PSC200: Lecture 2

Matthew Blackwell

9/10/2012

Terminology: Data and statistics

What is data? (variables, observations)

- Datasets are in a rectangle.
- What are the rows of the dataset called? observations
  - Various kinds of observations: survey respondents, countries, states, elections, months, or years.
  - We denote the number of observations as n.
- What are the columns of the dataset called? variables.
  - Examples: age, percent vote for Obama, gender.
  - We often use x to refer to a variable. So x_i is the response for observation i on variable x.

Types of variables

- **Nominal**: responses are unordered categories (religion, marital status, name)
- **Ordinal**: responses are ordered categories, distance between values has no intrinsic meaning (Survey responses: strongly agree/agree/disagree/stronly disagree)
- **Interval/continuous**: responses ordered and distance matters (age, years of education, percent vote for Obama).

What is a statistics?

- A **statistic** is a way of reducing the dimensionality or complexity of a group of observations.
- If we were infinitely smart, we could look at a dataset and know everything about it. But we are not that smart so we need simple summaries of the data to tell us what is going on.
- Most of the things we are going to learn in this class will be statistics (mean, median, mode, standard deviation, etc).

The center of the data

- We often want to know something about the **central tendency** of the data, which gives us a sense of the typical observation:
    - The average test score, the most common response to a survey question, etc.

Mean

- The mean is simply the average of the observations:

$$\bar{x} = \frac{\sum x_i}{n}$$

- Example, incomes ($1,000s): 30, 60, 60, 70, 80. $\bar{x} = \frac{30+60+60+70+80}{5} = 60.$
- The number value of the observation must matter. Marital status (1 = Single, 2 = Married, 3 = Divorced, 4 = Widowed) is nominal and so has arbitrary numbers. Thus, the mean doesn't mean much here.
- The mean is highly senstive to **outliers** (which is value very far from the rest of the data):
    - If we add a millionaire to the income data: 30, 60, 60, 70, 80, 1000, then $\bar{x} = \frac{30+60+60+70+80+1000}{6} = 216.67.$

Median

- If we arranged all the observations in ascending (or descending) order, the median is the value in the middle of this list (half of the values are above this number and half are below).
- Income example: (30, 60, 60, 70, 80). Half are below (30, 60) and half are above (70, 80).
- For even numbers (no exact middle), we take the average of the two middle values. Thus, income with the millionaire: (30, 60, 60, 70, 80, 1000). We have two middle values (60 and 70) so we take their average: $\frac{60+70}{2} = 65.$
- The median is more resistant to changing by outliers: it only goes to 65 when we add the millionaire.

Mode

- The most common value among the observations.
    - How many of you are seniors/juniors/sophmores?
    - Frequency distribution.
    - Is this a nominal, ordinal, or continuous variable?
    - Which one is the mode?

Measuring spread

Range

- The range is the difference between the largest observed value and the smallest observed value.
- The range only tells us about the extremes of the data: think about the difference in age between the youngest student and the oldest student on campus.

Standard deviation

- The standard deviation or s is a measure of how spread out the data is.
- First, let's look a deviation: $(x_i - \bar{x})$. We want some idea of the "average" deviation, but it turns out that if we average the deviations, we always get 0, because the positive deviation cancel out the negative ones.
- Instead, we will have to use the squared deviations and then use a square root to get back the right units. Here is the formula:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

- You can think of the SD as the square root of the average squared deviations.
- Show examples of data with higher and lower SD.

Properties of the SD

- The SD is never negative. When would it be 0? When there is no variation/no deviations.
- The SD has the same units as the original variable.
- The SD gives us an indication of where most of the data fall.
- Empirical rule:
  - Around 68% of the data falls bewteen $\bar{x} - SD$ and $\bar{x} + SD$.
  - Around 95% of the data falls between $\bar{x} - 2 * SD$ and $\bar{x} - 2 * SD$.
  - Around 99% of the data falls between $\bar{x} - 3 * SD$ and $\bar{x} - 3 * SD$.

PSC200: Lecture 3

Matthew Blackwell

9/12/2012

Presidential approval data

- Have students download the dataset.
- Show approval dataset.
- What is the unit of observation of this data? poll result

Visualizing data

What is a Histogram

- Break the range of the data into equal sized bins.
- Draw a rectangle to have the height of the number of observations in that bin.

Histograms in R

- Basic use of `hist()` function.
- How to customize your histogram.
  - `breaks =` number of bins, no quotes
  - `col =` name of color, in quotes (see color guide)
  - `main =` main title, in quotes
  - `xlab =` x-axis label, in quotes

What is a boxplot?

- A way to summarize the information of histogram.
- The box region represents the middle 50% of the data.
- The lines extend out to include roughly 95% of the data.
- Any points outside the line are **outliers**.

Boxplots in R

- Shockingly, use the `boxplot()` command.

Summarizing data

means, medians, modes

- We can use the `summary()` function, which will give us a variety of measures such as the median, mean, minimum, maximum.

- We can also calculate the mean using the `mean()` function.

- We can calculate the median using the `median()` function.

standard deviation

- Calculate the standard deviation with `sd()`

Classwork

- First, I want you to subset the data to only be George Bush.

- Find the mean approval for Bush.

- Find the range that roughly contains 95% of his polls

Student evaluations

- Pick one volunteer

- Hand out evals.

PSC200: Lecture 4

Matthew Blackwell

9/17/2012

Z-scores

How big is big?

- We need to know if an observation is large or small relative to the rest of the distribution of data.
- But, as we know, different variables have different units, means, and amounts of spread. So, large in one dataset might be close the average in another.
- We need a set of common units that we can use in any dataset to tell if a value is far from the mean in either direction.

Calculating the Z-scores

- It is relatively easy to calculate the z-score for a given observation. Simply subtract the mean from the obsevation and divide by the standard deviation (s):

$$z = \frac{x_i - \bar{x}}{s}$$

- The z-score is no longer in the units of the variable (such as dollars or inches). It is now in terms of standard deviations. So, if an observation has a z-score of -2, it is 2 SDs below the mean.

Examples of Z-scores

- Jane works at the Starbucks on campus, where she has a tip jar. She's been keeping track of her daily tips and has calculated that her mean daily tip is $1.56 with a standard deviation of 20 cents. Today, she received $1.86 in tips. Calculate the z-score.

$$z = \frac{186 - 156}{20} = \frac{30}{20} = 1.5SD$$

- What about a day with $0.56?

$$z = \frac{56 - 156}{20} = \frac{-100}{20} = -5$$

Normal distribution

What is a distribution?

- A probability distribution is a function that describes the probability of different events happening. Like a histogram.
- The distribution tells us where there is more and less probability of an observation falling. If we know the distribution of a variable, we know where we should expect observations to fall.
- For example, we might have a distribution of waiting times for the bus (draw skewed distribution). It tells us that most buses come in very little time, but a few buses come quite late.
- There are many different kinds of distributions and they all have special properties: binomial, poisson, exponential, gamma, etc. But we are going to focus on one.

The Normal distribution

- (Draw normal curve) The normal distribution should be fairly familiar to you. It is sometimes called a "bell-curve" distribution (we won't call it that).
- The Normal distribution has three key properties:
  - It is **unimodel**: it only has one peak at the mean.
  - It is **symmetric** around the mean.
  - It is **everywhere positive**: all values from negative infinity to positive infinity have some positive probability of happening. But there the chances of something more than 2 or 3 SDs away from the mean is very unlikely.
- There are two values that can affect the shape of the Normal distribution: its mean and its standard deviation. Once we know these two values, we know everything about the Normal distribution for a given variable.
- There is one special kind of Normal distribution: the standard normal distribtuion. It has mean 0 and standard devation 1. The z-scores we calculated above are on the same scale as the standard normal. Note that with a standard normal, a 1 means 1 SD above the mean; -2 means 2 SD below the mean, etc: same as z-scores.

Normal tables

- We often need to know how likely some range of values are in a Normal distribution. We can figure this out.
  - First, we take our variable and standardize it to z-scores.
  - Then, we can take that z-score and see how likely it would be to get value or greater in a standard normal distribution. To figure that out, we can look in the back of our book.
- How do we get the probability of values less than some z?
  - It is the same the probability of being greater than $-z$, due to the symmetry of the Normal.

- How do we get the probability of values between two numbers?
  - $mean + 2 * SD, mean - 2 * SD$
  - Figure out how much is below and above the region (0.0228 above, 0.0228 below when $z = 2$). So the total probability of being outside this region is 0.0456.
  - To get the probability of being inside the region, we need to understand that the probabiliy of the entire line (from $-\infty$ to ∞) is 1. Thus, if we break up the line into to two parts, inside the region (A) and outside the region (B), the probability of these two events must add up to 1. Pr(A) + Pr(B) = 1. Since we know that Pr(B) = 0.0456, we can use that to find Pr(A) = 1 - 0.0456 = 0.954.

Central Limit Theorem

- The reason the Normal distribution is so important is that it has a special place in statistics. The CLT is deceptively simple, but very powerful:

  The sums and means of measurements tend to have an approximately normal distribution. This approximation gets more normal as more measurements are added to the sum or mean.

- This is huge: no matter what the distribution of the original variable, its mean or sum tends to be normally distributed. That means if we have some variable that is just 0/1 (like gender), it couldn't possibly be normal. But if we take the mean of a bunch of gender measurements, that mean will follow an approximate normal distribution.
- Let's say we were to take a survey of people and ask them if they are going to vote to re-elect President Obama. The number of people who responded that they will vote for Obama will be normally distributed because it is the sum of many small measurements—each individual's choice to vote or not. Thus, if we were to collect many different samples and take the number of Obama voters in each sample and then plot a histogram of all of those numbers, it would look like a normal distribution.
- Height is also normally distributed because it is the sum of many small measurements—the length of your leg bones, spine, skull, your nutrition, and so on.

PSC200: Lecture 5

Matthew Blackwell

9/19/2012

Central Limit Theorem (Demo)

Fulton County Data

- This is data from Fulton County, GA. It has every registered voter and whether or not they voted.
- Notice that the turnout variable is 0/1. No way it could be Normally distributed, right?
- I wrote a little script that will take lots of different samples of Fulton County and calculate the proportion of people that turned out. Then, we'll plot those sample means in a histogram to see what it looks like.
- Even early on you can see that the CLT is working.

