Introduction to Statistical Analysis

Mark Dunning and Sarah Vowler

Last modified: 06 Nov 2015

Contents

1 Introduction

In this practical, we will use several 'read-life' datasets to demonstrate some of the concepts you have seen in the lectures. We will guide you through how to analyse these datasets in Shiny and the kinds of questions you should be asking yourself when faced with similar data.

To answer the questions in this practical we will be using apps that we have developed using the ***Shiny*** add-on for the *R* statistical package. **R** is a freely-available open-source software that is popular within academic and commercial communities. The functionality within the software compares favourably with other statistical packages (SAS, SPSS and Stata). The downside is that **R** has a steep learning-curve and requires a basic familiarity with command-line software. To ease the transition we have chosen to present this course using a series of online tools that will allow you to perform statistical analysis without having to worry about learning R. At the same time, the R code required for the analysis will be recorded in the background. You will therefore be able to repeat the analysis at a later date, or pass-on to others. As you gain familiarity with R through other courses, you will see how the code generated by Shiny can be adapted to your own needs.

The datasets you will need for this practical should be ***downloaded and unzipped now***

2 T-tests practical

2.1 The effect of disease on height

A scientist knows that the mean height of females in England is 165cm and wants to know whether her patients with disease X have heights that differ significantly from the population mean - we will use a one-sample t-test to test this. The data are contained in the file **diseaseX.csv** and can be analysed online at:-

http://bioinf-rstud001:3838/OneSampleTest/

a) What are your null and alternative hypotheses?

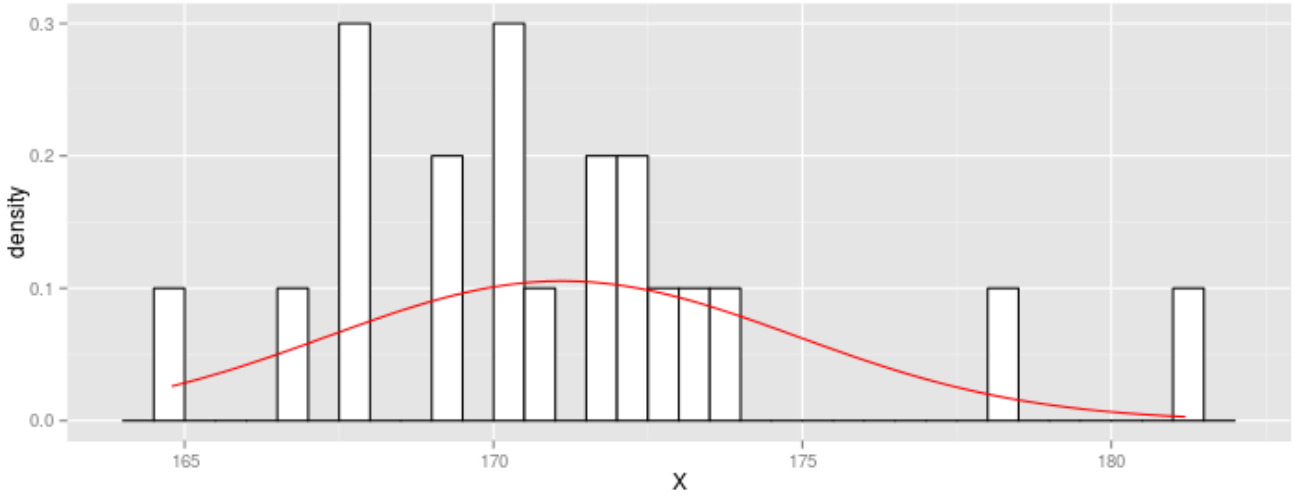
```
## Null hypothesis: The mean height of female patients with disease X = 165cm
##       (the population mean for females)
```

```
## Alternative hypothesis: The mean height of female patients with disease X != 165cm
##       (the population mean for females)
```

To import the file `diseaseX.csv` into ***Shiny*** you will need to select the `Choose File` option and navigate to where the course data are located on your laptop. You can use the ***The data*** tab to check that the data has be imported correctly.

b) A histogram of the `Height` variable will be automatically generate for you. To view it, click on the ***Data Distribution***

Do the data look normally distributed? Based on the histogram, is the one-sample t –test appropriate?




```
## In this case the data look normally distributed. Therefore, the one-sample t-test is appropriate.
```

c) We are interested in knowing whether the mean height in our sample of patients with disease X is different from that of the general population. Perform a **one-sample t-test**

Remember to change the value of ***True mean***.

