Numerical explorations with R

Waithaka Michael

February 14, 2016

Introduction

- Before analysing any study data, it is common practice to explore the data, to get a broad idea about the phenomenon we are studying.
- Summary statistics are then of interest, such as mean,variability, frequencies.
- It is also of interest to know if missing data are present.
- R provides these measures through the use of the function **summary()**.

.....*Recap managing data*

- Read in the data
- Datasets in R are typically stored as data frames, which have a matrix structure
- Observations are arranged as rows and variables, either numerical or categorical, are arranged as columns

Import the dataset

```
data <- read.csv("data/bwmal.csv")
```

Get the dimension of the dataset

```
dim(data)

## [1] 791  12
```

.....*Recap managing data*

Explore variable names of the dataset

```
names(data)

##  [1] "X"          "matage"    "mheight"  "gestwks"  "sex"
##  [6] "bweight"    "smoke"     "pfplacen" "parity"   "workload"
## [11] "matagegp" "gestcat"
```

The dataset at a glance

```
head(data)  #Returns the first six rows of dataset; tail(data)

##   X matage mheight gestwks sex bweight smoke pfplacen
## 1 1     26   1.575      40   0    3.11     0        0
## 2 2     23   1.529      40   0    2.65     0        0
## 3 3     18   1.540      40   1    3.41     0        0
## 4 4     25   1.581      40   1    2.99     0        0
## 5 5     25   1.555      40   1    3.16     0        0
## 6 6     21   1.561      40   1    2.82     0        0
```

.....*Recap managing data*

Explore the structure of the dataset

```
str(data)
```

```
## 'data.frame': 791 obs. of  12 variables:
##  $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ matage  : int  26 23 18 25 25 21 20 19 32 23 ...
##  $ mheight : num  1.58 1.53 1.54 1.58 1.55 ...
##  $ gestwks : int  40 40 40 40 40 40 41 38 40 41 ...
##  $ sex     : int  0 0 1 1 1 1 1 1 0 0 ...
##  $ bweight : num  3.11 2.65 3.41 2.99 3.16 ...
##  $ smoke   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pfplacen: int  0 0 0 0 0 0 0 1 0 0 ...
##  $ parity  : int  3 1 0 2 1 1 0 0 6 0 ...
##  $ workload: int  0 0 0 0 1 1 1 1 0 0 ...
##  $ matagegp: int  3 3 1 3 3 2 2 1 4 3 ...
##  $ gestcat : int  2 2 2 2 2 2 2 2 2 2 ...
```

.....Recap managing data

Viewing data contents of a variable

We can access variables directly by using their names, using the object $ variable notation

```
data$sex
```

```
##   [1] 0 0 1 1 1 1 1 1 0 0 0 1 1 0 1 0 1 0 0 1 1 0 1 1 1 1 1
##  [28] 1 1 0 1 1 1 1 0 0 0 1 0 1 0 0 0 1 1 0 1 0 0 0 0 0 1 0
##  [55] 0 1 1 1 1 0 1 1 0 1 0 0 1 0 0 0 0 0 1 1 1 0 0 1 0 1 1
##  [82] 1 1 0 1 1 1 1 1 1 0 0 1 1 0 1 1 0 0 1 1 0 0 1 1 0 1 0
## [109] 0 1 0 1 0 1 1 1 0 0 0 0 1 1 1 1 0 0 0 1 1 0 1 1 0 0 0
## [136] 1 0 1 0 0 1 0 0 0 1 1 1 1 1 0 1 0 0 1 0 1 0 0 1 1 0 0
## [163] 0 1 1 0 0 1 1 0 0 1 0 0 0 1 0 1 0 1 1 1 0 1 0 0 0 1 1
## [190] 1 0 1 0 0 0 1 0 0 1 0 0 1 1 0 0 0 0 0 0 1 0 0 1 0 0 1
## [217] 1 0 1 0 0 0 1 0 0 1 0 1 0 1 1 1 1 1 1 0 1 0 0 1 0 0 0
## [244] 1 1 0 0 0 0 1 1 1 0 1 0 1 0 1 1 1 1 1 0 0 1 1 1 0 1 1
## [271] 1 1 1 0 0 0 0 0 1 0 0 1 0 0 1 1 0 1 0 0 1 1 0 0 0 0 0
## [298] 0 1 0 0 1 1 0 0 0 0 0 1 1 0 0 1 0 0 0 1 1 0 1 1 1 1 0
## [325] 0 1 1 0 1 1 1 1 1 0 0 0 1 1 0 1 1 0 0 0 0 1 0 0 1 0 0
```

