Basic Statistics and Graphics I

C.-Y. Kuo, T. Eiting, M. Rosario

09.28.2012

1 Introduction

1.1 Welcome

This document covers the first part of doing basic statistics and graphics in R without installing any external packages. It includes descriptive statistics, checking for normality of data with statistical and graphical methods, performing t-tests and nonparametric two-group comparisons.

1.2 Getting started

It's always good to make sure that the workspace is empty when you start R. We also need to set our working directory. Then we can read in the data file.

```
# clear workspace
rm(list = ls())

# set working directory
setwd("D:/R")
# make sure the working directory is set up right
getwd()

## [1] "D:/R"


# read in data file CYAnoles.csv
anoles <- read.csv("CYAnoles.csv", header = T, na.string = "")
# make sure the data file was read in correctly
head(anoles)

##     species number tail perch.site pheight pdiameter pangle m5m
## 1 gundlachi    1.1    y      trunk     316      83.0     90   3
## 2 gundlachi    2.0    y   building     250      10.5     90   0
## 3 gundlachi    3.0    y      trunk      79       6.0     10   0
## 4 gundlachi    5.0    y      trunk     110      54.0     90   0
## 5 gundlachi    7.0    y      trunk     298      48.5     90   2
## 6 gundlachi    8.0    y      trunk     178     106.0     90   0
##   time.moving dm5m fid     tactics dmbs
## 1           3 25.0  76         run   76
## 2           0  0.0 126 squirelling    5
## 3           0  0.0 320 squirelling   48
## 4           0  0.0 114         run   25
## 5           2  9.1   0 squirelling    5
## 6           0  0.0  52 squirelling   83
```

2 Basic statistics and graphics

After importing data into R, we can begin to explore the data. In this section we follow the steps of how statistical analyses are routinely performed by starting from descriptive statistics, checking for normality of the data and data transformation. At the end of this section, we will compare the means of a variable between two groups using both parametric and nonparametric methods.

2.1 Descriptive statistics and graphics

First, we would like to know a few things about the variable of interest. For example, what are the means, standard deviation, ...etc?

```
# attach dataframe
attach(anoles)
# get the mean, median, variance, standard deviation, and the
# quantiles for the variable pdiameter
mean(pdiameter, na.rm = T)

## [1] 57.19

median(pdiameter)

## [1] 48

var(pdiameter)

## [1] 2624

sd(pdiameter)

## [1] 51.22

quantile(pdiameter)

##     0%    25%    50%    75%   100%
##   3.80  15.95  48.00  78.12 198.00
```

