Which statistical tests should I use?

Allison Shorten,[1] Brett Shorten[2]

10.1136/eb-2014-102003

[1]Yale School of Nursing, Yale University, New Haven, Connecticut, USA
[2]Informed Health Choices Trust, New South Wales, Australia

Correspondence to:
Dr Allison Shorten, Yale School of Nursing, Yale University, Yale West Campus, 400 West Campus Drive, Orange, CT 06477, USA;
allison.shorten@yale.edu

Statistical tests can be powerful tools for researchers. They provide valuable evidence from which we make decisions about the significance or robustness of research findings. Statistical tests are a critical part of the answers to our research questions and ultimately determine how confident we can be in the evidence to inform clinical practice. In addition to analysing data to answer research questions, readers of research also need to understand the underlying principles of common statistical tests. It is helpful to know whether statistical tests have been applied in the right way, at the right time with the right data.

There are dozens of statistical tests for researchers to choose from. The statistical tests we use help us gather evidence on which we will either accept or reject our stated null hypotheses[1] and therefore make conclusions about the findings of an experiment or exploration. The best way for researchers to ensure they are using the right statistical tests is to consult a specialist in statistics when research is being planned and before data is collected. Mapping out the analysis is an important step in research planning. Statistical tests should not be used as a substitute for good research design or to attempt to correct serious flaws in data.

We will use a hypothetical example to outline some common statistical tests we could use to answer common research questions. Let's say we developed a new 'mobile phone app' for patients recently diagnosed with diabetes, with the aim improving patient knowledge about healthy food choices. A sample of patients attending a diabetic clinic was carefully selected according to sound principles[2] and 200 patients were randomised into two groups. One group (n=100) would use the 'mobile phone app' over a period of 3 months and the other group would be the control (n=100), receiving their usual care in the clinic. Let's say we also developed a 15-item knowledge test consisting of true/false questions, based on the content in the 'mobile app' to assess levels of knowledge about healthy food choices. We administered this test to all patients both before and after the 3-month testing period. Now we will explore how different statistical tests can help us as we answer some key research questions that could have been asked by the researchers as they were planning the study. Note that different texts and statistical software packages sometimes use slightly different names for a given statistical test.

Question 1: Did patients have any baseline knowledge about healthy food choices before the study started?

What can we test? Suppose we wish to test whether the patients had any baseline knowledge of healthy food choices. If not, we would expect them, on average, to score 7.5 out of 15 on a true/false test (ie, 50%) even if they were guessing. So, for this test, the null hypothesis is that, for the population of all patients, baseline knowledge is 0 (mean=7.5) and the alternative is that patients already have some knowledge regarding healthy eating (mean >7.5).

Appropriate test: *one sample t test.*

Purpose of the test: this test is used to compare a single sample mean with a hypothesised population value.

What the results might look like? If the mean score of all 200 patients was higher than 7.5 with associated p value <0.05, then there is evidence that patients had at least some background knowledge of healthy eating.

Question 2: Did patients using the 'mobile app' score better on their knowledge test than those receiving usual care?

What can we test? We can test the hypothesis that mean knowledge scores were different between the patients who used the 'mobile app' and those receiving usual care, against the null hypothesis that mean scores were the same for both groups of patients.

Appropriate test: *independent samples t test.*

Purpose of the test: to compare means from two samples, to test whether they are significantly different or could plausibly have come from populations with the same mean value.

What the results might look like? If 200 patients completed the knowledge test after 3 months and the mean scores out of 15 were 11.30 for the 'mobile app' group and 9.05 for the control group and if the t statistic result was, say, 6.48 and associated p value was 0.000, this indicates that the difference observed between sample means was too large to be simply due to chance. We can be confident that the knowledge scores for patients in the 'mobile app' group were statistically greater than for the control group. A reported 95% CI of 1.57 to 2.93 for this test would indicate that it is 95% likely that the true population difference in mean knowledge scores in favour of the 'mobile app' group is between 1.57 and 2.93 questions.

Question 3: Did patients without the "mobile app" experience any improvements in knowledge over time?

What can we test? We can test the claim that, even without the 'mobile app' patient knowledge of healthy food choices improved over the period of 3 months, perhaps due to the effects of their routine clinic care. We can create a new variable defined as (postintervention score minus preintervention score). For example, a patient in the control group who scored 10 on the quiz at the start of the study and 12 at the end of 3 months would have the value of 12 −10=2 for this new variable.

Editor's choice
Scan to access more free content

Appropriate test: *paired samples t test.*

Purpose of the test: to compare two sample means where the data represent two observations of the same variable from each respondent (typically before and after scores or measurements).

What might the results look like? Suppose the mean difference in knowledge scores for the 100 patients who did not have the 'mobile app' was 0.43, with a calculated t statistic of 2.15 and p value of 0.034. This would indicate that the difference was statistically significant at the conventional significance level of 0.05. Therefore we would be confident that there was some improvement in knowledge among patients receiving usual care. If the reported 95% CI for this test was 0.03 to 0.82, it would suggest that we can be 95% confident that, in the overall population, the increase in correct responses would be in the range of 0.03–0.82 questions, or less than one question on the test.

Question 4: Did the education level of patients have an influence on patient preintervention level of knowledge (ie, before they started using the 'mobile app')?

What can we test? We could use this test if we wanted to examine the effect of level of education on preintervention knowledge scores, where education was measured in three categories: secondary school, diploma/certificate and bachelor degree or higher.

Appropriate test: *analysis of variance.*

Purpose of the test: to compare three or more means, in order to test whether some or all means are significantly different or whether all could plausibly have come from populations with the same mean value.

What might the results look like? Suppose the mean preintervention knowledge scores out of 15 for the three groups according to the level of education were 7.74, 8.70 and 9.56 respectively; the calculated F statistic was 9.83, with an associated p value of 0.000. This would mean there is convincing evidence that mean knowledge scores vary by level of education. Note that this does not mean that all three means are significantly different from each other. A range of further tests are available in this area. For example, another popular test called the Tukey test could be used to compare mean scores to see if, for example, secondary and degree or higher (7.74 vs 9.56) scores were statistically different.

Question 5: Could other factors have had a confounding effect on the differences observed in levels of knowledge for the patients?

What can we test? Suppose we are concerned that what we think has been a positive effect of the 'mobile app' has actually been confounded by unmeasured differences between the two groups involving English literacy levels, due to some patients being born overseas. We could use linear regression techniques to control for possible confounding factors such as whether patients have non-English speaking background. This way, after controlling for patients' history, we could better assess whether the improvements in knowledge were due to the 'mobile app'.

Appropriate test: *linear multiple regression analysis.*

Purpose of the test: regression analysis is a large and often complex field in its own right, with many different uses. In this type of study we might use regression models to allow us to control for a range of confounding factors that may affect the validity of the results obtained.

What the results might look like? If we used a linear multiple regression model and found that the regression estimate (known as a coefficient) was +2.29 it would tell us that, after controlling for place of birth, patients who received the 'mobile app' were estimated to get 2.29 more questions correct, on average, than patients in the control group. Note that it is common for a larger range of possible confounders to be included in a linear multiple regression model and this is just an example of one.

Remember, consult an expert to help select the right statistical tests for your research questions and gather the best evidence to inform future clinical practice.

Competing interests None.

References

1. **Shorten** A, Shorten B. Hypothesis testing and p values: how to interpret results and reach the right conclusions. *Evid Based Nurs* 2013;**16**:36–7.
2. **Shorten** A, Morley C. Selecting the sample. *Evid Based Nurs* 2014:**17**;32–3.