

Randomisation tests in R – clam shells

You will need the two samples of clam shells.

The samples

We have collected clam shells (cockles) from two sites in Malaysia, in Johor and Sarawak. We want to know if they are the same size or if one population has longer shells than the other. The shells in each population are marked with a letter, “J” or “S”.

You have samples of shells from each population, 10 shells from Sarawak and 8 from Johor (the Johor population is a bit small for everyone to take 10).

You will measure the mean length (the longest axis) of the shells, calculate the difference between the samples, and then do a randomisation test to see if the difference could be due to sampling error.

The simplest way to measure the mean length for one sample is to lay the shells end-to-end on a sheet of graph paper, measure the total length, and divide by the number of shells.

Do this for the two samples and record the total lengths and mean lengths for each sample, and the difference in means.

Could the difference in mean lengths be due simply to sampling error? Perhaps in reality the mean lengths at the two sites are the same and we have – just by chance – taken big shells from one site and smaller shells from the other site.

A simple randomisation test

If the difference between the two samples is really only due to chance, we could shuffle the 18 shells, pull out 8 at random, compare these 8 and the 10 left behind, and get a similar difference in means as for the actual samples.

Mix up the 18 shells thoroughly, then divide the pile of shells into a group of 8 and a group of 10. Measure the total length for each group. (Or just measure the total for the group of 8 and subtract from the overall total to get the length for the group of 10.) Calculate the difference in means and compare it with the difference between the actual Sarawak and Johor samples.

We are looking for randomised results which are *more extreme* than the actual observed result. So if the observed difference in means was 2.5mm, a randomised difference greater than 2.5mm would be more extreme, and a randomised difference of -2.5 mm would *also* be more extreme, as we don't know which of the sites might really have larger clam shells.

One randomised value doesn't allow us to draw a proper conclusion, we need to do it several times.

Mix the 18 shells again, pull out 8, measure the total length and calculate the difference in means; note whether the result is more extreme than the observed difference in means.

Do this many times – at least 10 times, more if you have time – and calculate the proportion which were more extreme than the original samples.

Based on the results of the randomisations, do you think the actual samples from Sarawak and Johor are typical of random draws from a common pool, or are there real differences between the two samples?

- If none of the random values are more extreme than the actual results, you would suspect there was something special about the Sarawak and Johor samples, ie. a real difference in size between the populations of clams.
- If several of the random values are more extreme, you would conclude that the differences in the samples are just due to chance.

Only 10 to 20 random draws are not enough to make firm inferences; we need to do several hundred. That would be tedious to do by shuffling and measuring the shells, but can be done very quickly in R.

A randomisation test in R

Now we need to measure the individual shells and put the values into R.

Measure the length of each shell and put them into two vectors in R, one for each of the samples (use the “J” and “S” marks to sort out the shells again); call these vectors something like ‘shells.J’ and ‘shells.S’.

Calculate the sum and the mean for each sample; these should be about the same as you measured before (not quite the same because of rounding errors).

Store the difference in means in an object called (say) ‘mean.diff’.

Now we’ll mix up the 18 measurements and draw out 8 at random:

Combine the two sample vectors into a single vector with the ‘c’ function [eg. `c(shells.J, shells.S)`]; call this something like ‘shells.all’. Also store the sum of the lengths for all the shells in an object, say ‘sum.all’.

Use the ‘sample’ function to draw a random sample of 8 from the combined vector [`sample(shells.all, 8)`] and calculate the sum – you can do this in one step with `sum(sample(shells.all, 8))`.

That’s just one random draw. We can use a *for loop* to do this many times (eg. 1000 times) and see how many values are more extreme than the actual value for the Johor sample. Before running the loop we have to create an object to store the results – I called this ‘result’. So the code looks like this:

```
result <- NULL
for(i in 1:1000) {
  sum1 <- sum(sample(shells.all, 8))
  sum2 <- sum.all - sum1
  result[i] <- sum2/10 - sum1/8
}
```

For each iteration of the loop, we draw a random sample of 8 from the ‘shells.all’ vector and calculate the total length (sum1); we calculate the length of the rest (sum2) by subtracting sum1 from sum.all; we calculate the means (sum1/8 and sum2/10) and store the difference in ‘result’.

Run the loop to generate 1000 random samples from the ‘shells.all’ vector and calculate the difference in means for each.

Check ‘result’ by displaying it as a histogram [`hist(result)`]; the values should be grouped around zero.

See how many of the values in ‘result’ are more extreme than the observed value in ‘mean.diff’.

Remember that “more extreme” means greater than mean.diff or less than –mean.diff, ie. we ignore the minus signs. We can do that with the R function ‘abs’.

`abs(result) >= abs(mean.diff)` will give us a vector of 1000 TRUEs and FALSEs, TRUE if the value in ‘result’ is more extreme than ‘mean.diff’. We use ‘sum’ to add up the TRUEs, and divide by 1000 to get the proportion: `sum(abs(result) >= abs(mean.diff))/1000`

This proportion is an estimate of the *p*-value, the probability of getting a difference in means more extreme than the observed difference just because of sampling error.

Most scientists agree that if the *p*-value is less than 0.05 (5%), the “just sampling error” hypothesis is not plausible, and they interpret the result as evidence that there’s a real difference between the two populations than the samples came from.

Did you find evidence that there is a difference between the shells from Sarawak and Johor?

Compare with other people in the group: did everyone come to the same conclusion? (Remember that all your samples were drawn from the same populations, so you might expect everyone to agree on whether there is a difference or not.)

Reference

Manly, B F J 1997. *Randomization, bootstrap and Monte Carlo methods in biology*. Chapman and Hall / CRC