SOS (Simplified Ophthalmic Statistics) Part 3: Which statistical test should I use (if any)?

A short series by **Catey Bunce** and **Tafadzwa Young-Zvandasara** for ophthalmic trainees.

P<0.05 is a statement that brings joy to many researchers. Arguably this is because inclusion of such a statement may increase the chance of acceptance for publication. Whilst statisticians and non-statisticians are united in trying to change this culture, cultural change takes time. It is therefore likely that many reading this article will be doing so in the hope that they may learn the skills to generate such statements.

In part 1 of this series we mentioned two types of statistical methods – inferential and descriptive. In part 2 of the series we gave you guidance on how best to describe your data (descriptive statistical methods). Here we cover tests of hypotheses (inferential statistical methods). We remind you that the reason you use statistical methods is to convert data into meaningful information to address important questions.

P values are generated by tests of significance and there are many different types. The tests work in a similar fashion and may also be described as hypothesis tests because they operate in a framework which involves declaring the current belief or null hypothesis and an alternative belief or hypothesis. The test computes a test statistic based upon the observed data and then determines by reference to a specific statistical distribution (different for different tests) a P value – the probability of observing as or more extreme data as that observed under the null hypothesis by chance alone. If the P value is less than a certain threshold (often set at 0.05) you may declare statistical significance and if not you may state that your results are not statistically significant at that threshold. In making this declaration, however, it is important to acknowledge that you may be making one of two mistakes:

a. A **type I** error where you say something is statistically significant when it is not.

b. A **type II** error where you say that something is not statistically significant whereas in reality it is.

The chance of making a type 1 error is called **alpha** and this is the threshold of significance. If we declare statistical significance to be a P value of 0.05, we are saying that the chance of making a type 1 error is 5%.

The chance of making a type II error is called beta and this depends on the effect size and sample size. Because of this dependence on sample size, when analysing big data even very small effect sizes may be declared statistically significant and similarly if you have very little data then you are unlikely to get a statistically significant result even where there is a large effect size. Some effect sizes matter clinically, others do not, for example, a difference between groups in intraocular pressure (IOP) of 1mmHg might not matter clinically, whilst 10mmHg might. ***Statistical significance is not the same as clinical significance.*** A non-significant P value does not mean that there is not a clinically relevant difference because beta depends also on sample size.

Which test?

Knowing the type of data that you have, whether the data are related and how many groups you have will guide you to the most appropriate statistical test (there may be more than one!) **What is paramount, however, is that you know what question you are trying to answer.**

- Types of questions may be to see whether groups differ on average to each other, perhaps one group treated with one drug and another with another drug?
- Are you seeing whether there are more adverse events in patients treated with one drug than in patients treated with another drug?
- Are you seeing whether there is a relationship between variables – does one tend to increase as another increases (perhaps intraocular pressure and age)?

Ideally, you will be clear about the primary research question your study is answering but sometimes this is not the case. You might, for example, have inherited a project from someone else or your supervisor might provide you with an interesting data set to "see if you can find something interesting here".

If there is a single primary research question (is IOP higher on patients treated with latanoprost than in patients treated on placebo?), here are some suggestions of how to approach this:

1. Step one is to think about the type of data that you have. From part 1 of this series we know that IOP is a continuous measure.
2. Step two would be to see if the data is normally distributed (see part 2 of this series).
3. Step three would be to think about whether or not the groups you are comparing (e.g. patients with latanoprost and patients without) are related in any way to each other.

IOP is typically normally distributed and if the patients are not the same patients, we would use an **unpaired t test**. If the groups we were comparing were the same patients (e.g. treated first with latanoprost and then placebo) we would use a **paired t test**.

If our outcome measure was continuous but skewed (as is sometimes the case with logMAR visual acuity) we would either apply

Table 1: Helpful things to determine which statistical test to use.

Research question	Is IOP higher in patients treated with latanoprost than in patients treated with placebo?	Are adverse events similar in patients treated with bevacizumab, ranibizumab and aflibercept?
Variable type(s)	IOP (continuous)	Adverse events (dichotomous)
Relatedness of data	Unpaired (different patients treated with latanoprost than placebo)	Unpaired (different patients treated with different anti-VEGF agents)
Number of groups	Two	Three

a transformation to normalise our dataset or use a non-parametric test. For more information on these, see the Ophthalmic Statistics notes series – papers 1, 9 and 10 [1,2,3].

The non-parametric equivalent to the paired t test is the **Wilcoxon signed rank test** whilst the non-parametric equivalent to the unpaired t test is the **Wilcoxon rank-sum test**, which is also rather confusingly known as the ***Mann-Whitney U*** test, the Mann-Whitney-Wilcoxon (MWW) test or indeed the Wilcoxon-Mann-Whitney test!

If we are comparing a continuous outcome with more than two groups we might use **ANOVA** for a continuous outcome and the **Kruskal Wallis test** for a skewed variable.

Suppose we now have two groups to compare but our outcome measure is categorical (has the patient suffered an adverse event, has the patient responded well to treatment?) In these situations again we need to think about whether data are paired or not. If paired we would use **McNemar's test** and if not we would use the **Chi square test** or **Fisher's exact test**.

Suppose now that we are not interested in seeing whether groups differ to each other but instead want to see whether a variable is related to another. For this we would use correlation and, if the data are normal, we use the **Pearson Correlation coefficient**, whilst if not we might use **Spearman's rank correlation. (Not used for assessing agreement between methods of measurement.)**

Should we use tests of normality?

