Solutions 1

Jumping Rivers

Simple Hypothesis Testing

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

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

1. We conducted an experiment and collected the data in the tables above. This data set isn't paired.[1]

   a) Input the data into [2]. Combine the two data sets into a single data frame.

```
## Data for question 1 Easier using Excel and export
## as CSV
x = c(78.64, 79.01, 79.57, 79.52, 80.71, 79.95, 78.5,
    79.1, 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)

## Suppose you have two separate data files. Here is
## some code that will help ## you combine them.
## First we read in the separate files:
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)
```

[1] I intentionally didn't make the data available for download so you would have to think about how to enter the data. You could enter it either Excel and import or directly into R.

[2] Here I would suggest input the data into Excel and using `read_csv()`

```
##   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 = -7.5603, df = 19.743, p-value = 3e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.308393 -1.876607
## sample estimates:
## mean in group 1 mean in group 2
##        79.79833        82.39083
```

How does this answer compare with your intuition?

d) 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 = -7.5603, df = 22, p-value = 1.489e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.30365 -1.88135
## sample estimates:
## mean in group 1 mean in group 2
##        79.79833        82.39083
```

2. Suppose we are interested whether successful business executives are affected by their zodiac sign. We have collected 4265 samples and obtained the following data

```
data(zodiac, package = "jrAnalytics")
head(zodiac)
```

```
##     sign count
## 1  Aries   348
## 2 Taurus   353
## 3 Gemini   359
## 4 Cancer   357
## 5    Leo   350
## 6  Virgo   355
```

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

```
x = zodiac$count
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] 355.4167 355.4167 355.4167 355.4167 355.4167
##  [6] 355.4167 355.4167 355.4167 355.4167 355.4167
## [11] 355.4167 355.4167
```

c) The formula for calculating the residuals [3] is given by

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

Which residuals are large?

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

```
##  [1] -0.39340499 -0.12818814  0.19007207
##  [4]  0.08398534 -0.28731825 -0.02210140
##  [7]  0.19007207  0.61441903 -0.55253510
## [10]  0.34920218 -0.65862184  0.61441903
```

One way ANOVA tables

1. A pilot study was developed to investigate whether music influenced exam scores. Three groups of students listened to 10 minutes of Mozart, silence or heavy metal before an IQ test. The results of the IQ test are as follows

Mozart	109	114	108	123	115	108	114
Silence	113	114	113	108	119	112	110
Heavy Metal	103	94	114	107	107	113	107

a) Construct a one-way ANOVA table. Are there differences between treatment groups?

```
x1 = c(109, 114, 108, 123, 115, 108, 114)
x2 = c(113, 114, 113, 108, 119, 112, 110)
x3 = c(103, 94, 114, 107, 107, 113, 107)
dd = data.frame(values = c(x1, x2, x3), type = rep(c("M",
    "S", "H"), each = 7))
m = aov(values ~ type, dd)
summary(m)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## type         2  193.1   96.57   3.401 0.0559 .
## Residuals   18  511.1   28.40
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## The p value is around 0.056. This suggests a
## difference may exist.
```

[3] These residuals are called Pearson residuals. Hint: use `str(m)` to extract the residuals.

b) Check the standardised residuals of your model.

```
plot(fitted.values(m), rstandard(m))
## Residual plot looks OK
```

c) Perform a multiple comparison test to determine where the difference lies.

```
TukeyHSD(m)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = values ~ type, data = dd)
##
## $type
##           diff        lwr       upr     p adj
## M-H  6.5714286 -0.6981512 13.841008 0.0804419
## S-H  6.2857143 -0.9838655 13.555294 0.0970627
## S-M -0.2857143 -7.5552941  6.983865 0.9944700
```