Práctica 1

Fundamentos de la Ciencia de Datos

Samuel Aós Paumard,
Enrique Coronado Barco,
Carmen Martínez Estévez,
Alberto Martínez Ortega

November 3, 2020

1 Ejericio 1

Realización de un ejercicio en clase con ayuda del profesor en el que se va a realizar un análisis con R de descripción de Datos aplicando todos los conceptos vistos en el tema. Para realizar el ejercicio vamos a utilizar dos ficheros de datos:

1.1

El primer fichero de datos será de tipo .txt, es decir, de texto plano, y estará formado por los datos de los satélites menores de Urano1 que hemos utilizado en la descripción teórica del tema. Lo denominaremos satelites.txt. El objetivo es obtener, utilizando R, los valores de las mismas magnitudes cuyo valor hemos calculado de forma manual.

Comenzamos con la carga de datos del fichero satelites.txt de la misma forma que vimos durante la clase.

```
> satelites<-read.table("satelites.txt")
```

Del conjunto de datos extraemos los valores de los radios para su tratamiento, asi como el número de valores que encontramos en esta columna de la tabla de satélites.

```
> Radio<-satelites$Radio
> Radio
```

```
 [1] 13 16 22 33 29 42 27 34 20 30 20 15
```

```
> size_1<-length(Radio)
```

Calculamos las frecuencias para los valores de los radios.

```
> frec_abs_1<-table(Radio)
> frec_abs_1
```

```
Radio
13 15 16 20 22 27 29 30 33 34 42
 1  1  1  2  1  1  1  1  1  1  1
```

```
> frec_abs_acum_1<-cumsum(frec_abs_1)
> frec_abs_acum_1

13 15 16 20 22 27 29 30 33 34 42
 1  2  3  5  6  7  8  9 10 11 12

> frec_rel_1<-table(Radio)/size_1
> frec_rel_1

Radio
       13         15         16         20         22         27         29
0.08333333 0.08333333 0.08333333 0.16666667 0.08333333 0.08333333 0.08333333
       30         33         34         42
0.08333333 0.08333333 0.08333333 0.08333333

> frec_rel_acum_1<-cumsum(frec_rel_1)
> frec_rel_acum_1

       13         15         16         20         22         27         29
0.08333333 0.16666667 0.25000000 0.41666667 0.50000000 0.58333333 0.66666667
       30         33         34         42
0.75000000 0.83333333 0.91666667 1.00000000
```

Calculamos la media y la mediana, mínimos y máximos de los valores, el rango, la desviación tipica y la varianza.

```
> media_1<-mean(Radio)
> media_1

[1] 25.08333

> mediana_1<-median(Radio)
> mediana_1

[1] 24.5

> minimo_1<-min(Radio)
> minimo_1

[1] 13

> maximo_1<-max(Radio)
> maximo_1

[1] 42

> rango_1<-range(Radio)
> rango_1

[1] 13 42

> desv_tip_1<-sqrt((sd(Radio)^2)*(size_1-1)/size_1)
> desv_tip_1
```

```
[1] 8.47996

> varianza_1<-(var(Radio)*(size_1-1))/size_1
> varianza_1

[1] 71.90972
```

Para acabar con el tratamiento de los datos calculamos los cuartiles.

```
> cuart1_1<-quantile(Radio,0.25)
> cuart1_1

25%
 19

> cuart2_1<-quantile(Radio,0.5)
> cuart2_1

 50%
24.5

> cuart3_1<-quantile(Radio,0.75)
> cuart3_1

  75%
30.75

> cuart4_1<-quantile(Radio,1)
> cuart4_1

100%
  42

> cuant54_1<-quantile(Radio,0.54)
> cuant54_1

 54%
26.7
```

1.2

El segundo fichero de datos será de tipo .sav, es decir, un fichero de datos procedente de SPSS. Se denomina cardata.sav y estará formado por datos de automóviles, como su consumo en mpg (millas por galón), cilindrada, aceleración, año de fabricación, modelo, etc.

```
> install.packages("haven")
> library("haven")      # Librería para la tabla de valores estadísticos
> dataset <- read_sav("cardata.sav")
```