What is the mean height in your sample? What is your value of t? What is the p-value? How do you interpret the p-value?

```
##
##  One Sample t-test
##
## data:  X
## t = 202.38, df = 19, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  169.3428 172.8822
## sample estimates:
## mean of x
##  171.1125
```

```
## Mean height in sample = 171.1cm (95% CI: 169.3-172.9)
## t = 7.23,
## df = 19,
## p  <= 0.0001.
##
## Under the null hypothesis, the probability of observing a t-statistic as extreme as 7.23,
## is very small P(t <= 7.23 | t >= 7.23) < 0.0001). Therefore, there is strong evidence
## to reject the null hypothesis in favour of the alternative hypothesis. There is strong evidence
## to suggest that the mean height in female patients with disease X is different to the
## population mean height of females of 165cm.
```

2.2 Biological processes duration

In the file **bp_times.csv**, we have the durations of a biological process for two samples of wild-type and knock-out cells (times in seconds). We are interested in seeing whether there is a difference in the durations for the two types of cells – we shall use an **independent t-test** to compare the two cell-types.

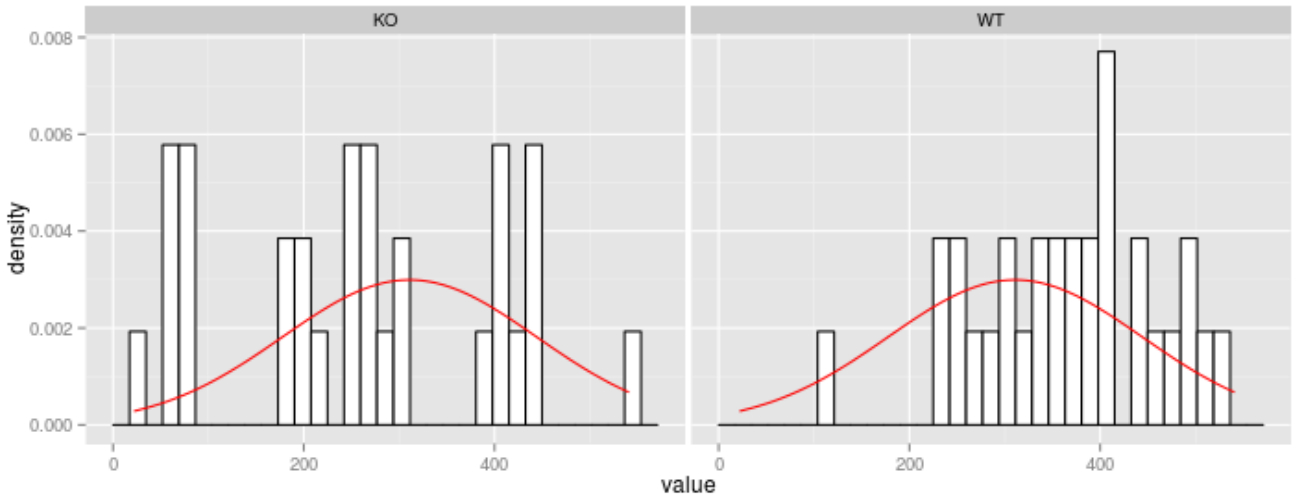
These data can be analysed online at http://bioinf-rstud001:3838/TwoSampleTest/

a) What are your null and alternative hypotheses?

```
## Null hypothesis: there is no difference in the duration of the
## biological process for the two cell types.
##
## Alternative hypothesis: there is a difference in the duration of the
## biological process for the two cell types.
```

Import the data using ***Choose File*** as before. Make sure that the ***1st column is a factor?*** checkbox is ticked.

b) Histograms to compare the two groups will be created for you automatically. Do the data look normally distributed for each cell-type? Is the independent t-test appropriate.

