CASE STUDY

Introducing the logic of hypothesis tests though randomisation tests

Paul Hewson , School of Computing, Electronics and Mathematics, Plymouth University, Plymouth, UK. Email: paul.hewson@plymouth.ac.uk

Abstract

There has been a lot of interest in the use of randomisation tests as a pedagogic alternative to hypothesis tests (Zieffler, 2012), although proposals to use randomisation tests in research are far from new (e.g. Hooton, 1991) with Good (2000) being an updated classic text in this area. This article will present a classroom activity that demonstrates the randomisation tests as a means of understanding several of the concepts around hypothesis testing in a manner that is as friendly as possible for maths-phobic and indeed computer-phobic students.

Keywords: Hypothesis test, randomisation test, threshold concept.

1. Activity

1.1. Background

The activity presents the randomisation test in the context of a designed and controlled experiment. Random allocation of treatments in designed experiments is a key concept that should be introduced or reinforced early on any statistical course. Hence we can readily appreciate the importance of random allocation. In order to work within a classroom (and not a laboratory) setting, we conduct a “thought experiment”.

1.2. Activity

The first stage in the activity is to ask a research question that could be answered with suitable human volunteers. As chocolate often serves as an inducement for volunteers to come forward we usually conduct some experiment on the stress relieving effects of chocolate (or rum truffles, or cola, or other snack). We need to discuss the ways of measuring stress (we posit a fictional biochemical test for cortisol as we can freely obtain fictional measurements and print them out). Having obtained volunteers we need to establish a working theory (that the snack has an effect) and then formulate a null hypothesis. The null hypothesis can usefully be explored in both an informal way (the snack has no effect) as well as a formal one (the population mean levels of stress measure cortisol are the same for treatment and control groups).

The student volunteers are obtained and are subsequently randomised to two groups (the treatment group receving the rum truffles or similar and the control group receiving the placebo) by means of a coin toss. We can then obtain the experimental data.

Having obtained data from both treatment group and control group we can examine the summary statistics. The fake data we have presented is as follows (Cortisol mmol l^{-1}):

Control: 4.95, 3.61, 5.15, 2.22

Treatment: 0.90, 3.16, 0.57, 0.15

Figure 1. Students are lined up on stage “wearing” their cortisol measurements.

The sample sizes are small (to make the stage production manageable in terms of explaining to non-statistians the sampling distribution of a test statistic under the null distribution) but computer animation can run the data through with larger samples later. We can first dwell on the obtained data; usually a student will notice that the data overlap (there is a value of 2.22 in the control group which is lower than the 3.16 in the treatment group. So we can make it very clear that we are looking at averages and not individuals. We can also at this stage discuss a test statistic. If our null hypothesis were true, we can usually prompt students to realise they would expect the difference in group means to be zero. At this stage we can introduce ideas of sampling error. Individuals have different levels of stress hormone (and we may introduce other errors in the testing regime). We need to know whether the difference we see is large enough that we can rule out random error. At this point we can explore designs that may block out the error.

The next step in the procedure is to introduce the idea of the randomisation test. If the null hypothesis is true, the only reason we got readings of 4.95,…2.22 mmol l^{-1} in the control group is because we allocated those individuals to the control group (because our null states that the treatment has no effect). We could therefore shuffle the subjects randomly between groups and get a fictional test statistic.

Figure 2. Students are lined up on stage having been shuffled.

To reinforce the idea that the measurements are a feature of the students the “least stressed” student has been given a surfboard and the “most stressed” student a yellow jacket. The key point is to explore the concept that if we believe our null hypothesis, we believe we would have got these data if we had allocated the students to the two groups. This is a key idea, indeed Meyer and Land (2003) give this idea of a sampling distribution as an exemplar threshold concept. Indeed, we often avoid engaging students with this concept as it is indeed rather troublesome. However, it is such a key concept underpinning many later ideas that I believe it is worth the effort of exploring carefully. So far in the classroom illustration, we are trying to convey the idea that **if our null hypothesis is true** we can use the data to give us some idea how much our test statistic can vary. The well stated advantage of randomisation tests is that we don’t need to pick our test statistic to match a theoretical distribution and need rather few assumptions about the data generating mechanism.

It is sensible at this stage to introduce ideas of the number of shuffles we might need in order to have a reasonable understanding of the distribution of the test statistic under the null. To some extent this is linked in with the discussion of the sample size of the experiment. However, it is usually straightforward to persuade students that a larger sample size is needed and that it would be helpful to have a computer doing the shuffling. We therefore move attention to a computer animation of a randomisation test of a larger dataset. Figure 3 presents one of the later stages in the animation, after we have shuffled the data 1000 times. The aim of the animation is to use the same data, but with larger datasets. We shuffle the data 1000 times and build up a histogram of values we obtain in this way.