.....Recap managing data

Viewing specific cell contents
To access a certain entry, we most commonly use **object[row,column]**

```
data[2, 3]

## [1] 1.529
```

Viewing specific variable contents
all data in variable 5 (sex)

```
data[, 5]

##   [1] 0 0 1 1 1 1 1 1 0 0 0 1 1 0 1 0 1 0 0 1 1 0 1 1 1 1 1
##  [28] 1 1 0 1 1 1 1 0 0 0 1 0 1 0 0 0 1 1 0 1 0 0 0 0 0 1 0
##  [55] 0 1 1 1 1 0 1 1 0 1 0 0 1 0 0 0 0 0 1 1 1 0 0 1 0 1 1
##  [82] 1 1 0 1 1 1 1 1 1 0 0 1 1 0 1 1 0 0 1 1 0 0 1 1 0 1 0
## [109] 0 1 0 1 0 1 1 1 0 0 0 0 1 1 1 1 0 0 0 1 1 0 1 1 0 0 0
## [136] 1 0 1 0 0 1 0 0 0 1 1 1 1 1 0 1 0 0 1 0 1 0 0 1 1 0 0
## [163] 0 1 1 0 0 1 1 0 0 1 0 0 0 1 0 1 0 1 1 1 0 1 0 0 0 1 1
```

.....*Recap managing data*

Viewing specific row/observation contents - all data in row 5

```
data[5, ]
```

```
##   X matage mheight gestwks sex bweight smoke pfplacen
## 5 5     25   1.555      40   1    3.16     0        0
##   parity workload matagegp gestcat
## 5      1        1        3       2
```

Data in a range - all data in rows 2 and 3, columns 2 and 3

```
data[2:3, 2:3]
```

```
##   matage mheight
## 2     23   1.529
## 3     18   1.540
```

....back to the function *summary()*

- This function returns some basic summary statistics, which differ according to the class of the objects that are considered.
- In particular R distinguishes between:
  - numerical vectors: mean,minimum, maximum and quartiles are calculated,
  - factors: frequencies are calculated,
  - character vectors: just the class of the object is returned,
  - ... *just try for the rest (but be critical towards the output!)*.

The function *summary()*

```
# summary statistics for continuous variables using the
# function summary()
mydata = data[, c(1:4, 6)]
summary(mydata)

##        X             matage          mheight     
##  Min.   :  1.0   Min.   :13.00   Min.   :1.352  
##  1st Qu.:198.5   1st Qu.:20.00   1st Qu.:1.506  
##  Median :396.0   Median :23.00   Median :1.544  
##  Mean   :396.0   Mean   :23.78   Mean   :1.543  
##  3rd Qu.:593.5   3rd Qu.:27.00   3rd Qu.:1.580  
##  Max.   :791.0   Max.   :46.00   Max.   :1.750  
##     gestwks         bweight    
##  Min.   :28.00   Min.   :0.78  
##  1st Qu.:38.00   1st Qu.:2.58  
##  Median :39.00   Median :2.90  
```

Specific functions to summarize the data

- Enables us to see the main characteristics of data before any formal modeling or hypothesis testing
- Particular techniques depends on the type of variable: Continuous or categorical
  - Continuous eg. matage, mheight, gestwks, bweight, parity
  - Categorical eg. smoking status, sex, pfplacen, workload, matagegp, gestcat

Examples of data explorations: Continuous variables

```
min(data$mheight)

## [1] 1.352

max(data$mheight)

## [1] 1.75
```

Some data explorations: Continuous variables

```
mean(data$mheight)

## [1] 1.543273

var(data$mheight)

## [1] 0.002884892

sd(data$matage)

## [1] 5.139645

median(data$matage)

## [1] 23
```

More data explorations using function *apply()*

Produce the defined summary statistic for continous variables

```
(mydata.mean = apply(mydata, MARGIN = 2, FUN = mean))

##          X    matage   mheight   gestwks   bweight
## 396.000000 23.782554  1.543273 38.988622  2.900354

(mydata.median = apply(mydata, MARGIN = 2, FUN = median))

##       X  matage mheight gestwks bweight
## 396.000  23.000   1.544  39.000   2.900

(mydata.quantiles = apply(mydata, MARGIN = 2, FUN = sd))

##           X      matage     mheight     gestwks     bweight
## 228.4863234   5.1396448   0.0537112   1.6369536   0.5108436
```

Exploring categorical variables

Summarize single categorical variable

```
# freq table for the factor variables
(freq.table.sex = table(mydata2$sex))

##
##   0   1
## 381 410

(freq.table.smoke = table(mydata2$smoke))

##
##   0   1
## 724  67
```