Florida 2000 election data

- The butterfly ballot in Palm Beach County, Florida caused lots of confusion for people. The way the ballot was set up, it alledgely confused some Al Gore voters into voting for Pat Buchanan.
- We're going to check if Palm Beach really had a very different pattern of voting than the other counties to see if the butterfly ballot really caused any problems.

Looking at the data

- We have the election day and absentee returns for Pat Buchanan, along with his share of the vote on election day and in the absentee ballots.

Buchanan election day fraction of the vote

- First, let's plot a histogram of the Buchanan's election day totals.
- Next, Let's calculate z-scores for each row in the data using the formula we used last week.
- Let's try it with `scale()` now.
- Note that Palm Beach doesn't exactly have a huge z-score for election day fraction. Does that mean we can exonerate the butterfly ballot?

Buchancan absentee voters

- Not so fast. Instead of looking at just the election day totals, we can use the data more effectively. It turns out that absentee voters in Palm Beach County did not use the butterfly ballot, only the voters on election day.
- So, how might that help us? We might take a look at the difference between election day voters and absentee voters to see if there was a big increase for Pat Buchanan on election day. If there was a big jump, then the butterfly ballot could be the culprit.
- First, create a new variable called `eday.inc` that is the difference between Buchanan's election day and absentee vote fractions.
- Next, create a z-score for that variable.
- Is Palm Beach unusual? How likley was this increase, roughly?

Other explanations

- We would like to say if the butterfly ballot **caused** the unusual increase in the vote for Buchanan, but can we? Are there any alternative explanations?

PSC200: Lecture 6

Matthew Blackwell

9/24/2012

Quick Review

- What have we learned so far: how to summarize data (find the mean, median, mode, standard deviation, z-score) and describe distributions (the mean and standard deviation of the normal distribution).
- Describing the data we have is useful, but what we would like to do is learn about data that we don't have.
- Every variable you can think of has a distribution. We are going to learn about those distributions using samples.

Samples

Terminology

- We have been implicitly talking about samples the entire class so far, but we want to define terms a bit to speak more concretely about their usefulness for statistics.
- The **population** is the set of people/countries/observations that we are trying to make inferences about.
  - Every student at UofR is an example of a population.
- The **population mean** is the average value of a variable in the population. We will represent this as μ.
  - The average height of students at UofR
- The **population standard deviation** (σ) is the standard deviation of a variable in the population.
  - The average height of students at UofR
- A **sample** is a subset of the population that we take in some way. We often take a **random sample**, where we select observations randomly (such as drawing names out of a hat).
  - This class is a sample of UofR students. Is it random?
- The **sample mean** is the average value of a variable in a subset of the population. We call this $\bar{x}$
  - The average height of people in this class.

- The **sample standard deviation** is the standard deviation of a variable in a subset of the population. We call this s.
  - The standard deviation of heights in this class.
- A **census** is a sample that includes every single observation in the population.
  - If we were to ask every student at UofR their height.
- In general, it is cost-prohibitive to do a census, so almost always work with samples. Our goal is to learn about the population parameters (mean μ and standard deviation σ) from our sample. The Central Limit Theorem actually provides the connection between our sample mean and the population mean.

Properties of samples

- We want to use samples to learn the truth about some population, but first it is useful to know the properties of samples. That is, if we know what the population distribution looks like, what should the sample look like? This will help us next week when we turn the questions around.
- Remember our mantra that **the sums and means of random variables tend to be normally distributed**. Thus, no matter what the original distribution of the data, we know that the mean or sum of a sample from that data will have a normal distribution.
- Not only this, but we actually know **which** normal distribution that the mean and sum follow.
  - The mean/expected value of the distribution of the mean is μ.
  - The standard error of the distribution of the mean is $\frac{\sigma}{\sqrt{n}}$.
  - The mean/expected value of the distribution of the sum is $n\mu$.
  - The standard error of the distribution of the sum is $\sqrt{n}\sigma$.
- Thus, the sample mean follows a normal distribution that is centered around the true, population mean. So our best guess about the population mean is the sample mean.

Standard deviation of dummy variables

- We know how to calculate the standard deviation for variables using the standard deviation formula, but it turns out that we don't need that much when we have a **dummy variable**, which is what we call a variable that can either be 0 or 1.
- The proportion of 1's is also the mean of a dummy variable. We sometimes refer to this as p.
- For a dummy variable, we only need to know the proportion of 1's, which we refer to as p (note that this is also the mean of the dummy variable). Thus, we can use this shortcut formula when we have a dummy variable:

$$s = \sqrt{p(1-p)}$$

- Thus, we only need to know the proportion of 1s to be able to know the mean and standard deviation of the variable.

Example

- Let's say we fill a bathtub with 1000 red marbles and 1000 blue marbles. Our variable might be $x_i = 1$ for red marbles and $x_i = 0$ for blue marbles. What kind of variable is this?
- First, what is the true proportion of red marbles? (mean of x = μ = 0.5)
- If we were take a sample of 50 marbles from the bathtub, what would distribution would the mean of the sample $\bar{x}$ be? (Normal)
- What percent of the sample would we expect to be red marbles? (50%)
- How many of the marbles would you expect to be red? (25)
- What is the standard deviation of this variable? ($\sqrt{0.5 \times (1 - 0.5)} = 0.5$)
- What is the standard error of the sample mean/proportion? ($\frac{0.5}{\sqrt{50}} = \frac{0.5}{7.07} = 0.071$)
- How does this help us? It tells us where we would expect to see most of the sample proportions fall. We know that 95% of them should fall within 2 SEs of the mean: $0.5 \pm 2 \times 0.071 = [0.429, 0.571]$

PSC200: Lecture 7

Matthew Blackwell

9/26/2012

Fulton County Revisited

- Let's review where we have been: we want to learn about the properties of samples from known populations.

- Last week, we looked at the Fulton County data and tried to show that when we take a sample of 100 registered voters in Fulton County and take the proportion of that sampled that turned out to vote, we see that the proportion followed a Normal distribution.

- Now from last class, we know exactly what kind of Normal distribution we should expect from this example. We know that in the population of Fulton County registered voters, the proportion of those who turned out is 0.44. Remember we call this μ or p.

- First, since we know the population mean, we need to calculate the population SD, σ. We can use the dummy variable shortcut: $\sigma = \sqrt{p \times (1-p)} = \sqrt{0.44 \times 0.56} = 0.496$.

- Now we want to know where we expect the sample mean to be, on average. We know that it should be around the true mean, 0.44.

- How much spread will there be around the true mean? To know that, we need to calculate the standard error (SE). We can just apply the formula: $\frac{\sigma}{\sqrt{n}} = \frac{0.496}{\sqrt{100}} = 0.0496$

- Let's actually go through the process of drawing a bunch of random samples, calculating the sample mean in each, plotting the histogram of the sample means, and calculate the average and SE of those sample proportions.

Box models

- How do we figure out what the population mean (μ) and population standard devation (σ)? We could use the formulas from last week on the entire population, but that might be unwieldy.

- When there are only a few different responses, we can use something called a box model to summarize the population distribution. The basic idea is that we put each population value in the box in the same proportion as in the population:

$$\underline{|0\ 1\ 2\ 2\ 2|}$$

- In this example, the population would be $\frac{1}{5}$ 0's, $\frac{1}{5}$ 1's, and $\frac{3}{5}$ 2's.

- What is nice aobut this is that the mean and standard deviation of the box model is the same as the mean and standard deviation of the population.
- Thus, in the above example, the population mean is $\frac{0+1+2+2+2}{5} = 1.4$.
- Let's calculate the SD:

$$\sqrt{\frac{(0-1.4)^2+(1-1.4)^2+(2-1.4)^2+(2-1.4)^2+(2-1.4)^2+(2-1.4)^2}{5-1}} =$$

$$\frac{1.96+0.16+0.36+0.36+0.36}{4} =$$

$$\frac{3.2}{4} = 0.8$$

Exit polls

- This is a dataset of states in the 2000 presidential election, reporting the results of an exit poll and the actual results from the election.
- Here we have the result of a sample, along with the true population values. First, let's create a variable that is the true proportion that voted for Bush in 2000 in each state.
- Next, we can use R to figure out what the population SD is for each state. How do we do that?
- Now that we have the population SD and the sample size n, we can calculate the standard error for each of our polls. We can use the formula straight in R.
- Now that we have the sample proportion, the population proportion, and the standard error, we can use the z-score of the proportion to see if the exit poll results we see are unusual compared to the true mean. What should expect? We should expect a nice normal distribution. Some above the their true means and some below their true means, with not many above 2 or below -2.
- What do we end up seeing? Almost all of them are negative and some are quite large. What might this tell us about these exit polls?

PSC200: Lecture 8

Matthew Blackwell

10/1/2012

The probability of sample means and sums

- Let's expand on what we covered a bit last week. Last week, we learned that if we knew the population mean and the population standard deviation, then we would also know the distribution of the sample mean or the sample sum in a sample from that population. Let try to go over an example of why we might want to know that.

Suppose that Mac works at a pizza place and he makes an average of 20 pizzas in an hour, with a standard deviation of 2.5 pizzas. Mac's boss randomly chooses 100 hours to watch Mac make pizza to check his productivity. During these 100-hour productivity checks, Mac's boss fires anyone that makes fewer than 1950 pizzas. What is the probability that Mac will be fired after his productivity review?

- We know that the population mean here is $\mu = 20$ and the population SD is $\sigma = 2.5$. From last week we know that we expect the sample sum to have a normal distribution with men $n \times \mu = 100 \times 20 = 2000$ and $SE = \sqrt{n} \times \sigma = \sqrt{100} \times 2.5 = 25$. Thus, we know that the distribution of the sample sum is distributed normally with a center at 2000 with an SE of 25, which means that roughly 95% of the distribution falls between $\mu \pm 2 \times SE = (1950, 2050)$.
- We need to know what the probability is of Mac being fired. We know he's going to be fired when the sum is less than 1950. We have a rough idea of how likely that is, but if we convert it to a z-score, it is much easier to figure out the exact probability of being less than 1950 in the distribution we just drew.
- We can just apply the z transformation formula, slightly adjusted for the sample sum, which we refer to as y here:

$$\frac{y - n\mu}{SE} = \frac{1950 - 2000}{25} = -2$$

We know that if we want the probability Pr$(Z < -2)$, we can (by the symmetry of the Normal distribution), just look up the probability Pr$(Z > 2)$, which is in the back of the book. Remember, we can do this because there is just as much area in the upper tail of the normal as the lower tail of the normal.