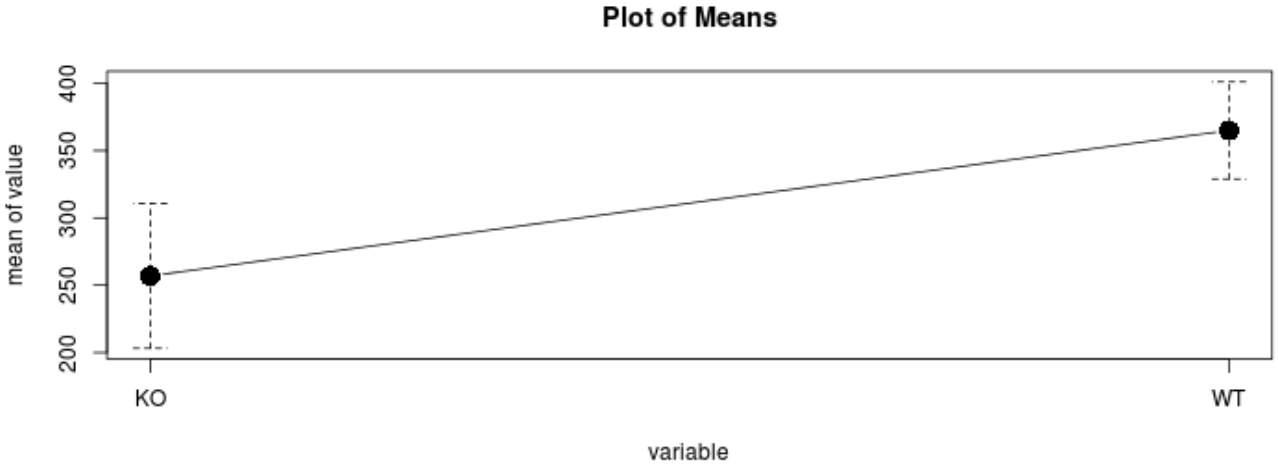



```
## The data do appear to be approximately normally distributed as we could
## easily draw a bell shape over each of the two histograms. The independent t-test is appropriate
```

c) Use the *Numerical statistics* analysis to compute descriptive statistics for each group.

```
## $KO
##      mean       sd   IQR 0% 25% 50%   75% 100%  n
##  256.8333 143.9413 219.5 22 177 256 396.5  541 30
```

```
##
## $WT
##      mean       sd    IQR  0%    25% 50% 75% 100%  n
##  365.0333 96.6335 118.75 118 310.25 374 429  530 30
```



Given the distribution of your data, which statistics might you report to summarise your data? Look at and compare the 95% confidence intervals of the mean durations of the two cell-types. Do they overlap?

```
## The normality assumption seems reasonable so the mean and standard deviation
## (or standard error or 95% CI) provide a good summary of the data.
## If the data were skewed, the median and interquartile range would be more meaningful.
## The confidence intervals do not overlap.
```

d) In order to apply the correct statistical test, we need to test to see if the variances of the two groups are comparable. This is tested for us automatically in the Shiny app. Click the ***Histogram*** tab to see the result.

What do you conclude from the p-value of this test. How does it influence what test to use?

```
## The test yields a p-value of 0.03568, which is sufficient evidence
##     to reject the null hypothesis that the variances of the two groups are the same.
##     Therefore we should apply Welch's correction.
```

e) Use the appropriate test to compare the durations of the two groups.

