Notes on the scripts to calculate two-sample randomisation tests

David Colquhoun

First download R and install it

Then download R Studio and install it in a directory/folder named something convenient like R-progs.

Dorothy Bishop has provided download instructions, and a short introduction to R, at http://www.slideshare.net/deevybishop/learning-r-while-exploring-statistics

Put the scripts into that directory, start R Studio and open the script.

R Studio Notes

Lines that begin with # are comments, not code.

Variable names often start with my (a curious convention in R thats intended to avoid clashes with pre-defined variables) the rest of the name is more informative.

Two programs are provided

two_sample_rantest.R -uses random sampling with replacement to generate the randomisation. This was used to generate Figure 1

2-sample-rantest-exact.R. Lists all possible allocations of a sample of 10 from 20 observations. The number of combinations is given by ncom=choose(20,10) = 187,456.

It takes a very short time to look at all of them.

The results are essentially the same with both programs.

All you have to do is to set the inputs, which are gathered together between

#START INPUTS and #END OF INPUTS

(1) For two_sample_rantest.R type in the number of re-samplings to be done (overwrite the value of *mynsim*=100000 if you want a different number).

(2) Define the name of the output text file (which will appear in the same folder as the program): *outfile* = . At present it’s called rantest2.txt but not that this will be overwritten

each time the program is run, unless the name is changed.
(3) Define the data: two independent samples in *mysampA* and *mysampB*. At present the data are as in Student's original 1908 paper (see Table 9.2.4 in http://www.dcscience.net/Lectures_on_biostatistics-ocr4.pdf). Any other values can be used, though at present it's limited to cases where the number of values in each sample is the same.
(4) At present *myn* = 10. If the number of values in each sample is not 10, then change this.

The program puts all 20 observations into a single array (*allobs*) and selects from this, 10 observations at random (without replacement). It does this *mynsim* times and plots the distribution of the *mynsim* differences between means.
In the case of 2-sample-rantest-exact.R the program calculates all 187,456 possible samples of 10 observations selected from 20, so *mynsim* does not need to be specified.

This procedure generates the distribution of differences between means that would be expected if the two treatments were identical. In that case, the response of each person is a characteristic only of the person, and not on which drug they were given.

The programs also calculate the conventional *t* test on the same data, and prints the results to the output file. Unlike the randomization test, the *t* test assumes that the data in each sample are normally-distributed, with the same variance for each sample. In this case, the result is very similar to that from the randomization test.

Now you are ready to run the script. Hit ctrl-alt-R and the script will run. Alternatively highlight the whole script (ctrl-A) and click the run button (the run button runs the highlighted sections of the script).

Any graph that you want to keep must be saved manually.

The output file (eg rantest2.txt) appears in the same directory in which the script resides. Here is an example of the text file (the numbers will differ slightly when you repeat the test, but if you do a large enough number of simulations, the variability should small).