- We could look up 2 in this case, but we can also use what we know about the Normal. That is, we know that 95% of the data is between -2 and 2 for a standard normal, so that means each tail much have 2.5% of the data. Thus, the probability of Mac being fired is 0.025.

Learning about populations from samples

- Let's say we get a completely new sample from a population we know nothing about. What's our best guess about where the true mean of the population is? This is a much more practical direction since we rarely, if ever, know what the true population looks like. Plus, if we did know that, why would we want to take samples from that population?

- Since we know from last week that the expected value of a sample mean is just the population mean, we can say that our best guess as to population mean is the sample mean. This gives us what we call a "point estimate" which is really useful.

- We would like to say something more, though. We would like to express our uncertainty about where the population mean might be. Thus, we will calculate a range of values that are plausible for where the true mean might be. In fact, we will be calculating a *confidence interval*, which is always prefaced by a number, such as a *95% confidence interval*.

- We are turning our ship, but most of the mechanics of what we are doing will stay the same. We will still have averages, standard errors, the normal distribution, etc.

Confidence intervals

- A confidence interval is simply: point estimate $\pm$ margin of error. The size of the margin of error is related to the confidence level of the CI.

- Here is exact formula for an $100 \times (1 - \alpha)$% confidence interval for the sample mean:

$$CI_{(1-\alpha)} = \bar{x} \pm z_{\frac{\alpha}{2}} SE$$

Where $z_{\frac{\alpha}{2}}$ is the z-score at which the Normal table would put $\alpha/2$ above that value. Thus, for a 95% confidence interval, we have $\alpha = 0.05$, so that we need to find the z-score that corresponds to $0.05/2 = 0.025$. We can look this up on the normal table: 1.96.

- Thus, the 95% confidence interval is:

$$CI_{95} = \bar{x} \pm 1.96 SE$$

We can also calculate the 90 or 80 percent confidence intervals:

$$CI_{90} = \bar{x} \pm 1.64 SE$$

$$CI_{80} = \bar{x} \pm 1.28 SE$$

- But, wait, you say, how do we know the SE if we don't know anything about the population? Don't we need to know σ to calculate the SE?

- It turns out that in large samples we can use the *sample standard deviation* (s) in place of the population standard deviation when calculating the the SE:

$$SE = \frac{s}{\sqrt{n}} \quad \text{(sample mean)}$$

$$SE = \sqrt{n}s \quad \text{(sample sum)}$$

Opinion polls

- Let's go through an example of how to use confidence intervals. Suppose Ines wants to enter a primary for a local office with 100,000 registered voters, but only wants to do so if she has a good chance of winning. So she hires a polling firm, which takes a simple random sample of 2,500 voters. In the sample, 1,328 support Ines.
- Thus, the sample proportion in this case $\bar{x} = \frac{1328}{2500} = 0.53$.
- Now, Ines is very excited about this and thinks she has won the race already. What is the problem here? There is some probability that she got this proportion just by random chance. It could be that, in the population, she has less than 0.5 of the votes, but this happened to be a strange sample. How can we figure out what a plausible range of values is for the true proportion? That's right, a confidence interval.
- If we want to calculate a 95% confidence interval, we already have two of the things we need: $\bar{x} = 0.53$ and $z_{\frac{\alpha}{2}} = 1.96$. All that's left is to calculate the standard error.
- Again, remember that we don't have the population SD in this case, but we can just use the sample SD in its place. And, since voting for Ines or not is a dummy variable, we can use the dummary variable shortcut: $s = \sqrt{p(1-p)} = \sqrt{0.53 \cdot (0.47)} \approx 0.5$.
- Thus, the standard error is $SE = \frac{0.5}{\sqrt{2500}} = \frac{0.5}{50} = 0.01$, or one percentage point. Thus, we can now fill in the confidence interval:

$$CI_{95} = \bar{x} \pm 1.96SE = 0.53 \pm 1.96 \cdot 0.01 = 0.53 \pm 0.0196 \approx [0.51, 0.55]$$

- Thus, we can have 95% confidence that Ines will win the election!
- Next time we'll talk more about exactly how to interpret a confidence interval.

PSC200: Lecture 9

Matthew Blackwell

10/3/2012

Confidence interval for the mean

Review of CIs

- Let's review confidence intervals. Why confidence intervals? Remember:

$$\text{sample mean} = \text{population mean} + \text{chance error}$$

Thus, we know that our sample mean is not *exactly* the same as our population mean and we want to give a plausible range of values around the sample mean where the true mean could be.

- Last week we learned that the sample mean has a normal distribution centered around the true mean (μ) with a SE equal to $\sigma/\sqrt{n}$. Given that, we know that roughly 95% of the time the chance errors above will be between -2 SE and 2 SE. That is, we know that chance error greater than 2 standard errors from the population mean (in either direction) are rare. Thus, we know that roughly 95% of the time, the sample mean will be within roughly 2 SEs of the true mean. But it could be above the true mean or below the true mean, we don't know. The confidence interval quantifies this procedure.
- So, once we have our sample mean, we can say, "Well, I know that this sample mean is unlikely to be more than 2 SEs above the population mean and it is also unlikely to be 2 SEs below the population mean". Thus, we can take the interval that is 2 SEs away from the sample mean as a reasonable confidence interval, given our data.
- If we wanted to be more or less confident, we could take a higher or lower number of SEs.
- Could we increase the SEs in the confidence interval to get an 100% confidence interval? It turns out that we cannot, since there is always a possibility of an extremely large chance error.

Example of CI for the mean

Suppose the mayor of Rochester comes to us and asks us to find out the average income for the households in Rochester. He gives us enough money to randomly sample 400 households and we calculate a mean income of $27,000 with a standard deviation of $14,000. The mayor, having taken PSC200 in the past, is unsatisfied with point estimates and asks for a 95% confidence interval around this mean.

- What do we need? First we need to recall the formula for the confidence interval:

$$CI_{95} = \bar{x} \pm 1.96SE$$

- So, in this case, we already have $\bar{x}$ and we already know that $z_{\frac{\alpha}{2}} = 1.96$. So all we need is the SE.

- Since this is a sample mean, we know that the $SE = \frac{\sigma}{\sqrt{n}}$, but all we have is the sample standard deviation. Remember, though, that in large samples (roughly over 100), we can substitute s for σ and proceed as if nothing happened. Thus, we have

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{14000}{\sqrt{400}} = \frac{14000}{20} = 700$$

- Now we can calculate our condifence interval:

$$CI_{95} = \bar{x} \pm 1.96SE \approx \$27,000 \pm 2 \cdot \$700 = (\$25600, \$28400)$$

Interpretation of confidence intervals

What does 95% confident mean?

- It's strange to say we are 95% confident about some interval containing the truth, since we know that the the truth is either in our interval or it isn't. What does this confidence interval actually mean?

- A 95% confidence interval should contain the true mean/proportion 95% of the time. That is, if we were to draw repeated samples and do our whole procedure over and over again, 95 times out of 100, our 95% confidence interval would contain the true mean/proportion.

- Thus, you might say that the confidence interval gives us a 95% chance of being "right". Remember, though the **chances are in the sampling procedure, not the parameter**. So, we do not say "there is a 95% chance that the population mean is in this interval" because the population mean is fixed at some value. It's the confidence interval that moves from sample to sample. Thus, we talk about the probability that the confidence interval will "cover" or contain the truth. (Draw a couple of CIs).

Example with Fulton County

- This is the Fulton County data (again) and this time we are going to look at turnout, but we are going to look at a confidence interval for the true proportion of people who turned out to vote.

- In class, we looked at a script that would draw repeated samples and

- What happens when we increase the size of our samples? What should happen to the size of the confidence intervals?

How big of a sample size do I need?

- Let's look at the formula for the CI again, this time, let's plug in the value for the standard error:

$$CI_{95} = \bar{x} \pm 1.96SE = \bar{x} \pm \frac{s}{\sqrt{n}}$$

- One thing that jumps out at us is that the size of confidence interval is inversely proportional to the sample size. That is, if we increase the sample size, we will decrease the size of the margin of error and thus the confidence interval. We saw this in the Fulton County demonstration.
- Often we might want to know how to ensure we get a certain width confidence interval. We could change the level of the CI, but suppose we wanted to keep the level constant. Well, we can write down the formula for the margin of error:

$$M = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

- Since we can't change σ and we don't want to change the level (which would change z), we can only adjust n, the sample size. We can rearrange these to solve for n:

$$n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{M} \right)^2$$

- Now, we can plug in the z we need for our confidence level, the population SD, the desired margin of error, and get back the sample size we would need to get that margin of error.
- Note that we would have to know what the population standard deviation is in this case, but we haven't even drawn a sample yet. We have no idea! Well, one thing we can do is provide a reasonable guess. We know that most of the data is within two SDs of the mean, so if we can think of the rough upper and lower bounds on the range of the data, we can simply divide that range by 4 to get a rough guess of the SD.

Example

Let's say we wanted to know many UR students we would have to sample to get an estimate of the average height with a 1-inch margin of error on a 90% interval. What sample size would we need?

- First, we know that for a 90% confidence interval, we have $z_{0.05} = 1.64$. We also know that $M = 1$. Now we just need to know σ.
- Let's think about the range of values we think heights on campus take. Let's say that most people are between 4'10" (58 inches) and 6'8" (80 inches). Given that we think that this represents rare events, we probability think these are 2 SDs away from the mean in either direction. That mean the difference between 80 inches and 58 inches (22 inches) is a rough guess of 4 SDs. Dividing that by 4 gives us our guess of the SD: 22/4 = 5.5 inches.
- Now we can plug this in and see what our sample size would have to be:

$$n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{M} \right)^2 = \left(\frac{1.64 \cdot 5.5}{1} \right)^2 \approx \left(\frac{9}{1} \right)^2 = 81$$

Thus, we would need to take a sample of at least 81 students to get a margin of error that is 1 inch or smaller.