Is a Welch's correction needed? What is your value of t? What is the p-value? How do you interpret the p-value? Is this in agreement with the 95% confidence intervals?

```
##
##  Welch Two Sample t-test
##
## data:  value by variable
## t = -3.4183, df = 50.727, p-value = 0.00125
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -171.75411  -44.64589
## sample estimates:
## mean in group KO mean in group WT
##         256.8333         365.0333
```

```
## When we run the independent (unpaired) t test, a formal comparison of the variances
## between the two groups is automatically run. The corresponding p-value is 0.00125 which indicates
```

```
## evidence to reject the null hypothesis of equal variances between the two groups.
## Therefore, Welch's correction is needed. We can also look at the histograms and summary statistics
## to see whether we might need to use the Welch's correction. We can see that the standard deviation
## (i.e. spread of data) in each of our groups is quite different (96.63 for WT and 143.94 for KO).
## The widths of the base of our histograms are also different, though not by a very large amount.
## Both of these suggest a Welch's correction may be needed.
```

2.3 Blood vessel formation

In blood plasma cancer, there is an increase in blood vessel formation in the bone marrow. A stem cell transplant can be used as a treatment for blood plasma cancer. The bone marrow micro vessel density was measured before and after treatment for 7 patients with blood plasma cancer.

We are interested in seeing whether there is a decrease in the bone marrow micro vessel density after treatment with a stem cell transplant. We will use a paired two-sample t-test to compare the before and after bone marrow micro vessel densities.

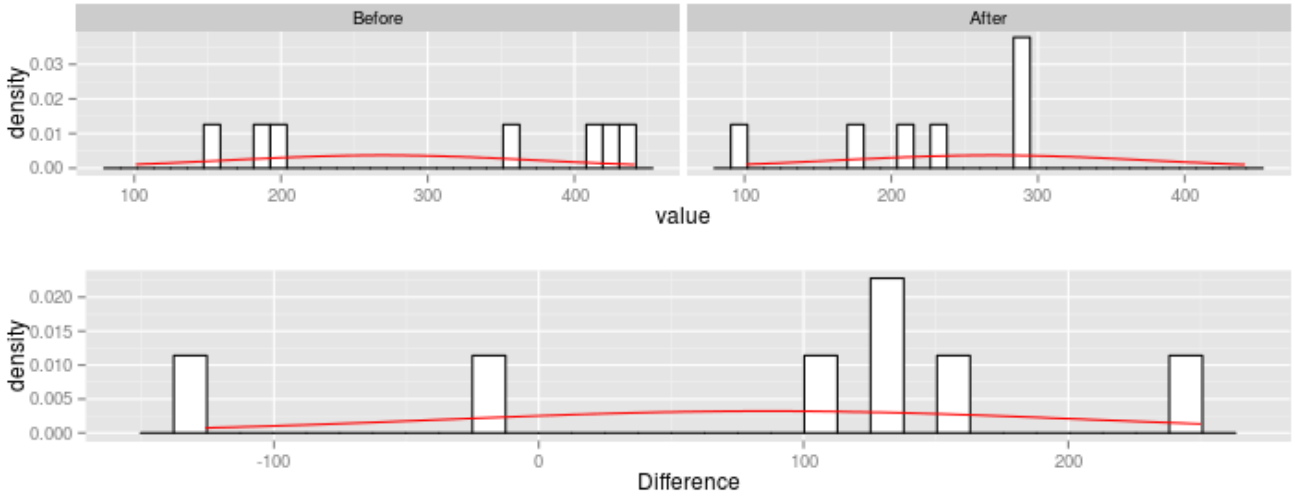
The data are contained in the file **bloodplasmacancer2.csv**. These data can be analysed online at http://bioinf-rstud001:3838/TwoSampleTest/

a) What are your null and alternative hypotheses?

```
## Null hypothesis: the bone marrow micro vessel density after treatment is greater
## than or equal to the bone marrow micro vessel density before treatment.
##
## Alternative hypothesis: the bone marrow micro vessel density after treatment
## is less than the bone marrow micro vessel density before treatment.
```

Import the data and create a column of differences (after-before).