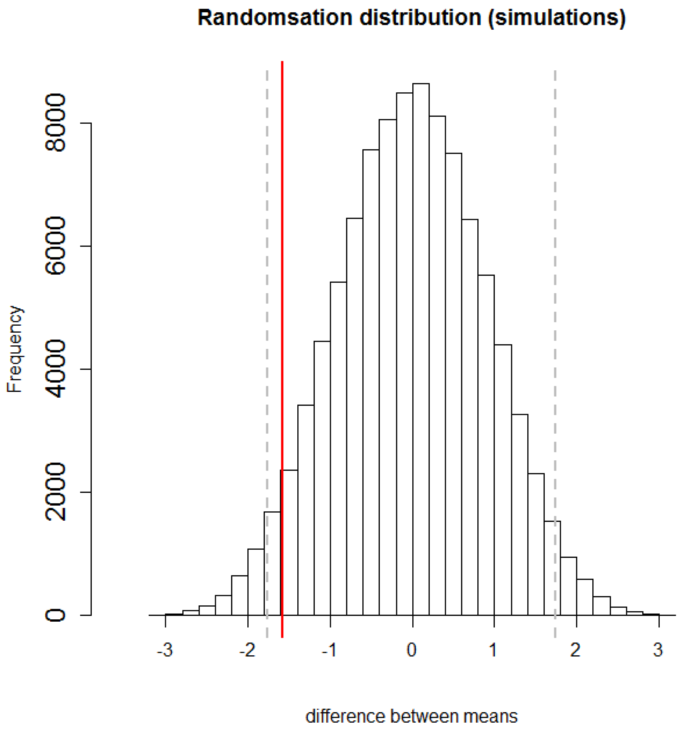
Here is an example of rantest2.txt, and the graph output.

```
Randomisation distribution using random samples
INPUTS
number of resamplings =  1e+05
number obs per sample =  10
sample A 0.7 -1.6 -0.2 -1.2 -0.1 3.4 3.7 0.8 0 2
sample B 1.9 0.8 1.1 0.1 -0.1 4.4 5.5 1.6 4.6 3.4

 OUTPUTS
mean for sample A=  0.75
mean for sample B =  2.33
Observed difference between means (A-B) -1.58
SD for sample A) =  1.78901
SD for sample B) =  2.002249
mean and SD for randomisation dist =  -0.0002256    0.9042624
quantiles for ran dist (0.025, 0.975) -1.76 1.74

Area equal to less than observed diff 0.04221
Area equal to or greater than minus observed diff 0.0401
Two-tailed P value 0.08231

Result of t test
P value (2 tail) 0.07918671
confidence interval -0.203874 3.363874
```



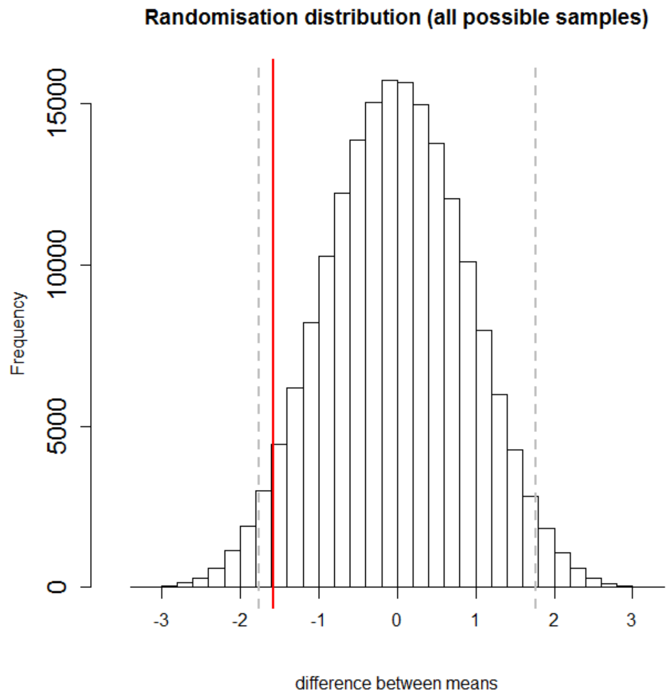
The output file from the version that lists all possible samples, e.g. rantest-exact2.txt, looks like this (it should be the same every t ime the program is run).

```
Randomisation test: exact calculation all possible samples

INPUTS: exact calculation: all possible samples
Total number of combinations =  184756
number obs per sample =  10
sample A 0.7 -1.6 -0.2 -1.2 -0.1 3.4 3.7 0.8 0 2
sample B 1.9 0.8 1.1 0.1 -0.1 4.4 5.5 1.6 4.6 3.4

 OUTPUTS
mean for sample A=  0.75
mean for sample B =  2.33
Observed difference between means (A-B) -1.58
SD for sample A) =  1.78901
SD for sample B) =  2.002249
mean and SD for randomisation dist =  -7.787562e-17    0.9024436
quantiles for ran dist (0.025, 0.975) -1.76 1.76
Area equal to orless than observed diff 0.04072398
Area equal to or greater than minus observed diff 0.04072398
Two-tailed P value 0.08144796

Result of t test
P value (2 tail) 0.07918671
confidence interval -0.203874 3.363874
```


Randomisation distribution (all possible samples)
Frequency
difference between means