PSC200: Lecture 9

Matthew Blackwell

10/3/2012

An example start

- Let's suppose there are two people that are arguing about the upcoming presidential election. One person, Nully, claims that the true percentage of people that will vote for Obama is exactly 50%. That is, he says there will be a tie. His friend Alto think there's no way it is a tie and claims that it must be that either Romney or Obama are ahead: he claims that the true proportion of people voting for Obama is not 50%.
- To settle this, they hire a company to sample of size 500 from the population of likely voters and see what proportion of them will vote for Obama. They find that (among people that are going to vote for Romney or Obama) 52% say they will vote for Obama and the pollster reports a sample standard deviation of 15.3 percentage points.
- Nully says hah! It is clear that the propotion is really 0.5 and this is just random error, there is a standard deviation of 0.153 of course!
- Alto says, hold on, but we need to look at the SE, not the SD. Because the SE tells us how far the sample proportion should be from the true proportion. In this case, he calculates the $SE = \frac{s}{\sqrt{n}} = \frac{0.153}{\sqrt{500}} \approx 0.7$. Alto then says, hah! you thought this could be due to random chance, but if the true were 0.5, then the value we got would be 3 SEs above that proportion:

$$\frac{52 - 50}{0.7} \approx 3$$

- Why is this important? Nully asks. Alto responds by telling him that if his hypothesis is true and the true value is 0.5, then the sample we got was very unusual! It is unlikely to be due to randome chance. Something else is going on. It is either that we got an extremely abnormal sample or we should reconsider Null's hypothesis.

Hypotheses

- The **null hypothesis** is the statement that a population parameter (like the population mean) takes a certain value. This corresponds to the idea that any observed difference between the observed value and the hypothesized value is due to chance. (We sometimes write this as $H_0 : \mu = a$, where a is the hypothesized value.)

- In the above example the null hypothesis was the true proportion of likely voters that were going to vote for Obama was 0.5.
- The **alternative hypothesis** is another statement about the population. It states that the the true value of the parameters is not the hypothesized value as in the null hypothesis. Thus, the alternative hypothesis states that the difference between the observed value and the null hypothesis is real and not just due to random chance. (We sometimes write this as $H_0 : \mu \neq a$, where a is the hypothesized value.)
- In the above example, the alternative hypothesis was that true proportion was **not** 0.5.

Test statistics and significance levels

- A **test statistic** is a measure of the difference between the observed data and what we expected from the null hypothesis.
- The test statistic gets bigger (in absolute value) as the observed data looks unusual compared the null hypothesis. So big test statistics cast doubt on the null hypothesis. There are many different kinds of tests, but we are going to focus on a few in this class.
- A **z-test** is the number of SEs away an observed mean or sum is from the expected value (where the expected value comes from the null hypothesis). Here is the formula:

$$z = \frac{\text{observed} \; - \; \text{expected}}{SE} = \frac{\bar{x} - \mu_0}{SE}$$

- In this case we are looking at a z-test for the mean. Here, $\bar{x}$ is the mean that we calculated in the sample and SE is just the standard error we calculated from the sample. Th μ_0 is the value from the null hypothesis. (From the example above, we would have $\mu_0 = 0.5$.)
- A **p-value** of a test is the chance of getting a test statistic this big (or bigger), if the null hypothesis is right. The p-value is **not** the chance of the null hypothesis being right.
- A result is **statistically significant** if the p-value is less than 0.05. This is somewhat arbitrary, but it is what most disciplines agree on.

Confidence intervals and tests

- A conclusion from a $(1 - \alpha)$% confidence interval are the same as conclusions from a hypothesis test at the α%-level.
- What does that mean? It means that any value outside of a $(1 - \alpha)$% confidence interval would be rejected by a α-level test.
- Thus, for instance, if we test the null hypothesis of 50, as in the first example, and we reject that at the 5% level (that is, the p-value is below 0.05), then we also know that 50 would not be in the 95% confidence interval.

PSC200: Lecture1 11

Matthew Blackwell

10/15/2012

Difference in means

- A study finds that freshman at public universities work 10.2 hours a week for pay, on average, with an SD of 8.5 hours. At private universities, the average is 8.1 hours and the SD is 6.9 hours. Each of these results comes from independent samples of size 1000. Could this be due to random chance?

What is the difference in means?

- Up to now, we have been learning about the population mean from a single population. Average height of UR students, for example. Very often we are interested in knowing is two different samples come from the same population with the same mean. For instance, we might want to know if men's and women's heights come from the same distribution.
- We can take two different samples (say men and women) and look if there is a difference in their means, but remember that this difference could be due to random chance. Therefore, we want some way of seeing if the difference in means that we calculate could be due to simple random chance and the Central Limit Theorem.
- Our best guess about the difference in population means is simply the difference in sample means:

$$\bar{x}_{1-2} = \bar{x}_1 - \bar{x}_2$$

- This is the difference between the sample means and it's an estimator for the difference in population means:

$$\mu_{1-2} = \mu_1 - \mu_2$$

- This is all very similar to how we used the sample mean to learn about the population mean. The difference in means is just a different population parameter.
- For the problem above, we can easily calculate the difference in means:

$$\bar{x}_{1-2} = 10.2 - 8.1 = 2.1 \text{ hours}$$

SE for a difference in means

- The difference in sample means comes from samples, so it must vary from sample to sample, just like the mean (remember the Fulton county example). How much will it vary from sample to sample? We can quantify that with the standard error of the difference in means:

$$SE_{1-2} = \sqrt{SE_1^2 + SE_2^2}$$

- Where SE_1 is the standard error from the first sample (say, men) and SE_2 is the standard error from the second sample (say, women). If we had continuous variables, we could plug in the formulas that we know for these standard errors:

$$SE_{1-2} = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$$

- If we were comparing the means of two dummy variables (which would mean we were comparing proportions), we can plug in our dummy variable shortcut:

$$SE_{1-2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- The SE_{1-2} tells us how much we should expect the difference in means to vary from sample to sample. That is, we know that the distribution of sample difference in means, should center on the true difference in means with a normal distribution around it. The SE_{1-2} tells us how spread out that normal distribution should be.

- Let's calculate the SE_{1-2} for the example:

$$\begin{aligned}
SE_{1-2} &= \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}} \\
&= \sqrt{\frac{8.5^2}{1000} + \frac{6.9^2}{1000}} \\
&= \sqrt{\frac{72.25}{1000} + \frac{47.61}{1000}} \\
&= \sqrt{0.07225 + 0.04761} \\
&= \sqrt{0.12} \approx 0.346
\end{aligned}$$

CI for a difference in means

- Now that we know how to calculate the standard error for a difference, we can easily calculate a confidence interval for the true difference:

$$CI_{95} = \bar{x}_{1-2} \pm 1.96 SE_{1-2}$$

- What would it mean if 0 were in this confidence interval? It would mean that the two population means being identical could have happened by random chance. If 0 is not in the confidence interval, then we can be confident that the true difference in means is not 0 and the two samples from different populations.
- Let's calculate a CI for this our example:

$$CI_{95} = \bar{x}_{1-2} \pm 1.96 SE_{1-2} = 2.1 \pm 1.96 \times 0.35 \approx 2.1 \pm 0.7 = [1.4, 2.8]$$

Hypothesis test for a difference in means

- We can take a hypothesis testing approach to differences as well. The value we will often want to test is the null hypothesis that the two means are the same, or that the true difference in means is zero:

$$H_0 : \mu_1 - \mu_2 = 0$$

- It turns out that we can use the z-test as we did last week:

$$z = \frac{\text{observed } - \text{ expected}}{SE_{1-2}} = \frac{\bar{x}_{1-2} - 0}{SE_{1-2}} = \frac{\bar{x}_{1-2}}{SE_{1-2}}$$

- Thus, we can divide the observed difference in means by the standard error for the difference in means and that gives us how many SEs away from 0 our observed difference is, if there was in fact no difference between the groups. We can then use the Normal table to look up the p-value.
- In the above example:

$$z = \frac{\bar{x}_{1-2}}{SE_{1-2}} = \frac{2.1}{0.35} \approx 6$$

We don't even need to look this up in the Normal table, we know the p-value for that is very close to zero.

PSC200: Lecture1 12

Matthew Blackwell

10/17/2012

Get out the vote

- One question political scientists are very intesested in is this: how can we increase the probability that someone will turn out to vote? That is, how do we effectively get out the vote.

- We could take independent samples of people who were contacted by the campaigns and those who weren't and use the difference in means methods above to figure out if campaign contact increases turnout. What is the problem with this? Selection bias: campaigns are more likely to contact some voters rather than others.

- Instead, if we want to get at the causal effect of GOTV attempts, we need to randomly assign messages to voters.

- In the primary election in Michigan in 2006, a group of political scientists conducted a turnout experiment where they randomly sent mailers to voters. The goal was to see if social pressure could induce people to turnout to vote. By randomly sending messages, we can be sure that the voters aren't being selected strategically and that those that receive the messages are similar to those who do not.

- The treatment group (which was 360 households) in this case received a mailer that reported how they and their neighbors voted, using the publicly available voter records. Thus, the treated group could see if they were keeping up with other people in their neighborhood. The control group (which was 1890 households) received no mailers.

- The turnout among the treated group was $\bar{x}_T = 0.37$ and the turnout among the control group was $\bar{x}_C = 0.30$. Obviously there is a difference, but we want to make sure that this difference isn't just due to random chance! What can we do? Use the difference in means methods above to calculate a confidence interval.

- First we calculate the difference in means:

$$\bar{x}_{1-2} = \bar{x}_T - \bar{x}_C = 0.07$$

- Now, we need to calculate the standard error. Since these are proportions and they are based on dummy variables, we can use our dummy variable shorcut:

$$\begin{aligned}
SE_{1-2} &= \sqrt{\frac{\bar{x}_T(1-\bar{x}_T)}{n_T} + \frac{\bar{x}_C(1-\bar{x}_C)}{n_C}} \\
&= \sqrt{\frac{0.37 \cdot 0.63}{360} + \frac{0.30 \cdot 0.70}{1890}} \\
&= \sqrt{\frac{0.233}{360} + \frac{0.21}{1890}} \\
&= \sqrt{0.00064 + 0.00012} \\
&= 0.028
\end{aligned}$$

- With the observed difference and the standard error, we can calculate the confidence interval:

$$\begin{aligned}
CI_{95} &= \bar{x}_{1-2} \pm 1.96 SE_{1-2} \\
&= 0.07 \pm 1.96 \cdot 0.028 \\
&= 0.07 \pm 0.054 \\
&= [0.016, 0.124]
\end{aligned}$$

- Since 0 is not in the confidence interval, we can conclude that this difference is not just due to random chance. This is a likely a real difference due to receiving social pressure to vote. *(Note: we simplified things here a bit. In the real paper, the sample sizes are 100 times bigger. Reducing them just makes the math a little clearer.)*

Gender and views on the Iraq war

- Scott Sigmund Gardner wrote a paper for the *American Political Science Review* called "The Multiple Effects of Causalities on Public Support for War: An Experimental Approach" that looked, in part, on the views of men and women on whether or not the Iraq war was a mistake. You can download the `iraq.RData` dataset from the course webpage. This dataset has a variable for `female` which is 1 for females and 0 for males. In addition, there is a varaible `mistake` which is 1 when the respondent thinks the Iraq war was a mistake and 0 when they think it was not a mistake.