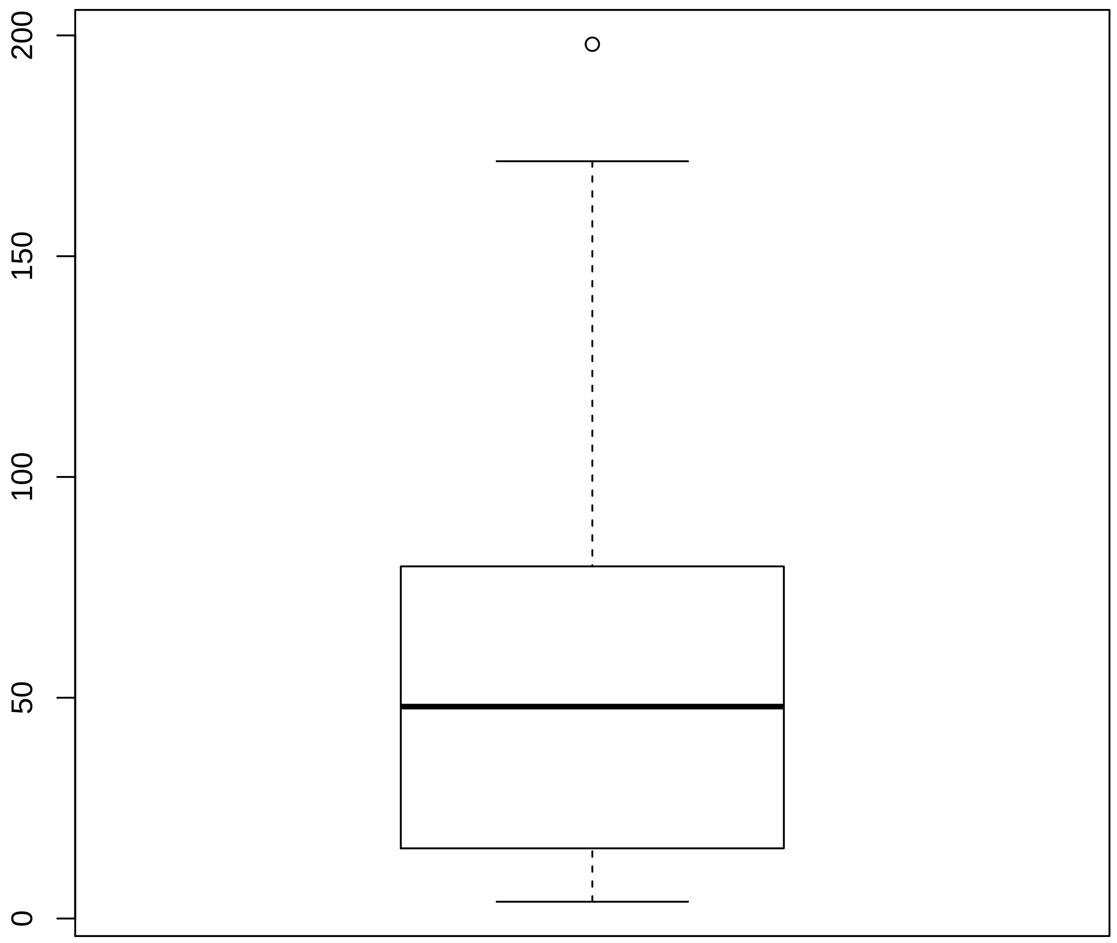
We can also use the function "summary" to get the values of mean, median and the quantiles all at once.

```
summary(pdiameter)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.8    16.0    48.0    57.2    78.1   198.0
```

(You might have noticed that we did not demonstrate how to obtain mode and standard error. The reason is that there is no built-in functions to obtain them in R. We have to write custom functions for that purpose.)

We can also examine the central tendency and dispersion of our data using boxplots.

```
# creating a boxplot for pdiameter
boxplot(pdiameter)
```



