18.4: Other Ways of Doing Inference

A different sense in which this book is incomplete is that it focuses pretty heavily on a very narrow and old-fashioned view of how inferential statistics should be done. In Chapter 10 I talked a little bit about the idea of unbiased estimators, sampling distributions and so on. In Chapter 11 I talked about the theory of null hypothesis significance testing and p-values. These ideas have been around since the early 20th century, and the tools that I've talked about in the book rely very heavily on the theoretical ideas from that time. I've felt obligated to stick to those topics because the vast majority of data analysis in science is also reliant on those ideas. However, the theory of statistics is not restricted to those topics, and – while everyone should know about them because of their practical importance – in many respects those ideas do not represent best practice for contemporary data analysis. One of the things that I'm especially happy with is that I've been able to go a little beyond this. Chapter 17 now presents the Bayesian perspective in a reasonable amount of detail, but the book overall is still pretty heavily weighted towards the frequentist orthodoxy. Additionally, there are a number of other approaches to inference that are worth mentioning:

- **Bootstrapping** Throughout the book, whenever I've introduced a hypothesis test, I've had a strong tendency just to make assertions like "the sampling distribution for BLAH is a t-distribution" or something like that. In some cases, I've actually attempted to justify this assertion. For example, when talking about χ^2 tests in Chapter 12, I made reference to the known relationship between normal distributions and χ^2 distributions (see Chapter 9) to explain how we end up assuming that the sampling distribution of the goodness of fit statistic is χ^2. However, it's also the case that a lot of these sampling distributions are, well, wrong. The χ^2 test is a good example: it is based on an assumption about the distribution of your data, an assumption which is known to be wrong for small sample sizes! Back in the early 20th century, there wasn't much you could do about this situation: statisticians had developed mathematical results that said that "under assumptions BLAH about the data, the sampling distribution is approximately BLAH", and that was about the best you could do. A lot of times they didn't even have that: there are lots of data analysis situations for which no-one has found a mathematical solution for the sampling distributions that you need. And so up until the late 20th century, the corresponding tests didn't exist or didn't work. However, computers have changed all that now. There are lots of fancy tricks, and some not-so-fancy, that you can use to get around it. The simplest of these is bootstrapping, and in it's simplest form it's incredibly simple. Here it is: simulate the results of your experiment lots and lots of time, under the twin assumptions that (a) the null hypothesis is true and (b) the unknown population distribution actually looks pretty similar to your raw data. In other words, instead of assuming that the data are (for instance) normally distributed, just assume that the population looks the same as your sample, and then use computers to simulate the sampling distribution for your test statistic if that assumption holds. Despite relying on a somewhat dubious assumption (i.e., the population distribution is the same as the sample!) bootstrapping is quick and easy method that works remarkably well in practice for lots of data analysis problems.
- **Cross validation** One question that pops up in my stats classes every now and then, usually by a student trying to be provocative, is "Why do we care about inferential statistics at all? Why not just describe your sample?" The answer to the question is usually something like this: "Because our true interest as scientists is not the specific sample that we have observed in the ***past***, we want to make predictions about data we might observe in the ***future***". A lot of the issues in statistical inference arise because of the fact that we always expect the future to be similar to but a bit different from the past. Or, more generally, new data won't be quite the same as old data. What we do, in a lot of situations, is try to derive mathematical rules that help us to draw the inferences that are most likely to be correct for new data, rather than to pick the statements that best describe old data. For instance, given two models A and B, and a data set X you collected today, try to pick the model that will best describe a new data set Y that you're going to collect tomorrow. Sometimes it's convenient to simulate the process, and that's what cross-validation does. What you do is divide your data set into two subsets, X1 and X2. Use the subset X1 to train the model (e.g., estimate regression coefficients, let's say), but then assess the model performance on the other one X2. This gives you a measure of how well the model *generalises* from an old data set to a new one, and is often a better measure of how good your model is than if you just fit it to the full data set X.
- **Robust statistics** Life is messy, and nothing really works the way it's supposed to. This is just as true for statistics as it is for anything else, and when trying to analyse data we're often stuck with all sorts of problems in which the data are just messier than they're supposed to be. Variables that are supposed to be normally distributed are not *actually* normally distributed, relationships that are supposed to be linear are not *actually* linear, and some of the observations in your data set are almost certainly junk (i.e., not measuring what they're supposed to). All of this messiness is ignored in most of the statistical theory I developed in this book. However, ignoring a problem doesn't always solve it. Sometimes, it's actually okay to ignore the mess, because some types of statistical tools are "robust": if the data don't satisfy your theoretical assumptions, they still work pretty well. Other types of statistical tools are not robust: even minor deviations from the theoretical assumptions cause them to break.

Robust statistics is a branch of stats concerned with this question, and they talk about things like the "breakdown point" of a statistic: that is, how messy does your data have to be before the statistic cannot be trusted? I touched on this in places. The mean is *not* a robust estimator of the central tendency of a variable; the median is. For instance, suppose I told you that the ages of my five best friends are 34, 39, 31, 43 and 4003 years. How old do you think they are on average? That is, what is the true population mean here? If you use the sample mean as your estimator of the population mean, you get an answer of 830 years. If you use the sample median as the estimator of the population mean, you get an answer of 39 years. Notice that, even though you're "technically" doing the wrong thing in the second case (using the median to estimate the mean!) you're actually getting a better answer. The problem here is that one of the observations is clearly, obviously a lie. I don't have a friend aged 4003 years. It's probably a typo: I probably meant to type 43. But what if I had typed 53 instead of 43, or 34 instead of 43? Could you be sure if this was a typo? Sometimes the errors in the data are subtle, so you can't detect them just by eyeballing the sample, but they're still errors that contaminate your data, and they still affect your conclusions. Robust statistics is a concerned with how you can make *safe* inferences even when faced with contamination that you don't know about. It's pretty cool stuff.