De este conjunto de datos extraemos los valores de millas por galon (mpg) y, justo despues, para evitar los posibles errores estadísticos provenientes de valores nulos, limpiamos el conjunto de ellos.

```
> mpg<-dataset$mpg
> mpg<-mpg[!is.na(mpg)]
> size_2<-length(mpg)
```

Tras esto volvemos a repetir las operaciones del ejercicio anterior.
Calculamos las frecuencias para los valores de los radios.

```
> frec_abs_2<-table(mpg)
> frec_abs_2
```

```
mpg
15.5 16.2 16.5 16.9   17 17.5 17.6 17.7 18.1 18.2 18.5 18.6 19.1 19.2 19.4 19.8
   1    1    1    1    2    1    2    1    2    1    1    1    1    3    2    1
19.9 20.2 20.3 20.5 20.6 20.8 21.1 21.5 21.6   22 22.3 22.4   23 23.2 23.5 23.6
   1    4    1    2    2    1    1    1    1    1    1    1    2    1    1    1
23.7 23.8 23.9   24 24.2 24.3   25 25.1 25.4 25.8   26 26.4 26.6 26.8   27 27.2
   1    1    2    1    1    1    1    1    2    1    1    1    2    1    4    3
27.4 27.5 27.9   28 28.1 28.4 28.8   29 29.5 29.8 29.9   30 30.4 30.7 30.9   31
   1    1    1    3    1    1    1    1    1    2    1    2    1    1    1    3
31.3 31.5 31.6 31.8 31.9   32 32.1 32.2 32.3 32.4 32.7 32.8 32.9   33 33.5 33.7
   1    1    1    1    1    3    1    1    1    2    1    1    1    1    1    1
33.8   34 34.1 34.2 34.3 34.4 34.5 34.7   35 35.1 35.7   36 36.1 36.4   37 37.2
   1    2    2    1    1    1    2    1    1    1    1    5    2    1    3    1
37.3 37.7   38 38.1   39 39.1 39.4 40.8 40.9 41.5 43.1 43.4   44 44.3 44.6 46.6
   1    1    4    1    1    1    1    1    1    1    1    1    1    1    1    1
```

```
> frec_abs_acum_2<-cumsum(frec_abs_2)
> frec_abs_acum_2
```

```
15.5 16.2 16.5 16.9   17 17.5 17.6 17.7 18.1 18.2 18.5 18.6 19.1 19.2 19.4 19.8
   1    2    3    4    6    7    9   10   12   13   14   15   16   19   21   22
19.9 20.2 20.3 20.5 20.6 20.8 21.1 21.5 21.6   22 22.3 22.4   23 23.2 23.5 23.6
  23   27   28   30   32   33   34   35   36   37   38   39   41   42   43   44
23.7 23.8 23.9   24 24.2 24.3   25 25.1 25.4 25.8   26 26.4 26.6 26.8   27 27.2
  45   46   48   49   50   51   52   53   55   56   57   58   60   61   65   68
27.4 27.5 27.9   28 28.1 28.4 28.8   29 29.5 29.8 29.9   30 30.4 30.7 30.9   31
  69   70   71   74   75   76   77   78   79   81   82   84   85   86   87   90
31.3 31.5 31.6 31.8 31.9   32 32.1 32.2 32.3 32.4 32.7 32.8 32.9   33 33.5 33.7
  91   92   93   94   95   98   99  100  101  103  104  105  106  107  108  109
33.8   34 34.1 34.2 34.3 34.4 34.5 34.7   35 35.1 35.7   36 36.1 36.4   37 37.2
 110  112  114  115  116  117  119  120  121  122  123  128  130  131  134  135
37.3 37.7   38 38.1   39 39.1 39.4 40.8 40.9 41.5 43.1 43.4   44 44.3 44.6 46.6
 136  137  141  142  143  144  145  146  147  148  149  150  151  152  153  154
```

```
> frec_rel_2<-(table(mpg)/size_2)
> frec_rel_2
```

```
mpg
       15.5        16.2        16.5        16.9          17        17.5
0.006493506 0.006493506 0.006493506 0.006493506 0.012987013 0.006493506
       17.6        17.7        18.1        18.2        18.5        18.6
0.012987013 0.006493506 0.012987013 0.006493506 0.006493506 0.006493506
```

```
      19.1        19.2        19.4        19.8        19.9        20.2
0.006493506 0.019480519 0.012987013 0.006493506 0.006493506 0.025974026
      20.3        20.5        20.6        20.8        21.1        21.5
0.006493506 0.012987013 0.012987013 0.006493506 0.006493506 0.006493506
      21.6          22        22.3        22.4          23        23.2
0.006493506 0.006493506 0.006493506 0.006493506 0.012987013 0.006493506
      23.5        23.6        23.7        23.8        23.9          24
0.006493506 0.006493506 0.006493506 0.006493506 0.012987013 0.006493506
      24.2        24.3          25        25.1        25.4        25.8
0.006493506 0.006493506 0.006493506 0.006493506 0.012987013 0.006493506
        26        26.4        26.6        26.8          27        27.2
0.006493506 0.006493506 0.012987013 0.006493506 0.025974026 0.019480519
      27.4        27.5        27.9          28        28.1        28.4
0.006493506 0.006493506 0.006493506 0.019480519 0.006493506 0.006493506