- We can take various approaches to figuring out if there is a difference between men and women on this issue. First, we can just use the basic tools of R (`subset`, `mean`, `sd`) to calculate the difference in means and the SE of that difference. In this case we just apply the formulas from last week to the current data.

- Once we have the difference in means and the standard error, we can calculate a z-test using the formula from the last time:

$$z = \frac{\bar{x}_{1-2}}{SE_{1-2}}$$

- Once we have this test statistic, we could use the table in the back of our book to calculate the p-value, which is the probability of getting a test statistic this extreme or more extreme. It turns out we can also use R to get the same quantity, using the `pnorm` function. The `pnorm` function takes in a z value and gives the probability of being *less than that value* on a standard Normal distribution. This is the opposite of what the table in the back of the book gives us. To get them to match, we use the following structure: `pnorm(..., lower.tail = FALSE)`, where `...` is the z-value.

- When we use the `pnorm` function with a z-test, remember that we have to first take the absolute value of the z-value (so that we are getting the correct tail) and then double the result. This is because the p-value is the probability of being more extreme than our observed z-value in either direction. If $z = -1.5$, then we want to know the probability of being less than -1.5 *and* greater than 1.5. Since these two values are the same, we can just find the probability of $z > 1.5$ and double that probability.

- Let's say that instead of performing a hypothesis test, we wanted to calculate a confidence interval. Remember that we need to find the critical values $z_{\alpha/2}$. We have memorized some, like 1.96 for a 95% confidence interval, but we might want other CIs. We can use R to calculate these critical values using the `qnorm` function. Again, R switches around from the back of the book, so we'll use `qnorm(..., lower.tail = FALSE)`, where we will plug in $\alpha/2$ where `...` is.

- We can think of all of this as the "manual" way of doing things. There is a quicker way to do a test and it's called `t.test()`. If we pass two subsets of the data, it will perform a difference in means test along with calculating a confidence interval. You'll notice this is called `t.test()` and not `z.test()`. It turns out that there is a very similar test to the z-test that is called the t-test. It differs from the z-test in small samples and does better there, but in large samples it is exactly the same z-test.

PSC200: Lecture 13

Matthew Blackwell

10/29/2012

Relationships between variables

- A *relationship* exists between one variable and another when we can use one variable to predict the value of another. For instance, we might want to predict the outcome of the election: specifically, what percent of the vote will President Obama receive? We might see if the variables that measure the economic state of country can help us predict Obama's share of the vote.
- Note, though, that just because there is a relationship between two variables does not mean that there is a causal relationship between them. That is, just because we observe a pattern (the sun rises when I wake up) does not mean that this is causal (my waking up does not cause the sun to come up). We will talk more about causality in a few weeks.
- Last time we talked about the relationship between two variables, but it was a certain type of variables: the difference in means between two groups. In general, we were looking at explaining the variation in one variable with another variable that was categorical (men vs. women, private vs. public schools, etc).
- This is pretty limiting, though. We would like to find the relationship between different sorts of variables. In particular, what happens if we want to know how a variable varies with a continuous variable. Not just men versus women, but now we want to know how income varies with years of education.
- The first thing we could do it dichotomize the problem—create two groups (finished college or not) and run a difference in the mean incomes. But this is unsatisfying: what about other changes in education?

Some terminology

- The variable we are trying to explain is called the *dependent variable* or the *outcome variable*, or y. Our goal is to find out how this variable changes when other variable change.
- The variables that are causing the changes in the dependent variable are called *independent variables* or *explanatory variables* or just x. These variables explain the variation in the outcome variable.

Scatterplots

- A scatterplot shows the relationship between two variables. For each observation, we plot the point on the graph that represents that observation's value of each variable.

- Note that for scatterplots we put the variable we are trying to explain (the dependent variable) on the y-axis and the variable that is doing the explaining (the independent variable) on the x-axis.

- For instance, let's say we wanted to know the relationship between unemployment at election time and the share of the vote that goes to the incumbent. We would put unemployment on the x-axis and the vote share on the y-axis. Then for 2008, we had 5.5% unemployment and President Bush won 50.7% of the vote, so we would put a dot at $(5.5, 50.7)$. Then we could do this for all the years and get a sense for the relationship between the economy and the vote.

- Scatterplots are useful for investigating the relationship between two variables. We can see if there is a strong relationship (that one variable accurately predicts the other variable) or if there is just noise.

Bivariate Regression

Using a line to describe relationships

- It is often useful to summarize the relationship between two variables with a line. Remember the equation for a line:

$$y = mx + b$$

- We can estimate the exact line that best summarizes the relationship between y and x, which we write in a slightly different form that above. We right the true *regression line* as:

$$y = \alpha + \beta x$$

- The α and β are the true intercept and slope of the *population regression line*. That is, these are the true parameters of the population, which we will never observe directly. These are similar to the true population mean and the true difference in means from previous weeks. These are all different features of the population we might want to estimate in our sample.

Ordinary Least Squares

- Rmember that we wanted to get estimates of the true mean and true difference in means from our samples? We used the sample mean and the sample difference in means to estimate these quantities. To estimate the paramaters (intercept and slope) of a linear regression, we are going to use a machinery called "ordinary least squares" (OLS).

- First, what does this mean? Basically, we will feed two variables into OLS and it will return an estimate of the intercept and the slope. We refer to these estimates as $\hat{\alpha}$ for the intercept and $\hat{\beta}$ for the slope. We can use these estimated values to get our best guess about y based on x, which we call $\hat{y}$ (y-hat):

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

- The difference between the true value of y and our prediction is called the *residual*.

$$\text{residual } = y_i - \hat{y}$$

- How does OLS pick the slope and intercept of the regression line? It minimizes the sum of the squared residuals. That is, it minimizes this quantity:

$$(y_i - \hat{y})^2$$

- Why does it minimize the squared residuals instead of the just the residuals?

PSC200: Lecture 15

Matthew Blackwell

11/5/2012

The linear regression model

- Last time we talked about investigating the relationships between variables and we talked about how we can summarize many relationships with a line. This helps us find out how one variable changes with another variable.

- Example: we might want to know the relationship between percent turnout (y) in a state and the percent ads that are negative in that state (x).

- Today, we are going to talk about how we evaluate and interpret regression. Typically with a regression we are going to ask three questions. Does a relationship have statistical significance? Does it have causal significance? And does it have substantive signifiance?

- To start, let's talk about what the linear regression model is. We said that it summarizes the relationship between two variables as a line, but we can be more specific.

- The linear regression model describes the mean of y *conditional* on values of x. So, with the exmaple of turnout and negativity, a linear regression shows how the average turnout changes as we change negativity. This generalizes the difference in means we looked at earlier. Now, we can see the difference in the mean of y between any two values of x.

Assumptions

- We have a random sample of the population.

- The mean of y is related to x by a linear equation: $y = \alpha + \beta x$

- The conditional distribution of y at each value of x is Normal.

- The conditional SD σ of y is identical at each x-value.

Statistical Significance

- Statistical significance is the same question we asked with difference in means. We want to know if the slope we estimate could have been due to random chance. Before we do that, it is useful to know how we interpret the slope and the intercept.

Interpretation of the coefficients

- The value of the intercept is the estimated mean of the dependent variable when the independent variable is zero. For example, if we are looking at the relationship between turnout and negative ads, this would be the mean turnout when there are no negative ads.
- The value of the slope has this interpretation: for a one-unit change in x, we estimated a $\hat{\beta}$ change in y. For example, if the slope on the negative ads variable was 3, we would say: "a one percentage point increase in negative ads leads to a 3 percentage point increase in turnout."
- If the true slope β is 0, then there is no relationship between the dependent variable and the independent variable. That would mean that no matter how many negative ads were on TV, the average turnout would always be the same.

Standard Errors and Confidence intervals

- Remember that we have estimated the slope and the intercept from a sample. We would like to get a sense for the uncertainty of these estimates just like we uncertainty for the sample mean, the sample proportion, and the sample difference in means.
- Remember, the SE of an estimate tells us how much the estimates vary from sample to sample. In fact, it turns out that our estimates, $\hat{\beta}$ and $\hat{\alpha}$ are just sums and means of the observed data. This implies that, like the regular sample mean and the difference in means, the slope and intercept estimates are normally distributed.
- We can use the SE to perform a hypothesis test. With differences in means, we took a certain null hypothesis and tried to find evidence against it. Then, the null hypothesis was that there is no difference in means between the two groups ($H_0 : \mu_1 - \mu_2 = 0$). This is the same as saying there is no relationship between the groups and the variable we are investigating.
- With regression, we want to set up a similar null hypothesis: that there is no relationship between the independent variable and the dependent variable. When does that happen? When the slope is zero: $H_0 : \beta = 0$. Thus, the alternative hypothesis is $H_A : \beta \neq 0$.
- With this null hypothesis, we can calculate a p-value that tells us, *if the null hypothesis were true, how likely would a slope this big or bigger be?* To start this, we calculate a t-test (which is exactly the same as a z-test):

$$t = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

- This test is almost exactly the same as the z-test from earlier except in small ($n < 50$) samples. That is, most of our intuitions will stay the same. If we get a t-test of more than 2, we know that this must have been a fairly unusual sample. To be more precise, we can look at the p-value, which is given by our output from R.
- We say that a slope or relationship is **statistically significant** if the p-value is below 0.05. This is usually indicated with stars or astericks in the regression table.
- R can also calculate confidence intervals and we can see if 0 falls in the confidence interval.

