Basic Descriptive Statistics using R

shyamjith c

14 October, 2023

Introduction

- Descriptive statistics (in the broad sense of the term) is a branch of statistics aiming at summarizing, describing and presenting a series of values or a dataset.
- Descriptive statistics is often the first step and an important part in any statistical analysis.
- It allows to check the quality of the data and it helps to "understand" the data by having a clear overview of it.
- If well presented, descriptive statistics is already a good starting point for further analyses.

Types of Descriptive Summary

There exists many measures to summarize a dataset. They are divided into two types:

- location measures and
- dispersion measures

Working with Toy Dataset

As a first step load the data set to R:

```
dat <- iris # load the iris dataset and renamed it dat
```

```
head(dat) # first 6 observations
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Structure of a dataset

```
str(dat) # structure of dataset
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Basic summary statistics

`min, max, mean, median, range, IQR, quantiles`

```r
print("Minimum")
```

```
## [1] "Minimum"
```

```r
min(dat$Sepal.Length)
```

```
## [1] 4.3
```

```r
median(dat$Sepal.Length)
```

```
## [1] 5.8
```

```r
quantile(dat$Sepal.Length, c(0.25,0.5,0.75)) # three quartile
```

```
## 25% 50% 75%
## 5.1 5.8 6.4
```

Standard deviation and variance

The standard deviation and the variance is computed with the `sd()` and `var()` functions:

```r
sd(dat$Sepal.Length)
```

```
## [1] 0.8280661
```

```r
var(dat$Sepal.Length)
```

```
## [1] 0.6856935
```

```r
sqrt(var(dat$Sepal.Length))
```

```
## [1] 0.8280661
```

Tip: to compute the standard deviation (or variance) of multiple variables at the same time, use `lapply()` with the appropriate statistics as second argument:

```r
lapply(dat[, 1:4], sd)
```

```
## $Sepal.Length
## [1] 0.8280661
##
## $Sepal.Width
## [1] 0.4358663
##
## $Petal.Length
## [1] 1.765298
##
## $Petal.Width
## [1] 0.7622377
```

Five point Summary

```r
summary(dat)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Group-wise summary

```r
by(dat, dat$Species, summary)
```

```
## dat$Species: setosa
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.300   Min.   :1.000   Min.   :0.100
##  1st Qu.:4.800   1st Qu.:3.200   1st Qu.:1.400   1st Qu.:0.200
##  Median :5.000   Median :3.400   Median :1.500   Median :0.200
##  Mean   :5.006   Mean   :3.428   Mean   :1.462   Mean   :0.246
##  3rd Qu.:5.200   3rd Qu.:3.675   3rd Qu.:1.575   3rd Qu.:0.300
##  Max.   :5.800   Max.   :4.400   Max.   :1.900   Max.   :0.600
##        Species
##  setosa    :50
##  versicolor: 0
##  virginica : 0
##
##
##
## ------------------------------------------------------------
## dat$Species: versicolor
##   Sepal.Length    Sepal.Width     Petal.Length   Petal.Width          Species
##  Min.   :4.900   Min.   :2.000   Min.   :3.00   Min.   :1.000   setosa    : 0
##  1st Qu.:5.600   1st Qu.:2.525   1st Qu.:4.00   1st Qu.:1.200   versicolor:50
##  Median :5.900   Median :2.800   Median :4.35   Median :1.300   virginica : 0
##  Mean   :5.936   Mean   :2.770   Mean   :4.26   Mean   :1.326
##  3rd Qu.:6.300   3rd Qu.:3.000   3rd Qu.:4.60   3rd Qu.:1.500
##  Max.   :7.000   Max.   :3.400   Max.   :5.10   Max.   :1.800
## ------------------------------------------------------------
## dat$Species: virginica
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.900   Min.   :2.200   Min.   :4.500   Min.   :1.400
##  1st Qu.:6.225   1st Qu.:2.800   1st Qu.:5.100   1st Qu.:1.800
##  Median :6.500   Median :3.000   Median :5.550   Median :2.000
##  Mean   :6.588   Mean   :2.974   Mean   :5.552   Mean   :2.026
##  3rd Qu.:6.900   3rd Qu.:3.175   3rd Qu.:5.875   3rd Qu.:2.300
##  Max.   :7.900   Max.   :3.800   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    : 0
##  versicolor: 0
##  virginica :50
##
##
```

```
##
```

Coefficient of variation

The coefficient of variation can be found by computing manually (remember that the coefficient of variation is the standard deviation divided by the mean):

```
sd(dat$Sepal.Length) / mean(dat$Sepal.Length)
```

```
## [1] 0.1417113
```

Mode

```
tab <- table(dat$Sepal.Length) # number of occurrences for each unique value
sort(tab, decreasing = TRUE) # sort highest to lowest
```

```
##
##   5 5.1 6.3 5.7 6.7 5.5 5.8 6.4 4.9 5.4 5.6   6 6.1 4.8 6.5 4.6 5.2 6.2 6.9 7.7
##  10   9   9   8   8   7   7   7   6   6   6   6   6   5   5   4   4   4   4   4
## 4.4 5.9 6.8 7.2 4.7 6.6 4.3 4.5 5.3   7 7.1 7.3 7.4 7.6 7.9
##   3   3   3   3   2   2   1   1   1   1   1   1   1   1   1
```

Takeaway

- In R programming, basic descriptive statistic functions are simple and exactly same as in statistical defintions.