
R for data analysis

3

3 steps to Basic data analysis

1. Reading in data

- `read.table()`
- `read.csv()`, `read.delim()`

2. Analysis

- Manipulating & reshaping the data
- Any maths you like
- Plotting the outcome
 - High level plotting functions (covered tomorrow)

3. Writing out results

- `write.table()`
- `write.csv()`

A simple walkthrough

Exemplifies 3 steps to R analysis

- 50 neuroblastoma patients were tested for NMYC gene copy number by interphase nuclei FISH
 - Amplification of NMYC correlates with worse prognosis
 - We have count data
 - Numbers of cells per patient assayed
 - For each we have NMYC copy number relative to base ploidy
- We need to determine which patients have amplifications
 - (i.e. $>33\%$ of cells show NMYC amplification)

Step 1.

Read in the data

Patient	Nuclei	NB_Amp	NB_Nor	NB_Del
1	42	0	34	8
2	40	3	30	7
3	56	6	50	0
4	42	5	37	0
5	32	1	30	1
6	70	10	53	7
7	65	3	58	4
8	40	4	31	5
9	60	0	54	6
10	61	0	57	4
11	43	13	29	1

This data is a tab delimited text file
Each row is a record, each column is a field
Columns are separated by tabs in the text.

We need to read in the results table and assign it to an object (rawData)

```
rawData <- read.delim("08_NBcountData.txt")
rawData[1:10,]    # View the first 10 rows to ensure import is OK
                  # Note data frame contains a patient index column
```

If the data had been comma separated values, then sep=",",

```
read.csv("08_NBcountData.csv")
?read.table for a full list of arguments
```

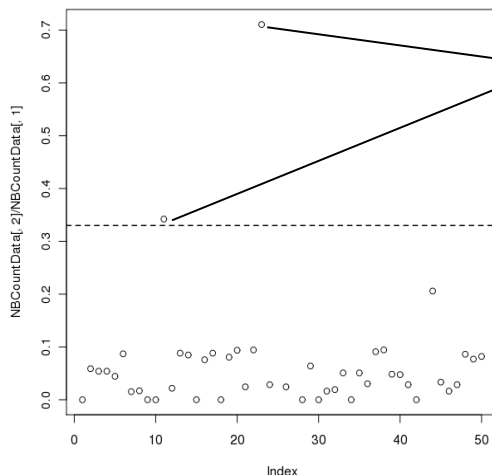
08_NBcountData.R
(script commands)

08_NBcountData.txt
(data file)

Step 2.

Analysis (reshaping data & maths)

- Our analysis involves identifying patients with $> 33\%$ NB amplification
 - `prop <- rawData$NB_Amp / rawData$Nuclei` # create an index of results
 - `amp <- which(prop > 0.33)` # Get sample names of amplified patients
- We can plot a simple chart of the % NB amplification
 - `plot(prop, ylim=c(0,1.2))`
 - `abline(h=0.33,lwd=1.5,lty=2)`



These 2 samples are amplified (11 & 23)

Step 3.

Outputting the results

- We write out a data frame of results (patients > 33% NB amplification) as a 'comma separated values' text file
 - `write.csv(rawData[amp,], file="selectedSamples.csv") #`
`Export table, file name = selectedSamples.csv`
 - Files are directly readable by Excel and Calc
- Its often helpful to double check where the data has been saved
 - Use get working directory function
 - `getwd() # print working directory`

Data analysis exercise:

Which samples are near normal?

- Patients are near normal if:

`(NB_Amp/Nuclei <0.33 & NB_Del ==0)`

- Modify the condition in our previous code to find these patients
- Write out a results file of the samples that match these criteria, and open it in a spreadsheet program

Solution to NB normality test

Basic data analysis

```
> norm <- which( prop < 0.33 & rawData$NB_Del==0)
```

```
> norm
```

```
[1] 3  4  7 15 20 24 36 37 42 47
```

```
> write.csv(rawData[norm,], "My_NB_output.csv")
```