Review on Basics of R (cont.) + Paired-Samples t Test

1. Objectives

- Basic graphical and tabular methods
- Editing graphs
- Paired-samples t-test

2. Basic tabular methods

Remember that before we apply a graphical or statistical method on a variable that is to be treated as a categorical variable, we should be sure that it has been converted into a factor.

We can convert a character variable into a factor by the **factor**() function. For example,

```
data1 <- read.table("mtcars.csv",header=TRUE,sep=",",quote="\"",
stringsAsFactors = FALSE) #Read data into R
data1$am <- factor(data1$am,ordered=FALSE,levels=c(0,1),labels =
c("automatic","manual")) #Convert into categorical variable
```

Exercise 1:

a. Explain why the conversion of the **`data1$am`** variable is necessary in the above code.
b. Let's check whether the **`data1$am`** variable has been converted into a factor correctly.
c. Go back to the code for importing a text file. What happens if we use **`stringAsFactors = TRUE`**? Try this:

```
data2 <- read.table("mtcars.csv",header=TRUE, sep=",",quote="\"",
stringsAsFactors=TRUE)
```

Now, let's create a frequency table. We can try:

```
am.table <- table(data1$am)
prop.table(am.table)
prop.table(am.table)*100
```

To create a contingency table, use the following format of the **table()** function:

```
tableName <- table(row variable, column variable)
```

Exercise 2: Create a contingency table named **gearVSam.table2** showing the relationship between gear and am. Are you happy with the output? Let's discuss how to improve it.

3. Basic graphical methods

3.1 Simple bar graph

Based on the frequency table produced previously, we can now produce a simple bar graph. The following listing shows different ways to plot a bar graph.

```
am.table <- table(data1$am)
barplot(am.table)
```

- `barplot(am.table, main="Bar graph of Transmission", xlab="Types of Transmission", ylab="Frequency",ylim=c(0,20))`
- `barplot(am.table, main="Bar graph of Transmission", xlab="Types of Transmission", ylab="Frequency", horiz=TRUE)`
- `barplot(am.table, col="skyblue",main="Bar graph of Transmission ", xlab=" Types of Transmission ", ylab="Frequency")`

3.2 Clustered bar graph

Let's type in the following code:

- `trans.vs.gear<-table(data1$am,data1$gear)`
- `trans.vs.gear`
- `barplot(trans.vs.gear, beside=TRUE)`
- `barplot(trans.vs.gear, col=c("red", "yellow"), beside=TRUE,ylim=c(0,20))`

Exercise 3: For the above clustered bar graph

a. Add a title for the graph and labels for the two axes.
b. Use different colors for the bars
c. Convert **gear** variable to a factor before producing a contingency table and clustered bar graph and observe the difference.

3.3 Stacked bar graph

The code below will produce a stacked bar graph

- `barplot(trans.vs.gear,col=c("red","yellow"))`

We can use the **spineplot()** function to produce a spine plot, a generalized version of the stacked bar graph. Let's observe how the spine plot differs from the previous stacked bar graph.

- `spineplot(trans.vs.gear, col=c("blue", "green", "pink"))`

3.4 Stem and leaf display

- mpg <- data1$mpg
- stem(mpg)

3.5 Histogram

- `hist(mpg)`
- `hist(mpg,breaks=5,col="red")`
- `hist(mpg, freq=FALSE, breaks=5,col="red")`

3.6 Boxplot

The following command is to work with boxplot (for numerical data):

- `boxplot(data1$mpg)`
- `boxplot.stats(data1$mpg)`
- `boxplot(data1$mpg ~ data1$gear)`

4. Editing graphs

4.1 Adding title and axis labels

The function **`title()`** adds title and axis labels to a graph. The general format is:

```
title(main="my title", sub="my sub-title", xlab="x-axis label", ylab="y-axis label")
```

The **`title()`** function works with the currently active graph.

4.2 Adding a box outside the graph

Use the **`box()`** function

5. Paired-samples t Test

We can use the following code to conduct a paired-samples t test to see if the population mean difference is not zero:

```
t.test (y1, y2, paired=TRUE, alternative = …,conf.level=0.95)
```

where $\mathbf{y_1}$ and $\mathbf{y_2}$ are numeric vectors for the two matched groups and **`conf.level`** argument allows us to specify the confidence level of the reported CI.

Exercise 4. Load the **GolfScores.csv** dataset. The dataset contains scores of the first and final rounds for a sample of 20 golfers who completed in PGA tournaments. Suppose you would like to determine if the mean score for the first round of a PGA Tour event is significantly different than the mean score for the fourth and final round. Use R to generate the test output. Use $\alpha = 0.1$.

a) What is the mean difference between in scores for the two rounds? For which round is the sample mean score lower?
b) What is the p-value? Was the mean score significantly different for the two rounds?
c) What is the 90% confidence interval estimate for the difference between two population means? Does this CI support your conclusion in part (b) (Does the interval include 0)?
d) Remember that in practice we have to check assumptions for each test we perform. Is the data distribution for the paired differences reasonably normal?

