Tutorial 2: Descriptive Statistics and Exploratory Data Analysis Answers Sheet

Rob Nicholls – nicholls@mrc-lmb.cam.ac.uk

MRC LMB Statistics Course 2014

Task 1

1.

```
library(boot)
```

2.

```
hist(paulsen$y,breaks=30,freq=FALSE)
```

3.

```
lines(density(paulsen$y),col="red")
```

Task 2

1.

```
hist(rivers,breaks=30,freq=FALSE)
```

2.

```
lines(density(rivers),col="red")
```

3.

```
abline(v=mean(rivers)->x,col="orange")
abline(v=median(rivers)->y,col="blue")
```

4.

`(x-y)/(x+y)`

The relative difference is ≈ 0.164.

`x/y`

The mean is ≈ 1.39 times (or 39%) larger than the median.

Task 3

1.

```
hist(rivers,breaks=30,freq=FALSE)
```

2.

```
lines(density(rivers),col="red")
```

3.

```
x=quantile(rivers,prob=c(0.05,0.95))
abline(v=x,col="blue")
```

Task 4

1.

```
hist(rnorm(1000)->x,breaks=30)
```

2.

```
y=mean(x)+2*sd(x)*c(-1,1)
```

3.

```
c(sum(x<y[1]),sum(x<y[2]))/length(x)
```
These values are close to the 2.5% and 97.5% quantiles.

4.

Repeating the calculations yields results fluctuating around the 2.5% and 97.5% quantiles.

Task 5

1.

```
hist(faithful$eruptions)
```
The distribution is bimodal.

2.

```
qqnorm(faithful$eruptions)
qqline(faithful$eruptions)
```
The data are not Normally distributed.

3.

```
qqnorm(abs(faithful$eruptions-3))
qqline(abs(faithful$eruptions-3))
```

According to visual inspection alone, the transformed data appear to be approximately Normally distributed.

Without very good reason to do so (based on the mechanics of the system), there is no good reason why such a transformation would be applied to such a dataset. It is sometimes useful to transform data so that the distribution displays preferable characteristics (e.g. being able to be considered Normally distributed), although doing so is only valid for monotonic transformations – i.e. so that the original data could be recovered without any information loss.

In the example, the `abs(faithful$eruptions-3)` transformation is artificially applied in order to make the data appear Normally distributed. However, it would not be possible to recover the original data after transformation. Therefore, the application of such a transformation would not be a statistically valid protocol.

Task 6

1.

```
boxplot(PlantGrowth$weight)
```

2.

```
boxplot(PlantGrowth$weight~PlantGrowth$group)
```

Task 7

1.

```
boxplot(chickwts$weight)
```

2.

```
boxplot(chickwts$weight~chickwts$feed)
```

There is visual evidence for differences between the groups. Overall, 'Horsebean' seems to result in particularly light chickens, and 'Casein' and 'Sunflower' seem to result in particularly heavy chickens, on average.

Task 8

1.

```
boxplot(DNase$density)
```

2.

```
boxplot(DNase$density~DNase$conc)
```

There appears to be a strong positive non-linear relationship between optical density and concentration.

Task 9

1.

```
library(boot)
```

2.

```
plot(calcium)
```

3.

```
cor(calcium)
```

4.

```
cor(calcium,method="spearman")
```

There is not a very large difference between the Pearson (0.87) and Spearman (0.91) correlation coefficients. The lack of difference may be due to (1) the relationship being strong and relatively linear; and (2) the relationship not being monotonic. The fact that the Spearman correlation coefficient is larger than the Pearson might be explained to some degree by the non-linearity of the relationship.

Task 10

1.

```
library(boot)
```

2.

```
plot(survival)
```

3.

```
cor(survival)
```

4.

```
cor(survival,method="spearman")
```

The Spearman (-0.91) correlation coefficient is considerably stronger than the Pearson (-0.68). This is due to the strong non-linear and semi-monotonic nature of the relationship.

5.

```
a = log(1/survival$surv)
plot(a∼survival$dose)
```

6.

```
cor(a,survival$dose)
cor(a,survival$dose,method="spearman")
```

Both Pearson and Spearman correlation coefficients change from negative to positive. The Pearson correlation becomes much stronger in magnitude after the transformation – this is because the original data displayed a negative non-linear correlation, whilst the transformed data exhibit a positive linear correlation. In contrast, the Spearman correlation coefficient adopts the exact same value, except for the sign change – this is because the ordering is perfectly preserved, but reversed.

Task 11

1.

```
cor(iris[iris$Species=="virginica",1:4])
cor(iris[iris$Species=="setosa",1:4])
cor(iris[iris$Species=="versicolor",1:4])
```

or better:

```
x=levels(iris$Species)
cor(iris[iris$Species==x[1],1:4])
cor(iris[iris$Species==x[2],1:4])
cor(iris[iris$Species==x[3],1:4])
```

or better:

```
for(i in 1:length(levels(iris$Species)->x)){
print(x[i])
print(cor(iris[iris$Species==x[i],1:4]))
}
```

2.