b) Plot a histogram of the differences. Do the data look normally distributed? Is the paired t –test appropriate?




```
## From this histogram it is difficult to tell whether the differences between
##   the densities before and after treatment are normally distributed. In situations
##   like this, we may need to draw on the experience of similar sets of measurements.
```

c) We are interested in seeing whether there is a decrease in the bone marrow micro vessel density after treatment with a stem cell transplant. Is this a one-tailed or two-tailed test?

```
## One-tailed as we are only interested in a decrease.
```

```
## Usually a two-sided test is preferred unless there is a strong argument
## for a one-sided test. In this case our treatment is only considered to be effective
## if we see a reduction in the bone marrow micro vessel density after treatment.
## Observing an increase in bone marrow density after treatment would lead to the same
## action/conclusion as if no difference had been observed - the treatment might be
## dropped from the research programme (but bear in mind here,
## the sample size is small and so only large differences may be detected).
```

d) Compare the durations before and after values. Ensure you select the one- or two-tailed test as appropriate. What is the mean difference? What is your value of t? What is the p-value? How do you interpret the p-value?

```
##
##  Paired t-test
##
## data:  value by variable
## t = 1.8425, df = 6, p-value = 0.05749
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -4.709347       Inf
## sample estimates:
## mean of the differences
##                86.14286
```

```
## Under the null hypothesis, the probability of observing a t-statistic as extreme as 1.84, is 0.06,
## slightly greater than 0.05, our nominal significance level. Our result is borderline.
## There is insufficient evidence to reject the null hypothesis.
## Therefore, we might conclude that there is an association of a decrease in bone marrow micro vessel
## density after treatment with bone marrow transplant. It is important to note the small sample size here
## Studying just 7 patients means we will only be able to detect large differences.
```

2.4 Birth-weight of twins

Dr D. R. Peterson of the Department of Epidemiology, University of Washington, collected the data found in file **`twins.csv`**. It consists of the birth-weights of each of 20 dizygous twins. The hypothesis to be tested is that the SIDS child of each pair has a lower birth-weight.

a. Construct the null and alternative hypothesises
b. Decide on the level of significance to be used and whether the test should be one-sided or two-sided.
c. Carry out both the sign and Wilcoxon signed rank tests on the data. Do both tests draw the same conclusion about the data? Which test is the most appropriate?

2.5 Vitamin D levels

The file **`vitd.csv`** contains data on vitamin D levels for subjects with fibrosis.

a. Use the Mann-Whitney U and median tests to compare vitamin D levels between those with and without fibrosis. Interpret the results from both tests. Do both tests reach the same conclusion?
b. Which test is the more appropriate?

3 Tests for categorical variables

3.1 Nucleotide frequency

In **Table 1**, we have the frequencies of the four nucleotides in two sequences. We are interested in comparing the nucleotide proportions of the two sequences.

	A	C	G	T	Total
Sequence 1	273.00	233.00	236.00	258.00	1000.00
Sequence 2	281.00	246.00	244.00	229.00	1000.00
Total	554.00	479.00	480.00	487.00	2000.00

Table 1: Nucleotide frequencies for two sequences

a) What are your null and alternative hypotheses?

```
## Null hypothesis. there is no association between sequence number and nucleotide.
##
## Alternative hypothesis. there is an association between sequence number and nucleotide.
```

We can analyse these data online in the Shiny app, by modifying the contents of the ***Enter your data as a table*** box. Columns need to be separated by a '-' character, and rows by a '|'. You can check that you have entered the data correctly by looking at ***The data*** tab. You do not need to calculate row or column totals.

Note that you do not need to enter the totals.

What is your value of your Chi-squared statistic and its corresponding p-value? How do you interpret the result?

```
##
##  Pearson's Chi-squared test
##
## data:  .Table
## X-squared = 2.3286, df = 3, p-value = 0.5071
```

```
## Under the null hypothesis, the probability of observing a Chi-squared statistic
## as extreme as 2.3286, is 0.5071. There is no evidence to reject the null hypothesis.
## Therefore, there is no evidence of an association between sequence number and nucleotide.
```

3.2 Disease association

Table 2 gives the frequencies of wild-type and knock-out mice developing a disease thought to be associated to the absence of the knock-out gene.

