RNSH SERT Institute

Surgical | Education | Research | Training

GUIDES TO UNDERTAKING RESEARCH

3.1 Statistical Analysis –
A Random Sample of Thoughts

Statistical analysis is an enormous subject which cannot be seriously addressed in a short article, but below are some significant points to consider when designing, executing and analysing a study.

Statistical thinking needs to be done before the study begins

Researchers often think of data as needing statistical analysis after it has been collected, but statistical modelling and analysis should be considered early in studying the process of design of the study. The study should be designed at the outset to give the best chance that the data obtained will be informative. In the same way, no-one would consider using measurement equipment that was inaccurate or returning random data. Usually in medical projects the most important consideration is statistical power, summarised below.

The legendary RL Fisher, 1938:

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."*

What is your level of statistical training?

Any researcher that has a good grasp of normal distributions, probability, p-values, t-tests, ANOVA, chi-square tests and non-parametric testing has a good enough understanding to run simple analyses and should be at a level to discuss analyses and planning with a data manager or other statistically trained person. A researcher without those skills needs to tread very carefully, and will need an expert mentor to help with this aspect of the project, and to provide training. A researcher with relatively high skills, such that that Cox's proportional hazards model, Kolmogorov-Smirnov tests, Bayes' theorem and Kaplan-Meier curves hold no fears, then that researcher is better equipped. Even so it is always worthwhile to confer with peers and mentors about the methods and outcomes of the project analysis.

Statistical significance does not always mean medical significance

If patient weight is monitored as part of a study of a weight reduction therapy and that therapy causes a statistically significant decrease of 10g, no one would care. Large datasets in particular can sometimes give highly significant results when a statistical test is applied but the observations may not be meaningful if they only show up tiny changes that imply the intervention has negligible (if measurable) clinical effects.

Confidence in significant differences

When applying a t-test to compare the effects of a surgical intervention and sham intervention on levels of serum IL-101 (an imaginary hormone), three types of outcome are possible:

1) Treated patients have significantly higher IL-101 compared to sham treated people.
2) Treated patients have significantly lower IL-101 than the sham group.
3) No significant difference in IL-101 is seen between surgical and sham groups.

In the third outcome mean serum IL-101 levels in the treatment group lie inside a grey zone, the confidence interval, that lies around the mean of the sham operated group. If the treatment group mean lands in that zone it means that either:

- the intervention had *no effect* - the apparent difference seen is due to random variation, or
- The effects of the surgery was just *too small to see* because it was swamped by natural variation (a.k.a. noise) in the data.

The latter is a false negative, and means the *power* (or ability of the study to see an effect) may be insufficient, just as a magnifying glass cannot help someone see a single bacterium. We cannot distinguish lack of effect from lack of study power; in the same way if we can't see bacteria we don't know whether that bacterium is truly absent or whether we need better magnification.

So a statistically significant result proves an effect exists?

Evidence for an effect yes, proof not really. When the surgery *does* result in a significantly changed IL-101 level it means there is only *a low probability* that the data (or data more extreme) arose due to random chance –but not impossible. How low the probability is depends the level of rigour needed, but both <0.05 or <0.01 are common. Statistical inference can be a slippery thing; always involving layers of interpretation and depending on the right statistical model and method being used.

With great power comes great reproducibility

From the above, it follows we can only show a significant difference if the clinical effect is large enough that it is not overwhelmed by the data noise. One way to reduce random data variation or noise is to increase the number of patients measured in the study. If we do that then the surgery effect measurements may escape the grey zone, so that an effect (if it one exists) can be seen. This increase in number of measurements causes an increasing in the '*power'* of the experiment. This is the power to spot a small but true difference between groups. If a study with high power should avoid the error of claiming an effect when there is none (a false positive) and of failing to spot an effect when there really is one (a false negative).

Power can be estimated by calculation beforehand, and is commonly used to determine the number of patient measurements that will be needed in a study. Sufficient power is required in a study design for research ethics and funding applications. Insufficient power means that the study is a waste of time and of patients' time, hence has ethical and cost implications

The hidden horrors of multiple testing

When comparing mean values of two groups (say, placebo and a drug treatment) we may perform a significance test, such as a t-test. This give us a p-value representing the probability that difference between the groups (or a bigger difference) is due to random chance. We conventionally accept a p-value <0.05 as significant, though 0.01 is more rigorous.

However, if we make two comparisons at the same time (e.g., comparing placebo with drug A and with drug B) and if *both comparisons* give a p-value a smidge below 0.05 then the chance that one or both of the two results is actually due to random chance is about 9.75%, which is no good at all. This is because multiple comparisons mean that p-values must be combined and adjusted so that we are not fooled by a false positive result. A very common method for this is the Bonferroni adjustment.

In sum, multiple comparisons give a very different outcome to single comparison. It is like flipping a coin – the chance of getting a head is 50% with one flip once but flipping 20 times makes it near 100%.

Torturing the data till it confesses what you want

With large datasets it is possible to perform significance testing on many parameters, and if nothing looks good you may look for significance in a subset of the data. For example, if the test of a drug finds no effect then you might notice that if you exclude patients over 80 year old it looks more promising; this may reduce the p-value to below 0.05. This procedure is bad, it is called data torture and it is horrible, but surprisingly popular. A hypothesis test must be designed before the data is obtained, not after the researcher had a chance to squint at it. That said, such *post hoc* subset analyses may suggest a useful idea for a new hypothesis for testing in later studies, which is fine.

A good source of information about these and related statistical sins is found in the short classic *How to Lie with Statistics* by Darrell Huff (Norton & Co, 1954) if you can locate a copy.

But is it normal?

Most statistical tests (e.g., t-tests) assume that the data follows a normal distribution, so when it is plotted out the data has the distribution of a bell curve. This can be tested mathematically, which is less trouble than plots and more objective (after all, how bell-like is your curve?) but it if turns out your data does not follow a normal distribution then t-tests will tell you lies. There are ways around this problem, the most straightforward of which is to use a non-parametric statistical test that does not assume normal distribution of data. Thus, instead of a t-test use a Mann-Whitney test. Bear in mind, however, the power of these tests is less, so they may return a false negative result.

Statistics: not the only way.

Statistical analysis is not the only way to interrogate your data. There is an emerging field of machine learning which can spot patterns in the data when statistical analysis cannot. This can be worth considering, but it is not for the faint hearted and is generally only useful for very large datasets. It should also be noted that that this approach gives outcomes that are hypothesis-generating rather than analytical in nature.

*Presidential address to the Indian Statistical Congress, 1938. *Sankhya* 4, 14-17. A common quote, but Prof. Fisher was presumably unaware of any female statisticians in 1938.

Authorship: Julian Quinn ***Version*** 2.1 (Nov 2020)
Thanks to Thomas Hugh and Richard Piper for reviewing and critiquing this article.

Royal North Shore Hospital, DIVISION of SURGERY and ANAESTHESIA