1517 1437 21810 575 800
9661 906 656 193 683 ...
##  $ Provincia                        : Factor w/ 8 levels
"Almeria","Cádiz",..: 1 1 1 1 1 1 1 1 1 ...
```

```
str(Datos.Andalucia$Provincia)
```

```
##  Factor w/ 8 levels "Almeria","Cádiz",..: 1 1 1 1 1 1 1 1 1 ...
```

```
Datos.Andalucia$Provincia
```

```
##   [1] Almeria Almeria Almeria Almeria Almeria Almeria Almeria Almeria
##   [9] Almeria Almeria Almeria Almeria Almeria Almeria Almeria Almeria
##  [17] Almeria Almeria Almeria Almeria Almeria Almeria Almeria Almeria
##  [25] Almeria Almeria Almeria Almeria Almeria Almeria Almeria Almeria
##  [33] Almeria Almeria Almeria Almeria Almeria Almeria Almeria Almeria
##  [41] Almeria Almeria Almeria Almeria Almeria Almeria Almeria Almeria
##  [49] Almeria Almeria Almeria Almeria Almeria Almeria Almeria Almeria
##  [57] Almeria Almeria Almeria Almeria Almeria Almeria Almeria Almeria
##  [65] Almeria Almeria Almeria Almeria Almeria Almeria Almeria Almeria
##  [73] Almeria Almeria Almeria Almeria Almeria Almeria Almeria Almeria
##  [81] Almeria Almeria Almeria Almeria Almeria Almeria Almeria Almeria
##  [89] Almeria Almeria Almeria Almeria Almeria Almeria Almeria Almeria
##  [97] Almeria Almeria Almeria Almeria Almeria Almeria Cádiz   Cádiz
```

```
## [105] Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz
## [113] Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz
## [121] Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz
## [129] Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz
## [137] Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz   Cádiz
## [145] Cádiz   Cádiz   Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba
## [153] Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba
## [161] Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba
## [169] Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba
## [177] Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba
## [185] Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba
## [193] Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba
## [201] Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba
## [209] Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba Córdoba
## [217] Córdoba Córdoba Córdoba Córdoba Córdoba Granada Granada Granada
## [225] Granada Granada Granada Granada Granada Granada Granada Granada
## [233] Granada Granada Granada Granada Granada Granada Granada Granada
## [241] Granada Granada Granada Granada Granada Granada Granada Granada
## [249] Granada Granada Granada Granada Granada Granada Granada Granada
## [257] Granada Granada Granada Granada Granada Granada Granada Granada
## [265] Granada Granada Granada Granada Granada Granada Granada Granada
## [273] Granada Granada Granada Granada Granada Granada Granada Granada
## [281] Granada Granada Granada Granada Granada Granada Granada Granada
## [289] Granada Granada Granada Granada Granada Granada Granada Granada
## [297] Granada Granada Granada Granada Granada Granada Granada Granada
## [305] Granada Granada Granada Granada Granada Granada Granada Granada
## [313] Granada Granada Granada Granada Granada Granada Granada Granada
## [321] Granada Granada Granada Granada Granada Granada Granada Granada
## [329] Granada Granada Granada Granada Granada Granada Granada Granada
## [337] Granada Granada Granada Granada Granada Granada Granada Granada
## [345] Granada Granada Granada Granada Granada Granada Granada Granada
## [353] Granada Granada Granada Granada Granada Granada Granada Granada
## [361] Granada Granada Granada Granada Granada Granada Granada Granada
## [369] Granada Granada Granada Granada Granada Granada Granada Granada
## [377] Granada Granada Granada Granada Granada Granada Granada Granada
## [385] Granada Granada Granada Granada Granada Huelva  Huelva  Huelva
## [393] Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva
## [401] Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva
## [409] Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva
## [417] Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva
## [425] Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva
## [433] Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva
## [441] Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva
## [449] Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva
## [457] Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva  Huelva
## [465] Huelva  Huelva  Huelva  Huelva  Jaén    Jaén    Jaén    Jaén
## [473] Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén
## [481] Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén
## [489] Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén
## [497] Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén
```

```
## [505] Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén
## [513] Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén
## [521] Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén
## [529] Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén
## [537] Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén
## [545] Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén
## [553] Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén    Jaén
## [561] Jaén    Jaén    Jaén    Jaén    Jaén    Málaga  Málaga  Málaga
## [569] Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga
## [577] Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga
## [585] Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga
## [593] Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga
## [601] Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga
## [609] Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga
## [617] Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga
## [625] Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga
## [633] Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga
## [641] Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga
## [649] Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga
## [657] Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga  Málaga
## [665] Málaga  Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [673] Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [681] Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [689] Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [697] Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [705] Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [713] Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [721] Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [729] Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [737] Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [745] Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [753] Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [761] Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla Sevilla
## [769] Sevilla Sevilla
## Levels: Almeria Cádiz Córdoba Granada Huelva Jaén Málaga Sevilla
```

