Introduction to Statistical Analysis using R commander

Sarah Vowler and Mark Dunning *

3rd November 2014

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 `Rcmdr` 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 the `R commander` plugin for the `R` statistical package. `R` is a freely-available open-source software that is popular within academic and commercial communities.[1]. 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 GUI that will allow you to perform statistical analysis without having to worry about learning R. At the same

*Acknowledgements: Sarah Dawson and Deepak Parashar

[1] A New York Times article on the emergence of R `http://tinyurl.com/ktw7g5b`

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 R commander can be adapted to your own needs.

2 Loading R commander

The first step is to load `RStudio`. There should be an icon for this on your Desktop.

Once loaded, type the following in the 'Console' in the bottom-left where the 'blinking' cursor is;

```
library(Rcmdr)
```

and press Return. A new window should be launched.

You could try exploring some of the available menu options. This practical will not comprehensively cover the usage of R commander, and certainly not `R`. We recommend the following for further documentation about R commander.

http://cran.r-project.org/doc/contrib/Karp-Rcommander-intro.pdf

3 T-tests practical

3.1 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*.

a) What are your null and alternative hypotheses?

Solution: *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 $\neq$ 165cm (the population mean for females).

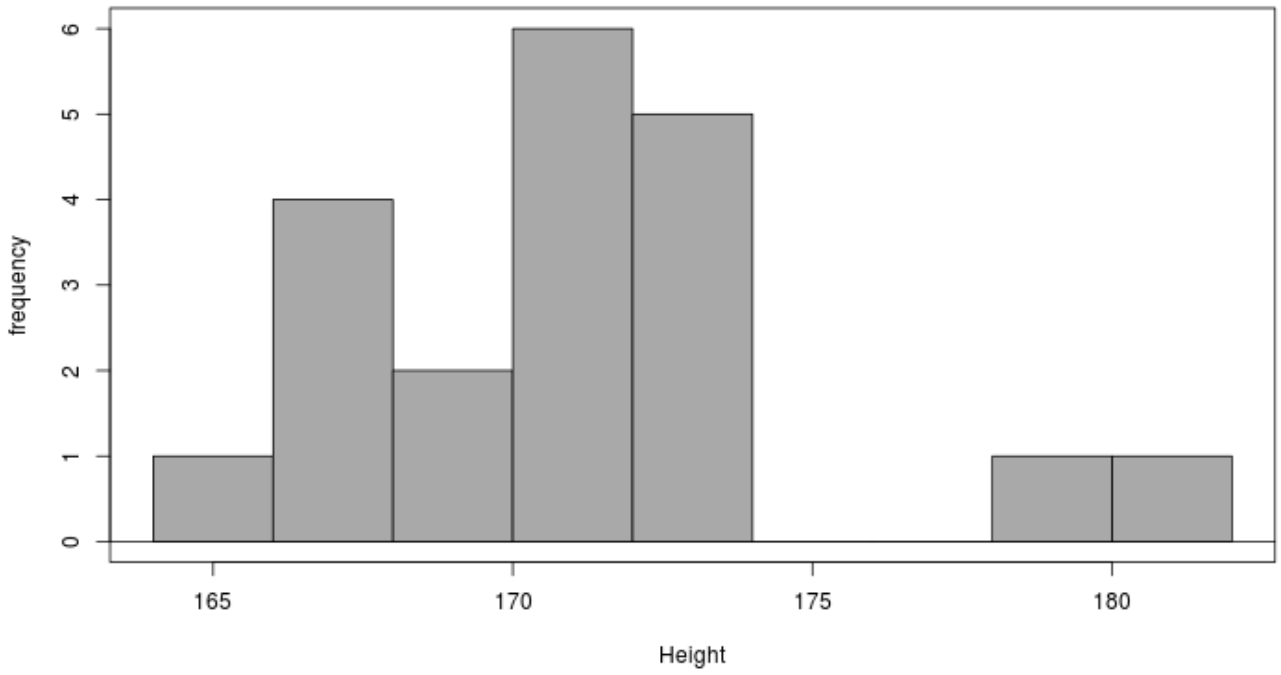
To import the file `diseaseX.csv` into `R commander` you will need to follow the menus.
`Data` → `Import data` → `from text file, clipboard, or URL...`

The screen that follows allows you to specify various properties of the file that we are trying to import. In this case, we need to select `Commas` as the field separator. You can use the `View data set` button to check that the data has be imported correctly.

b) Create a histogram of the `Height` variable using

`Graphs` → `Histogram`

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



Solution: *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** to test this by selecting

`Statistics` → `Means` → `Single-sample t-test`

Remember to change the value of `Null hypothesis: mu =` .

