Research Methods Statistics Tutorial R

Background

In this tutorial you will do some exploratory analysis of data from METABRIC (Molecular Taxonomy of Breast Cancer International Consortium).

The data were originally used in a study that looked at the patterns of molecules inside tumours. Their conclusion was that 'breast cancer' is in fact at least ten different diseases, each with its own molecular fingerprint, and each with different weak spots. You can read about it here:

https://pubmed.ncbi.nlm.nih.gov/27161491/

All the analysis that the authors reported in the paper is reproducible using R - you can find their scripts on github https://github.com/cclab-brca/mutationalProfiles.

In this tutorial we will only look at the clinical data. These data have been downloaded from cbioportal.org.

The learning objective is for you to be able to apply simple exploratory analysis methods to real data using R.

Analysis

The METABRIC data consist of clincal characteristics coupled with gene expression data. In this tutorial we look at the clinical characteristics. This is similar to a first stage in prognostic modelling, to understand the existing factors, before proceeding to investigate the new ones and how they relate to the existing ones.

Before you start

1. Create an account with RStudio Cloud https://rstudio.cloud/
2. Log in
3. Load this tutorial by selecting "New project" > "New project from Git Repository", and enter **https://github.com/brentnall/canm937-r-2023**
4. Wait for project to load, then RStudio will load. The data are now loaded into the cloud. Follow instructions in the worksheet below.

(Note that the code in the worksheet is also available in the file `tutorial-code.R` that has loaded. You can work through without copying and pasting by selecting the line with the code to run and hitting Ctrl + Space).

Tasks

1. Load the data

The clinical data have been downloaded, and saved in the project (`brca_metabric_clinical_data.tsv`). Please load these data into your R session using `read.delim()` and inspect what has been loaded using `head()`.

Code

The authors saved this as a tab-separated file. This is like a commar separated file, but instead of a commar a tab is used to separate the columns. It can be read in Excel, and R. To do so in R one method is to use the function `read.delim()` to load it (previously we have seen `read.csv()` for csv files). `read.delim()` is more general, and allows for columns to be separated by any special character - you just need to say *what* the separator is. For tab-separated data `"\t"` tells it that it is a tab.

```
## 1. Load data
mydta = read.delim("brca_metabric_clinical_data.tsv", sep="\t")

## 2. Show first few rows, list column names
head(mydta)
```

Question: Using your output, state what are the first five columns called?

2. Calculate summary statistics

Using the data you have loaded calculated summary statistics on each column using the function `summary()`.

Code

```
summary(mydta)
```

Question: Using your output, identify the mean and median age at diagnosis

3. Plot a boxplot of age at diagnosis

Code

```
boxplot(mydta$Age.at.Diagnosis, xlab="Age")
```

Question: What is the interquartile range of age from the boxplot? Check your answer with the summary statistics output.

4. Plot a histogram of tumour size

Code

```
hist(mydta$Tumor.Size, xlab="Tumour size", main="Tumour size Histogram")
```

Question: Is tumour size approximately normally distributed? Justify using your histogram.

5. Is tumour size associated with lymph node positivity? Scatter chart.

Finally, we look at the association between tumour size and lymph nodes. It is expected that larger tumours are more likely to have spread to other parts of the body, and so the two should be positively correlated. We examine their association next.

Code

```
plot(mydta$Tumor.Size, mydta$Lymph.nodes.examined.positive, log="x")
```

This plots size vs nodes, and nodes vs size. We use the `log="x"` argument to put the x-axis on a log scale. This is sometimes useful for insepction of the plots, particularly when some data is a long way from the bulk.

Question: Do you think there is an association from looking at the plot?

6. Is tumour size associated with lymph node positivity? Correlation coefficient

Code

Calculate a Spearman correlation coefficient.

```
mydta$LNpos =  mydta$Lymph.nodes.examined.positive >0

cor.test(mydta$Tumor.Size, mydta$Lymph.nodes.examined.positive, method="spearman")
```

```
## Warning in cor.test.default(mydta$Tumor.Size,
## mydta$Lymph.nodes.examined.positive, : Cannot compute exact p-value with ties
```

Question: Why did I suggest Spearman, not Pearson correlation?

Question: Interpret the results of the analysis for the question above

6. Is tumour size associated with lymph node positivity? Show a boxplot of tumour size by lymph node positivity

Code

We can also plot the distribution of tumour size by nodal status as follows.

First let us define node positive vs negative in the data frame, and use this to get summary statistics of tumour size by nodal statue (neg/pos).

```
boxplot(split(mydta$Tumor.Size, mydta$LNpos))
```

Question: Do you think there is a real difference in median tumour size by node positivity based on the boxplot? Give reasons

7. Is tumour size associated with lymph node positivity? Calculated summary statistics of tumour size by lymph node positivity

Code You can do this using `tapply()`:

```
tapply(mydta$Tumor.Size, mydta$LNpos, summary)
```

Another way would be to use `subset()`

```
mysize_pos = subset(mydta$Tumor.Size, mydta$LNpos)
mysize_neg = subset(mydta$Tumor.Size, !mydta$LNpos)
summary(mysize_pos)
summary(mysize_neg)
```

Question: What is the mean tumour size by nodal status (postive/negative) based on the output?

8. Is tumour size associated with lymph node positivity? Do a Wiloxon test of tumour size by lymph node positivity

Code

To test the hypothesis that mean tumour size is the same in both (independent) groups one might use a t-test, or a non-parametric Wilcoxon test.

```
wilcox.test(mysize_pos, mysize_neg)
```

Question: Interpret the output. Why did I not use a t-test?

9. Are the percentage of tumours sized more than 20mm different by nodal status? Calculate a confidence interval on the difference.

Code

One might sometimes be interested in particular cutpoints. For tumour size, 20mm is a cutpoint used in tumour staging, so we look at that next.

First we calculate the number sized more than 20mm positive by nodal status

```
npos = c(sum(mysize_neg>20, na.rm=TRUE), sum(mysize_pos>20, na.rm=TRUE))

ntot = c(sum(!is.na(mysize_neg)), sum(!is.na(mysize_pos)))

## number size >20mm
npos

## number with size available
ntot

## proportion
npos / ntot
```

Then we can carry a test of the difference in two proportions (based on a z-test) as follows

```
prop.test(npos, ntot)
```

Question: What is a 95%CI on the difference in proportions

If time left

Keep exploring! eg. Consider trying to fit a linear regression using code from the first two tutorials.

Troubleshooting RStudio Cloud for the tutorial

If you cannot run above via RStudio Cloud for the tutorial, then an alternative way is to follow this link:

http://mybinder.org/v2/gh/brentnall/canm937-r-2023/main?urlpath=rstudio

Note however, that you cannot save anything using this route - any changes you make will be lost. Therefore, for doing your own analysis where you need to save your results and scripts please use RStudio Cloud (or install on your own computer).