```
which.max(Datos.Andalucia$Provincia)
```

```
## [1] 666
```

```
which.min(Datos.Andalucia$Provincia)
```

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

2.2.2 Obtener la distribución de frecuencias absolutas y relativas.

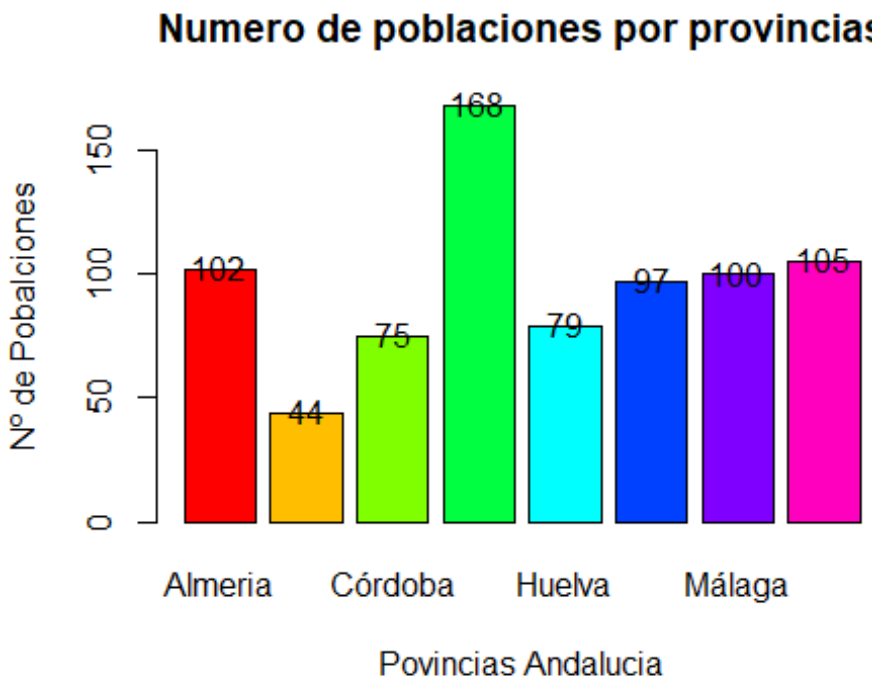
```
Frec.abosolutas <- table(Datos.Andalucia$Provincia)
Frec.abosolutas
```

```
##
## Almeria   Cádiz Córdoba Granada  Huelva    Jaén  Málaga Sevilla
##     102      44      75     168      79      97     100     105
```

```
Frec.rel <- prop.table(Frec.abosolutas)
Frec.rel

##
##    Almeria      Cádiz    Córdoba    Granada     Huelva       Jaén
## 0.13246753 0.05714286 0.09740260 0.21818182 0.10259740 0.12597403
##     Málaga    Sevilla
## 0.12987013 0.13636364

Frec.rel <- round(Frec.rel*100, 2)
Frec.rel

##
## Almeria    Cádiz Córdoba Granada  Huelva    Jaén  Málaga Sevilla
##   13.25     5.71    9.74   21.82   10.26   12.60   12.99   13.64
```

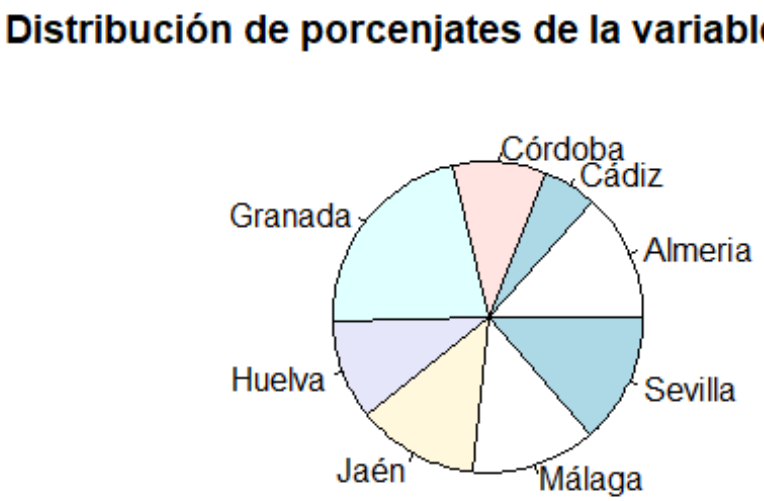
2.2.3 Un diagrama de barras con las frecuencias absolutas

```
barras.frec.abosolutas <-barplot(Frec.abosolutas, col = rainbow(8),
xlab="Povincias Andalucia", ylab = "Nº de Pobalciones")
text(barras.frec.abosolutas,Frec.abosolutas + 1,labels=Frec.abosolutas,
xpd = TRUE)
title(main = "Numero de poblaciones por provincias")
```

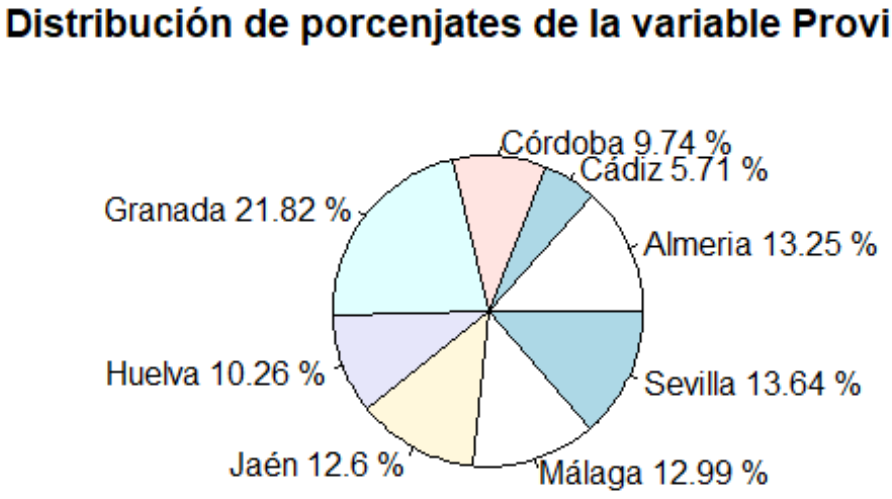


2.2.4. Un diagrama de sectores con las frecuencias relativas en porcentajes de esta variable tipo factor.