18.4.1 Miscellaneous topics

- **Missing data** Suppose you're doing a survey, and you're interested in exercise and weight. You send data to four people. Adam says he exercises a lot and is not overweight. Briony says she exercises a lot and is not overweight. Carol says she does not exercise and is overweight. Dan says he does not exercise and refuses to answer the question about his weight. Elaine does not return the survey. You now have a missing data problem. There is one entire survey missing, and one question missing from another one, What do you do about it? I've only barely touched on this question in this book, in Section5.8, and in that section all I did was tell you about some R commands you can use to ignore the missing data. But ignoring missing data is not, in general, a safe thing to do. Let's think about Dan's survey here. Firstly, notice that, on the basis of my other responses, I appear to be more similar to Carol (neither of us exercise) than to Adam or Briony. So if you were forced to guess my weight, you'd guess that I'm closer to her than to them. Maybe you'd make some correction for the fact that Adam and I are males and Briony and Carol are females. The statistical name for this kind of guessing is "imputation". Doing imputation safely is hard, but important, especially when the missing data are missing in a systematic way. Because of the fact that people who are overweight are often pressured to feel poorly about their weight (often thanks to public health campaigns), we actually have reason to suspect that the people who are not responding are more likely to be overweight than the people who do respond. Imputing a weight to Dan means that the number of overweight people in the sample will probably rise from 1 out of 3 (if we ignore Dan), to 2 out of 4 (if we impute Dan's weight). Clearly this matters. But doing it sensibly is more complicated than it sounds. Earlier, I suggested you should treat me like Carol, since we gave the same answer to the exercise question. But that's not quite right: there is a systematic difference between us. She answered the question, and I didn't. Given the social pressures faced by overweight people, isn't it likely that I'm *more* overweight than Carol? And of course this is still ignoring the fact that it's not sensible to impute a *single* weight to me, as if you actually knew my weight. Instead, what you need to do it is impute a range of plausible guesses (referred to as multiple imputation), in order to capture the fact that you're more uncertain about my weight than you are about Carol's. And let's not get started on the problem posed by the fact that Elaine didn't send in the survey. As you can probably guess, dealing with missing data is an increasingly important topic. In fact, I've been told that a lot of journals in some fields will not accept studies that have missing data unless some kind of sensible multiple imputation scheme is followed.
- **Power analysis** In Chapter 11 I discussed the concept of power (i.e., how likely are you to be able to detect an effect if it actually exists), and referred to power analysis, a collection of tools that are useful for assessing how much power your study has. Power analysis can be useful for planning a study (e.g., figuring out how large a sample you're likely to need), but it also serves a useful role in analysing data that you already collected. For instance, suppose you get a significant result, and you have an estimate of your effect size. You can use this information to estimate how much power your study actually had. This is kind of useful, especially if your effect size is not large. For instance, suppose you reject the null hypothesis $p<.05$, but you use power analysis to figure out that your estimated power was only .08. The significant result means that, if the null hypothesis was in fact true, there was a 5% chance of getting data like this. But the low power means that, even if the null hypothesis is false, the effect size was really as small as it looks, there was only an 8% chance of getting data like the one you did. This suggests that you need to be pretty cautious, because luck seems to have played a big part in your results, one way or the other!
- **Data analysis using theory-inspired models** In a few places in this book I've mentioned response time (RT) data, where you record how long it takes someone to do something (e.g., make a simple decision). I've mentioned that RT data are almost invariably non-normal, and positively skewed. Additionally, there's a thing known as the speed-accuracy tradeoff: if you try to make decisions too quickly (low RT), you're likely to make poorer decisions (lower accuracy). So if you measure both the

accuracy of a participant's decisions and their RT, you'll probably find that speed and accuracy are related. There's more to the story than this, of course, because some people make better decisions than others regardless of how fast they're going. Moreover, speed depends on both cognitive processes (i.e., time spend thinking) but also physiological ones (e.g., how fast can you move your muscles). It's starting to sound like analysing this data will be a complicated process. And indeed it is, but one of the things that you find when you dig into the psychological literature is that there already exist mathematical models (called "sequential sampling models") that describe how people make simple decisions, and these models take into account a lot of the factors I mentioned above. You won't find any of these theoretically-inspired models in a standard statistics textbook. Standard stats textbooks describe standard tools, tools that could meaningfully be applied in lots of different disciplines, not just psychology. ANOVA is an example of a standard tool: it is just as applicable to psychology as to pharmacology. Sequential sampling models are not: they are psychology-specific, more or less. This doesn't make them less powerful tools: in fact, if you're analysing data where people have to make choices quickly, you should really be using sequential sampling models to analyse the data. Using ANOVA or regression or whatever won't work as well, because the theoretical assumptions that underpin them are not well-matched to your data. In contrast, sequential sampling models were explicitly designed to analyse this specific type of data, and their theoretical assumptions are *extremely* well-matched to the data. Obviously, it's impossible to cover this sort of thing properly, because there are thousands of context-specific models in every field of science. Even so, one thing that I'd like to do in later versions of the book is to give some case studies that are of particular relevance to psychologists, just to give a sense for how psychological theory can be used to do better statistical analysis of psychological data. So, in later versions of the book I'll probably talk about how to analyse response time data, among other things.
