Science Discussion Series: The importance of sample size in science and how to talk about sample size. : science

Summary: Most laymen readers of research do not actually understand what constitutes a proper sample size for a given research question and therefore often fail to fully appreciate the limitations or importance of a study's findings. This discussion aims to simply explain what a sample size is, the consequence of too big or too small sample sizes for a given research question, and how sample size is often discussed with respect to evaluating the validity of research without being too technical or mathematical.

It should already be obvious that very few scientific studies sample whole population of individuals without considerable effort and money involved. If we could do that and have no errors in our estimations (e.g., like counting beads in a jar), we would have no uncertainty in the conclusions barring dishonesty in the measurements. The true values are in front of you for to analyze and no intensive data methods needed. This rarely is the case however and instead, many theatres of research rely on obtaining a sample of the population, which we define as the portion of the population that we actually can measure.

Defining the sample size

One of the fundamental tenets of scientific research is that a good study has a good-sized sample, or multiple samples, to draw data from. Thus, I believe that perhaps one of the first criticisms of scientific research starts with the sample size. I define the sample size, for practical reasons, as the number of individual sampling units contained within the sample (or each sample if

multiple). The sampling unit, then, is defined as that unit from which a measurement is obtained. A sampling unit can be as simple as an individual, or it can be a group of individuals (in this case each individual is called a sub-sampling unit). With that in mind, let's put forward and talk about the idea that a proper sample size for a study is that which contains enough sampling units to appropriately address the question involved. An important note: sample size should not be confused with the number of replicates. At times, they can be equivalent with respect to the design of a study, but they fundamentally mean different things.

The Random Sample

But what actually constitutes an appropriate sample size? Ideally, the best sample size is the population, but again we do not have the money or time to sample every single individual. But it would be great if we could take some piece of the population that correctly captures the variability among everybody, in the correct proportions, so that the sample reflects that which we would find in the population. We call such a sample the “perfectly random sample”. Technically speaking, a perfect random sample accurately reflects the variability in the population regardless of sample size. Thus, a perfect random sample with a size of 1 unit could, theoretically, represent the entire population. But, that would only occur if every unit was essentially equivalent (no variability at all between units). If there is variability among units within a population, then the size of the perfectly random sample must obviously be greater than 1.

Thus, one point of the unending discussion is focused on what sample size would be virtually equivalent to that of a perfectly random sample. For intuitive reasons, we often look to sample as many units as possible. But, there’s a catch: sample sizes can be either too small or, paradoxically, too large for a given question (Sandelowski 1995). When the sample size is too small, redundancy of information becomes questionable. This means that the estimates obtained from the sample(s) do not reliably converge on the true

value. There is a lot of variability that exceeds that which we would expect from the population. It is this problem that's most common among the literature, but also one that most people cling to if a study conflicts with their beliefs about the true value. On the other hand, if the sample size is too large, the variability among units is small and individual variability (which may be the actual point of investigation) becomes muted by the overall sample variability. In other words, the sample size reflects the behavior and variability of the whole collective, not of the behavior of individual units. Finally, whether or not the population is actually important needs to be considered. Some questions are not at all interested in population variability.

It should now be more clear why, for many research questions, the sample size should be that which addresses the questions of the experiment. Some studies need more than 400 units, and others may not need more than 10. But some may say that to prevent arbitrariness, there needs to be some methodology or protocol which helps us determine an optimal sample size to draw data from, one which most approximates the perfectly random sample and also meets the question of the experiment. Many types of analyses have been devised to tackle this question. So-called power analysis (Cohen 1992) is one type which takes into account effect size (magnitude of the differences between treatments) and other statistical criteria (especially the significance level, alpha [usually 0.05]) to calculate the optimal sample size. Others also exist (e.g., Bayesian methods and confidence intervals, see Lenth 2001) which may be used depending on the level resolution required by the researcher. But these analyses only provide numbers and therefore have one very contentious drawback: they do not tell you how to draw the sample.

Discussing Sample Size

Based on my experiences with discussing research with folks, the question of sample size tends not to concern the number of units within a sample or across multiple samples. In fact, most people who pose this argument, specifically to dismiss research results, are really arguing against how the

researchers drew their sample. As a result of this conflation, popular media and public skeptics fail to appreciate the real meanings of the conclusions of the research. I chalk this up to a lack of formal training in science and pre-existing personal biases surrounding real world perceptions and experiences. But I also think that it is nonetheless a critical job for scientists and other practitioners to clearly communicate the justification for the sample obtained, and the power of their inference given the sample size.

I end the discussion with a point: most immediate dismissals of research come from people who associate the goal of the study with attempting to extrapolate its findings to the world picture. Not much research aims to do this. In fact, most don't because the criteria for generalizability becomes much stronger and more rigorous at larger and larger study scales. Much research today is focused on establishing new frontiers, ideas, and theories so many studies tend to be first in their field. Thus, many of these foundational studies usually have too small sample sizes to begin with. This is absolutely fine for the purpose of communication of novel findings and ideas. Science can then replicate and repeat these studies with larger sample sizes to see if they hold. But, the unfortunate status of replicability is a topic for another discussion.

Some Sources


```
Lenth 2001 (http://dx.doi.org/10.1198/000313001317098149)
Cohen 1992 (http://dx.doi.org/10.1037/0033-2909.112.1.155)
Sandelowski 1995 (http://onlinelibrary.wiley.com/doi/10.1002/nur.4770180
```


An example of too big of a sample size for a question of interest.

A local ice cream franchise is well known for their two homemade flavors, serious vanilla and whacky chocolate. The owner wants to make sure all 7 of his parlors have enough ice cream of both flavors to satisfy his customers, but also just enough of each flavor so that neither one sits in the freezer for too

long. However, he is not sure which flavor is more popular and thus which flavor there should be more of. Let's assume he successfully surveys every person in the entire city for their preference (sample size = the number of residents of the city) and finds out that 15% of the sample prefers serious vanilla, and 85% loves whacky chocolate. Therefore, he decides to stock more whacky chocolate at all of his ice cream parlors than serious vanilla.

However, three months later he notices that 3 of the 7 franchises are not selling all of their whacky chocolate in a timely manner and instead serious vanilla is selling out too quickly. He thinks for a minute and realizes he assumed that the preferences of the whole population also reflected the preferences of the residents living near his parlors which appeared to be incorrect. Thus, he instead groups the samples into 7 distinct clusters, decreasing the sample size from the total number of residents to a sample size of 7, each unit representing a neighborhood around the parlor. He now found that 3 of the clusters preferred serious vanilla whereas the other 4 preferred whacky chocolate. Just to be sure of the trustworthiness of the results, the owner also looked at how consistently people preferred the winning flavor. He saw that within 5 of the 7 clusters, there was very little variability in flavor preference meaning he could reliably stock more of one type of ice cream, but 2 of the parlors showed great variability, indicating he should consider stocking equitable amounts of ice cream at those parlors to be safe.