```
sectores.frec.rel <- pie(Frec.rel,labels = names(Frec.rel),main =
"Distribución de porcenjates de la variable Provincia")
```




```
sectores.frec.rel.por <- pie(Frec.rel,labels =
paste(names(Frec.rel),Frec.rel,"%"),main = "Distribución de porcenjates
de la variable Provincia")
```



¿Qué provincia tiene más municipios?
Granada con 168 municipios

¿Cual tiene menos?
Cadiz con 44 municipios

¿Qué porcentaje representa en cada caso?
Grandda un 21.82% y Cádiz un 5.71%

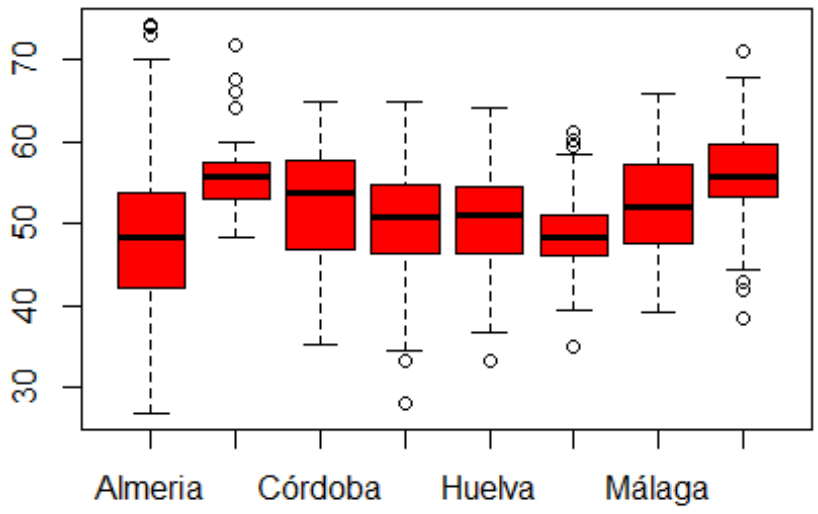
2.3. Obtener un resumen descriptivo de la variable tasa de actividad de 2001 que incluya:

Parámetros de posición.

```
summary(Datos.Andalucia$Tasa.actividad.2001)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.92   46.73   51.88   51.44   56.22   74.21

fivenum(Datos.Andalucia$Tasa.actividad.2001)

## [1] 26.920 46.720 51.875 56.220 74.210
```

Dispersión

```
plot (Datos.Andalucia$Provincia, Datos.Andalucia$Tasa.actividad.2001,
type = "p", col = "red")
```



Asimetría

```
skewness(Datos.Andalucia$Tasa.actividad.2001,na.rm = TRUE)

## [1] -0.09917887
```

Curtosis

```
kurtosis(Datos.Andalucia$Tasa.actividad.2001,na.rm = TRUE)