2.2 Checking the normality of data

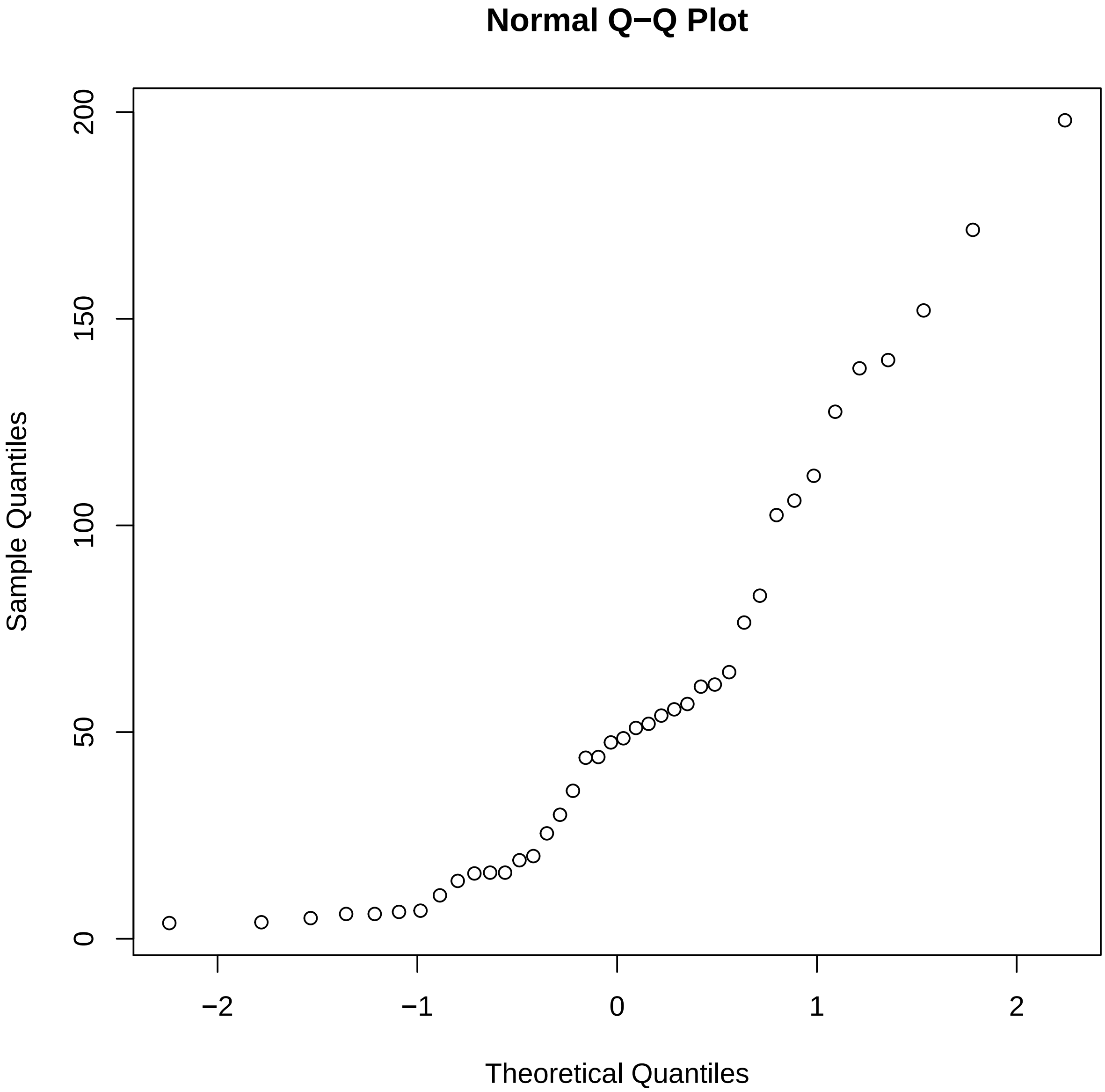
It is crucial to check if the data points are normally distributed before performing any statistical analyses. There are various statistical and graphical methods that allow us to assess the normality of our data. The statisitcal test we will introduce here is the Shapiro-Wilk Normality Test.

```
# Use Shapiro-Wilk test on pdiameter
shapiro.test(pdiameter)

##
##  Shapiro-Wilk normality test
##
## data:  pdiameter
## W = 0.8757, p-value = 0.0004035
```

You are testing for a significant *difference* from normality. Since the p-value is way lower than 0.05, pdiameter is very likely not normally distributed.

Another way to check for normality is using a graphical method called Q-Q plot (Q-Q stands for quantile versus quantile).

```
# Use Q-Q plot on pdiameter
qqnorm(pdiameter)
```



