# The Randomization test (or shuffle test)

*Michael Wood (*Michael.wood@port.ac.uk*)*

*22 October 2012*

(Introduction and links to electronic versions of this document and the other parts at http://woodm.myweb.port.ac.uk/stats. The data in the tables, and the figures, are in the spreadsheet, at http://woodm.myweb.port.ac.uk/stats/StatNotes.xls. There is a video at http://youtu.be/Uyub9cYimWw .

*Although the randomization test is a very useful and flexible method, it is not widely used. I am includingExcel spreadsheet at* http://woodm.myweb.port.ac.uk/nms/diffofmeanstestnd.xls.

The table below shows the marks obtained by ten students in an exam.

***Table 1. Exam marks and sex of ten students***

| Mark | Sex |
|---|---|
| 37 | F |
| 46 | F |
| 56 | F |
| 49 | F |
| 78 | F |
| 50 | M |
| 55 | M |
| 81 | M |
| 55 | M |
| 53 | M |

The average (mean) of the marks for the female (F) students is 53.2%, whereas for the males (M) it is 58.8%. The difference is 53.2%-58.8% or –5.6%. On average the males did 5.6% better.

However this is just a small group of students. With another group of students we may get a different answer. Can we be sure that males really do better overall?

*Quick question 1. What do you think the answer is? Does this evidence prove the point, or could it be a fluke?*

Let's now do a hypothesis test. The ***null hypothesis*** is that whether a student is male or female has no relationship to the mark they are likely to get. There are no systematic differences between males and females, and the average mark for ***all*** male students (who might have taken the exam) is the same as the average for ***all*** female students. The actual difference we have observed in our small sample of ten students is a 5.6% difference in the average marks of five males and five females. How likely is this to have occurred if this null hypothesis is true?

The idea is to work out the *p* value by a simple computer simulation.

The spreadsheet diffofmeanstestnd.xls (at http://woodm.myweb.port.ac.uk/nms/diffofmeanstestnd.xls) will do this simulation. Go to the Sample sheet, key the data in starting from Cell C7 (or copy and paste from Table 1). Obviously Exam marks goes in the number variable column, and Sex is the Group variable. Now key F in cell C3 and M in Cell D3 to tell the spreadsheet what the two groups are. (Or go to http://woodm.myweb.port.ac.uk/stats/diffofmeanstestTable1.xls for a version with the data entered.)

The actual marks obtained by the female students were 37, 46, 56, 49 and 78. However, if we assume that the null hypothesis is true and there are no systematic differences between males and females, the females are equally likely to get ***any*** of the marks in the list. **This means we can simulate what is likely to happen in a group of ten students by shuffling the marks at random – an example of a process known as *resampling*.** The first shuffle (resample) was:

***Table 2. First Shuffle of the data in Table 1***

| Mark | Sex |
|---|---|
| 78 | F |
| 46 | F |
| 55 | F |
| 37 | F |
| 56 | F |
| 53 | M |
| 81 | M |
| 55 | M |
| 49 | M |
| 50 | M |

The marks here are just the same as the marks in the previous table, but they are arranged in a different order. If the null hypothesis is right, this order is just as likely as the original order.

It's now easy to work out the average of the "female" marks (54.4%) in this shuffle, and of the "male" marks (57.6%), and the difference between them (-3.2%).

To do your own shuffle, go to the Single resample sheet and Press F9. For obvious reasons this is very unlikely to be the same as Table 2.