## [1] 0.1241295
```

Histograma

```
hist(Datos.Andalucia$Tasa.actividad.2001, breaks = 10, freq = TRUE, main
= "Historigrama de la Tasa de actividad del 200,",xlab = "Tasa 2001",ylab
= "Frecuencias", col = "lightblue", border = "blue")
```

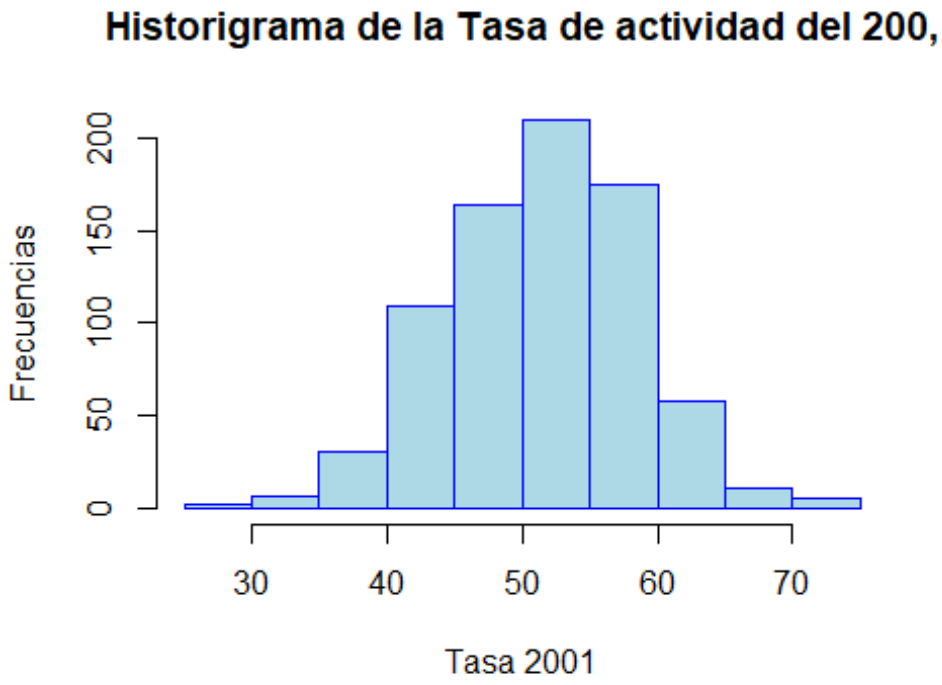
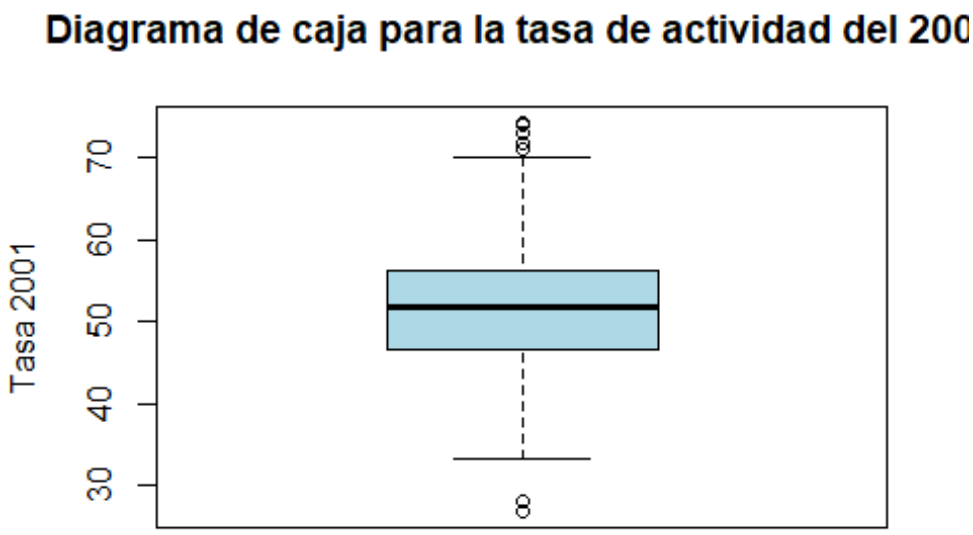


Diagrama de caja.

```
boxplot(Datos.Andalucia$Tasa.actividad.2001, main="Diagrama de caja para
la tasa de actividad del 2001",ylab="Tasa 2001", col="lightblue")
```



En función de este resumen, contestar a las siguientes preguntas:

2.3.1. ¿Cuál es la tasa media de actividad de los municipios andaluces?

51.44

¿Crees que este valor es adecuado para representar la Tasa de Actividad de los municipios andaluces durante 2001?

Si

2.3.2. ¿Cómo valoras la homogeneidad de los valores de la tasa de actividad en los municipios andaluces? ¿Qué parámetro elegirías para representar la dispersión de la Tasa de Actividad de 2001?

La varianza y la desviación típica.

2.3.3. ¿En ese sentido, qué municipios andaluces destacan significativamente del resto (como atípicos) por su alta tasa de actividad y por su baja tasa de actividad?

Sevilla por alta y Almeria por baja.

¿Se te ocurre alguna explicación al respecto?

2.3.4. ¿Cómo valoras la simetría de la distribución de frecuencias?

---

3. DISTRIBUCIONES DE PROBABILIDAD

3.1. Consideremos una variable aleatoria que sigue una distribución B (15, 0.33). Se pide:

3.1.1. ¿Qué valor de la variable deja por debajo de sí el 75% de la probabilidad?

```
qbinom(0.75, 15, 0.33)
```

```
## [1] 6
```

3.1.2. Calcular el percentil 95% de la distribución.

```
qbinom(0.95, 15, 0.33)
```

```
## [1] 8
```

3.1.3. Obtener una muestra de tamaño 1000 de esta distribución,

