R Handout: Descriptive Statistics

This handout will introduce a basic use of R with an example describing a data. This pdf file and the R program are provided on my Github:

https://github.com/bolus123/R-handout/tree/master/DescriptiveStatistics

1. Load a data. Here, I will only show how to read a CSV file from my Github. You can also load a data from different types of data source, such as Excel, SQL server and etc.. Because I am not going to cover all of them, if you are interested in other methods, please see:

   https://cran.r-project.org/doc/manuals/r-devel/R-data.html

   This CSV file contains a data monitoring a manufacturing process about Automobile Engine Piston Rings [1]. Theoretically, each column is identically and independently distributed with each other.

   - R code

   ```
   # Load table from my Github
   add <- 'https://raw.githubusercontent.com/
   bolus123/R-handout/master/DescriptiveStatistics/example.csv'
   data <- as.matrix(read.csv(file = add))
   data
   ```

2. Describe this data with graphs

   - R code

   ```
   # Graph this data
   # histogram for the whole data with 20 breaks
   hist(data, breaks = 20)

   #boxplot for the whole data
   boxplot(as.vector(data))
   ```

```
#boxplot for each column
boxplot(data)
```

3. Describe this data with statistics

   - R code

   ```
   # Basic statistics
   mean(data) #grand mean
   colMeans(data) # means for each column
   rowMeans(data) # means for each row

   var(as.vector(data)) # grand variance
   var(data) #covariance matrix

   # percentiles including 1%, 5%, 10%, 25%,
   # 50%, 75%, 90%, 95% and 99%
   quantile(data
   , c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99))
   ```

4. Fit a model. Suppose we guess this data is following a normal distribution, and then fit a model.

   - R code

   ```
   # fit a model based on the univariate normal distribution
   # for the whole data
   mu <- mean(data) #grand mean
   sigma <- sqrt(var(as.vector(data))) #standard deviation

   # histogram for the whole data with 20 breaks
   hist(data, breaks = 20, freq = FALSE, ylim = c(0, 40))
   curve(dnorm(x, mean = mu, sd = sigma), add = T, col = 'blue')
   ```

5. Check the normality. We need to verify our normal assumption, because there is no guarantee that we are right. Here, I will show a verification by a Q-Q plot.

- R code

```
# check the normality (Q-Q plot)
# we need to have the empirical quantile
# and the theoretical quantile based
# on the empirical probability

# 1. we need to know the whole sample size
n <- dim(data)[1] * dim(data)[2]

# 2. sort the data and this is our empirical quantile
e.q <- sort(data)

# 3. calculate the Empirical cdf
e.p <- 1:n / n

# 4. find out the theoretical quantile
t.q <- qnorm(e.p, mean = mu, sd = sigma)

# 5. draw a Q-Q plot
plot(e.q, t.q, xlab = 'Empirical'
, ylab = 'Theoretical', main = 'Q-Q plot')
# reference line
points(c(0, 100), c(0, 100), type = 'l', col = 'blue')
```

■References References

[1] Douglas C. Montgomery, *Introduction to Statistical Quality Control*, Wiley, NJ, 6th edition, 2009.