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:  Height
## t = 7.2293, df = 19, p-value = 7.293e-07
## alternative hypothesis: true mean is not equal to 165
## 95 percent confidence interval:
##  169.3428 172.8822
## sample estimates:
## mean of x
##  171.1125
```

Solution: Mean height in sample = 171.1cm (95% CI: 169.3-172.9) $t = 7.23, df = 19, p \leq 0.0001$.

Under the null hypothesis, the probability of observing a t-statistic as extreme as 7.23, is very small $(P(t \leq 7.23 | t \geq 7.23) < 0.0001)$. 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.

3.2 Biological processes duration

In the file `wt_ko_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.

a) What are your null and alternative hypotheses?

Solution: *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 the same menu sequence as before

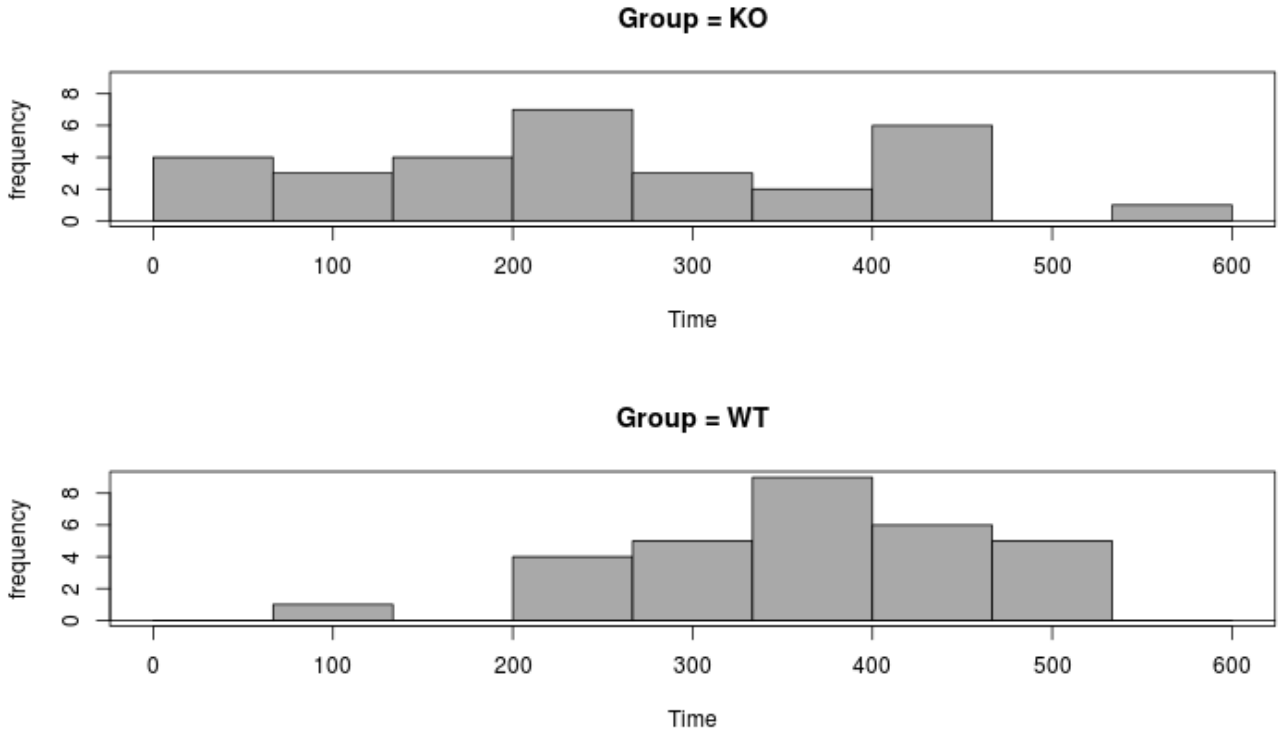
`Data` → `Import data` → `from text file, clipboard, or URL..`

b) Create histograms to compare the two groups; `WT` and `KO`. You will need to use the Histogram option as before

`Graphs` → `Histogram`

and select the `Plot by: Group` option.

Do the data look normally distributed for each cell-type? Is the independent t-test appropriate.



Please note that the histogram by group option is not available in Rcmdr version 2.1.0. You could try and use the Boxplot instead to compare distributions

Solution: *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.

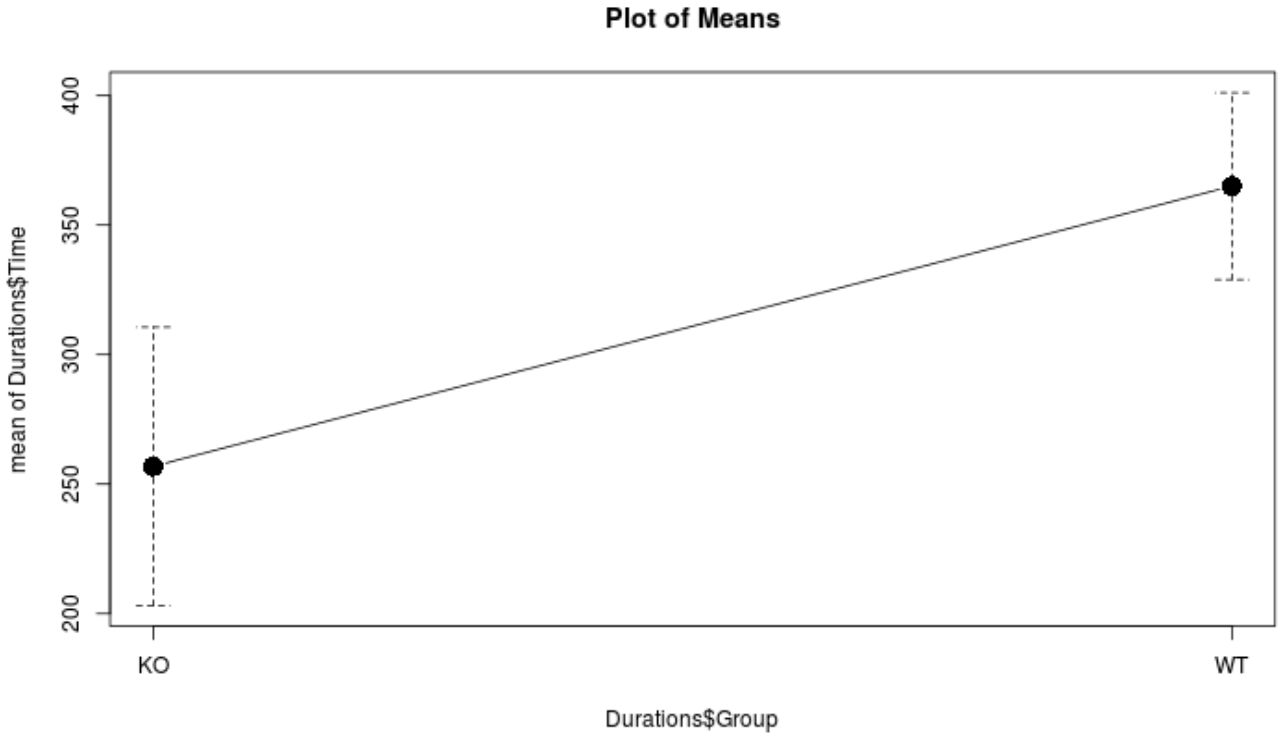
`Statistics` → `Summaries` → `Numerical summaries`

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.

`Graphs` → `Plot of means`

Do they overlap?