```
rbinom (1000,15,0.33)
```

```
##    [1]  5  7  5 10  7  4  5  2  6  6  5  6  2  2  7  2  3  7  6  4  4
8  2
##   [24]  4  5  5  5  5  5  3  4 10  5  5  7  2  6  4  6  3  6  3  3  1
3  6
##   [47]  2  5  4  2  4  5  7  4  4  6  5  5  5  6  2  7  4  3  4  5  5
6  5
##   [70]  5  4  8  3  5  5  6  4  5  7  5  7  2  5  5  4  5  4  6  7  8
5  7
##   [93]  5  6  4  3  2  3  6  5  7  6  6  5  3  5  5  6  3  6  3  4  6
5  4
##  [116]  6  6  5  8  5  5  1  8  4  8  6  4  7  5  6  6  4  6  2  9  5
5  4
##  [139]  1  5  3  4  1  4  5  5  5  5  5  6  3  4  6  5  3  5  3  6  8
2  4
##  [162]  5  3  4  5  6  2  9  3  8  5  6  4  4  2  6  6  7  7  2  4  7
5  7
##  [185]  1  3  4  4  3  9  5  6  3  7  6  7  5  5  5  7  3  4  9  6  9
4  8
##  [208]  6  3  5  5  5  5  5  7  3  6  7  6  5  6  4  6  5  5  5  4  4
5  7
##  [231]  4  5  7  6  7  2  4  1  4  3  7  8  7  8  5  5  5  4  4  7  5
2  6
##  [254]  3  5  5  6  3  5  6  3  5  2  5  6  3  5  4  2  4  5  6  3  9
5  4
##  [277]  8  4  5  7  5  5  2  7  4  3  6  5  3  3  5  6  5  7  7  3  4
2  9
##  [300]  4  1  6  5  4  3  6  5  2  5  4  4  4  4  4  5  5  9  6  2  2
5  8
##  [323]  7  3  8  3  7  7  9  4  6  2  5  2  5  4  6  4  4  6  7  4  2
5  9
##  [346]  1  4  5  6  3  3  6  3  3  6  5  4  4  4  3  4  8  2  1  4  3
6  4
```

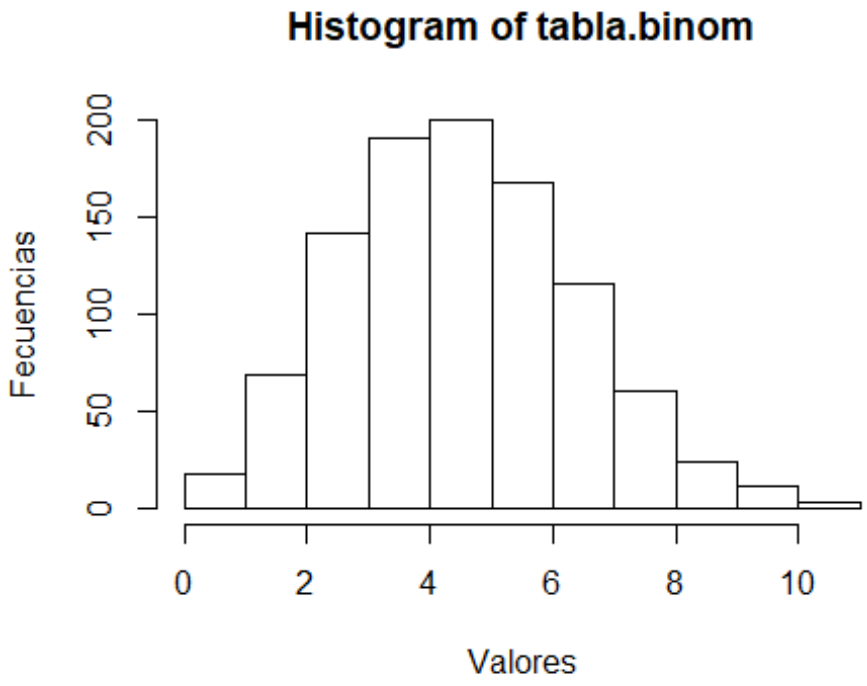
```
##  [369]  7  8  6  3  4  3  5  4  4  4  6  6  8  5  5  6  5  3  9  4  3
6  7
##  [392]  5  3  6  6  6  3  5  4  5  3  3  2  6  5  6 10  4  9  3  4  6
4  5
##  [415]  7  3  2  6  6  5  6  6  7  5  5  4  5  6  6  5  4  6  3  4  3
3  6
##  [438]  8  1  4  5  6  5  3  2  4  2  6  3  4  4  2  5  7  4  6  6  2
5  6
##  [461]  6  6  6  6  6  5  3  4  3  5  6  3  9  2  6  5  3  2  3  5  7
4  3
##  [484]  7  1  6  6  3  9  5  5  8  8  6  3  6  6  6  6  4  3  3  5  8
6  6
##  [507]  7  5  4  4  4  4  5  7  2  7  6  5  3  6  3  6  3  5  5  7  4
4  5
##  [530]  4  6  3  5  1  3  5  8  5  3  6  6  5  6  5  4  6  4  4  9  5
1  4
##  [553]  4  4  6  7  3  5  6  4  5  5  7  8  6  8  3  8  5  4  7  7  6
4  3
##  [576]  4  3  3  9 10  7  3  6  5  4  7  6  4  3  4  6  1  6  5  5  5
7  3
##  [599]  7  3  6  6  7  6  7  9  4  4  4  7  3  4  7  5  5  6  4  6  5
4  3
##  [622]  6  7  4  6  3  7  6  4  4  5  5  4  5  4  7  4  7  4  4  4  4
5  2
##  [645]  8  4  6  7  4  2  6  6  4  6  3  5  4  6  3  7  7  3  4  3  6
7  6
##  [668]  5  6  4  4  6  3  2  7  9  7  6  3  3  5  4  5  6  8  6  5  4
5  3
##  [691]  8  3  4  7  5  7  3  2  5  6  4  6  6  5  4  4  5  4  1  6  4
5  2
##  [714]  5  8  8  5  4  7  6  6  1  3  4  1 10  9  5  3  5  4  7  5  5
8  5
##  [737]  4  4  6  6  4  4  3  6  3  4  5  4  4  7  5  7  5  8  2  3  6
5  5
##  [760]  6  6  3  5  6  5  6  7  3  6  3  4  3  8  5  3  4  6  3  4  5
6  6
##  [783]  9  6  3  9  5  6  5  4  5  4  3  6  5  7  5  5  4  3  7  6  9
2  7
##  [806]  3  4  6  4  8  6  7  9  9  4  2  7  6  4  7  5  5  5  3  3  7
6  3
##  [829]  5  3  6  4  2  6  4  4  5  3  4  7  3  7  6  5  2  8  5  5  3
6  6
##  [852]  4  5  6  4  5  4 11  5  6  7  5  4  7  4  7  4  6  4  5  2  4
7  4
##  [875]  4  1  5  5  6  7  3  4  3  5  4  5  2  3  7  4  7  5  5  4  6
3  4
##  [898]  7  6  7  5  5  1  6  5  8  6  6  5  6  7  6  3  3  2  7  6  8
6  2
##  [921]  4  6  6  5  4  1  2  5  4  3  8  5  6  2  5  4  7  2  2  7  4
4  8
```