      28.8          29        29.5        29.8        29.9          30
0.006493506 0.006493506 0.006493506 0.012987013 0.006493506 0.012987013
      30.4        30.7        30.9          31        31.3        31.5
0.006493506 0.006493506 0.006493506 0.019480519 0.006493506 0.006493506
      31.6        31.8        31.9          32        32.1        32.2
0.006493506 0.006493506 0.006493506 0.019480519 0.006493506 0.006493506
      32.3        32.4        32.7        32.8        32.9          33
0.006493506 0.012987013 0.006493506 0.006493506 0.006493506 0.006493506
      33.5        33.7        33.8          34        34.1        34.2
0.006493506 0.006493506 0.006493506 0.012987013 0.012987013 0.006493506
      34.3        34.4        34.5        34.7          35        35.1
0.006493506 0.006493506 0.012987013 0.006493506 0.006493506 0.006493506
      35.7          36        36.1        36.4          37        37.2
0.006493506 0.032467532 0.012987013 0.006493506 0.019480519 0.006493506
      37.3        37.7          38        38.1          39        39.1
0.006493506 0.006493506 0.025974026 0.006493506 0.006493506 0.006493506
      39.4        40.8        40.9        41.5        43.1        43.4
0.006493506 0.006493506 0.006493506 0.006493506 0.006493506 0.006493506
        44        44.3        44.6        46.6
0.006493506 0.006493506 0.006493506 0.006493506

> frec_rel_acum_2<-cumsum(frec_rel_2)
> frec_rel_acum_2

      15.5        16.2        16.5        16.9          17        17.5
0.006493506 0.012987013 0.019480519 0.025974026 0.038961039 0.045454545
      17.6        17.7        18.1        18.2        18.5        18.6
0.058441558 0.064935065 0.077922078 0.084415584 0.090909091 0.097402597
      19.1        19.2        19.4        19.8        19.9        20.2
0.103896104 0.123376623 0.136363636 0.142857143 0.149350649 0.175324675
      20.3        20.5        20.6        20.8        21.1        21.5
0.181818182 0.194805195 0.207792208 0.214285714 0.220779221 0.227272727
      21.6          22        22.3        22.4          23        23.2
0.233766234 0.240259740 0.246753247 0.253246753 0.266233766 0.272727273
      23.5        23.6        23.7        23.8        23.9          24
0.279220779 0.285714286 0.292207792 0.298701299 0.311688312 0.318181818
```

```
      24.2        24.3          25        25.1        25.4        25.8
0.324675325 0.331168831 0.337662338 0.344155844 0.357142857 0.363636364
       26        26.4        26.6        26.8          27        27.2
0.370129870 0.376623377 0.389610390 0.396103896 0.422077922 0.441558442
      27.4        27.5        27.9          28        28.1        28.4
0.448051948 0.454545455 0.461038961 0.480519481 0.487012987 0.493506494
      28.8          29        29.5        29.8        29.9          30
0.500000000 0.506493506 0.512987013 0.525974026 0.532467532 0.545454545
      30.4        30.7        30.9          31        31.3        31.5
0.551948052 0.558441558 0.564935065 0.584415584 0.590909091 0.597402597
      31.6        31.8        31.9          32        32.1        32.2
0.603896104 0.610389610 0.616883117 0.636363636 0.642857143 0.649350649
      32.3        32.4        32.7        32.8        32.9          33
0.655844156 0.668831169 0.675324675 0.681818182 0.688311688 0.694805195
      33.5        33.7        33.8          34        34.1        34.2
0.701298701 0.707792208 0.714285714 0.727272727 0.740259740 0.746753247
      34.3        34.4        34.5        34.7          35        35.1
0.753246753 0.759740260 0.772727273 0.779220779 0.785714286 0.792207792
      35.7          36        36.1        36.4          37        37.2
0.798701299 0.831168831 0.844155844 0.850649351 0.870129870 0.876623377
      37.3        37.7          38        38.1          39        39.1
0.883116883 0.889610390 0.915584416 0.922077922 0.928571429 0.935064935
      39.4        40.8        40.9        41.5        43.1        43.4
0.941558442 0.948051948 0.954545455 0.961038961 0.967532468 0.974025974
       44        44.3        44.6        46.6
0.980519481 0.987012987 0.993506494 1.000000000
```

