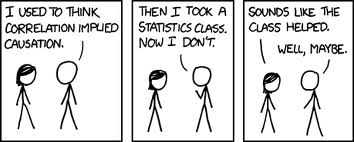
David Youngberg

Acct/Busa/Econ 222—Bethany College

**Lecture 03: Research Design**

1. Scatter diagrams
   1. The first step in any research project is finding data (this sometimes occurs even before you know what you want to investigate).
   2. The second step is determining your approach to the data.
   3. A *scatter diagram* indicates how two (or more, if you are feeling daring) values relate to each other.
   4. This is a good starting point for exploring relations between data. It also may be a good diagram to include in your final paper.
   5. Gapminder ([www.gapminder.org](http://www.gapminder.org)) is an excellent resource to explore relations between different variables. The website employs data from all over the world to various sophisticated scatter plots. The raw data are available in Excel format.
   6. You’ll notice on Gapminder that you can express a variable on a linear (lin) or logarithmic (log) scale.
      1. A linear scale means each unit is some previous unit plus a fixed value. For example: 10; 20; 30; 40; 50; etc.
      2. A logarithmic scale means each unit is some previous unit *times* a fixed value. For example: 10; 100; 1,000; 10,000; etc
      3. For values with a wide range (especially ones skewed right) logarithmic scales are a better visual choice.
2. Of Correlation and Causation
   1. The scatter diagram graphically illustrates if two variables are correlated, or that if one value changes the other will change in a predictable way.
      1. Positively correlated means the values change in the same direction, such as “time studied” and “grade earned.”
      2. Negatively correlated means the values change in the opposite direction, such as “time partying” and “grade earned.”
   2. Of course, *correlation does not mean causation*. Just because it looks like two variables run together doesn’t mean they do. Two other things could be going on:
      1. *Reverse causation*—when the dependent and independent are confused (Greater CO2 emissions cause people to earn more?)
      2. *Confounding variable*—variable which correlates with both independent and dependent variables (Does a greater portion of agricultural workers lower infant mortality?)
   3. At the same time, *correlation is evidence for causation*. The internet is full of people, upon seeing a strongly correlated pair of variables, dismissing any possibility that the two are connected because it doesn’t mean causation. But it should give you reason to pause. And if you can find an explanation *why* one would cause another, you’re in good standing.
      1. Hedge fund manager and blogger James Altucher dismisses evidence of higher earning potential thanks to a college degree by invoking the tired mantra.[[1]](#footnote-1) Of course, there are good reasons to think college causes higher earnings such as credentials, signaling, and skill building.
   4. You need a narrative—some sort of reason—why one thing can cause another. In the comic on the left, it makes sense the male character knows “correlation doesn’t mean causation” is because the statistic course would emphasize such thinking. If he learned, since the class, that North Korea is an oppressive dictatorship which puts disgruntled citizens into death camps, then that’s probably a coincidence. North Korea politics aren’t covered in (most) statistic courses.





1. Probability
   1. Express probability as a decimal or fraction
   2. “Mutually exclusive” means you add the probabilities.
      1. If both events can’t occur at the same time (mutually exclusive), “or” means you add them as well.
      2. Otherwise, add them and then subtract the product of the two.
      3. Example: Probability of drawing either a queen or a heart card isn’t (4/52)(13/52) because you have to include the card that’s both. What’s the actual probability?
   3. “Independent” means you multiply the probabilities.
      1. If the outcome of one event doesn’t affect the likelihood of the other (independent), “and” means you multiply them as well.
      2. Otherwise, multiply by the conditional probability (see below).
      3. Example: Probability of drawing two aces isn’t (4/52)(3/52) because if you draw one, there’s only 3 cards left. What’s the actual probability?
   4. One of the most interesting applications of probability is the Monty Hall paradox.
2. Monty Hall paradox
   1. Suppose before you were three doors. Behind one of these doors is a car. Behind each of the other two is a goat. Pick a door.
   2. Before the door’s opened, suppose someone opens one of the two other doors. This person always opens a door with a goat behind it.
   3. You are then offered to change your selection. Do you? Does it matter?
   4. Here’s the paradox: it doesn’t seem to matter what you do, but it matters a great deal. The strategy here is always the same: you switch.
      1. When I first heard this paradox, I didn’t think it would matter. It took a while before I believed it did.
      2. By the way, the name of this paradox comes from a game show starring Monty Hall. He put contestant after contestant in this scenario.
   5. *Conditional probability* is the probability one event will occur given another event has occurred.
      1. The probability I’m carrying an umbrella on any random day is low. But the probability I’m carrying an umbrella given it rained today is much higher than the nonconditional. The probability I’m carrying an umbrella given that it snowed today is much lower than the nonconditional.
   6. Suppose you always swap:
      1. **If** you pick a car the first time, swapping means you never get the car.
      2. **If** you pick a goat the first time, swapping will get you always get the car.
   7. Suppose you never swap:
      1. **If** you pick a car the first time, staying means you always get the car.
      2. **If** you pick a goat the first time, staying means you never get the car.
   8. Since the probability of picking the car is \_\_\_\_% and the probability of picking the goat is \_\_\_%, swapping is always the better idea.
      1. This is because the host eliminated one of the undesirable choices for you; he gave you information.
      2. Still don’t believe me? Then let’s look at a series of tests.