Causal Significance

- We'll be brief, because we have more to say about this in a few weeks.
- Causal significance requires you to think of *why* our independent variable might (or might not) *cause* the dependent variable to change.
- By saying that x causes y, we are saying that if we were to force x to take a different value, y would also change.
- Not all relationships are causal: some happen because a third variable causes both the indepdent and dependent variables. For instance, there is a positive relationhip between sales of ice cream cones and accidental drowning deaths. But this is not because ice cream causes drowning, but rather because the weather (heat waves, specifically) cause people to buy ice cream **and** go for more swims (which results in more accidents).
- We need to think of any of these lurking variables that might cause these **spurious relationships**.

Substantive Significance

- Are the estimated effects large in pratice? Are they are enough to care about. For instance, what if we estimated $\hat{\beta} = 0.005$ for the regression of turnout on percent of ads that are negative. That means, for a one percentage point increase in negativity, we would see a 0.005 percentage point increase in turnout. That's fairly small! We can think of how small that is by thinking about the largest possible change in negativity: from 0% negative ads to 100% negative ads. If $\hat{\beta}$ is a one-unit change, then $100 \times \hat{\beta}$ is a 100-unit change. Therefore, going from 0% to 100% would only lead to a 0.5 percentage point increase in the average turnout. Going from no negative ads to all negative ads would only increase turnout from (let's imagine) 35.2% to 35.7%. That's pretty small!
- In the above example, we looked at a large change in x (traversing its entire range) and saw it produced a small change in y. Sometimes, we might look at other sorts of changes.
- We know that the standard deviation (SD) of a variable is similar to the average deviation from the mean. Instead of a one-unit change in x (which might be quite small), we may want to consider a "typical" change in x: a one-SD change.

PSC200: Lecture 15

Matthew Blackwell

11/5/2012

Interpretting Regression Output From R

- Today, we're going to learn how to interpret regression output R. To do so, we are going to look at a fairly small dataset that looks at the relationship between economic performance and how leftist a government is.
- Last week we learned how to produce a scatterplot with `plot()`. We also learned how to store a model using `lm()`. It turns out we can use the `summary()` function on the output from `lm()` to give us a regression table.
- In the regression table, we can see the slope and the intercept, along with the t-tests and p-values for each estimate. We can look at the p-values to determine if our slope or intercept at statistically significant. Our rule is that if the p-value is below 0.05, then we can reject the null hypothesis of no effect and say the relationship is significant.
- Alternatively, we could look at confidence intervals. To do this, we can just use the `confint()` function on the regression output. By default this gives us a 95% confidence interval, but we can change that with the `level` argument (we can set it to 0.99 for instance to get a 99% confidence interval). With this, we can see if 0 is in a 95% confidence interval. If it is not, we can say that the relationship is statistically significant.

Causal Significance

- We'll be brief, because we have more to say about this in a few weeks.
- Causal significance requires you to think of *why* our independent variable might (or might not) *cause* the dependent variable to change.
- By saying that x causes y, we are saying that if we were to force x to take a different value, y would also change.
- Not all relationships are causal: some happen because a third variable causes both the indepdent and dependent variables. For instance, there is a positive relationhip between sales of ice cream cones and accidental drowning deaths. But this is not because ice cream causes drowning, but rather because the weather (heat waves, specifically) cause people to buy ice cream **and** go for more swims (which results in more accidents).
- We need to think of any of these lurking variables that might cause these **spurious relationships**.

Substantive Significance

- Are the estimated effects large in pratice? Are they are enough to care about. For instance, what if we estimated $\hat{\beta} = 0.005$ for the regression of turnout on percent of ads that are negative. That means, for a one percentage point increase in negativity, we would see a 0.005 percentage point increase in turnout. That's fairly small! We can think of how small that is by thinking about the largest possible change in negativity: from 0% negative ads to 100% negative ads. If $\hat{\beta}$ is a one-unit change, then $100 \times \hat{\beta}$ is a 100-unit change. Therefore, going from 0% to 100% would only lead to a 0.5 percentage point increase in the average turnout. Going from no negative ads to all negative ads would only increase turnout from (let's imagine) 35.2% to 35.7%. That's pretty small!

- In the above example, we looked at a large change in x (traversing its entire range) and saw it produced a small change in y. Sometimes, we might look at other sorts of changes.

- We know that the standard deviation (SD) of a variable is similar to the average deviation from the mean. Instead of a one-unit change in x (which might be quite small), we may want to consider a "typical" change in x: a one-SD change.

PSC200: Lecture 17

Matthew Blackwell

11/11/2012

Substantive Significance

- Are the estimated effects large in pratice? Are they are enough to care about. For instance, what if we estimated $\hat{\beta} = 0.005$ for the regression of turnout on percent of ads that are negative. That means, for a one percentage point increase in negativity, we would see a 0.005 percentage point increase in turnout. That's fairly small! We can think of how small that is by thinking about the largest possible change in negativity: from 0% negative ads to 100% negative ads. If $\hat{\beta}$ is a one-unit change, then $100 \times \hat{\beta}$ is a 100-unit change. Therefore, going from 0% to 100% would only lead to a 0.5 percentage point increase in the average turnout. Going from no negative ads to all negative ads would only increase turnout from (let's imagine) 35.2% to 35.7%. That's pretty small!

- In the above example, we looked at a large change in x (traversing its entire range) and saw it produced a small change in y. Sometimes, we might look at other sorts of changes.

- We know that the standard deviation (SD) of a variable is similar to the average deviation from the mean. Instead of a one-unit change in x (which might be quite small), we may want to consider a "typical" change in x: a one-SD change.

Adding a dummy variable to a regression

- Relationship between education and income: it might be the case that there is a relationship but we would also imagine that women, on average, have lower incomes than men at any level of education. How do see if this is the case? It turns out we can add a dummy variable for gender to out regression.

$$y = \alpha + \beta_1 x + \beta_2 d$$

- We can formulate predicted values, just as before:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x + \hat{\beta}_2 d$$

- We also estimate these values using the same machinery: OLS, which is going to minimize the sum of the squared residuals:

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- How does this affect the lines that we have? Well, now we essentially have two regression lines. One when $d = 1$ and another when $d = 0$.
- Note that we can calcuate each of these regression lines by plugging the d value in. When $d = 0$ (for men), the regression line is just the familiar formula with α as the intercept and β_1 as the slope.

$$(d = 0) : y = \alpha + \beta_1 x + \beta_2 \times 0 = \alpha + \beta_1 x$$

- When $d = 1$ (for women), we have an extra β_2 that increases our line at every value of x. We can think of this as an increase to the intercept:

$$(d = 1) : y = \alpha + \beta_1 x + \beta_2 \times 1 = \alpha + \beta_1 x + \beta_2 = (\alpha + \beta_2) + \beta_1 x$$

- Thus, a dummy variable in the regression will allow for different intercept for different groups. This represents the difference in means between the groups, conditional on x. Thus, this extends our difference in means discussion from earlier in the class. β_2 here represents the difference in average income between men and women who have the same level of education. β_1 represents the effect of education on income within levels of gender.
- Why might we include a dummy variable? It turns out that it might save us from finding spurious correlations. Let's think about the relationship between height and math achievement among kids in 5th grade and 8th grade. Kids take a test which tells them what grade level their math skill are at. We might see a positive relationship overall—taller children have higher math scores. But, once we introduce a dummy variable for 5th grade versus 8th grade, this effect goes away. Within grade leve, there is no relationship between height and math achievement.
- This is an example of *Simpson's paradox*: a relationship that holds in general might not hold in all subgroups.

Adding continuous variables to a regression

- Instead of a dummy variable, we might want to control for a continuous variable. Let's look example of the effect of development in terms of GDP per capita on civil conflict, as measured by the number of deaths due to political violence (civil wars, riots, coups, etc). We might find that higher GDPs are associated with more conflict. We might want to make sure that this isn't due to the resource curse—countries with mineral and oil wealth are more likely to have conflict (because there's a prize to fight over) and more likely to have higher incomes. Thus, we might want to add another variable to our regression: percent of GDP coming from natural resources.
- It turns out we can add a continuous variable in the same way as a dummy variable:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2$$

- Now, x_1 and x_2 are both continuous variables. Before we had two regression lines, one for $d = 0$ and one for $d = 1$. Now, though, we have an infinite number of regression lines for each possible value of x_2. So it's not feasible to draw the regression lines anymore.
- We can interpret β_1 and β_2 in a similar fashion as we do above:
  - β_1 is the effect of x_1 when x_2 is held fixed.
  - β_2 is the effect of x_2 when x_1 is held fixed.

PSC200: Lecture 18

Matthew Blackwell

11/14/2012

Adding continuous variables to a regression

- Instead of a dummy variable, we might want to control for a continuous variable. Let's look example of the effect of development in terms of GDP per capita on civil conflict, as measured by the number of deaths due to political violence (civil wars, riots, coups, etc). We might find that higher GDPs are associated with more conflict. We might want to make sure that this isn't due to the resource curse—countries with mineral and oil wealth are more likely to have conflict (because there's a prize to fight over) and more likely to have higher incomes. Thus, we might want to add another variable to our regression: percent of GDP coming from natural resources.
- It turns out we can add a continuous variable in the same way as a dummy variable:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2$$

- While x_2 is another independent variable, we will often call x_2 a **control variable**, especially if it is not the main indepdent variable we are interested in.
- Now, x_1 and x_2 are both continuous variables. Before we had two regression lines, one for $d = 0$ and one for $d = 1$. Now, though, we have an infinite number of regression lines for each possible value of x_2. So it's not feasible to draw the regression lines anymore.
- We can calculate SEs and confidence intervals in the same way we did before and they have the same interpretation. A SE tells us how much we expect the estimate of the coefficient ($\hat{\beta}_1$ or $\hat{\beta}_2$) to vary from sample to sample. The formulas change slightly when you add another variable, but R will calculate all of this for us. Thus, our procedure for determining if a $\hat{\beta}_1$ is statistically significant is exactly the same as with one variable.
- We can interpret β_1 and β_2 in a similar fashion as we have before, with one exception:
  - Holding x_2 constant, a one unit change in x_1 leads to a β_1 change in y.
  - Holding x_1 constant, a one unit change in x_2 leads to a β_2 change in y.
- Why is holding things fixed (or "controlling for" them) important? Because it can remove spurious relationships. Remember the relationship between ice cream sales and drowning deaths. We said that we didn't think that ice cream sales **causes** drowning deaths, but that the weather probably causes both. If we were to control for or hold fixed the weather (by, say, making the comparison only when it's cold or only when it's hot), we would see that the spurious relationship would disappear: the relationship is caused by fluctuations in the weather. If the weather was always the same, we wouldn't see the relationship anymore.