Calculamos la media y la mediana, mínimos y máximos de los valores, el rango, la desviación típica y la varianza.

```
> media_2<-mean(mpg)
> media_2

[1] 28.79351

> mediana_2<-median(mpg)
> mediana_2

[1] 28.9

> desv_tip_2<-sd(mpg)
> desv_tip_2

[1] 7.37721

> var_2<-var(mpg)
> var_2

[1] 54.42323

> minimo_2<-min(mpg)
> minimo_2
```

```
[1] 15.5

> maximo_2<-max(mpg)
> maximo_2

[1] 46.6

> rango_2<-range(mpg)
> rango_2

[1] 15.5 46.6
```

Para acabar con el tratamiento de los datos calculamos los cuartiles.

```
> cuart1_2<-quantile(mpg,0.25)
> cuart1_2

  25%
22.55

> cuart2_2<-quantile(mpg,0.5)
> cuart2_2

 50%
28.9

> cuart3_2<-quantile(mpg,0.75)
> cuart3_2

   75%
34.275

> cuart4_2<-quantile(mpg,1)
> cuart4_2

100%
46.6

> cuant54_2<-quantile(mpg,0.54)
> cuant54_2

54%
 30
```

2 Ejericio 2

Desarrollo por parte de cada grupo del enunciado y la solución de un ejercicio en el que se realice un análisis con R de descripción de Datos introduciendo modificaciones sobre el ejercicio hecho en clase (por ejemplo: los datos se leen desde un fichero generado con Excel o los ficheros que hay en un directorio se listan con la función dir().

Las mejoras implementadas para esta segunda parte del ejercicio son la carga de datos desde un fichero excel y la presentación de los datos de una forma más ordenada.

Se pretende estudiar, mediante los datos contenidos en un fichero *distancias.xlsx*, las distancias recorridas por una muestra de la población española de las distancias recorridas en sus viajes a lo largo del verano. Con estos datos y junto a un conjunto de nuevas librerías se llevará a cabo el estudio. Para llevar a cabo las mejoras se requiere de las siguientes librerías:

- `readxl:` para leer del fichero Excel los datos
- `pastecs:` para realizar los cálculos de distintos valores como media y mediana y mínimos y máximos.
- `dplyr:` para trabajar con tablas y poder concatenar columnas con su función mutate.

```
> #Instalamos las librerías necesarias
> install.packages("readxl")  #Lectura de fichero Excel
> install.packages("pastecs") #Realiza calculos (media, mediana, ...)
> install.packages("dplyr")   #Mutate - concatencación de columnas
> #Importamos las librerías instaladas previamente
> library("readxl")
> library("pastecs")
> library("dplyr")
```

