4.3: Independence

Learning Objectives

- Most statistical tests assume that you have a sample of independent observations, meaning that the value of one observation does not affect the value of other observations. Non-independent observations can make your statistical test give too many false positives.

Measurement variables

One of the assumptions of most tests is that the observations are independent of each other. This assumption is violated when the value of one observation tends to be too similar to the values of other observations. For example, let's say you wanted to know whether calico cats had a different mean weight than black cats. You get five calico cats, five black cats, weigh them, and compare the mean weights with a two-sample *t*–test. If the five calico cats are all from one litter, and the five black cats are all from a second litter, then the measurements are not independent. Some cat parents have small offspring, while some have large; so if Josie the calico cat is small, her sisters Valerie and Melody are not independent samples of all calico cats, they are instead also likely to be small. Even if the null hypothesis (that calico and black cats have the same mean weight) is true, your chance of getting a P value less than 0.05 could be much greater than 5%.

A common source of non-independence is that observations are close together in space or time. For example, let's say you wanted to know whether tigers in a zoo were more active in the morning or the evening. As a measure of activity, you put a pedometer on Sally the tiger and count the number of steps she takes in a one-minute period. If you treat the number of steps Sally takes between $10:00a.m.$ and $10:01a.m.$ as one observation, and the number of steps between $10:01a.m.$ and $10:02a.m.$ as a separate observation, these observations are not independent. If Sally is sleeping from $10:00$ to $10:01$, she's probably still sleeping from $10:01$ to $10:02$; if she's pacing back and forth between $10:00$ and $10:01$, she's probably still pacing between $10:01$ and $10:02$. If you take five observations between $10:00$ and $10:05$ and compare them with five observations you take between $3:00$ and $3:05$ with a two-sample t–test, there a good chance you'll get five low-activity measurements in the morning and five high-activity measurements in the afternoon, or vice-versa. This increases your chance of a false positive; if the null hypothesis is true, lack of independence can give you a significant P value much more than 5% of the time.

There are other ways you could get lack of independence in your tiger study. For example, you might put pedometers on four other tigers—Bob, Janet, Ralph, and Loretta—in the same enclosure as Sally, measure the activity of all five of them between $10:00$ and $10:01$, and treat that as five separate observations. However, it may be that when one tiger gets up and starts walking around, the other tigers are likely to follow it around and see what it's doing, while at other times all five tigers are likely to be resting. That would mean that Bob's amount of activity is not independent of Sally's; when Sally is more active, Bob is likely to be more active.

Regression and correlation assume that observations are independent. If one of the measurement variables is time, or if the two variables are measured at different times, the data are often non-independent. For example, if I wanted to know whether I was losing weight, I could weigh my self every day and then do a regression of weight vs. day. However, my weight on one day is very similar to my weight on the next day. Even if the null hypothesis is true that I'm not gaining or losing weight, the non-independence will make the probability of getting a P value less than 0.05 much greater than 5%.

I've put a more extensive discussion of independence on the regression/correlation page.

Nominal variables

Tests of nominal variables (independence or goodness-of-fit) also assume that individual observations are independent of each other. To illustrate this, let's say I want to know whether my statistics class is more boring than my evolution class. I set up a video camera observing the students in one lecture of each class, then count the number of students who yawn at least once. In statistics, 28 students yawn and 15 don't yawn; in evolution, 6 yawn and 50 don't yawn. It seems like there's a significantly ($P = 2.4 \times 10^{-8}$) higher proportion of yawners in the statistics class, but that could be due to chance, because the observations within each class are not independent of each other. Yawning is contagious (so contagious that you're probably yawning right now, aren't you?), which means that if one person near the front of the room in statistics happens to yawn, other people who can see the yawner are likely to yawn as well. So the probability that Ashley in statistics yawns is not independent of whether Sid yawns; once Sid yawns, Ashley will probably yawn as well, and then Megan will yawn, and then Dave will yawn.

Solutions for lack of independence

Unlike non-normality and heteroscedasticity, it is not easy to look at your data and see whether the data are non-independent. You need to understand the biology of your organisms and carefully design your experiment so that the observations will be independent. For your comparison of the weights of calico cats vs. black cats, you should know that cats from the same litter are likely to be similar in weight; you could therefore make sure to sample only one cat from each of many litters. You could also sample multiple cats from each litter, but treat "litter" as a second nominal variable and analyze the data using nested anova. For Sally the tiger, you might know from previous research that bouts of activity or inactivity in tigers last for 5 to 10 minutes, so that you could treat one-minute observations made an hour apart as independent. Or you might know from previous research that the activity of one tiger has no effect on other tigers, so measuring activity of five tigers at the same time would actually be okay. To really see whether students yawn more in my statistics class, I should set up partitions so that students can't see or hear each other yawning while I lecture.

For regression and correlation analyses of data collected over a length of time, there are statistical tests developed for time series. I don't cover them in this handbook; if you need to analyze time series data, find out how other people in your field analyze similar data.