<http://www.youtube.com/watch?v=o_djTy3G0pg>

1. Testing
   1. Suppose you’re interested which of your employees are on drugs. This is very rare, of course, but if they work with heavy machinery or sensitive information it’s particularly important that they’re clean.
   2. Suppose national surveys reveal that **one out of every 100 people** who work in your industry use drugs.
   3. The problem is that no drug test is 100% accurate. Still, most can get close. Suppose you use a test which is 90% accurate, or **its sensitivity and specificity are 90%**:
      1. *Sensitivity*—describes the ability to detect a positive state. For every 10 drug users, 9 will get a (correct) positive test result. Subtracting sensitivity from one tells you the chance of getting a false negative.
      2. *Specificity*— describes the ability to a detect negative state. For every 10 non-users, 9 will get a (correct) negative test result. Subtracting specificity from one tells you the chance of getting a false positive.
      3. For simplicity, I’m assuming the sensitivity and specificity are the same but it doesn’t have to be that way.
      4. In practice, there is often a trade-off between the two. A test that is made to be very sensitive often means its specificity decreases. For example, a sensitive metal detector will not only pick up more threats (true positives) but also detect more false positives such as cell phones and belt buckles.
   4. Let’s begin by making a *truth table*, or a table of probabilities with all possible outcomes. Here there are two variables—the drug use of the subject and the result of the test—each with two outcomes—“drug user” or “not drug user” and “positive” or “negative.”

|  |  |  |
| --- | --- | --- |
| *Subject* | *Test Result* | *Pr()* |
| User | Positive |  |
| User | Negative |  |
| Not User | Positive |  |
| Not User | Negative |  |

* + 1. For each combination, what is the probability that combination will occur? (Remember how we calculate probabilities!)
  1. We can now determine how likely it is for someone—regardless of actual drug use—to test positive. How do we do that?
  2. Finally, what is the likelihood that a positive test indicates you are a user?

1. Bayes Rule
   1. Before testing, you had a 1% chance of being a user and a 99% of not being a user. Note that, after testing, a positive result increased your likelihood of being a user and decreased your likelihood of not being a user.
   2. This is an example of *Bayes Rule*—a way to describe how one should adjust their beliefs to account for evidence.
   3. To better understand Bayes, let’s first understand some terminology.
      1. *Pr(A)*—the probability event A will occur.
      2. *Pr(A|B)*—the probability event A will occur assuming event B happens.
      3. *~* —this symbol means “not.” Pr(~A) means “the probability event A will not occur.”
      4. Employing the example from Part II, let’s let **U** be “employee is a user” (so ~U is “employee is not a user”) and **P** be “the test came back positive for drugs” (so ~P is “the test is not positive, or came back negative for drugs).
   4. Bayes’ Theorem is as stated:
      1. Note why the denominators are equivalent. If A happens, how likely is B? If A doesn’t happen, how likely is B? Adjusting for the likelihood A happens, and you get the probability that B happens. This last step is similar to how we solved the Simpson Paradox; it’s a weighted average.
   5. We want to know “what is the probability that someone who tested positive for drugs is actually a user?” Or, given that the test is positive, how likely is it that they actually use drugs? Or, what is Pr(U|P)?
      1. Note the result is as before: 8.3%. Only 8.3% of positive results are actually drug users!
   6. Bayes Rule thus tells of two interrelated things:
      1. When you receive new information (such as the results of a drug test), it tells you how much to adjust your estimation of the truth.
      2. It reminds you that because no test is 100% accurate, its results should not be weighed too heavily, especially if it is testing for something rare (since the false positives will overwhelm the true positives).

**Lab Section**

1. Chapter 5
   1. *Permutation*—selecting X units from Y possibilities, where different orders are treated as different possibilities.
      1. N is the number of options.
      2. k is the number of “slots” you can put those options in.
   2. *Combination*—selecting X units from Y possibilities, where different orders are treated as the same possibility.
      1. N is the number of options.
      2. k is the number of “slots” you can put those options in.
   3. There is no “Bayes Rule” function in Excel. To calculate Bayes Rule, just type it out starting with the “equals” sign; don’t forget the parenthesis.
      1. It is sometimes important, however, to know how to format cells to increase decimal points. Right click the cell and select “Format Cells…”.
      2. On the “Number” tab select “Number” under Category.
      3. You can increase or decrease the number of decimal places. You may also add commas (great for permutations and combinations!).
      4. You can also select it as a Percent under Category. It will automatically convert decimal expression to percent expression.
2. Homework
   1. Chapter 5: 1-4, page 65
   2. For a fifth question, answer the following:
      1. Consider a test for a type of cancer which is 99% specific and 95% sensitive. If 0.1% of the population have this type of cancer and you test positive, what the chance that you actually have the cancer?

1. http://www.jamesaltucher.com/2011/01/10-more-reasons-why-parents-should-not-send-their-kids-to-college/ [↑](#footnote-ref-1)