Para podeer leer los datos desde el Excel hemos implementado lo siguiente:

```
> # La función file.choose() abre un explorador de archivos
> # que permite elegir el fichero Excel deseado
> datosExcel<-read_excel(file.choose(),sheet="Hoja1")
> distancias<-datosExcel$dis
> size_3<-length(distancias)
> distancias

 [1] 252 288  90 114 460  88 598 501 512 531 539  58  80 432 463 137 461 408 600
[20] 248 511 253 460 217 377 470 429  61 417  90
```

Realizamos los cálculos de las frecuencias y los cuartiles como hicimos anteriormente

```
> frec_abs_3<-table(distancias)
> frec_abs_acum_3<-cumsum(frec_abs_3)
> frec_rel_3<-table(distancias)/size_3
> frec_rel_acum_3<-cumsum(frec_rel_3)
> cuart1_3<-quantile(distancias,0.25)
> cuart2_3<-quantile(distancias,0.5)
> cuart3_3<-quantile(distancias,0.75)
> cuart4_3<-quantile(distancias,1)
> cuant54_3<-quantile(distancias,0.54)
```

Para mostrar las frecuencias todas juntas y de una forma más visual hemos creado la siguiente tabla.

```
> frecuencias_3<-data.frame(table(distancias))
> frecuencias_3<-mutate(frecuencias_3,frec_abs_acum_3,frec_rel_3,frec_rel_acum_3)
> frecuencias_3
```

	distancias	Freq	frec_abs_acum_3	frec_rel_3	frec_rel_acum_3
1	58	1	1	0.03333333	0.03333333
2	61	1	2	0.03333333	0.06666667
3	80	1	3	0.03333333	0.10000000
4	88	1	4	0.03333333	0.13333333
5	90	2	6	0.06666667	0.20000000
6	114	1	7	0.03333333	0.23333333
7	137	1	8	0.03333333	0.26666667
8	217	1	9	0.03333333	0.30000000
9	248	1	10	0.03333333	0.33333333
10	252	1	11	0.03333333	0.36666667
11	253	1	12	0.03333333	0.40000000
12	288	1	13	0.03333333	0.43333333
13	377	1	14	0.03333333	0.46666667
14	408	1	15	0.03333333	0.50000000
15	417	1	16	0.03333333	0.53333333
16	429	1	17	0.03333333	0.56666667
17	432	1	18	0.03333333	0.60000000
18	460	2	20	0.06666667	0.66666667
19	461	1	21	0.03333333	0.70000000
20	463	1	22	0.03333333	0.73333333
21	470	1	23	0.03333333	0.76666667
22	501	1	24	0.03333333	0.80000000
23	511	1	25	0.03333333	0.83333333
24	512	1	26	0.03333333	0.86666667
25	531	1	27	0.03333333	0.90000000
26	539	1	28	0.03333333	0.93333333
27	598	1	29	0.03333333	0.96666667
28	600	1	30	0.03333333	1.00000000

A parte hemos creado otra tabla para mostrar la media, la varianza, la desviación típica, el rango, la mediana, el mínimo y el máximo.

```
> tabla_3<-stat.desc(distancias)[c("mean","var","std.dev","range",
+                                  "median","min","max")]
> names(tabla_3)<-c("Media","Varianza","Desviacion tipica","Rango",
+                   "Mediana","Minimo","Maximo")
> tabla_3
```

Media	Varianza	Desviacion tipica	Rango
338.1667	32621.7989	180.6151	542.0000
Mediana	**Minimo**	**Maximo**	
412.5000	58.0000	600.0000	

Por último, para mostrar los cuántiles hemos creado una última tabla.

```
> cuantiles<-c(summary(distancias),cuant54_3)
> cuantiles_3<-cuantiles[order(unlist(cuantiles))]
> cuantiles_3
```

```
   Min.  1st Qu.     Mean   Median      54%  3rd Qu.     Max.
58.0000 157.0000 338.1667 412.5000 424.9200 468.2500 600.0000
```