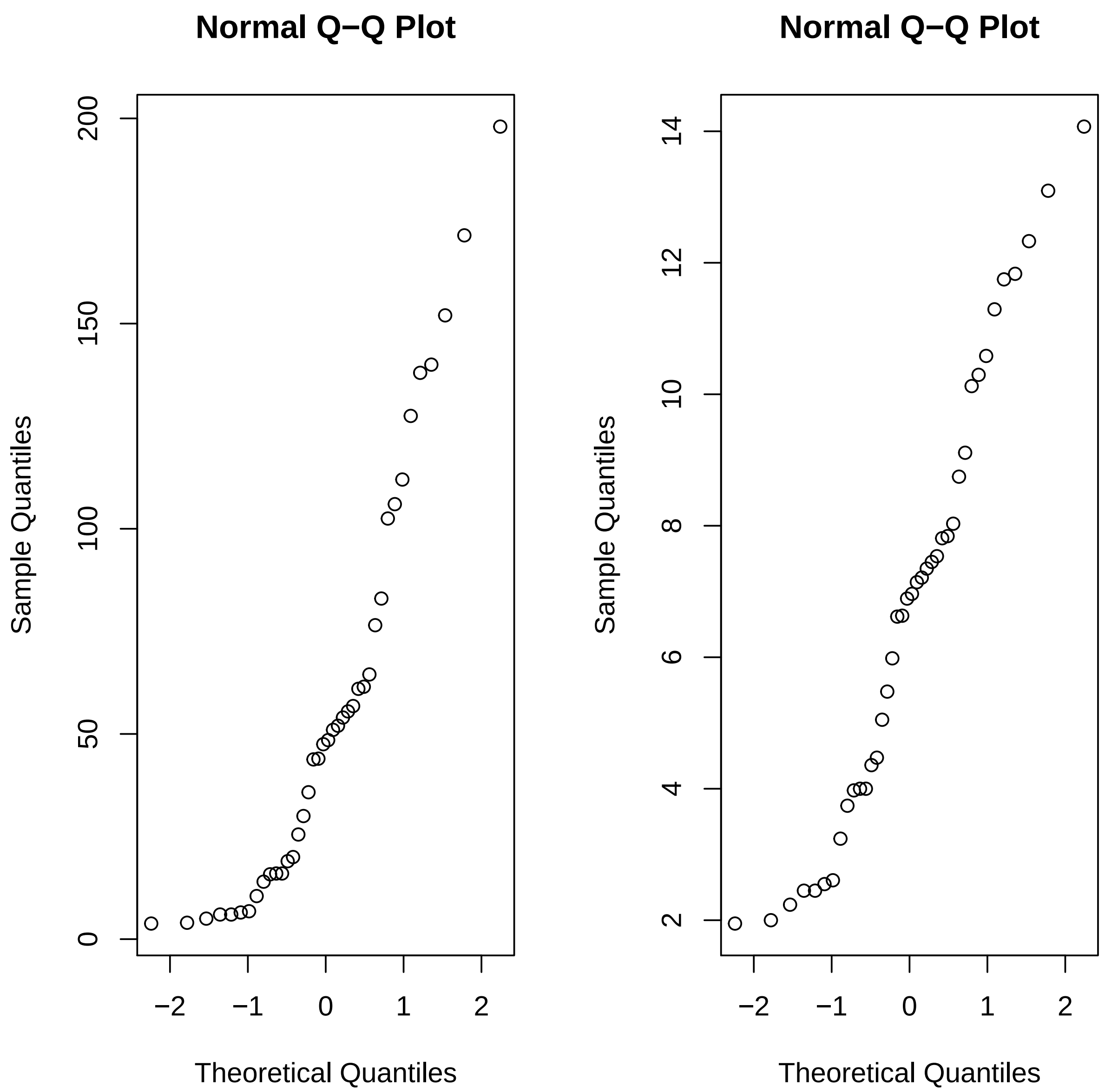
If the data is normally distributed, we should see a perfectly straight line. In other words, the more the data distribution deviates from normality, the less straight the line would be. From the Q-Q plot, again it seems that pdiameter is not normally distributed.