Note: To check the normality of a dataset, a histogram can be used (but a **QQ plot** is more useful). In case of small sample size, however, it is better to use the stem and leaf display and the qq plot to check if data is normally distributed. The R code for a qq plot is as follows.

- `qqnorm(data)#Compare quantiles of our data with theoretical normal quantiles`
- `qqline(data)# Add a line to a normal quantile-quantile plot passing through the first and third quartiles`

If the data is normally distributed, the data points should fall in a straight line. Departures from the line are indicative of a lack of normality.

The R output for this exercise is provided below. You are expected to write R code that produces the same output:

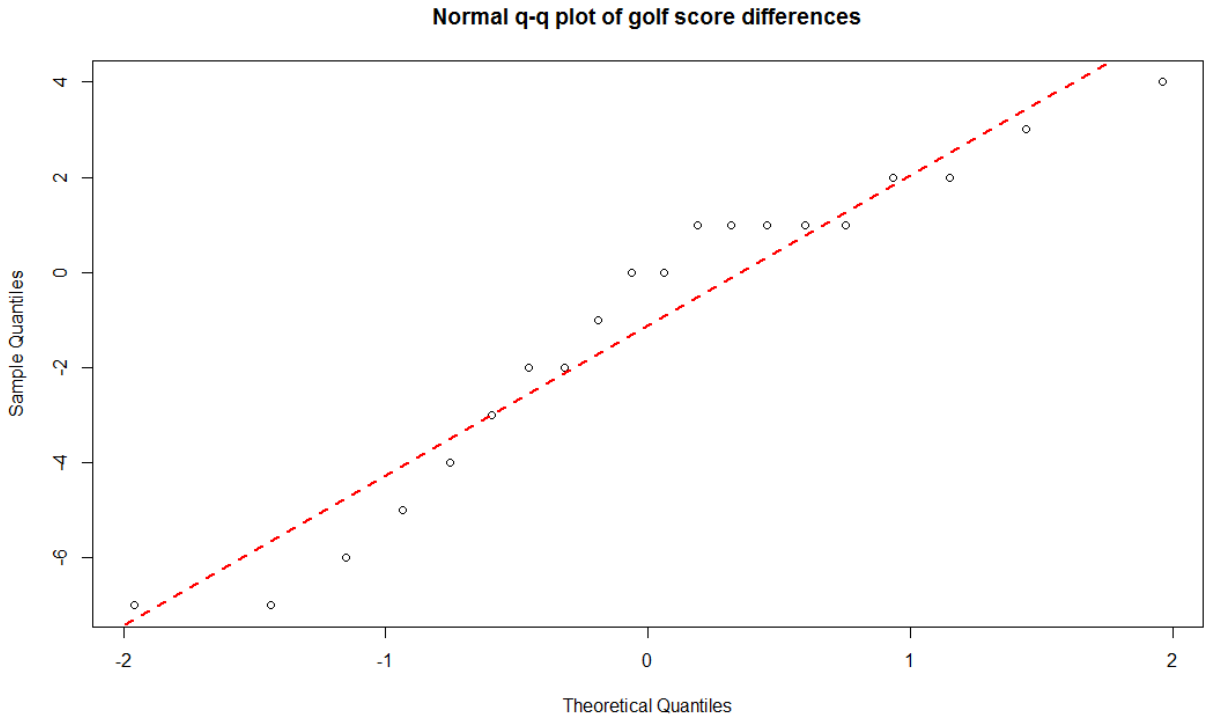
```
Paired t-test

data:  data3$First and data3$Final
t = -1.416, df = 19, p-value = 0.173
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -2.3322058  0.2322058
sample estimates:
mean of the differences
                  -1.05
```

Stem and Leaf Display of Golf Score Differences

```
The decimal point is 1 digit(s) to the right of the |

  -0 | 7765
  -0 | 43221
   0 | 00111112234
```



Note: If the assumption of normality is violated, the t test may provide misleading results (you should refer to the practical guidelines regarding how to use one-sample t-test in the Probability and Statistics course). In such cases, we should use a nonparametric test (to be taught later in this course).

Exercise 5. Load the **PriceChange.csv** dataset. In early 2009, the economy was experiencing a recession. The dataset contains data price per share of stock for a sample of 15 companies on January 1 and April 30 (The Wall Street Journal, May 1, 2009).

a. What is the change in the mean price per share of stock over the four-month period?
b. Provide a 90% confidence interval estimate of the change in the mean price per share of stock. Interprete the results.
c. How was the recession affecting the stock market? Use $\alpha = .1$

The R output for this exercise is provided below. You are expected to write R code that produces the same output:

```
Paired t-test

data:  data4$Jan and data4$Apr
t = 2.0043, df = 14, p-value = 0.06478
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 0.2970457 4.6029543
sample estimates:
mean of the differences
                   2.45
```

Stem and Leaf Display of Price Changes

```
The decimal point is 1 digit(s) to the right of the |

 -0 | 432211
  0 | 12344
  0 | 778
  1 | 2
```