	WT	KO	Total
Sequence 1	1.00	7.00	8.00
Sequence 2	9.00	3.00	12.00
Total	10.00	10.00	20.00

Table 2: Frequencies of wild-type and knock-out mice developing disease

a) What are your null and alternative hypotheses?

```
## Null hypothesis: there is no association between mouse type and disease X
##
## Alternative hypothesis: there is an association between mouse type and disease X
```

b) What are your expected frequencies?

```
##            WT KO
## Disease     4  4
## No Disease  6  6
```

Enter the data into the Shiny app as before;\

c) Select the **Fisher's exact test** option to compare the proportion of mice in each group that developed the disease. What is your p-value? How do you interpret the result?

```
##            WT KO
## Disease     4  4
## No Disease  6  6
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  .Table
## p-value = 0.01977
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0009621944 0.7209145117
## sample estimates:
## odds ratio
## 0.05788421
```

```
## p = 0.02. Under the null hypothesis, there is a small probability (p=0.02<0.05)
## of observing such an extreme distribution of the mice given the observed row and column totals.
## There is evidence to reject the null hypothesis in favour of the alternative hypothesis.
## There is evidence of an association between mouse type and disease X.
```

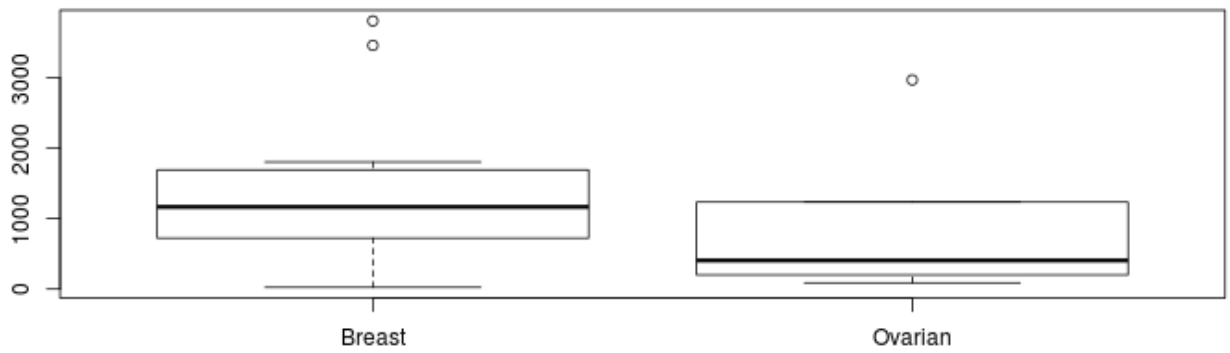
4 Choosing a test

In this section, we introduce several datasets and will invite you in groups to select a dataset and discuss what methods / tests you would use to analyse those data.

4.1 Dataset 1 `data1.csv`

Survival times of patients with Ovarian or Breast Cancer.

Is there a difference in survival time between the two diseases?



```
##
##  Wilcoxon rank sum test
##
## data:  Time by Disease
## W = 43, p-value = 0.3502
## alternative hypothesis: true location shift is not equal to 0
```

