Data Summarization in R

L. Torgo

ltorgo@dcc.fc.up.pt
Departamento de Ciência de Computadores
Faculdade de Ciências / Universidade do Porto

Oct, 2014

Motivation for Data Summarization?

- With big data sets it is hard to have an idea of what is going on in the data
- Data summaries provide overviews of key properties of the data
- Their goal is to describe important properties of the distribution of the values across the observations that were measured

Examples of Types of Summaries

- What is the “most common value” of a variable?
- What is the “variability” in the values of a variable?
- Are there “strange” / unexpected values in the data set?
    - Outliers
    - Unknown values

What is the “most common value” of a variable?

Statistics of location

- The **mean** (or sample mean)

$$\mu_X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The **median**
    - It is the value above (below) which there are 50% of the values in the data set
    - Usually calculated by sorting the values and peeking the value in the middle position
- The **mode**
    - It is the most common (more frequently occurring) value in a set of values
        - Note that the mode can be applied to categorical variables

Illustrations in R

```
library(DMwR)
data(algae)
mean(algae$oPO4)

## [1] NA

mean(algae$oPO4,na.rm=TRUE)

## [1] 73.59

median(algae$a2)

## [1] 3

centralValue(algae$season) # mode for nominal vars.

## [1] "winter"

centralValue(algae$Chla)   # median for numeric vars.

## [1] 5.475
```

Illustrations in R with dplyr

```
library(dplyr)
alg <- tbl_df(algae)
alg %>% summarise(avg.oPO4=mean(oPO4,na.rm=TRUE),
                  med.oPO4=median(oPO4,na.rm=TRUE),
                  cen.season=centralValue(season),
                  cen.Chla=centralValue(Chla))

## Source: local data frame [1 x 4]
##
##   avg.oPO4 med.oPO4 cen.season cen.Chla
## 1    73.59    40.15     winter    5.475
```

What is the “variability” of the values of a variable?

Statistics of variability or dispersion

- The **variance**

$$\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)^2$$

- The **standard deviation**

$$\sigma_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)^2}$$

- The **inter-quartile range**
  - It is the difference between the 3rd and 1st quartiles
    - The 1st quartile is the number below which there are 25% of the values
    - The 3rd quartile is the number below which there are 75% of the values
- The **range**
  - It is the difference between the maximum and minimum values

Illustrations in R

```
var(algae$NH4,na.rm=TRUE)

## [1] 3851585

sd(algae$a6)

## [1] 11.66

IQR(algae$Cl,na.rm=TRUE)

## [1] 46.84

quantile(algae$mnO2,na.rm=TRUE)

##      0%     25%     50%     75%    100%
##   1.500   7.725   9.800  10.800  13.400
```

```
quantile(algae$mxPH,na.rm=TRUE,
         probs=c(0.1,0.9))

##  10%  90%
## 7.34 8.70

fivenum(algae$a5)

## [1]  0.0  0.0  1.9  7.5 44.4

range(algae$a7)

## [1]  0.0 31.6
```

Illustrations in R with dplyr

```
library(dplyr)
alg <- tbl_df(algae)
alg %>% summarise(var.NH4=var(NH4,na.rm=TRUE),
                  sd.a6=sd(a6),
                  iqr.Cl=IQR(Cl,na.rm=TRUE))

## Source: local data frame [1 x 3]
##
##   var.NH4 sd.a6 iqr.Cl
## 1 3851585 11.66  46.84
```

Are there “strange” values in the data?

- Outliers
  - Informally, an outlier is a value that deviates so much from the other values as to arouse suspicions that it was generated by a different mechanism
  - A frequently used formal definition for an outlier is any value outside the interval,

    $$[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$$

    where $Q_1 (Q_3)$ is the 1st(3rd) quartile and IQR is the inter-quartile range
- Unknown values
  - In real-world applications we frequently have situations were the value of some variable in a certain observation is unknown
- On both cases we need to decide how to handle these situations
  - Remove the data?
  - Change somehow these values?
  - etc.

