Statistical modelling: practical 1 solutions

I've starred, *, some of the questions. This indicates that we didn't directly cover the material in the lecture. If you are particular interested in this statistical area, try the question. Otherwise, just move on.

Course R package

Installing the the course R package[1] is straightforward. First install `drat`:

[1] A package is an *add-on* or a *module*. It provides with additional functions and data.

```
install.packages("drat")
```

Then

```
drat::addRepo("rcourses")
install.packages("nclRmodelling", type="source")
```

This R package contains copies of the practicals, solutions and data sets that we require. To load the package, use

```
library("nclRmodelling")
```

1. Consider the following data set:

Method A					
78.64	79.01	79.57	79.52	80.71	79.95
78.50	79.10	81.98	80.09	80.29	80.22

(a) Input the data into R.[2]

[2] I intentionally didn't make the data available for download so you would have to think about how to enter the data.

```
##Data for question 1 & 2
## Easier using Excel and export as CSV
x = c(78.64,79.01, 79.57, 79.52, 80.71, 79.95, 78.50,
  79.10, 81.98, 80.09, 80.29, 80.22)
y = c(81.92, 81.12, 82.47, 82.86, 82.89, 82.45,
     82.51, 81.11, 83.07, 82.77, 82.38, 83.14)
dd = data.frame(x, y)
```

(b) Construct a boxplot and a histogram of the data.

```
##Graphics not shown
boxplot(dd$x)
hist(dd$x)
```

(c) Construct a q-q plot of the data.

```
qqnorm(dd$x)
qqline(dd$x)
```

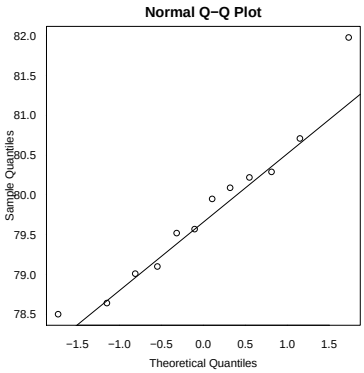

Figure 1: q-q plot from question 1c.

(d) What is the mean and standard deviation of this data.

```
##Mean & Sd
mean(dd$x)
```

```
## [1] 79.8
```

```
sd(dd$x)
```

```
## [1] 0.9716
```

(e) Carry out a one sample t-test at the 99% level, where

$$H_0 : \mu = 79 \quad \text{and} \quad H_1 : \mu \neq 79 \, .$$

```
##1-sample t-test
t.test(dd$x, mu=79, conf.level=0.99)
```

```
##
##  One Sample t-test
##
## data:  dd$x
## t = 2.8, df = 11, p-value = 0.02
## alternative hypothesis: true mean is not equal to 79
## 99 percent confidence interval:
##  78.93 80.67
## sample estimates:
## mean of x
##      79.8
```

(f) Calculate a 95% confidence interval for the population mean value.

```
##95% CI
t.test(dd$x, conf.level=0.95)$conf.int
```

```
## [1] 79.18 80.42
## attr(,"conf.level")
## [1] 0.95
```

(g) * Now carry out a `wilcox.test`. Compare the p-values of this test to the one sample t-test.

```
wilcox.test(dd$x)$p.value
```

```
## [1] 0.0004883
```

(h) * Use the `str` function to explore the output of the `wilcox.test` function.

(i) Imagine that this was a proper statistical analysis. Save your data as a `csv` file. Clean up your R script - commenting where necessary - and save it as a file. You should be able to open the file and reproduce your analysis.

2. ANOTHER EXPERIMENT (with different people) was carried out and the following data were collected

Method B					
81.92	81.12	82.47	82.86	82.89	82.45
82.51	81.11	83.07	82.77	82.38	83.14

(a) Input the data into R. Combine the two data sets into a single data frame.

```
## Suppose you have two separate data files. Here is some code that will help ## you combine them.
d1 = read.csv("Method1.csv")
d2 = read.csv("Method2.csv")

## In order to combine the data frames,
## they must have the same column names:
head(d1, 2)
```

```
##   value
## 1 78.64
## 2 79.01
```

```
head(d2, 2)
```

```
##   value
## 1 81.92
## 2 81.12
```

```
## We combine data frames using rbind (row bind)
d = rbind(d1, d2)

## Finally we create a new column to indicate the Method
## rep is the replicate function. See ?rep
d$Method = rep(1:2, each=12)
head(d, 2)
```

```
##   value Method
## 1 78.64      1
## 2 79.01      1
```

(b) Exploratory data analysis.

- Construct boxplots, histograms and q-q plots for both data sets. Work out the means and standard deviations. Before carrying out any statistical test, what do you think your conclusions will be? Do you think the variances are roughly equal? Do you think the data conforms to a normal distribution.