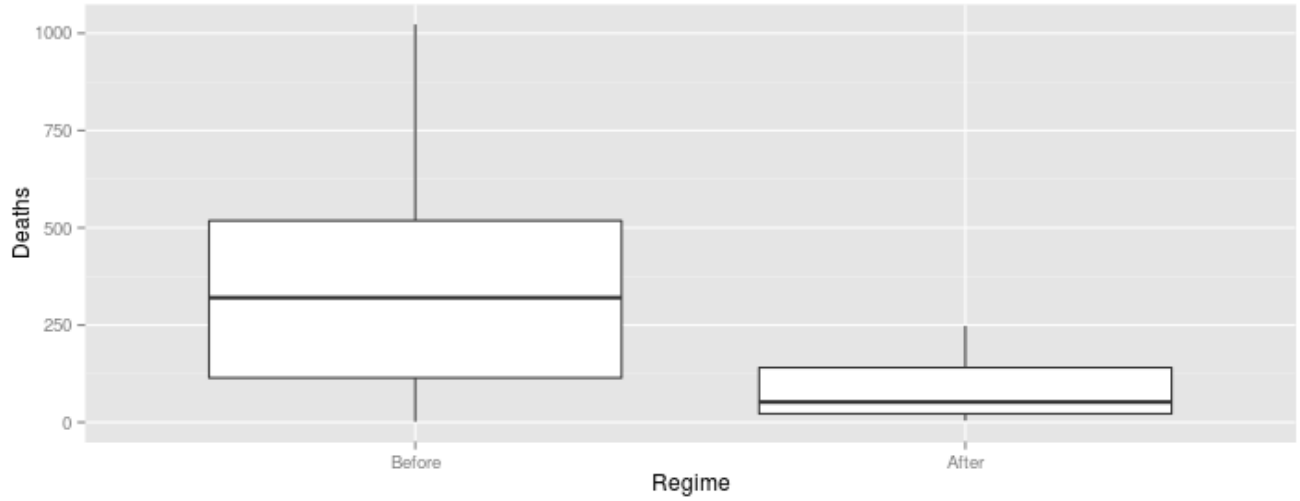
4.2 Dataset 2 `data2.csv`

In the history of data visualization, Florence Nightingale is best remembered for her role as a social activist and her view that statistical data, presented in charts and diagrams, could be used as powerful arguments for medical reform.

After witnessing deplorable sanitary conditions in the Crimea, she wrote several influential texts (Nightingale, 1858, 1859), including polar-area graphs (sometimes called "Coxcombs" or rose diagrams), showing the number of deaths in the Crimean from battle compared to disease or preventable causes that could be reduced by better battlefield nursing care.

Her Diagram of the Causes of Mortality in the Army in the East showed that most of the British soldiers who died during the Crimean War died of sickness rather than of wounds or other causes. It also showed that the death rate was higher in the first year of the war, before a Sanitary Commissioners arrived in March 1855 to improve hygiene in the camps and hospitals.

Do the data support the claim that deaths due to avoidable causes decreased after a change in regime?




```
##
##  Wilcoxon rank sum test
##
## data:  Deaths by Regime
## W = 106, p-value = 0.02593
## alternative hypothesis: true location shift is greater than 0
```

4.3 Dataset 3: Gene expression `data3.csv`

The expression level of a gene was measured in a breast cancer cohort of ER negative and positive patients.

Is the gene differentially-expressed between ER positive and negative patients?

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

```
##
## data:  Expression by ERStatus
## t = -38.746, df = 205.88, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.246953 -1.126198
## sample estimates:
## mean in group Negative mean in group Positive
##            -1.17388506             0.01269076
```

4.4 Dataset4: Sleep Data `data4.csv`

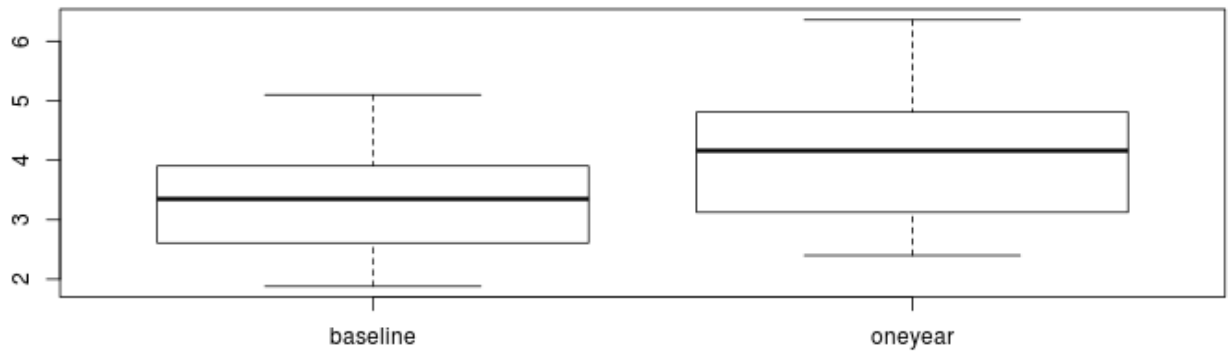
Data which show the effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients.