When the data deviates too much from normal distribution, it can be problematic to apply parametric statistics. One solution is data transformation. There are various options for data transformation. Which option is the most appropriate depends on the property of the data and is beyond the scope of this document. Here, we use square root transformation on pdiameter.

```
# square root transform pdiameter and make a new variable in the
# anoles.tail dataset
anoles$sqrt.pdiameter <- sqrt(pdiameter)
# detach and then re-attach anoles.tail so that the changes we made
# can take effect
detach(anoles)
attach(anoles)
# make sure the new variable is in the dataset
head(anoles)

##     species number tail perch.site pheight pdiameter pangle m5m
## 1 gundlachi    1.1    y      trunk     316      83.0     90   3
## 2 gundlachi    2.0    y   building     250      10.5     90   0
## 3 gundlachi    3.0    y      trunk      79       6.0     10   0
## 4 gundlachi    5.0    y      trunk     110      54.0     90   0
## 5 gundlachi    7.0    y      trunk     298      48.5     90   2
## 6 gundlachi    8.0    y      trunk     178     106.0     90   0
##   time.moving dm5m fid     tactics dmbs sqrt.pdiameter
## 1           3 25.0  76         run   76          9.110
## 2           0  0.0 126 squirelling    5          3.240
## 3           0  0.0 320 squirelling   48          2.449
## 4           0  0.0 114         run   25          7.348
## 5           2  9.1   0 squirelling    5          6.964
## 6           0  0.0  52 squirelling   83         10.296
```

We can use Shapiro-Wilk test or Q-Q plot again to see if square root transformed pdiameter conforms better to normal distribution.

```
# use Shapiro-Wilk test on square root transformed pdiameter
shapiro.test(sqrt.pdiameter)

##
##  Shapiro-Wilk normality test
##
## data:  sqrt.pdiameter
## W = 0.9515, p-value = 0.08526

# the Q-Q plot of untransformed vs. transformed pdiameter
par(mfrow = c(1, 2))
qqnorm(pdiameter)
qqnorm(sqrt.pdiameter)
```



It appears that the data transformation helped!

2.3 Two-group comparisons: t-test and Wilcoxon test

Now we would like to see whether the means of sqrt.pdiameter differ between lizards with and without tails. To do so, we can use a t-test.

```
# before conducting a t-test, we have to know whether the two
# groups to be compared have the same variance
var.test(sqrt.pdiameter ~ tail)

##
##  F test to compare two variances
##
## data:  sqrt.pdiameter by tail
## F = 0.8283, num df = 18, denom df = 20, p-value = 0.6925
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3311 2.1197
## sample estimates:
## ratio of variances
##             0.8283

# the var.equal argument specifies whether the two groups have the
# same variamnce; the paired augument specifies whether a paired
# t-test is performed; the alternative augument specifies whether
# it is a one-sided or two-sided test
t.test(sqrt.pdiameter ~ tail, var.equal = T, paired = F, alternative = "two.sided")

##
##  Two Sample t-test
##
## data:  sqrt.pdiameter by tail
## t = -0.5207, df = 38, p-value = 0.6056
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.758  1.630
## sample estimates:
## mean in group n mean in group y
##           6.485           7.049
```

If the data distribution deviates from normality even after data transformation, we can use a nonparametric test to compare the means between two groups, such as the Wilcoxon test.

```
wilcox.test(sqrt.pdiameter ~ tail)

## Warning:  cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  sqrt.pdiameter by tail
## W = 188.5, p-value = 0.7761
## alternative hypothesis: true location shift is not equal to 0
```

Judging from the p-values, lizards with and without tails do not differ in sqrt.pdiameter.

3 Exercises

3.1 In-class

1. Obtain the mean and standard deviation of the variable ”pheight” in the CYAnoles dataset. Make a boxplot to examine data distribution of pheight.

2. Use both Shapiro-Wilk test and Q-Q plot to see whether pheight is normally distributed. If not, square-root transform ”pheight” and add the new variable into the dataset. Write out the dataset (now with a new variable) as a csv file. PLEASE KEEP THIS NEW FILE! We will be using it next week.

3. Repeat 2 to test whether the square root transformed pheight is normally distributed.

3.2 Take-home

1. Based on the results of normality tests, use either a t-test or a Wilcoxon test to compare the means of square-root transformed ”pheight” between lizards with and without tails. Interpret the results.