The next shuffle (resample) I produced when I was writing this was

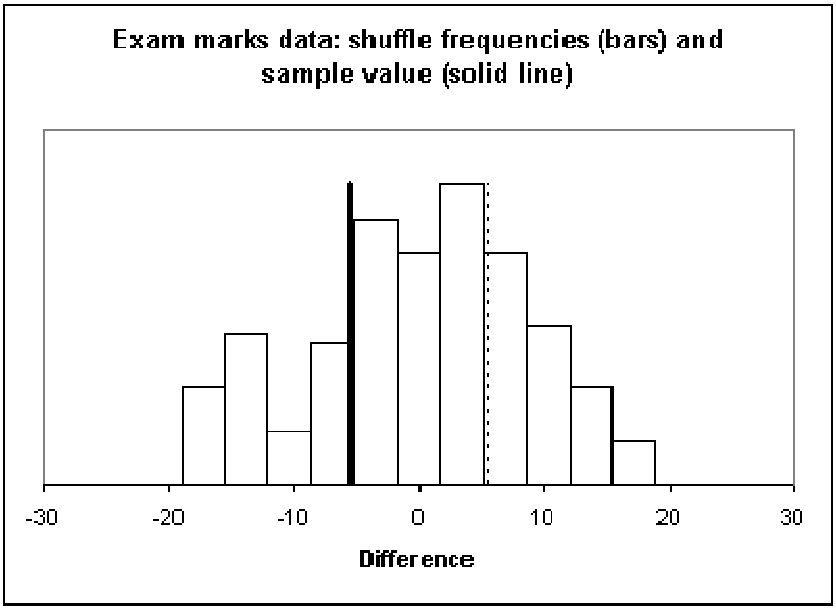
***Table 3. Second Shuffle of the data in Table 1***

| Mark | Sex |
|---|---|
| 78 | F |
| 55 | F |
| 50 | F |
| 81 | F |
| 55 | F |
| 46 | M |
| 56 | M |
| 53 | M |
| 49 | M |
| 37 | M |

*Quick question 2. What are average marks for the females and males in Table 3. What is the difference between them?*

If we do this lots of times, we will see how much difference there is likely to be between the average mark of five female students and five male students ***if the null hypothesis is true and there are no systematic differences between females and males***. The graph below is based on 200 shuffles (go to the Lots of resamples sheet to see this: the individual resamples are listed on the left of this sheet.)

*Figure 1. Resampling results based on data in Table 1*




**This graph shows how the results of all these shuffles compare with the actual data**. The histogram (bars) represents the differences between the average marks of females and males worked out from the

shuffles. All of these differences are between –20 and +20, with the commonest values clustered around zero. The bar on the left, for example, represents 11 shuffles with differences between –19 and –15.5, and the next bar represents 17 shuffles with differences between 15.5 and 12. Remember, these shuffles are based on the assumption that there are no consistent differences between males and females. This is what would happen if the null hypothesis were true.

The solid line represents the actual difference between females and males in the sample of data (-5.6). This is the result that was observed: on average, the females got 5.6 marks less than the males. The dotted line represents the situation where the females average 5.6 more than the males – this is the "opposite" of the real result.

This graph suggests that the actual difference could easily have arisen from the shuffling process, because the solid line is near the middle of the histogram. In other words the data is consistent with the null hypothesis. It is entirely plausible that there really are no systematic differences between males and females, and that the difference in the sample (-5.6) is just a matter of chance.

**More formally, we work out the probability of getting a difference as extreme as -5.6 *if the null hypothesis is true***. In this case, this means the probability of the difference being –5.6 or less, or +5.6 or more. This tells us the probability of our results (or similar results) occurring if the null hypothesis is true.

*Quick question 3. What do you think this probability is? (You should be able to see roughly what this is from the graph.*

**This probability is called a *p value* or *significance level***. In this case it is fairly large, indicating that it is entirely plausible that the data could have arisen from the null hypothesis.

This *p* value is *two tailed* because it includes probabilities from both *tails* (ends) of the distribution. Occasionally, *one-tailed p* values are used but the reasons for doing this can be a bit convoluted. In this case the one tailed p value is 26.5% (half of the two tailed *p* value). *If in doubt, always use two tails.*

