The first three steps in selecting an appropriate sample size (created 2009-07-20, updated 2010-01-15)

This page is moving to a new website.

I got an email last week from a client wanting to start a new research project looking at relationships between parenting beliefs and childhood behaviors. The description of the sorts of things to examine was quite elaborate, and it ended with the question "*how many families would we need to have any significant differences if they exist?*" Unfortunately, all the elaborate information provided did not include the information I would need to answer this question. Justifying a sample size usually involves three steps.

Step 1: Define your research hypothesis. Not all research requires a research hypothesis, but most research does, and until you define that hypothesis, it is impossible to make any progress on calculating an appropriate sample size. This particular email provided a fair amount of detail that could be used to derive a hypothesis, but no formal hypothesis was directly stated. I like to use the PICO format described in Evidence-Based Medicine to help people formulate a good research hypothesis. A research hypothesis will usually (but not always) have four elements:

- P: patient population. This is the group of patients that you want to examine.
- I: intervention. This is what you do to the group of patients that you think will help them improve
- C: comparison group. This is the group of patients without the intervention that you want to compare to.
- O: outcome. This is the variable that will indicate whether or not the intervention is successful.

Sometimes the intervention is not really something that you think will help the patients but rather an exposure that the patients have to endure that may produce some bad results.

A well-formulated hypothesis is important, because it tells the statistician what type of statistic is likely to be needed to test the hypothesis.

What do you do if you don't have a research hypothesis? In some research studies, the goal is exploratory. You don't have a formal hypothesis at the start of the study, but rather you are hoping that the data you collect will generate hypotheses for future studies. The path to selecting a sample size in these settings is quite different. Often you want to establish that the confidence intervals for some of the key descriptive statistics in these studies has a reasonable amount of precision.

Step 2: Find an estimate of the variability of your outcome measure. You've already done a literature review haven't you? If so, search through the papers in your review that used the same outcome measure that you are proposing in your study (the O in PICO). Ideally, the outcome measure will be examined in a group of patients that is close to the types of patients that you are studying (the P in PICO, or possibly the C in PICO). This is not always easy, and you will sometimes be forced to use a study where the patients are quite different from your patients. Don't fret too much about this, but make a good faith effort to find the most representative population that you can.

Some clients will raise an objection here and say that their research is unique, so it is impossible to find a comparable paper. It is true that most research is unique (otherwise it wouldn't be research). But what these people are worried about is that their intervention (the I in PICO) is unique. In these situations, the remainder of the hypothesis is usually quite mundane: the patients, the comparison group, and the outcome (P, C, and O in PICO) are all well studied. If you find a study where the P, C, and O match reasonably well, but the I doesn't, then you are probably going to get a good estimate of variation.

If there are major dissimilarities because this patient population (P) is very different than any previously studied patient population, or because the outcome measure (O) is newly developed by the researcher, then perhaps a pilot study would be needed to establish a reasonable estimate of variation.

Sometimes you can infer a standard deviation through general principles. If a variable is constrained to be between 0 and 100, it would be impossible, for example, for the standard deviation to be five thousand. There are formulas relating the range of a distribution to the standard deviation that can serve in a pinch if no other data is available. If your outcome measure is a proportion, for example, then the variation is related to the estimated proportion. Similarly, the variation in a count variable is related to the mean of the counts. Find a paper that establishes a proportion or average count in a control group similar to your control group and any competent statistician will be able to get an estimate of variation. In some situations, the amount of variation in a proportion or count is larger than would be expected by the statistical distributions (binomial and Poisson) traditionally associated with these measures. Still, a calculation based on binomial or Poisson assumptions is a reasonable starting point for further calculations.

Step 3: define the minimum clinically important difference.

The minimum clinically significant difference is the boundary between a difference so small that no one would adopt the new intervention on the basis of such a meager changer and a difference large enough to make a difference (that is, to convince people to change their behavior and adopt the new therapy).

Establishing the minimum clinically relevant difference is a tricky task, but it is something that should be done prior to any research study. It's not easy but this is something that you have to do for yourself. The clinically relevant difference is determined by medical experts and not by statisticians. Hey, I'm still trying to understand the difference between good and bad cholesterol; I wouldn't even be able to start thinking about how much of a change in cholesterol is considered clinically relevant. You might start by asking yourself "How much of an improvement would I have to see before I would adopt a new treatment?" Also, try talking with some of your colleagues. And look at the size of improvements for other successful treatments.

For binary outcomes, the choice is not too difficult in theory. Suppose that an intervention "costs" X dollars in the sense that it produces that much pain, discomfort, and inconvenience, in addition to any direct monetary costs. Suppose the value of a cure is kX where k is a number greater than 1. A number less than 1, of course, means that even if you could cure everyone, the costs outweigh the benefits of the cure.

For k>1, the minimum clinically significant difference in proportions is 1/k. So if the cure is 10 times more valuable than the costs, then you need to show at least a 10% better cure rate (in absolute terms) than no treatment or the current standard of treatment. Otherwise, the cure is worse than the disease.

It helps to visualize this with certain types of alternative medicine. If your treatment is aromatherapy, there is almost no cost involved, so even a very slight probability of improvement might be worth it. But Gerson therapy, which involves, among other things, coffee enemas, is a different story. An enema is reasonably safe, but is not totally risk free. And it involves a substantially greater level of inconvenience than aromatherapy. So you'd only adopt Gerson therapy if it helped a substantial fraction of patients. Exactly how many depends on the dollar value that you place on having to endure a coffee enema, a task that I will leave for someone else to quantify.

If there are side effects associated with the treatment that only occur in a fraction of the patients receiving the treatment, then the calculations are a bit trickier, but still possible in theory. It becomes more tricky still when different people place different monetary values on the risks and inconveniences of a new therapy.

For continuous variables, the minimum clinically significant difference could be defined as above. Define a threshold that represents "better" versus "not better" and then try to shift the entire distribution so that the fraction "better" under the new treatment is at least 1/k.

There have also been efforts to elucidate, through experiments, interviews, and other approaches, what the average person considers an important shift to be. For the visual analog scale of pain, for example, a shift of at least 15 mm is considered the smallest value that is noticeable to the average patient.

This page is an update of material originally published at my old website:

- http://www.childrens-mercy.org/stats/size/power.asp

This page was written by Steve Simon and was last modified on 2017-06-15. Need more information? I have a page with general help resources. You can also browse for pages similar to this one at Category: Sample size justification.