The previous assertion that Petal.Length and Petal.Width are highly positively correlated is misleading. When looking at individual species, it transpires that the within-species correlation between between Petal.Length and Petal.Width is much weaker. The correlation is strongest for Versicolor (0.79), being much weaker for Setosa (0.33) and Virginica (0.32). The reason for the higher positive correlation when combining the three species (0.96) is due to systematic differences in the Petal.Length and Petal.Width between the species. Indeed, it may be true that species with a larger Petal.Length would also tend to have a larger Petal.Width (although note that we could not draw such conclusions without testing hypotheses properly).

3.

Again, the previous assertion that Petal.Length and Sepal.Length are highly positively correlated is misleading. The correlation is strong for Versicolor (0.75) and Virginica (0.86), but is much weaker for Setosa (0.27). Consequently, there is not a strong correlation between Petal.Length and Sepal.Length for all species. The reason for the higher positive correlation when combining the three species (0.87) is again partially due to systematic differences between the species.

4.

The previous assertion that Petal.Length and Petal.Width are negatively correlated with Sepal.Width is highly misleading. Considering the within-species correlations, we see that both Petal.Length and Petal.Width are positively correlated with Sepal.Width, for all species. These correlations are strongest for Versicolor, slightly weaker for Virginica, and particularly weak for Setosa. The reason for the higher positive correlations when combining the three species is again due to systematic differences between the species (in this case Setosa appears to have a systematically lower Petal.Length and Petal.Width and systematically higher Sepal.Width than the other two species, although to draw such conclusions we would have to test such hypotheses properly).

5.

Again, the previous assertion that Sepal.Length and Sepal.Width are uncorrelated (but negatively correlated, if at all) is highly misleading. Considering the within-species correlations, we see that both Sepal.Length is positively correlated with Sepal.Width for all species. From these results, we may conclude that (1) these are not uncorrelated, and (2) they are not negatively correlated. Both of these statements disagree with the previous assertion. The reason for the difference in the observered correlation when combining the three species is again due to systematic differences between the species.

6.

This exercise demonstrates that it is important to interpret results carefully, as such analyses may be prone to error or misinterpretation. It is important to know exactly what questions you want to ask, and know exactly what conclusions can be drawn from different pieces of information (results). For example, when determining whether or not Petal.Length and Petal.Width are correlated, it is important to know whether you want to make general statements about the relationship between Petal.Length and Petal.Width over all plants in the study (thus ignoring species type), or investigate the correlations observed for each species separately.

Furthermore, it is important to explore data in detail, using different approaches, in order to gain maximal information. If there are different groupings in the data (e.g. different species) then can the distributions of some property be considered the same for each species? Also, can the relationships between different properties be considered the same for each species? Modelling such systems and consequently testing such hypotheses is an important part of statistics and data analysis.

The example also demonstrates how correlations can be achieved in cases where the compared data are heterogeneous, and exhibit non-linear relationships. In such cases, correlation coefficients should be interpreted carefully.

Task 12

1.

```
boxplot(CO2$uptake~CO$Type)
```

Visually, it does appear that there is a difference between the CO2 uptake rates for plants originating from the different regions.

2.

```
boxplot(CO2$uptake~CO2$Plant)
boxplot(CO2$uptake~CO2$Plant,col=c(rep("red",6),rep("blue",6)))
```

3.

```
boxplot(CO2$uptake~CO2$conc)
```
Higher concentrations tend to result in higher uptake. The relationship seems more pronounced at lower concentrations.

4.

```
boxplot(CO2$uptake~CO2$conc+CO2$Type)
```
This representation reinforces previous assertions.

5.

```
boxplot(CO2$uptake~CO2$Treat)
```
It appears that there are differences in the distributions, but whether or not the differences are significant is not clear.

6.

```
boxplot(CO2$uptake~CO2$Treat+CO2$Type)
boxplot(CO2$uptake~CO2$Treat+CO2$Type,col=rep(c("red","blue"),2))
```

7.

```
x = boxplot(CO2$uptake~CO2$Treat+CO2$Type+CO2$conc)
```

Looking at the contents of the box plot object `x` reveals the existence of the `names` attribute, which is a vector specifying which level combinations each of the 28 box plots correspond to, in the same order as displayed. This attribute would also have been revealed by the `summary` function. Manual inspection of this vector (i.e. `x$names`) reveals the pattern required to colour the box plots accordingly:
`boxplot(x,col=rep(c("red","blue","orange","green"),7))`

Given a particular value of conc (concentration), we can see that the effect of Plant (i.e. within-group variation) is small relative to the effect of Type or Treat. It seems that both Type and Treat have a substantial effect on uptake, at all concentrations. Certainly, given the Type, the nonchilled (red and orange) plants have systematically higher uptakes than the chilled (blue and green) plants. However, note that there are some nonchilled plants (orange) that have a systematically lower uptake than some chilled plants (blue); the difference between these being the Type. Furthermore, note that the uptake is systematically lower for Mississippi (orange and green) than for Quebec (red and blue). Consequently, we might conclude from visual analysis that changing the Type from Quebec to Mississippi is generally more inhibitive for uptake than changing Treatment or Plant. Such assertions would ordinarily be confirmed or disproven by modelling and testing hypotheses.