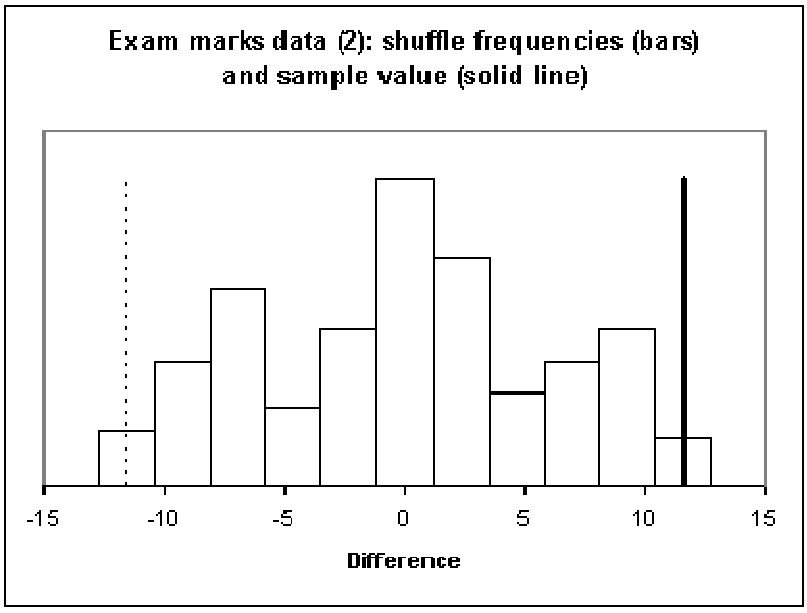
The next table shows a similar set of data – except that here the females do better, and the pattern looks more consistent.

***Table 4. A second set of data on exam marks and sex***

| Mark | Sex |
|---|---|
| 55 | F |
| 60 | F |
| 56 | F |
| 72 | F |
| 78 | F |
| 50 | M |
| 55 | M |
| 50 | M |
| 55 | M |
| 53 | M |

A similar process of 200 shuffles produced this graph:

*Figure 2: Resampling results based on Table 4*




*Quick question 4. What do you think the p value is here? What would you conclude from this?*

**You can also use this method for many other examples**: e.g. you might want to compare the profitability of two different types of companies, or the **proportions** of males and females who smoke (the first column in the Data sheet should be headed smoker, with 1 representing a smoker and 0 representing a non-smoker), or a measure of the performance of two different drugs in a medical trial.

## Exercise

The table below shows some data similar to the data used in the research which produced the customer service ratings (see http://woodm.myweb.port.ac.uk/stats/StatNotes3.pdf). You will need the spreadsheet at http://woodm.myweb.port.ac.uk/nms/diffofmeanstestnd.xls.

**Small sample of data comparing customer service in banks and building societies**

| Service rating | Institution |
|---|---|
| 6 | Bank |
| 9 | BS |
| 7 | Bank |
| 5 | BS |
| 2 | Bank |

Use this data to compare the service in banks and building societies. Key the data into the appropriate cells (C7:D11) in the Sample sheet of http://woodm.myweb.port.ac.uk/nms/diffofmeanstestnd.xls. Put the words Bank and BS in Cells C3 and D3. Now click on the Single resample sheet and press F9 to see how a resample works. Then click on Lots of resamples, and press F9 again.

Find the size of the difference and test the null hypothesis that there is no difference between the ratings. What *p* value does your test give? Explain what your results mean.

Now do the same with the two larger samples of data in the Part3data sheet in StatNotes.xls (http://woodm.myweb.port.ac.uk/stats/StatNotes.xls). How would you explain the different results?

It would also be interesting to try the same questions with SPSS (use compare means).

## Notes on answers to Quick questions

1 Intuitively, I feel this data could be a fluke. The next sample might give a different result. However, your intuition might give a different answer. Intuition is unreliable in this context, which is why we need a formal way of testing hypotheses.

2 This has a "female" average of 63.8%, and a "male" average of 48.2%. The male average is now ***less*** than the female, so this difference counts as positive: +15.6.

3 According to the spreadsheet 106 of the 200 shuffles produced results which were either -5.6 or less, or +5.6 or more. This gives a *p* value of 53%. You probably haven't managed to estimate this exactly, but your answer should be close to this.

4 The *p* value here is 2.5% because 5 of the 200 shuffle results were in the "tails" (extremes) of the distribution. This should be roughly obvious from the graph. This is far lower than the first example, indicating that the data is far less likely to have arisen if the null hypothesis were true. In other words, this supports the idea the null hypothesis is false, and that there is a real difference between males and females in their performance in this exam.

## Notes on answer to the Exercise

The *p* value for the small sample should be large (more than 50%) indicating that the difference could be due to chance. The *p* values for the bigger samples (Exercise 2) should be much smaller (more significant) indicating that the null hypothesis is much less likely so the evidence indicates a real difference. This should all be clear from the graphs. With SPSS the results should be similar but not identical.