EDA: Descriptive statistics

Note updated August 18, 2019. Not for sale :-)

Wan Nor Arifin
Unit of Biostatistics and Research Methodology,
Universiti Sains Malaysia.

Email: wnarifin@usm.my
Website: wnarifin.github.io

Contents

1 Descriptive Statistics

In this practical session, we use `cholest.sav` dataset. Now we give it a proper object name `cholest`,

```
library(foreign)
cholest = read.spss("cholest.sav", to.data.frame = T)
str(cholest)
```

```
## 'data.frame':    80 obs. of  5 variables:
##  $ chol    : num  6.5 6.6 6.8 6.8 6.9 7 7 7.2 7.2 7.2 ...
##  $ age     : num  38 35 39 36 31 38 33 36 40 34 ...
##  $ exercise: num  6 5 6 5 4 4 5 5 4 6 ...
##  $ sex     : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ categ   : Factor w/ 3 levels "Grp A","Grp B",..: 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "variable.labels")= Named chr  "cholesterol in mmol/L" "age in year" "duration of exercis
##   ..- attr(*, "names")= chr  "chol" "age" "exercise" "sex" ...
##  - attr(*, "codepage")= int 65001
```

In general, simple descriptive statistics can be obtained using `summary()` function,

```
summary(cholest)
```

```
##       chol            age           exercise          sex        categ
##  Min.   : 6.50   Min.   :28.00   Min.   :2.000   female:40   Grp A:25
##  1st Qu.: 7.60   1st Qu.:36.00   1st Qu.:4.000   male  :40   Grp B:33
##  Median : 8.30   Median :39.00   Median :4.000               Grp C:22
##  Mean   : 8.23   Mean   :39.48   Mean   :4.225
```

```
##  3rd Qu.: 8.80   3rd Qu.:43.25   3rd Qu.:5.000  
##  Max.   :10.00   Max.   :52.00   Max.   :6.000
```

The results depend on the variable type.

1.1 Central tendency and dispersion

For numerical variables, we can obtain the measures of central tendency (mean and median) and dispersion (standard deviation, SD and interquartile range, IQR). Now we obtain in pairs of mean (SD) and median (IQR),

Mean,

```
mean(cholest$chol)
```

```
## [1] 8.23
```

```
mean(cholest$age)
```

```
## [1] 39.475
```

Standard deviation, SD,

```
sd(cholest$chol)
```

```
## [1] 0.8386849
```

```
sd(cholest$age)
```

```
## [1] 5.128661
```

Median,

```
median(cholest$chol)
```

```
## [1] 8.3
```

```
median(cholest$age)
```

```
## [1] 39
```

and interquartile range, IQR,

```
IQR(cholest$chol)
```

```
## [1] 1.2
```

```
IQR(cholest$age)
```

```
## [1] 7.25
```

1.2 Proportions

For categorical variables, we want to obtain the count per group, proportions and percentages.

The count per group using `table()` function (we can also obtain the counts from `summary()` function as done before),

```
tab_sex = table(cholest$sex)
tab_categ = table(cholest$categ)
tab_sex
```

```
##
## female   male
##     40     40
```

```
tab_categ
```

```
##
## Grp A Grp B Grp C
##    25    33    22
```

The proportions,

```
prop.table(tab_sex)
```

```
##
## female   male
##    0.5    0.5
```

```
prop.table(tab_categ)
```

```
##
##  Grp A  Grp B  Grp C
## 0.3125 0.4125 0.2750
```

and to obtain the percentages, we multiply the proportions by 100,

```
prop.table(tab_sex)*100
```

```
##
## female   male
##     50     50
```

```
prop.table(tab_categ)*100
```

```
##
## Grp A Grp B Grp C
## 31.25 41.25 27.50
```

1.3 Statistics by groups

For numerical variables, we can obtain the statistics by groups (the categorical variables) using `by()` function. The syntax is `by(numerical_variable, categorical_variable, function)`.

Mean and SD for `chol` by `sex`,

```
by(cholest$chol, cholest$sex, mean)
```

```
## cholest$sex: female
## [1] 8.9275
## -------------------------------------------------------------------
## cholest$sex: male
## [1] 7.5325
```

```
by(cholest$chol, cholest$sex, sd)
```

```
## cholest$sex: female
## [1] 0.4551627
## -------------------------------------------------------------------
## cholest$sex: male
## [1] 0.4687066
```

1.4 Cross-tabulation

For categorical variables, it is important to be able to perform cross-tabulation to explore the count per cells for each combination of groups. Again, we use `table()` function.

For `sex` and `categ`, we obtain the basic cross-tabulation,

```
tab_sex_categ = table(Gender = cholest$sex, Category = cholest$categ)
tab_sex_categ
```

```
##         Category
## Gender   Grp A Grp B Grp C
##   female     0    18    22
##   male      25    15     0
```

Notice we can give headers ("Gender" and "Category") to groups in the table as shown above.

We can also easily obtain the proportions and percentages,

```
prop_sex_categ = prop.table(tab_sex_categ)
prop_sex_categ
```

```
##         Category
## Gender    Grp A  Grp B  Grp C
##   female 0.0000 0.2250 0.2750
##   male   0.3125 0.1875 0.0000
```

```
per_sex_categ = prop.table(tab_sex_categ)*100
per_sex_categ
```

```
##         Category
## Gender   Grp A Grp B Grp C
##   female  0.00 22.50 27.50
##   male   31.25 18.75  0.00
```

and add the marginal counts,

```
margin_sex_categ = addmargins(tab_sex_categ)
margin_sex_categ
```

```
##         Category
## Gender   Grp A Grp B Grp C Sum
##   female     0    18    22  40
##   male      25    15     0  40
##   Sum       25    33    22  80
```

and view the proportions and percentages again, including that of the marginal counts,

```
addmargins(prop_sex_categ)
```

```
##         Category
## Gender    Grp A  Grp B  Grp C    Sum
##   female 0.0000 0.2250 0.2750 0.5000
##   male   0.3125 0.1875 0.0000 0.5000
##   Sum    0.3125 0.4125 0.2750 1.0000
```

```
addmargins(per_sex_categ)
```

```
##         Category
## Gender    Grp A  Grp B  Grp C    Sum
##   female   0.00  22.50  27.50  50.00
```

```
##   male    31.25  18.75   0.00  50.00
##   Sum     31.25  41.25  27.50 100.00
```

References

Chongsuvivatwong, V. (2018). *EpiDisplay: Epidemiological data display package*. Retrieved from https://CRAN.R-project.org/package=epiDisplay

R Core Team. (2019). *Foreign: Read data stored by 'minitab', 's', 'sas', 'spss', 'stata', 'systat', 'weka', 'dBase', ...* Retrieved from https://CRAN.R-project.org/package=foreign

Revelle, W. (2019). *Psych: Procedures for psychological, psychometric, and personality research*. Retrieved from https://CRAN.R-project.org/package=psych