In part 2 of this series we advised using a histogram to assess whether or not data followed a normal distribution. There are statistical tests that can be used to examine whether or not there is evidence of departure from normality. If you have a very large data, these tests may indicate evidence of non-normality but not of a magnitude that will impact upon the statistical methods you use because many are robust to non-normality. If you have a small data set then running these tests can be problematic because they say there is no evidence of non-normality when actually the data are pretty skewed.

Whenever looking a statistical test, check that any assumptions necessary for its use are adhered to.

If you use a test which makes an assumption that is not adhered to by your data you may end up with an incorrect answer. Remember that in making a decision based upon a P value you may be making a type I error – you might get statistical significance where there is none or a type II error (not getting statistical significance when in reality there is a true difference between groups or a true relationship between variables).

Everyone makes mistakes. Fortunately many mistakes leave no lasting impact and we learn from the experience. Statistical errors in medicine can and do on occasion result in harm to patients – a message most eloquently championed by Professor Doug Altman [4]. Despite commenting upon this in the 1980s, statistical errors persist in medicine [5]. Prof Altman died in June of this year (2018) and this has left a huge gap in the applied statistical community. His legacy remains, however, and we can demonstrate support for him by checking assumptions, reading his notes in the BMJ and speaking out, albeit politely, when we see misuse of statistics in medicine. It is hoped that this series might in some way support the message that he championed.

This is absolutely not a comprehensive guide to every statistical hypothesis test that exists. Even if we were to attempt to do that it would be time sensitive since statistics, just like medicine, is an evolving science. New tests are developed perhaps because methodologists identify weaknesses in a test that is in current use or because someone develops a novel way of better using data.

Table 2: Examples of statistical tests and when to use them.

Outcome is continuous	
Parametric	Non-parametric
Paired – Paired T test Unpaired – Unpaired t test ANOVA – More than two groups	Paired – Wilcoxon signed rank test Unpaired – Wilcoxon rank-sum test / Mann-Whitney U test Kruskal Wallis – More than two groups, skewed data
Outcome measure is categorical	
	Paired data – McNemar's test UnPaired data – Chi square test or Fisher's exact test
For correlation / group differences	
Pearson Correlation coefficient	Spearman's rank correlation

Catey Bunce is Ambassador for the Royal Statistical Society, championing the message better data = better research = better healthcare.

PREVIOUS LEARNING

- Statistical methods attempt to convert data into meaningful information that might answer a research question that you have.
- After classifying, exploring and summarising your variables you might wish to run a hypothesis test to establish whether your data provide support towards or against a particular belief or hypothesis.

CURRENT LEARNING

- Different tests are used for different research questions.
- Flow charts exist to assist you in identifying the correct test and there are apps for phones that can help.
- Statistical significance is NOT the same as clinical significance, the chance of a type II error depends upon the effect size and sample size – large data sets may result in significance but the effect size is of little clinical value. Small data sets may result in non-significance even though the effect size observed is of clinical value.
- Statistical mistakes in medicine can harm. When using a statistical test, check that you are adhering to the assumptions made by that test. If you are reviewing someone else's work, check whether they mention having checked assumptions.

References

1. Bunce C, Patel KV, Xing W, et al; Ophthalmic Statistics Group. Ophthalmic statistics note 1: unit of analysis. *Br J Ophthalmol* 2014;**98(3)**:408-12.
2. Skene SS, Bunce C, Freemantle N, Doré CJ; Ophthalmic Statistics Group. Ophthalmic statistics note 9: parametric versus non-parametric methods for data analysis. *Br J Ophthalmol* 2016;**100(7)**:877-8.
3. Bunce C, Stephenson J, Doré CJ, Freemantle N; Ophthalmic Statistics Group. Ophthalmic statistics note 10: data transformations. *Br J Ophthalmol* 2016;**100(12)**:1591-3.
4. Altman DG. Statistics and Ethics in Medical Research. Misuse of statistics is unethical. *Br Med J* 1980;**281(6249)**:1182-4.
5. Smith R. Medical research – still a scandal. *The BMJ Opinion* 2014. https://blogs.bmj.com/bmj/2014/01/31/richard-smith-medical-research-still-a-scandal/ Last accessed October 2018.

Further reading: the Simplified Ophthalmic Statistics series

- Bunce C, Young-Zvandasara T. SOS (Simplified Ophthalmic Statistics) Part 1: An introduction to data – how do we classify it and why does it matter? *Eye News* 2018;**24(6)**:36-7.
- Bunce C, Young-Zvandasara T. SOS (Simplified Ophthalmic Statistics) Part 2: How to summarise your data and why it's a good idea to do so. *Eye News* 2018;**25(2)**:34-6.

Useful resources

- Presenting Medical Statistics' book website: http://medical-statistics.info
- NIHR Statistics Group: https://statistics-group.nihr.ac.uk/research/new-sections/
- https://www.ucl.ac.uk/drupal/site_child-health/sites/child-health/files/test_flow.pdf

SECTION EDITOR

Tafadzwa Young-Zvandasara,

Fellow VR Surgery, Christchurch, New Zealand.

E: tpzvandas@hotmail.com

AUTHOR

Dr Catey Bunce,

Reader in Medical Statistics, School of Population Health and Environmental Sciences, Faculty of Life Sciences and Medicine, King's College London, UK.