```
##  [944]  5  9  5  4  8  7  7  1  4 10  7  6  2  5  3  4  4  3  7  2  3
7  4
##  [967]  6  5  3  4  4  3  7  4  6  5  4  5  5  3  3  6  5  4  9  5  3
4  5
##  [990]  4  2  6  4  4  5  5  5  7  4  4
```

Representarla gráficamente las frecuencias observadas de cada valor de la distribución mediante un diagrama de barras

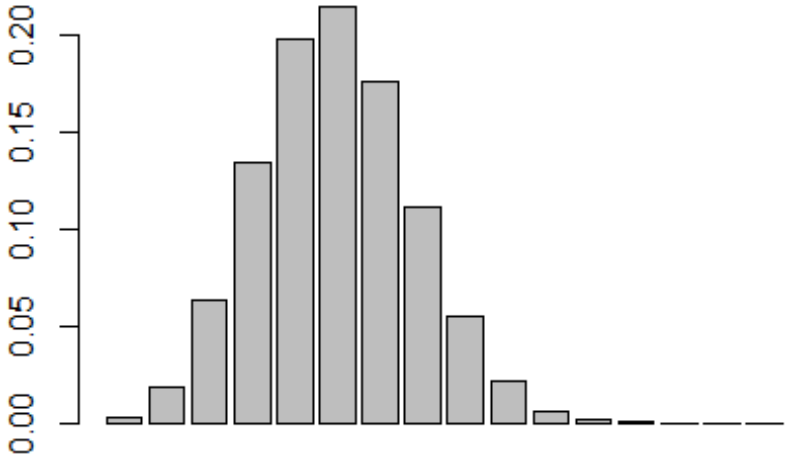
Comparar éste con las frecuencias esperadas según el modelo que genera los datos.

```
tabla.binom<-c(rbinom (1000,15,0.33))


hist(tabla.binom, xlab = "Valores", ylab = "Fecuencias")
```



Comparar éste con las frecuencias esperadas según el modelo que genera los datos.

```
barplot(dbinom(0:15,15,0.33))
```



3.2. Consideremos una variable aleatoria W con distribución N (250, 13). Se pide:

3.2.1. P [240 < W ??? 245.5]

```
pnorm(c(245.5),mean = 250,sd = 13)-pnorm(c(240),mean = 250,sd = 13)

## [1] 0.1437354
```

3.2.2. P [W ??? 256].

```
pnorm(256,13,250, lower.tail = F)

## [1] 0.1655253
```

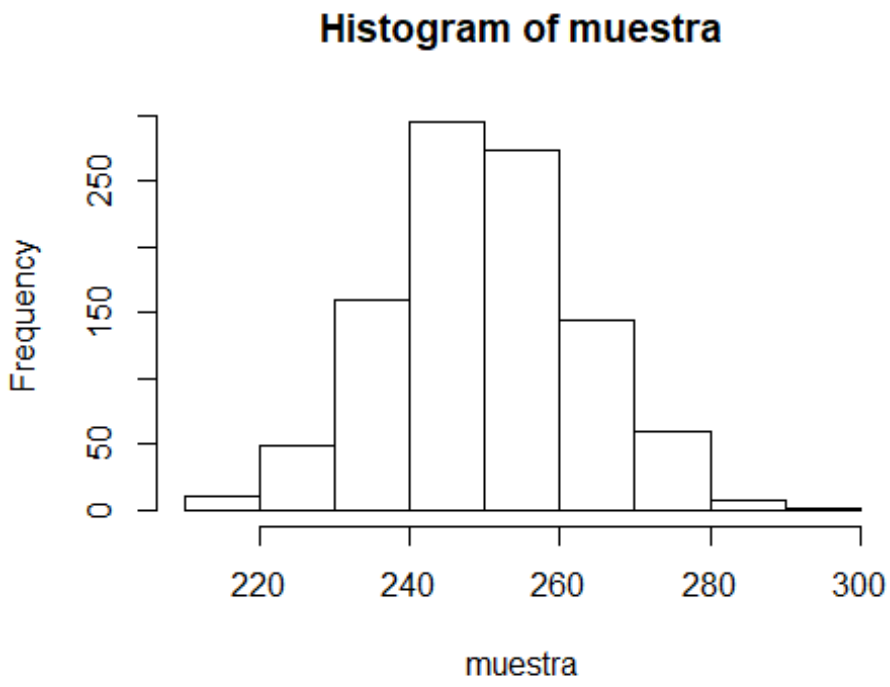
3.2.3. Si queremos desechar el 5% de valores más altos de la distribución y el 5% de valores más bajos, ¿con qué intervalo de valores nos quedaremos?