```
##        mean       sd    IQR  0%    25% 50%   75% 100% data:n
## KO 256.8333 143.9413 219.50  22 177.00 256 396.5  541     30
## WT 365.0333  96.6335 118.75 118 310.25 374 429.0  530     30
```



Solution: *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.

`Statistics` → `Variances` → `Two-variances F-test`

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

```
##
##  F test to compare two variances
##
## data:  Time by Group
## F = 2.2188, num df = 29, denom df = 29, p-value = 0.03568
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.056065 4.661663
## sample estimates:
## ratio of variances
##           2.218787
```

Solution: *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.

`Statistics` → `Means` → `Independent samples t-test`

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:  Time by Group
## 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
```

Solution: *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.*

3.3 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 `bloodplasmacancer.csv`.

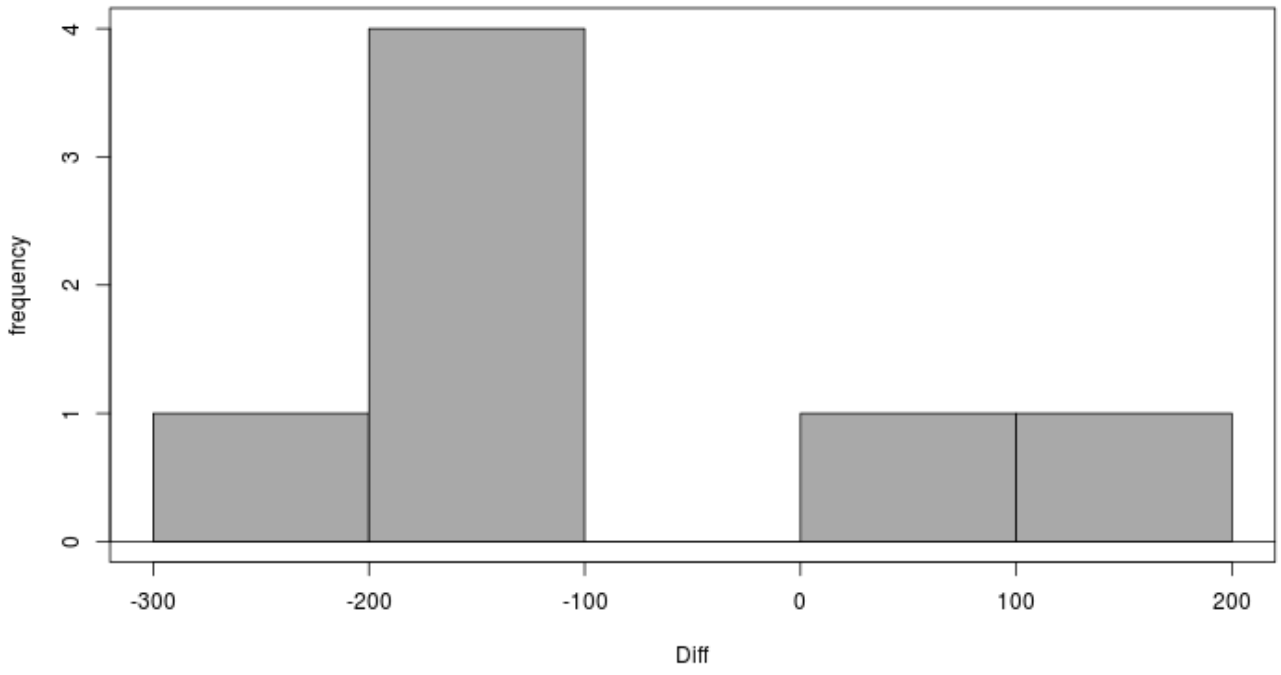
a) What are your null and alternative hypotheses?

Solution: *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).

`Data` → `Manage variables in active data set` → `Compute new variable`

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



Solution: *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?

Solution: **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:  Blood$Before and Blood$After
## 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
```

Solution: *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.*

3.4 Saving your work

For many people, one of the most important features of R is being able to track each step of the analysis and enable the holy grail of **Reproducible Research**. You have probably noticed that each time you select a menu option, the corresponding R code gets recorded within R commander. Not only does this provide a valuable document for yourself if you have to re-visit an analysis in the future, but you can pass you script on to someone else and they would be able to repeat the analysis. A reporting-writing mechanism `'markdown'` is also provided so that you can write comments about the analysis you have done.

a) Save your R commands to a file

`File` → `Save script as` →

b) Choose the `Generate HTML report` from the `R markdown` tab. Can you see how you might add your own details to the report? The report template can also be saved to disk

`File` → `Save R markdown file as` →

4 Tests for categorical variables

4.1 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) What are your null and alternative hypotheses?

	A	C	G	T	Total
Sequence 1	273	233	236	258	1000
Sequence 2	281	246	244	229	1000
Total	554	479	502	465	2000

Table 1: Nucleotide frequencies for two sequences

Solution: *Null hypothesis. there is* **no association** *between sequence number and nucleotide.*

Alternative hypothesis. there is an **association** *between sequence number and nucleotide.*

Enter the data from **Table 1** using;

`Statistics` → `Contingency tables` → `Enter and analyze two-way table`

You will need to set the number of rows and columns appropriately. **Note that you do not need to enter the totals.**. Clicking `Ok` will then perform the analysis.

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

```
##     1   2   3   4
## 1 273 233 236 258
## 2 281 246 244 229
##
##  Pearson's Chi-squared test
##
## data:  .Table
## X-squared = 2.3286, df = 3, p-value = 0.5071
```

Solution: $\chi^2 = 2.3285746, df = 3, p = 0.5070689$

Under the null hypothesis, the probability of observing a Chi-squared statistic as extreme as 2.3285746, is 0.5070689. There is **no evidence to reject the null hypothesis**. *Therefore, there is no evidence of an association between sequence number and nucleotide.*

4.2 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
Disease	1	7	8
No Disease	9	3	12
Total	10	10	20

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

a) What are your null and alternative hypotheses?

Solution: *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	Total
Disease			8
No Disease			12
Total	10	10	20

Solution: $Expected\ frequency = column\ total \times \frac{row\ total}{overall\ total}$

	WT	KO	Total
Disease	4	4	8
No Disease	6	6	12
Total	10	10	20

Enter the data as before;
`Statistics` → `Contingency tables` → `Enter and analyze two-way table`

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     1  7
## No Disease  9  3
##
##  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
```

Solution: *$p = 0.02$. 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.*