Steve Sawin
Statistics

Study Design

Steps in Study Design

1. **initial question** This is the question you want to consider. It is important that you know what your question is before you look at your data: "hunting expeditions" that look for patters in data after the fact are great, but the powerful techniques we will learn to draw conclusions do not work with them.

2. **precise formulation** Translate the initial question into a precise mathematical question. This means choosing the exact population under consideration, the variables, the parameters, and a precise statement about them which is the *Alternate Hypothesis* (see below) you will be testing.

   For example, if your initial question is "does TV rot your brain," you must pick a population to take the place of "you," (perhaps adult Americans, or students at Fairfield U.), you need variables that measure someone's TV watching and their level of brain rot (perhaps you would measure hours per week of TV watching, and score on some universal measure of intelligence, like a statistics final), and parameters and a question that captures the initial question precisely (perhaps the coefficient of correlation or slope between these variables, in which case the Alternate Hypothesis would be that the slope or coefficient is negative.

   Note that the precise formulation is generally significantly different from the original question, and it is an important issue how well the former does at capturing the content of the latter.

3. **sampling procedure/design** since you will generally not be able to gather information about the population of interest, you will take a sample that you hope is representative. The way to do this is to develop a *sampling procedure* that is *random.* That is, a procedure for picking the individuals in the sample which can in principle be repeated as often as you like, each time giving potentially different individuals, in such a way that in the long run each individual should show up in samples equally often (that's not quite true, in some case certain individuals get put into samples more often, but are weighted in the sample less in a way that accounts for this. this is not an issue we will cover). The gold standard of this is the *simple random sample*, in which a list of every individual in the population is made, and the correct number of individuals from the list is chosen by a random number generator, or pulling names from a hat, or other purely random procedure.

   It is important to note that a random procedure is not the same as haphazard. Procedures that rely on whim ("I stopped at a place that looked good and dug"), convenience ("I asked all my friends"), or consciously varying the procedure as you go ("we asked some people in dorms, some people at a soccer game, and some at lunch") are recipes for adding bias (see below) and for making precise analysis of the accuracy impossible.

In practice a simple random sample is impractical and hence rare. On the one hand studies often use more complex random samples. This is perfectly legitimate, and gives just as reliable results as long as the statistical calculations are modified in appropriate ways to account for the different procedure. On the other hand, often studies rely on procedures that are not random. This technically invalidates the statistical procedure, but is often unavoidable, and thus is frequently accepted. When it is unavoidable, the key question to ask is: Which individuals are more or less likely to be selected, and how are they likely to differ with respect to the variables being measured. If you are convinced that they will not differ in a significant way, you should remain aware that you may have failed to think of a significant effect that invalidates your findings, but it is generally considered appropriate to proceed with caution (and humility!).

4. **measurement** Measurement is the process of actually finding the values of the variables for each element of your sample. Sometimes this is straightforward, but often not. For example if your variable is the number of minutes per week a person spends watching television, does it count if they are watching TV and studying at the same time? If their roommate is watching TV and they are in the line of sight? Renting a video? Downloading clips of Comedy Central from a website? Maybe none of these decisions makes a big difference to your results, but you need to decide this ahead of time. And of course you could measure it by following the person around every waking moment with a clipboard, but that is not practical. If you just ask them based on memory, how accurate will their answer be and how will it depend on the form of the question?

5. **data analysis** Most of the course focuses on this step, in some ways the simplest. If everything else has been done well, all you need do is take the data from thew previous step, plug it into the appropriate statistical calculation, and read off the answer. The key here is that it is generally straightforward to answer precisely formulated question if it were about the sample, but the trick is deciding how likely it is to be telling you something about the population as a whole.

Bias and Variability

Always remember that when you design a study you are designing a *procedure* for doing a study that could be repeated indefinitely, and then you are implementing it once to get your actual data. Almost all confusions on this subject involving failing to distinguish between the procedure and the instance properly.