Multiple Regression

- We've seen regression with one or two variables now and nothing has broken yet, so maybe we can add more variables. But first, why might we do this? Again, if we want to make sure that the relationship we have estimated is not spurious, we want to control for variables that might cause the relationship. One control variable might not be enough.
- It turns out that we can more than just one variable. What would that look like in terms of our regression model:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- You can see that we now have k independent variables instead of 1 or 2. Each of these could be continuous or dummy.
- Again, the SEs and CIs will come from R, but they have the same interpretation as before.
- We have the same interpretation about the coefficients here as we did with 2 variables:
  - **Holding all the indepdent variables fixed**, a one unit change in x_1 leads to a β_1 change in y.
  - **Holding all the indepdent variables fixed**, a one unit change in x_2 leads to a β_2 change in y.
  - …and so on.

How well does this regression predict y?

- Using our skills of regression interpretation, we can see how any individual variable affects the dependent variable, but we might want to go further. We might want to know **how much does this explain the dependent variable?**
- One way to do to measure this would be to see how well we do at predicting y using our regression model. The problem is that we need something to compare it to. We should compare it to some baseline predictions: those based on no independent variables.
- That is, we can compare the predictions from our regression:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_2 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k \tag{1}$$

And compare these predictions with those from a model without using $x_1, x_2, \ldots, x_k$. Our best prediction about y without any covariates is just the sample mean: $\bar{y}$.

- Once we have these two predictions, we can compare the *prediction errors* from these two models. A prediction error is the difference between the actual vablue of y and the predicted value. For a regression we have called these *residuals*: $y - \hat{y}$.
- From this setup we can calculate a measure of how well our independent variables predict y. We call this measure R^2 or **R-squared**. It is a number between 0 and 1 that tells us the proportion of total variation in y that is explained by the independent varaibles.

- How is it calculated? We do this by calculating the *proportional reduction in error*. The error with no covariates is called the total sum of squares: $TSS = \sum(y - \bar{y})^2$. This is is a measure of the prediction error with no covariates. What is the prediction error for our regression? It is the sum of the squred residuals: $SSR = \sum(y - \hat{y})^2$. Note that the book calls this the sum of squared errors (SSE).
- Finally, we can calculate the R^2 by showing how much of the total variation (TSS) is reduced with the regression:

$$R^2 = \frac{TSS - SSR}{TSS} \tag{2}$$

- Note that we can compare different regressions (with the same y) by comparing their R^2. When we add an additional independent variable to the model, we can see how much it increases the R^2 to determine how important it is for predicting y.

Multiple regression in R

- Multiple regression in R is almost exactly the same as bivariate regression in R. We simply have to add our new variables to the `lm()` function:

```
> mod <- lm(turnout ~ rain + polltax + literacy, data = rain)
> summary(mod)

Call:
lm(formula = turnout ~ rain + polltax + literacy, data = rain)

Residuals:
    Min      1Q  Median      3Q     Max
-46.157  -7.035  -0.177   7.019  35.208

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) 65.48146    0.06547 1000.172  < 2e-16 ***
rain        -1.04233    0.30585   -3.408 0.000655 ***
polltax      7.80947    2.80870    2.780 0.005432 **
literacy     1.41006    0.27279    5.169 2.37e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.51 on 27397 degrees of freedom
Multiple R-squared: 0.001802,   Adjusted R-squared: 0.001693
F-statistic: 16.48 on 3 and 27397 DF,  p-value: 1.067e-10
```

- Now, each row of the coefficients table has (1) the estimated coefficient for that variable, (2) the SE for that estimate, (3) the t-statistics for that variable, and (4) the p-value for the hypthesis test for that coefficient being equal to 0.

- Now in the second to last line, you can also see the R^2, referred to as the `Multiple R-squared`.

PSC200: Lecture 19

Matthew Blackwell

11/19/2012

Association and Causation

- We have talked about causal significance up to this point, but we haven't specifically talked about what causality means or how we should think about it.
- Examples: democracy causes economic growth. Obama won because of Hurricane Sandy.

Why do we care?

- Knowing causal relationships helps explain what has happened, predict what will happen, and guide our choice of policy.

Association and its pitfalls

- When we establish a relationship using regression, there could be a number of different causal processes at work. It could be that X causes Y, that Y causes X, or that they are both caused by a third factor, Z. With democracy and growth, we might think that strong property rights leads to both economic growth and democracy.
- Beware the *cum hoc* fallacy: association does not imply causation.
- Pitfall #1: Spurious relationships. When there is an association between x and y, but no causation because a third variable, z, causes both x and y we say that the relationship is **spurious**. We call the z variable a **z-factor**, a **lurking variable**, or a **confounder**.
- With a spurious relationship, the relationship between x and y is a coincidence.
- Pitfall #2: *post-hoc* fallacy. x must precede y to cause it, but this isn't enough to conclude that x causes y.
- Pitfall #3: A lack of association does not mean no causation. If candidates that are behind in the polls tend to go negative and going negative helps them, then it will look like negativity has no effect.

Counterfactuals

- One definition that we will use in this course is a *counterfactual* definition of causality.

- Let's call $y(1)$ be the dependent variable when we force $x = 1$ and $y(0)$ be the value the dependent variable takes when we force $x = 0$. Then, the causal effect of x is:

$$\text{causal effect } = y(1) - y(0)$$

- What does this mean? Let's say we are talking about the effect of drinking coffee on my productivity this afternoon. So that $y(1)$ is how many pages I would write if I were to drink coffee. And $y(0)$ is how many pages I would write if I were to abstain from coffee. Sometimes we call $x = 1$ the **treatment** and $x = 0$ the **control**.
- Now, there's a fundamental problem here. This morning I drank coffee so that means we get to see $y(1)$: it's the number of pages I write. But we never get to observe $y(0)$ because that's counterfactual: I'll never know how many pages I would have written if I had not had coffee. This is called **the fundamental problem of causal inference**.
- We can never observe the same unit in two different scenarios. We have to try and find a comparable group that did not receive the treamtent and see how they did on the outcome.
- In this example, we might try to find someone exactly like me who did not drink coffee and compare our productivities. If I had a twin, this might be possible.
- Rarely, we can perform a **cross-over study**: compare my productivity on one day with coffee to my productivity on a day without coffee. Thus, I would use myself as the comparison group. I'm pretty similar to myself! But it might not work if the coffee doesn't wear off between the comparisons.

Assessing Causation in Regressions

- So we said that we want to find comparable groups that have different values of x to evaluate the causal effect of x. Why would it be the case that two groups are not comparable? When there is a confound/z-factor.
- To see this, think about the coffee example. Suppose I took all of my coffee drinking days and compared my productivity to when I don't drink coffee. These two groups might not be comparable because of z-factor: stress. On days that I am more stressed with work, I tend to drink more coffee **and** I tend to write more (because I am under a deadline). Thus, the coffee days and the non-coffee days are not comparable.
- In this case the relationship between coffee and productivity is a messy combination of the causal effect of coffee and the effect of stress.
- How do we avoid these kinds of problems? We could compare coffee and non-coffee days only with low stress. And then compare coffee and non-coffee days only with high stress. Among low stress days, the coffee and non-coffee days are comparable.
- Thus, we want to **control** for any potential confounder/z-factor. We want to make sure our effect is holding those confounders constant. How do we do that? Well, we could just run our regression on low-stress days and drop any day that is high stress. But this wouldn't work for continuous variables.

- To control for these potential z-factors, we need to include them in the regression. If the relationship is still statistically significant after we include those variables, we know that these weren't really z-factors and they cannot explain the relationship we see. If we cannot think of any more potential z-factors, then we can argue that the relationship we see is causal.

- Can we be sure that we have established causality?

PSC200: Lecture 20

Matthew Blackwell

11/26/2012

Substantive Significance Revisited

- How can we evaluate substantive significance in a more systematic way? Well, we might want to know how much a **typical** change in x increases or decreases y. Up until now, we have looked mostly at one-unit changes in x. We know that a one-unit change in x_1 leads to a β_1 change in y. But a one-unit change might not be that important. What if we were looking at the relationship between income (x) and happiness (y). Would we care about a one-unit increase increase in annual income? That is, would care about the difference between $50,000 and $50,001 annual incomes? No, of course not. Obviously with income, we care about larger changes than the straightforward one-unit change. But what is the "right" change that we care about? Will it be different in different datasets? Yes, it will be.
- We can think of these changes as spread. Why do we think of a one-dollar difference in annual income as being "small"? That's because it's small relative to the spread of incomes: incomes range from as $10,000 up to millions. Obviously, a dollar is small in that range.
- Instead of a one-unit change, which may or may not be important with different independent variables, we might want to look for a **typical** change in x. That is, we might look for a "reasonable" change in x. How do we define "reasonable" or "typical" change for different datasets?
- One measure of spread is the range, but this might be too much. If we wanted to know the substantive significance, changing income from it's minimum to its maximum might be moving it too much. Just like a dollar is too little, millions of dollars is too much.
- Another measure of spread that we can use from variable to variable is the standard deviation. This is actually a good measure of a typical change: we know that an interval of roughly 6 standard deviations (from -3 SD to 3 SD) contains almost all of the data. So, a one-SD change in x is a pretty reasonable "typical" change in x. At the same time, a one-SD change in y is a fairly large change.
- This gives us a good way to judge substantive significance: compare a one-SD change in x with the SD of y.

Review

- A confounder/z-factor/lurking variable is a variable that causes both x and y. Last time we talked about how we want to hold these confounders constant because if they are constant, we know that they no longer can cause both x and y.

Assessing Causation in Regressions

- So we said that we want to find comparable groups that have different values of x to evaluate the causal effect of x. Why would it be the case that two groups are not comparable? When there is a confound/z-factor.

- To see this, think about the coffee example. Suppose I took all of my coffee drinking days and compared my productivity to when I don't drink coffee. These two groups might not be comparable because of z-factor: stress. On days that I am more stressed with work, I tend to drink more coffee **and** I tend to write more (because I am under a deadline). Thus, the coffee days and the non-coffee days are not comparable.

