Statistical modelling: practical 1

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]

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

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

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

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

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

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

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

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

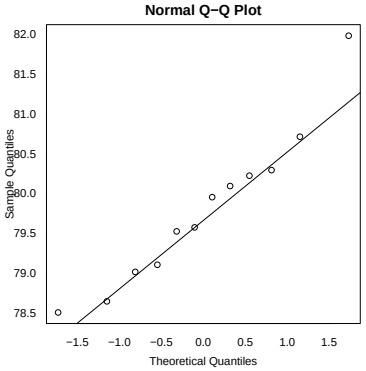

Figure 1: q-q plot from question 1c.

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

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

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

   (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.

   How does this answer compare with your intuition?

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

   (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 common distribution valid?

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

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

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?

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

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

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

   Which residuals are large?

[5] These residuals are called Pearson residuals.

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

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.

	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

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

(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?

(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?

Solutions

Solutions are contained within this package:

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