THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

DATA to insight
An Introduction to Data Analysis

Statistical tests

Chris Wild, University of Auckland

The randomisation tests you have seen and been using are a special case of what is known as **statistical significance tests**, or **statistical hypothesis tests**. (The situations we have worked with correspond to testing the hypothesis that "there are no true differences between the effects of the treatments".)

This optional article is primarily addressed at those of you who had already met statistical significance testing, in some shape or form, before starting this course.

Just as with confidence intervals, there are methods for performing statistical significance tests based upon mathematical theories, particularly theory based on the normal distribution. You may have heard of the *2-sample t-test* for testing for a difference between two group means, the *one-way analysis of variance F-test* for testing for differences between more than 2 group means, or the *Chi-square test* which can be used for testing for differences between proportions, and there are many more. These methods have a long history and became well-established long before modern computer power made computationally-intensive methods like the randomisation test a practical possibility.

These theoretical methods are based on sampling theory – they assume that the data was sampled from a population of interest. There is no meaningful sampling involved in most experimental studies. The experimental units are "convenience samples" – people or entities close to hand who perhaps consented to be experimented upon. For these situations randomisation tests are the gold standard and the theoretical methods are justified by how well they approximate a relevant randomisation test. That said, subject to "passing appropriate assumption checks", the theoretically-based methods referred to above usually do a good job. (And, for long established problems, "they got there first".)

The tail proportions we have been working with correspond to the *p*-values of significance testing theory (there is a complication about *"sided-ness"*that we will deal with at the end of this article). We will use the *p*-value language for the remainder of this article so that the lessons you learn are more transferable. There are direct equivalents of what we are doing in this course with experimental situations for sampling-from-populations contexts but we won't deal with them here as it would take too long to set those ideas up. All of the major messages below apply to the use of significance testing in sampling situations as well.

Interpreting P-values

In experimental situations a **large *p*-value** (large tail proportion) means that the luck of the randomisation quite often produces group differences as large or even larger than what we've got in our data. In this case the data provides **no evidence** that there are true treatment-group differences.

A **small *p*-value** means that the luck of the randomisation draw hardly ever produces group differences as large as we've got in our data. They are almost always smaller. In this case the data **does provide evidence** that there are true treatment-group differences.

Furthermore, **the smaller the *p*-value, the stronger the evidence** that true treatment differences exist.

When should we start to claim "evidence" of true differences?

It is quite common practice to start claiming evidence of true differences if the *p*-value is 5% or smaller. At that point the result is commonly said to be **"statistically significant"** .

If the *p*-value is greater than 5%, the group differences are then said to be **"nonsignificant"**. In this case, writers of research reports often mistakenly write, "There is no difference" between the treatment groups, or between the effects of the treatments. "Nonsignificance" does not mean this at all. It is much more like a "not proven" verdict. Not being able to prove that something exists is not at all the same as proving that it does not exist.

Statistical significance does not imply practical significance

Treatment differences are *statistically significant* if "the data provides evidence that a true difference exists."

Treatment differences are *practically significant* if they are big enough to have a real-world impact.

Statistical significance says nothing about the size of treatment differences. *To estimate the sizes of differences you need confidence intervals.*

It also says nothing about whether these treatment differences are of any practical importance. For that you need both to know something about the size of the differences and to know whether differences of that size would have a practical impact.

So when you read a news report about a new drug that *significantly improves cancer survival*, it does not generally mean that people will live a lot longer. It means that someone has done a significance test and found that people on the drug *live **detectably** longer*. If you read that some hazard makes *no difference* to cancer rates it usually doesn't mean there is no difference, it means someone has done a significance test and got a *p*-value larger than 5% - that they were *unable to "prove" that a true difference exists.*

Error rates and multiple testing

For the purposes of this section we will think in terms of operating significance testing by getting excited (believing we've found a true difference) whenever we see a *p*-value smaller than 5%.

If you do a large number of tests in situations where there are no true differences, then you will mistakenly say you've found a true difference for 5% of the tests you perform even though there are no true differences to be found. (This is called making **Type 1 errors**).

Let's think drug trials. Now nobody operates in situations where all the drugs they test are completely useless. But in medicine most bright ideas turn out not to work. Let's suppose that 2% of the drugs we test produce detectable improvements, while 98% are completely useless. In the course of a year, we test 1,000 drugs. We will end up claiming "true improvement" for the 20 good drugs (2% of 1,000) and 5% of the 980 (i.e. 49) useless drugs. So of the drugs that we claim are "true improvements", we have 20 real improvements and 49 false claims.

Now let's combine this with a historical practice in many research journals where only research with "significant results" gets published. It's not hard to see why there have been claims that "most published research is false". The development here is a gross oversimplification of reality but it does highlight an area for concern. (There is a good, if rather technical, discussion of these issues on the Simply Statistics blog).

What all of this does highlight is that multiple testing is another area where, "Here be dragons". (Here is a great xkcd comic about multiple testing).

Geneticists working in genome-wide association studies are much less excitable. They work in a context where they test for associations between a disease and huge numbers of genes - believing that only a handful are active. They just don't know what genes are in that handful. *P*-values of 5% or even 1% don't raise the slightest flicker of interest. They don't start getting excited until their *p*-values get smaller than about If they are less stringent most of their results cannot be replicated.

This isn't a bad working rule: *Don't take study "results" too seriously until they've been replicated by others*. (Independent replication doesn't just address Type-1-error problems, it also addresses problems of bias.)

Sided-ness

You may have heard of 1-sided versus 2-sided tests (equivalently 1-tailed versus 2-tailed tests). What's that all about? We've ignored the issue addressed by this dichotomy till now to keep the complexity levels down.

Mainly we've emphasised tail proportions as a measure of "closeness to the edge" of the re-randomisation distribution. Consider the last video and the comparison between the proportions staying cocaine free under desipramine and lithium (the last video). In the data, the desipramine proportion was larger than the lithium proportion by 0.33 (33%) and the tail proportion was 0.02 (1 in 50).

Referring to the 0.33, I said that the values from re-randomisation were almost always smaller than that. But if desipramine being *bigger* than lithium by 0.33 made me sit up and take notice, so equally would desipramine being *smaller* than lithium by 0.33 (the same difference but in the opposite direction).

So if I want to answer the question: "How unusual is 0.33 under re-randomisation?" - I shouldn't be looking just at the tail propoportion above 0.33, I should be looking at the sum of the tail proportion above 0.33 **plus** the tail proportion below -0.33 (that is where the "2-tailed" comes from). If the re-randomisation distribution is basically symmetric this will be about or 4%. Thus, re-randomisation gives us a difference smaller **in magnitude** (or absolute value) than 0.33 about 96% of the time. So this is a small refinement of the way we argued in the video.

See Also

- The Guardian: Can chocolate make you smarter? (And thinner? And healthier?)
- XKCD cartoon: Clickbait-corrected p-value
- 2019 Article in *Nature* about "Statistical Significance"
- Teaching statistics through inferential reasoning (MOOC)
  This course allows you to learn, along with colleagues from other (mostly American) schools, how to emphasize inferential reasoning in teaching statistics through posing different types of investigative questions.