```
w1 <- qnorm(((1-0.95)/2), 250, 13)
w1

## [1] 224.5205

w2 <- qnorm(((1-0.95)/2), 250, 13, lower.tail = F)
w2

## [1] 275.4795
```

3.2.4. Obtener una muestra de tamaño 1000 de la distribución, representar la función de densidad de esta distribución y compararla con el histograma de la muestra obtenida.

```
n<-1000
m<-250
```

```r
sigma<-13

muestra <- rnorm(n, m, sigma)
hist(muestra)
```



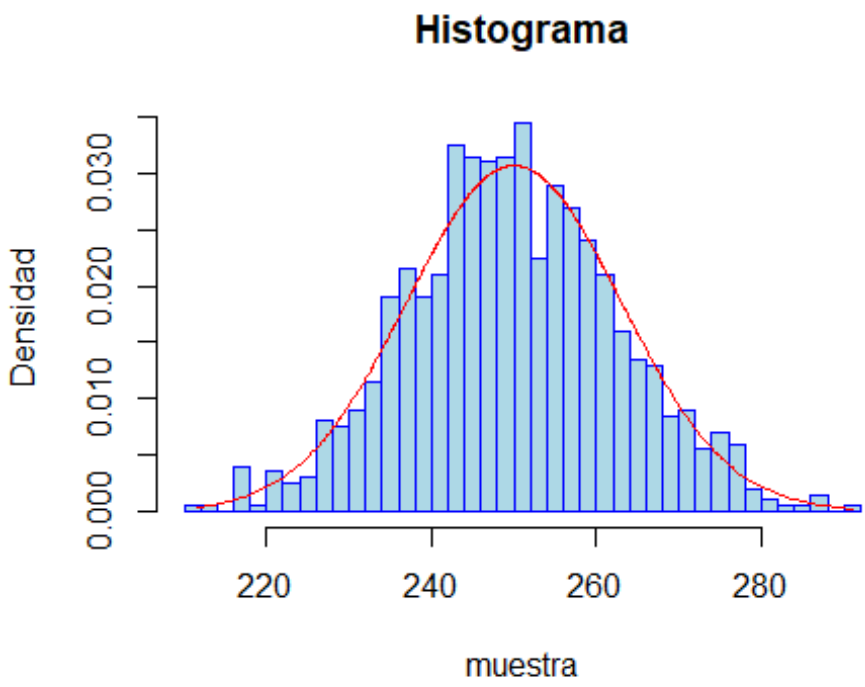

```r
mean(muestra)
```

```
## [1] 249.8254
```

```r
sd(muestra)
```

```
## [1] 12.88361
```

```r
int<-round(sqrt(n), 0)

hist(muestra, breaks=int, freq=F, xlab="muestra", ylab="Densidad",
main="Histograma",
     col="lightblue", border="blue")
lines(sort(muestra), dnorm(sort(muestra),m,sigma), type="l", col="red")
```



---

4. CONTRASTES DE HIPÓTESIS E INTERVALOS DE CONFIANZA

DESCRIPCIÓN DEL DATASET

Mediante una red de sensores se han recogido datos sobre la temperatura media diaria (ºC) en dos estaciones A y B durante 52 días. Los valores recogidos de la temperatura se encuentran en la hoja de datos "Temper" incluida en el fichero Temperatura.RData.

4.1. Cargar el fichero Temperatura.RData.

```
load("Temperatura.RData")
```

4.2. Crear dos nuevas variables, temp.A y temp.B, que contengan las temperaturas de las estaciones A y B, respectivamente.

```
table.A <- subset(Temper, (Estacion=="A"))
temp.A <- table.A$Temper
temp.A
```

```
##  [1] 23.10 22.15 23.87 23.62 23.42 23.10 22.81 22.70 22.47 22.45 22.44
## [12] 22.14 22.13 22.03 21.81 21.35 21.34 21.08 21.08 20.95 20.85 20.82
## [23] 20.61 20.56 20.52 20.35 20.29 20.28 20.12 20.11 20.04 20.01 19.97
## [34] 19.90 19.82 19.73 19.71 19.71 19.70 19.66 19.64 19.59 19.56 19.56
## [45] 19.44 19.42 19.36 19.25 19.13 19.08 19.08 19.02 19.01 18.73 18.69
```

```
## [56] 18.54 18.54 18.50 18.49 18.49 18.45 18.35 18.30 18.20 18.16 18.07
## [67] 17.75 17.58 17.32 17.29 16.96 16.93 16.90 16.83 16.37 16.37 16.21
```