Illustrations in R

```
boxplot.stats(algae$a4)

## $stats
## [1] 0.0 0.0 0.0 2.4 5.7
##
## $n
## [1] 200
##
## $conf
## [1] -0.2681  0.2681
##
## $out
##  [1] 44.6  6.8 11.5 28.8 13.4  7.6 11.0 11.3  6.8  6.6 12.7  8.3  6.2  7.7
## [15]  6.9  7.8

summary(algae$PO4)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     1.0    41.4   103.0   138.0   214.0   772.0       2
```

More Data Summaries

Global summary of the basic descriptive statistics of a data set:

```
summary(algae)
```

```
##     season        size         speed           mxPH            mnO2
##  autumn:40   large :45   high  :84   Min.   :5.60   Min.   : 1.50
##  spring:53   medium:84   low   :33   1st Qu.:7.70   1st Qu.: 7.72
##  summer:45   small :71   medium:83   Median :8.06   Median : 9.80
##  winter:62                           Mean   :8.01   Mean   : 9.12
##                                      3rd Qu.:8.40   3rd Qu.:10.80
##                                      Max.   :9.70   Max.   :13.40
##                                      NA's   :1      NA's   :2
```

...

...

More Data Summaries (cont.)

```
library(Hmisc)  # extra package, you need to install it
describe(algae)
```

```
## algae[, 1:5]
##
##  5  Variables      200  Observations
## ---------------------------------------------------------------------------
## season
##       n missing  unique
##     200       0       4
##
## autumn (40, 20%), spring (53, 26%), summer (45, 22%)
## winter (62, 31%)
## ---------------------------------------------------------------------------
## size
##       n missing  unique
##     200       0       3
##
## large (45, 22%), medium (84, 42%), small (71, 36%)
## ---------------------------------------------------------------------------
## speed
##       n missing  unique
##     200       0       3
##
## high (84, 42%), low (33, 16%), medium (83, 42%)
## ---------------------------------------------------------------------------
## mxPH
##       n missing  unique    Mean     .05     .10     .25     .50     .75
##     199       1      72   8.012   7.081   7.340   7.700   8.060   8.400
##     .90     .95
##   8.700   8.873
##
## lowest : 5.60 5.70 6.40 6.50 6.60, highest: 9.00 9.06 9.10 9.50 9.70
## ---------------------------------------------------------------------------
## mnO2
##       n missing  unique    Mean     .05     .10     .25     .50     .75
##     198       2      88   9.118   4.485   5.770   7.725   9.800  10.800
##     .90     .95
##  11.700  11.815
##
## lowest :  1.5  1.8  3.2  3.3  3.4, highest: 12.5 12.6 12.9 13.1 13.4
## ---------------------------------------------------------------------------
```

...
...

Conditional Summaries

```
apply(algae[,c('a1','a7')],2,max)

##   a1   a7
## 89.8 31.6

by(algae$a1,algae$season,summary)

## algae$season: autumn
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.65    8.50   17.70   24.00   86.60
## --------------------------------------------------------
## algae$season: spring
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     1.7     4.1    16.6    20.3    89.8
## --------------------------------------------------------
## algae$season: summer
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     1.7     8.7    16.1    25.5    64.2
## --------------------------------------------------------
## algae$season: winter
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     6.0    17.2    25.1    81.9
```

Hands on Summarization - the algae data set

Concerning the algae data set answer the following question:

1. Which season has more water samples?
2. What is the average value of `a5`?
3. What is the average value of `NO3`?
4. Check if there are unusually high values of `a2` and show the respective water samples.
5. Obtain a summary of the basic descriptive statistics of `a1` and `a4`, for each season of the year.
6. Try to obtain a table with the seasons ordered by decreasing average value of `NO3`. Hint: explore the capabilities of the function `aggregate()` that has similar objectives as the function `by()`. Also explore the function `order()`.