(c) Carry out a two sample t-test. Assume that the variances are unequal.

```
t.test(value ~ Method, data=d, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  value by Method
## t = -7.6, df = 20, p-value = 3e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.308 -1.877
## sample estimates:
## mean in group 1 mean in group 2
##           79.80           82.39
```

How does this answer compare with your intuition?

(d) Use the `var.test` function to test for unequal variances.[3]

[3] This function isn't in the notes. Look at the help file.

```
var.test(value ~ Method, data=d)
```

```
##
##  F test to compare two variances
##
## data:  value by Method
## F = 2, num df = 11, denom df = 11, p-value =
## 0.3
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.582 7.022
## sample estimates:
## ratio of variances
##              2.022
```

Since the $p = 0.2585$, this does not provide enough evidence to reject H_0 and so we don't have any reason to believe that the variances are unequal.

Does this correspond to your intuition?

(e) * Carry out a two sample t-test, assuming equal variances.

```
t.test(value ~ Method, data=d, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  value by Method
## t = -7.6, df = 22, p-value = 1e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.304 -1.881
## sample estimates:
## mean in group 1 mean in group 2
##           79.80           82.39
```

(f) * Now carry out a `wilcox.test`.

(g) * When carrying out the Wilcoxon test, we assume a common distribution. This assumption can be tested using the *Kolmogorov-Smirnov* test: `ks.test`.[4] Is the assumption of a com-

[4] Again this function isn't in the notes. Look at the help file.

mon distribution valid?

3. SUPPOSE WE are interested whether successful business executives are affected by their zodiac sign. We have collected 4265 samples and obtained the following data

Aries	Taurus	Gemini	Cancer	Leo	Virgo	Libra	Scorpio	Sagittarius	Capricorn	Aquarius	Pisces
348	353	359	357	350	355	359	367	345	362	343	367

Table 1: Zodiac signs of 4265 business executives

(a) Carry out a χ^2 goodness of fit test on the zodiac data. Are business executives distributed uniformly across zodiac signs?

```
x = c(348, 353, 359, 357, 350, 355, 359, 367, 345, 362, 343, 367)
m = chisq.test(x)
##Since p > 0.05 we can't accept the alternative hypothesis.
##However, the question is worded as though we can "prove" the Null
##hypotheis, which we obviously can't do.
```

(b) What are the expected values for each zodiac sign?

```
##expected values
(expected = m[["expected"]])
```

```
##  [1] 355.4 355.4 355.4 355.4 355.4 355.4 355.4
##  [8] 355.4 355.4 355.4 355.4 355.4
```

(c) The formula for calculating the residuals[5] is given by

$$\frac{\text{observed} - \text{expected}}{\sqrt{\text{expected}}}$$

[5] These residuals are called Pearson residuals.

Which residuals are large?

```
##Residuals
m[["residuals"]]
```

```
##  [1] -0.39340 -0.12819  0.19007  0.08399 -0.28732
##  [6] -0.02210  0.19007  0.61442 -0.55254  0.34920
## [11] -0.65862  0.61442
```

4. If you have brought your own data, feel free to try these techniques on this data.

5. THE University of Texas Southwestern Medical Center examined whether the risk of contracting Hepatitis C was related to tattoo use. The data from the study is summarised as follows:

Haley, R. and Fischer, P.R. 2001

(a) Carry out a χ^2 test to determine if the Hepatitis is related to tattoo status.

	Hepatitis C	No Hepatitis C	Total
Tattoo, Parlour	17	35	52
Tattoo, elsewhere	8	53	61
No tattoo	22	491	513
Total	47	579	626

Table 2: Counts of patients by their Hepatitis C status and whether they had a tattoo from a parlour, from elsewhere or had no tattoo at all.

```
h = c(17, 8, 22)
nh = c(35, 53, 491)
dd = data.frame(h, nh)
m = chisq.test(dd)
```

(b) When carrying out χ^2 tests, we should make sure that individual cells have expected values of at least five, otherwise the distributional assumptions may be invalid. What are the expected values of each cell. Which cells have an expected value less than five?

```
m[["expected"]]
```

```
##           h     nh
## [1,]  3.904  48.10
## [2,]  4.580  56.42
## [3,] 38.516 474.48
```

(c) Since some of the cells have expected values slightly less than five, we should ensure that these aren't driving the test statistic. Look at the test residuals. Which residuals are large? What should you do now?

```
##Some of the expected values are less then 5
##So consider combining cells.
```

Solutions

Solutions are contained within this package:

```
library("nclRmodelling")
vignette("solutions1", package="nclRmodelling")
```