```
table.B <- subset(Temper, (Estacion=="B"))
temp.B <- table.B$Temper
temp.B
```

```
##  [1] 20.19 24.63 23.32 23.21 22.73 22.69 22.59 22.59 22.37 22.35 22.31
## [12] 22.14 21.98 21.92 21.88 21.85 21.77 21.68 21.49 21.47 21.31 21.30
## [23] 21.29 21.22 21.18 20.80 20.75 20.74 20.40 20.40 20.34 20.34 20.28
## [34] 20.28 20.24 20.24 20.05 20.05 20.00 19.97 19.95 19.92 19.85 19.85
## [45] 19.76 19.67 19.65 19.63 19.62 19.48 19.47 19.41 19.25 19.17 19.08
## [56] 18.98 18.96 18.95 18.92 18.75 18.72 18.57 18.43 18.42 18.09 17.95
## [67] 17.86 17.83 17.73 17.71 17.60 17.49 17.42 17.41 17.29 17.17 17.14
## [78] 16.79 16.11
```

4.3. Da un intervalo de confianza para la temperatura media diaria de la estación A, al 95%, y a partir de éste indica si se puede admitir, y por qué, que la temperatura media diaria en dicha estación sea de 19ºC, con ese mismo nivel de confianza.

```
mean(temp.A)
```

```
## [1] 19.81766
```

```
test.tempo.A <- t.test(temp.A, alternative = "two.sided", conf.level =
0.95)
test.tempo.A
```

```
##
##  One Sample t-test
##
## data:  temp.A
## t = 93.167, df = 76, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  19.39401 20.24131
## sample estimates:
## mean of x
##  19.81766
```

4.4. Plantea un test de hipótesis que refleje la pregunta del apartado anterior y resuélvelo sin usar el intervalo de confianza (riesgo de 1ª especie 5%)

```
test.tempo.A.hipo <- t.test(temp.A, alternative = "two.sided", mu = 19,
conf.level = 0.95)
test.tempo.A.hipo
```

```
##
##  One Sample t-test
##
## data:  temp.A
## t = 3.844, df = 76, p-value = 0.0002496
## alternative hypothesis: true mean is not equal to 19
```

```
## 95 percent confidence interval:
##  19.39401 20.24131
## sample estimates:
## mean of x
##  19.81766
```

4.5. Determina si puede admitirse, con un riesgo de primera especie de 1%, que la temperatura media diaria es la misma en las dos estaciones. Plantea previamente el correspondiente contraste de hipótesis.

```
mean(temp.A)

## [1] 19.81766

mean(temp.B)

## [1] 20.00494

var.test.AB <- var.test(temp.A, temp.B, ratio = 1, alternative =
"two.sided", conf.level = 0.99)
var.test.AB

##
##  F test to compare two variances
##
## data:  temp.A and temp.B
## F = 1.0978, num df = 76, denom df = 78, p-value = 0.6825
## alternative hypothesis: true ratio of variances is not equal to 1
## 99 percent confidence interval:
##  0.6071355 1.9889245
## sample estimates:
## ratio of variances
##             1.0978
```

El p-valor es 0.6825, y es mayor que 0,1, por lo que podemos afirmar que la varianza de las temperaturas entre las estaciones no difiere, con un riesgo de 1ª especie del 1%.

4.6. Obtén un intervalo de confianza (99%) para la diferencia de temperaturas entre estaciones. ¿Aporta alguna información adicional al resultado obtenido en el apartado anterior?

```
mean.test.AB <- t.test(temp.A, temp.B, alternative = "two.sided", mu = 0,
paired = F, var.equal = T, conf.level = 0.99)
mean.test.AB

##
##  Two Sample t-test
##
## data:  temp.A and temp.B
## t = -0.64116, df = 154, p-value = 0.5224
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -0.9490744  0.5745257
```

```
## sample estimates:
## mean of x mean of y
##  19.81766  20.00494
```

El p-valor es 0.5224, y es mayor que 0,1, por lo que podemos afirmar que la temperatura medida de la primera estación no difiere de la segunda, con un riesgo de 1ª especie del 1%.

4.7. Se sabe que a lo largo de los 52 días, la estación A falló 5 días y la B 7 días. ¿Puede afirmarse con un nivel de confianza del 90% que la proporción de días fallados es la misma en las dos estaciones?

```
dias <- 52

fallo.A <- 5
fallo.B <- 7

no.fallo.A <- dias-fallo.A
no.fallo.B <- dias-fallo.B

tabla.fallos <- matrix(c(fallo.A, fallo.B, no.fallo.A, no.fallo.B), 2, 2)
tabla.fallos
```

```
##      [,1] [,2]
## [1,]    5   47
## [2,]    7   45
```

```
prop.test(tabla.fallos, alternative = "two.sided", conf.level = 0.9)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  tabla.fallos
## X-squared = 0.094203, df = 1, p-value = 0.7589
## alternative hypothesis: two.sided
## 90 percent confidence interval:
##  -0.16056582  0.08364275
## sample estimates:
##     prop 1     prop 2
## 0.09615385 0.13461538
```

Como p-valor es 0.7589, y es mayor que 0,1, podemos afirmar que la proporción días fallados es la misma en las dos estaciones, con un riesgo de 1ª especie del 10%.