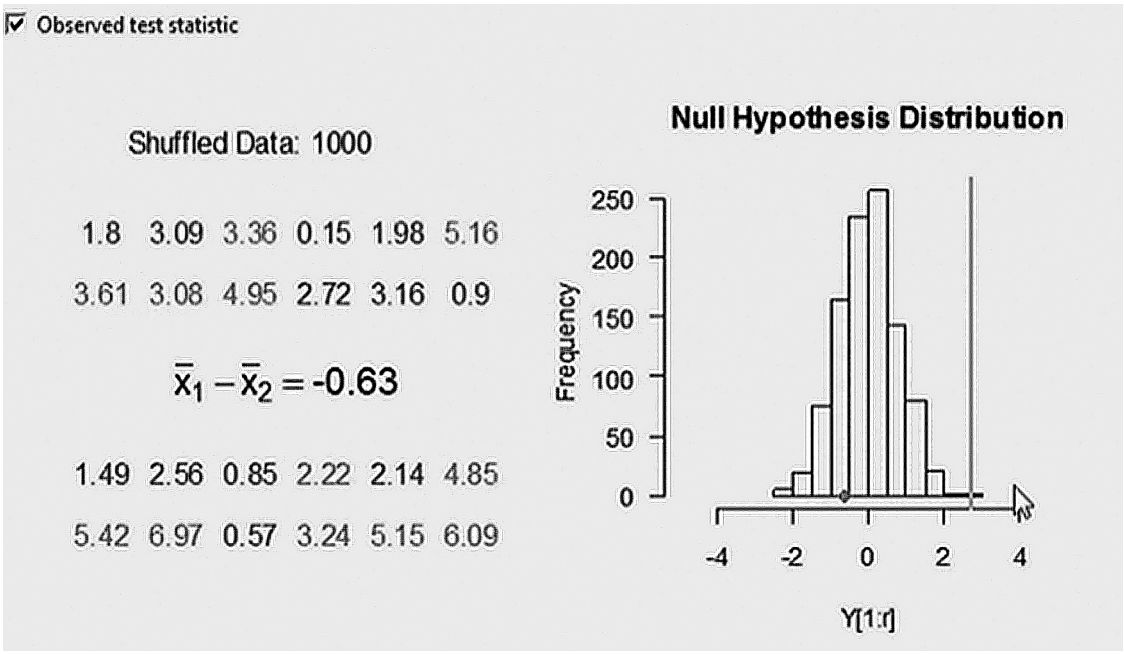

Figure 3. Computer animation of randomisation test.

The histogram therefore represents the sampling distribution of the test statistic for these data under the null hypothesis that the treatment has no effect (or more formally that both samples have been drawn from a population with the same population mean).

We can superimpose the original test statistic on these 1000 simulated values and then ask the question as to whether we believe the data we observed were indeed likely to have arisen from this null hypothesis distribution. Having hopefully rooted our students in the idea of the sampling distribution of the null we are free to explore several key ideas around hypothesis testing. Using a

simple Fisherian approach to hypothesis testing, we can approach the idea that if very few of the simulated replicates are larger than our observed test statistic, perhaps we will reject the null hypothesis. We can introduce the idea of tailedness. We can also introduce the idea that perhaps we should have used a better method for decision making and formulated an alternative hypothesis at the start of the experiment. Once using a computer simulation it is possible to consider the role of the size of the test quite carefully, and the implications for data sample size. We can explore more carefully what we mean by a p-value, and in particular, introduce the idea that if the null hypothesis is true we do indeed expect to make a decision to reject the null hypothesis incorrectly a proportion of times we are in this situation. In other words, we can contemplate simulations under scenarios we know the null hypothesis to be true and note that we will reject the null hypothesis a certain proportion of the time. We can also see that this approach to inference focuses attention on rejecting a null hypothesis and explain the logic behind non-statistically significant results only "failing to reject" the null.

1.3. Discussion

Nuzzo (2014) is just one of many papers that highlight the problems working professional scientists have with hypothesis testing. Responses to this paper vary from the suggestion we abandon hypothesis altogether (using either confidence intervals or Bayesian approaches). Inertia seems likely to prevent much movement in either direction for the foreseeable future. More conservative proposals suggest we focus on better education. The goal of this activity is to help to set out a sound conceptual framework for understanding hypothesis testing. It is hoped that this activity can at least make students clear that a p-value is something to do with the probability of data given hypothesis rather than the inverse, and highlight the way it is conditional on the sampling distribution of the very null hypothesis we wish to reject.

A YouTube clip of an early attempt to use this activity can be found here: https://www.youtube.com/watch?v=ESP0huKsKD0

2. References

Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer.

Hooton, J. W. L. (1991). Randomization tests: statistics for experimenters. *Computer Methods and Programs in Biomedicine*, 35, pp.43-51.

Meyer, J, and R. Land. (2003). *Threshold concepts and troublesome knowledge: linkages to ways of thinking and practising within the disciplines*. Edinburgh: University of Edinburgh.

Nuzzo, R. (2014). Statistical errors. *Nature* 506(7487), pp.150-152.

Zieffler, A. (2012). *Statistical Thinking: A Simulation Approach to Modeling*. Minneapolis: Catalyst Press.