A source of *bias* is any aspect of the study design which would cause, in the long run, the results of samples to differ in a consistent fashion from the true answer about the population. A source of *variability* is an aspect which cause individual samples to differ from the true answer about the population, but randomly, so that the differences average out in the long run. If you sample people by dialing phone numbers randomly, you will typically get few if any people without a phone or with an unlisted number, and get too many people who have multiple phone lines. Since that will be true in the long run, it is a source of bias. Of course some of the people you sample that way will be tall and some short, and in the sample you actually do there might be too many tall people, but in the long run that should average out. That is called a source of variability.

It is good to limit sources of variability when possible, because that increases the chance you will get conclusive results. So if height were really significant in your study, you might drop tall or short people randomly from your study to make the proportion of tall and short match the overall population. But variability is not a reason to distrust whatever conclusions you do draw from a statistical procedure, so it is not a deal-killer. Bias means you are not answering the question you set out to answer, and thus a bias can completely invalidate the study. Potential sources of bias are thus extremely important to identify, and to think about the effects of. You have thought of a potential source of bias if you can see the direction the problem would move the data in, otherwise it is probably a source of variability. So arguing that teens would underestimate the amount they have sex in a survey because of parental disapproval, or that they would overestimate because of peer disapproval, are both legitimate possible sources of bias (even though they could conceivably cancel out), but arguing people may not remember how often they do their laundry is just a source of variability unless you can argue they are likely misestimate consistently in one direction.

Sampling bias is bias introduced by the sampling procedure, which is to say any way in which some individuals are more likely than others to be picked in the sample, and differ in a consistent way in their answers to the variables. It includes *nonresponse bias,* where people who refuse to answer your questions are obviously underrepresented in your survey (one can almost always make an argument that such people would differ in some consistent way on your variables from those who agree to respond. To argue something is a source of sampling bias you must identify a group of individuals either more or less likely than others to be represented in the survey, and then identify a direction in which they would typically differ from others on the value of the variable.

Measurement bias is bias introduced when the values of the variable you measure do not accurately reflect the values you meant to be measuring (in a consistent way). Most often this will be because your are relying on people's honest responses, and they give an untruthful answer, as in the teens and sex question above. To argue something is a measurement bias, you must suggest why the responses would often differ in a particular direction from the true values of the variable.

If your precise formulation does not accurately reflect the original question, this can also be a source of bias. It has no name and is not generally talked about by textbooks because it is a little fuzzy, but it can be a huge source of bias.

If you are looking for a *causal relationship* between two variables and you do an *observational study* rather than an *experiment,* then any *lurking variable* is a source of bias. To identify something as a potential lurking variable you should offer a plausible reason why it might be associated with the response variable and influence the explanatory variable. Generally the second is the tricky bit.

The Null and Alternate Hypothesis

In science at least, studies frequently address yes or no questions. The logic here is particularly subtle and important, and the terminology a little scary. When you do a study you are gathering evidence for a particular proposition and assessing whether the evidence is convincing. That proposition is called the Alternate Hypothesis, and symbolized H_A or sometimes H_1. If you are assessing whether women as for directions more than men, the Alternate Hypothesis would be that they do. The Null Hypothesis or H_0 is the proposition you would presume until there is evidence otherwise. It is roughly the opposite of H_A, but for technical reasons tends to be highly specific. In the directions asking example H_0 would be that there is no difference (not quite the opposite, because we have left out the possibility that men ask more directions). If you can't decide which is Null and which Alternate, often the easiest way is to note that the Null generally involves an exact equality, or the assertion that there is no difference or no association or no effect, and the Alternate is generally an inequality, asserting that one thing is more or different from another, or that there is a difference or an association or an effect.

In the end, we will ask the question, "If the Null Hypothesis were true, how likely would it be for us to see the kind of results we did?" The answer to that question (a probability) is called the p-value. If the p-value is very small, our results are unlikely to have happened by chance so the Null Hypothesis is probably false, and we take this as evidence that the Alternate is true. If the p-value is large, than our results are perfectly consistent with the Null Hypothesis, so we do not take it as evidence for the Alternate Hypothesis.