Does the drug have an effect on the amount of sleep?

```
##
##  Paired t-test
##
## data:  sleep[, 1] and sleep[, 2]
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.4598858 -0.7001142
## sample estimates:
## mean of the differences
##                   -1.58
```

4.5 Dataset5: CD4 `data5.csv`

CD4 cells are carried in the blood as part of the human immune system. One of the effects of the HIV virus is that these cells die. The count of CD4 cells is used in determining the onset of full-blown AIDS in a patient. In this study of the effectiveness of a new anti-viral drug on HIV, 20 HIV-positive patients had their CD4 counts recorded and then were put on a course of treatment with this drug. After using the drug for one year, their CD4 counts were again recorded. The aim of the experiment was to show that patients taking the drug had increased CD4 counts which is not generally seen in HIV-positive patients.



```
##
##  Paired t-test
##
## data:  data[, 1] and data[, 2]
## t = -4.4908, df = 19, p-value = 0.0002504
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.1801882 -0.4298118
## sample estimates:
## mean of the differences
##                  -0.805
```

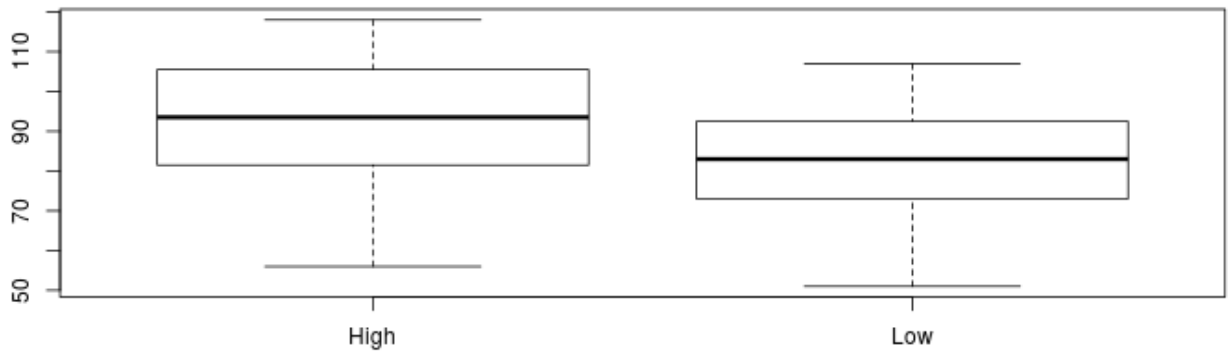
4.6 Dataset6: Birth Weight `data6.csv`

Risk Factors Associated with Low Infant Birth Weight

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bwt by smoke
## W = 5249.5, p-value = 0.006768
## alternative hypothesis: true location shift is not equal to 0
```

4.7 Dataset7: Weight gain in Rats `data7.csv`

The data arise from an experiment to study the gain in weight of rats fed on four different diets, distinguished by amount of protein (low and high) and by source of protein (beef and cereal).



4.8 Dataset8: Colon cancer `data8.csv`

The tumor and the normal counter-part samples were prospectively collected from 9 patients who underwent surgical resection at the INT-MI. Neoplastic samples were obtained from the central area of the neoplasia, avoiding to select necrotic material or transition zones with healthy mucosa. Samples of colonic healthy mucosa were resected at least 20 centimeters far from the neoplasia and distant from the surgical resection margins. Tissue samples were stored in liquid

nitrogen until RNA extraction. Total RNA was extracted from 10–20 mg of tumor samples and from 30–40 mg of normal samples.

The gene expression measurements for a paticular gene were extracted from the data.

Is this gene differentially-expressed between tumours and normals?