Chapter 6 Sample size, power, and efficient design

Answers to additional self-test questions

Q6.1 Why might we sometimes want to increase the power of an experiment, and why might we sometimes not?

The power of an experiment is the probability of detecting an effect when there is an effect to detect. Another way to say this is that the power is the probability of rejecting the null hypothesis when the null hypothesis is actually false. Yet another way to put this is that the power is one minus the probability of making a type II error.

Clearly an experiment with low power seems unappealing because we would go to a lot of work to perform the experiment with little chance that it will actually tell us anything very interesting. It will probably tell us that there is no evidence for rejecting the null hypothesis, but we won’t know if that is because the null hypothesis is actually true or because the experiment has low power to detect an effect that is really there.

We will not always strive for ever more power because there are diminishing returns on efforts made to increase power the higher the power already is, so we have to invest more and more in the experiment for further gains in power. At some point the ethical (or other) costs of further improvement in power will no longer be appealing.

Q6.2 How might we improve the power of an experiment?

Most simply we could increase the number of replicates (the sample size). We can also sometimes increase the effect size. For example, when we are looking to see if a drug has an effect relative to a placebo, we could increase the dosage of the drug. Lastly, we could reduce inherent variation. There are several ways to do that: if our drug or placebo is being administered to mice, we could strive to make sure the mice used in the experiment are as similar as possible and are kept in as standardized a way as possible. We could strive to make the measure of the effect of the drug or placebo as precise as we can, and to have a quantitative (interval ratio) measure of effect rather than a qualitative measure.

Q6.3 Can you think of an example where you would want power to be higher than 80%?

Imagine a company that makes some part of aeroplane engines wants to change the design of a certain critical component. The new design is expected to be just as reliable, but

Oxford University Press

cheaper. Our task in the experiment is to measure how long examples of components made to the two designs work before they malfunction. If our experiment detects no difference then the new design will be adopted. In this case, potentially thousands of lives rest on our experiment; if the new design is different from the old in terms of reliability then having a 20% chance of failing to detect that seems too large.

Q6.4 How can a pilot study help improve the power of your main experiment?

A pilot study can allow fine-tuning of measurement techniques; this should reduce imprecision in measurements and thus reduce inherent variation. The pilot study should also allow estimation of inherent variation, which can be used in a formal power analysis to help inform the design of the main experiment.

Q6.5 Imagine that we want to answer the question: '*Is the average height of male third year undergraduates in the engineering faculty at the University of Edinburgh different from the average height of equivalent students in the science faculty?*' How many would you sample in each faculty, given that there are about 300–400 males in each faculty?

Given that there are about 300–400 males in each population, measuring them all would be overkill. However, there are considerations that push us to thinking about having a large sample:

i) If there is a difference, then we expect it to be small.

ii) We expect quite substantial differences between individuals in the same population (i.e. within-group variation will be high).

iii) The measurement required from each individual sampled is quick and easy to take.

Overall, we think 100 would be very thorough, 50 would be just about right, and any less than 25 would seem like skimping.

Q6.6 If the populations were the same as the last question, but you wanted to test to see if there was a difference between these two populations in average number of books owned by a person, how would this influence the sample size that you would use?

Again, we expect a difference (if any) to be small, suggesting that a large sample would be useful. In comparison to the last question, we expect the within-population variance to be much higher for book ownership than height, suggesting that an even larger sample size would be required. Also, we'd expect errors in measurement to be much higher, again suggesting a larger sample size. Against this, collecting the information from an individual will require a greater amount of work, suggesting that, if possible, you'd like to get away with a smaller sample. This last point is of less concern than the other ones, so we'd say 50 still sounds like a good sample size, and 100 would not be unreasonable.

Q6.7 Explain in your own words the advantages of a larger sample size.

You want your sample to be representative of the population under study. The larger the sample size the more likely this is to be true. A smaller sample is more influenced by noise generated by the sampling process; a single unusual individual will have much less impact on a large sample than on a small one. For this reason, the power of statistical tests on a sample (i.e. the ability to reject the null hypothesis if this hypothesis is genuinely not true) increases with sample size.

Q6.8 Before you fly off on holiday, your plane is checked for metal fatigue. Which should you worry more about in this metal fatigue check: type I or type II errors?

A type I error occurs when the test suggests that metal fatigue has occurred when in fact it has not. The consequence of this is that your flight is delayed for between 30 minutes and 12 hours, while further checks are made or a replacement plane is found. A type II error occurs when there is actually metal fatigue but the test fails to detect it, the consequences of this are that you take off in a plane that is unsafe.

Hence, one should worry more about type II errors than type I in this case.