Exploring categorical variables

Cross-tabulation of two categorical variables: 2-Way Frequency Table

```
(mytable <- table(mydata2$sex, mydata2$smoke))

##
##         0   1
##    0 346  35
##    1 378  32
```

```
(mytable <- with(data, table(sex, smoke)))  #with command adds
able labels

##     smoke
## sex   0   1
##    0 346  35
##    1 378  32
```

Exploring categorical variables

Tables of marginal frequencies

```
# sex frequencies (summed over smoke)
margin.table(mytable, 1)

## sex
##   0   1
## 381 410

# smoking status frequencies (summed over sex)
margin.table(mytable, 2)

## smoke
##   0   1
## 724  67
```

Exploring categorical variables

Tables of proportions

```
100 * prop.table(mytable)  # cell percentages

##    smoke
## sex         0         1
##   0 43.742099  4.424779
##   1 47.787611  4.045512

100 * prop.table(mytable, 1)  # row percentages

##    smoke
## sex         0         1
##   0 90.813648  9.186352
##   1 92.195122  7.804878

# column percentages = 100*prop.table(mytable, 2)
```

Exploring categorical variables

Testing the independence of the row and column variable

```
chisq.test(mytable, correct = FALSE)  # chi-square test of indepedence

##
##  Pearson's Chi-squared test
##
## data:  mytable
## X-squared = 0.48614, df = 1, p-value = 0.4857

# summary(mytable) - chi-square test of indepedence
# chisq.test(mytable) - chi-square test of indepedence with
# Yates' continuity correction
```

Exploring categorical variables

3-Way Frequency Table : using xtabs

```
mytable <- xtabs(~sex + smoke + matagegp, data = mydata2)
ftable(mytable)  # print table

##           matagegp   1   2   3   4
## sex smoke
## 0   0               71 104  86  85
##     1                4   8   9  14
## 1   0               75 111 101  91
##     1                4   8   7  13
```

Exploring categorical variables

Log-linear models for 3-Way Frequency Table

```
library(MASS)
# Mutual Independence: sex, smoking status and matagegp are
# pairwise independent
loglm(~sex + smoke + matagegp, mytable)

## Call:
## loglm(formula = ~sex + smoke + matagegp, data = mytable)
##
## Statistics:
##                         X^2 df  P(> X^2)
## Likelihood Ratio  9.544296 10 0.4813403
## Pearson          10.060646 10 0.4351889
```

Exploring categorical variables

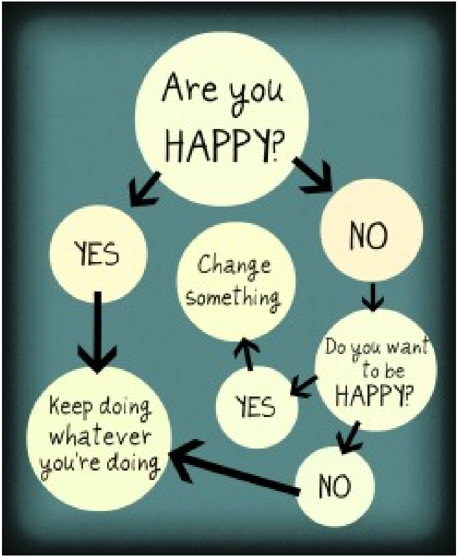
Log-linear models for 3-Way Frequency Table

```
# Conditional Independence: sex is independent of smoking
# status, given matagegp.
loglm(~sex + smoke + matagegp + sex * matagegp + smoke * matag
    mytable)

## Call:
## loglm(formula = ~sex + smoke + matagegp + sex * matagegp +
##     matagegp, data = mytable)
##
## Statistics:
##                       X^2 df  P(> X^2)
## Likelihood Ratio 0.762029  4 0.9434649
## Pearson          0.763251  4 0.9433058
```

Exploring categorical variables

Log-linear models for 3-Way Frequency Table

```
# No Three-Way Interaction
loglm(~sex + smoke + matagegp + sex * smoke + sex * matagegp +
    smoke * matagegp, mytable)

## Call:
## loglm(formula = ~sex + smoke + matagegp + sex * smoke + sex
##     matagegp + smoke * matagegp, data = mytable)
##
## Statistics:
##                        X^2 df  P(> X^2)
## Likelihood Ratio 0.2838056  3 0.9630449
## Pearson          0.2833175  3 0.9631349
```


Are you HAPPY?
YES
NO
Change something
Do you want to be HAPPY?
YES
NO
Keep doing whatever you're doing