- In this case the relationship between coffee and productivity is a messy combination of the causal effect of coffee and the effect of stress.

- How do we avoid these kinds of problems? We could compare coffee and non-coffee days only with low stress. And then compare coffee and non-coffee days only with high stress. Among low stress days, the coffee and non-coffee days are comparable.

- Thus, we want to **control** for any potential confounder/z-factor. We want to make sure our effect is holding those confounders constant. How do we do that? Well, we could just run our regression on low-stress days and drop any day that is high stress. But this wouldn't work for continuous variables.

- To control for these potential z-factors, we need to include them in the regression. If the relationship is still statistically significant after we include those variables, we know that these weren't really z-factors and they cannot explain the relationship we see. If we cannot think of any more potential z-factors, then we can argue that the relationship we see is causal.

- Can we be sure that we have established causality?

Randomized experiments and Causality

- When we run regressions how can we be sure that we have eliminated all the possible z-factors? There could always be lurking variables out there that cause both x and y.

- One way to eliminate z-factors is to randomly assign x. By randomly assign, I mean that for each unit in our study, we flip a coin and, if heads, the unit gets $x = 1$ and if tails the unit gets $x = 0$. We call this **random assignment** or **randomization**.

- Why is randomization so powerful? Because once we randomize, there are no differences between the "treated" group and the "control" group. Before, we might believe that a z-factor is causing whether people were in the treated group or the control group. Imagine negative advertising, for instance. We would believe that those campaigns that go negative (treated group), are different than those who remain positive (control group) on a number of differet attributes: more likely to be in close races, more likely to have opponents that have gone negative, etc.

- When we randomize, all of these differences disappear because the **only** thing that can affect being in the treated or control group is the randomization—the coin flip. Suppose we randomly assigned negativity to campaigns. Then, on average, just as many close races would be negative as lopsided races.

- Thus, when we randomize, there must be no confounders/z-factors/lurking variables because these have to cause both x and y. But since we know that nothing except the coin flip causes x, then there must be no z-factors.
- When we run a regression of x on y when x is randomized, then β will be just the causal effect of x and not due to any other confounders.
- Without randomization, it is very hard to definitively conclude that we have established causality. There is always the possibility that someone will say we missed some crucial z-factor. Randomization removes that possibility.

Examples of Randomized Experiments in Political Science

- Lab experiments
- Bring in some volunteers and randomly show them either positive or negative ads for either a real or ficitional election. Then measure how much they remember the ads, how likely they are to vote for either candidate, how they feel about either candidate.
- Game theory experiments. Have people come into the lab and have them play games from game theory, randomly assigning different conditions. For instance, we might have them play the divide the dollar game, which is a bargaining game. Player 1 goes first and makes an offer from 0 to 100 of a 100 cents. Player 2 can either accept the division and each gets their share or she can reject it and neither party gets a reward. Political scientists have run experiments where they randomly assign players to play against (fake) opponents either from their ethnic group or another ethnic group. Thus, we can get insight into whether ethnic gropus generate cooperation.
- Survey Experiments
- Framing. We can insight into how people make political decision by changing the frame that we give a story or question. For instance, we might ask about whether a Nazi rally should be allowed in a public space in Downtown Rochester. If we use a "Free Speech" frame, people might be more supportive of allowing the rally and if we use a "Hate Crimes" frame, people might be less supportive. We can randomly assign either the Free Speech or the Hate Crimes Frame to see what the effect of framing a question is on the answer.
- Field Experiments
- Have a candidate for office randomly assign their campaign strategy: negative/positive ads, national/local appeals.
- GOTV: randomly call, mail, text potential voters to see if we can increase turnout.

PSC200: Lecture 22

Matthew Blackwell

11/26/2012

Review of dummy variables in regression

- Remember back to when we had a bivariate regression and we wanted to add a single binary variable. How did that work? Here's the formula we looked at:

$$y = \alpha + \beta_1 x + \beta_2 d$$

- And remember how showed what that meant. We plugged in the different values of d to see how the regression line changed:

$$(d = 0) : y = \alpha + \beta_1 x + \beta_2 \times 0 = \alpha + \beta_1 x$$
$$(d = 1) : y = \alpha + \beta_1 x + \beta_2 \times 1 = \alpha + \beta_1 x + \beta_2 = (\alpha + \beta_2) + \beta_1 x$$

- When we draw these two lines, what do we see? We see that they are parallel. That is, the regression assumes that the effect of x is constant across the groups. Is this believable?

Different groups, different slopes

- We might not think that different groups have the same slopes. That is, there might be a *pooling problem*: different types of units are being pooled together in the same regression when they really have different effects. We might think that there is a different relationship between education and income for men and women. This could be due to different processes. It could be that there is a glass ceiling so that women are never promoted to higher paid position in companies whereas men with the same education are, but that this discrimination isn't as strong with lower education jobs. On the other hand, it could be that women are more likely to drop out of the labor force than men at all levels of education and these women receive no income, which brings down the average income for women.

- What do we do when we face this problem? One option is to run completely separate regressions for each of the groups. Create subsets of the data for each group in R and run our regression in each subset. This would allow each group to have a different slope for each variable. We can even have different sets of independent variables if we think a z-factor is only relevant in one of the groups.

- But this approach leaves something to be desired for two reasons. First (and we'll talk about this next), this would only work for dummy variables (or a categorical variable with a small number of categories). Second, it doesn't allow us to directly compare the slopes for each group.

- Let's say we were to run two regression, one for men and one for women and we compare the coefficient on education for each. It is highly unlikely they will be exactly the same, but even if they are different, we would like to know if that's a **real** difference between the two slopes or if they are pratically/statistically the same. Of course, this sounds a lot like our typical setup: we see a difference between the slopes and we want to know if that difference is big enough to be real or if is just do to random chance.
- To do this, we are going to introduce a new type of variable into our regression: an **interaction term**. An interaction term is a variable that we create by multiplying two other variables together. Here is how an interaction between a continuous variable and a dummy variable looks in a regression:

$$y = \alpha + \beta_1 x + \beta_2 d + \beta_3 (x * d)$$

- We can work through the regression line for x for each group. First, let's look at $d = 0$:

$$(d = 0) : y = \alpha + \beta_1 x + \beta_2 * 0 + \beta_3 (x * 0) = \alpha + \beta_1 x$$

- When $d = 0$, this is just a regression line with intercept α and slope β_1.
- Now, $d = 1$:

$$\begin{aligned}(d = 1) : y &= \alpha + \beta_1 x + \beta_2 * 1 + \beta_3 (x * 1) \\ &= \alpha + \beta_1 x + \beta_2 + +\beta_3 x \\ &= (\alpha + \beta_2) + (\beta_1 + \beta_3) x\end{aligned}$$

- Thus, with $d = 1$, we have a new regression line, with a different intercept, $(\alpha + \beta_2)$ (like before when just had a dummy variable), but now we also have a different slope, $(\beta_1 + \beta_3)$. Now, we can see that we can have different effects in each group.
- This is super cool! Now our groups have different slopes, but it's important to recognize the interpretation of each coefficient:
  - α: the intercept when $d = 0$.
  - β_1: the slope when $d = 0$.
  - β_2: the change in the intercept between $d = 0$ and $d = 1$.
  - β_3: the change in slope between $d = 0$ and $d = 1$. This is the coefficient on the interaction term and it measures how different the slopes are between the groups. If this is 0, then the slope for both groups are the same. Thus, we can formulate a 95% confidence interval for this coefficient and see if it is different from 0. If it is statistically significant, then we know that each group had a different slope: the effect varies between the groups.
- When we interpret the statistical significance of an interaction term, we use a slightly different phrasing than before. With other variables we say that a one-unit change in x leads to a β_1 unit change in y. Here we are going to use slightly different language. We say that the $d = 1$ group has an effect that is β_3 larger/smaller than the effect for $d = 0$.
- All of this stays exactly the same when we add more independent variables as in a multiple regression.

Varying slopes across a continuous variable

- What if the effect of one variable depends not on a dummy variable, but instead on a continuous variable?

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2)$$

- We can rewrite this to see how the interaction works:

$$\begin{aligned} y &= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2) \\ &= \alpha + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2 \end{aligned}$$

- Look carefully at how we wrote this, now the effect of a one-unit change in x_1 depends is $(\beta_1 + \beta_3 x_2)$. You can see that the effect of x_1 depends on the value of x_2. Thus, a one-unit change in x_2 leads to a β_2 unit change in y, but also a β_3 unit change in the **effect** of x_1.
- If x_1 is the main variable we care about, we can interpret the coefficients in the following way:
  - α: the intercept when $x_2 = 0$.
  - β_1: a one-unit change in x_1 leads to a β_1 unit change in y when $x_2 = 0$.
  - β_2: a one-unit change in x_2 leads to a β_2 unit change in y when $x_1 = 0$.
  - β_3: a one-unit change in x_2 leads to a β_3 unit change in the effect of x_1.
- Note that we could interpret β_3 in the opposite way:
  - β_3: a one-unit change in x_1 leads to a β_3 unit change in the effect of x_2.
- Again, we have to be careful in how we interpret our coefficients. The effects of x_1 and x_2 are no longer "holding all other variables constant" because they depend on each other.

PSC200: Lecture 23

Matthew Blackwell

12/5/2012

Interactions between two dummys

- Let's review what we did last time by talking about a special case of interactions: two dummy variables.

$$y = \alpha + \beta_1 d_1 + \beta_2 d_2 + \beta_3(d_1 * d_2)$$

- There are four possible ways a person could have these two dummy variables: $(0,0)$, $(1,0)$, $(0,1)$, $(1,1)$. We can see how each of these does to our regression:

$$\begin{aligned}
(0,0) &: y = \alpha + \beta_1 0 + \beta_2 0 + \beta_3(0 * 0) = \alpha \\
(1,0) &: y = \alpha + \beta_1 1 + \beta_2 0 + \beta_3(1 * 0) = \alpha + \beta_1 \\
(0,1) &: y = \alpha + \beta_1 0 + \beta_2 1 + \beta_3(0 * 1) = \alpha + \beta_2 \\
(1,1) &: y = \alpha + \beta_1 1 + \beta_2 1 + \beta_3(1 * 1) = \alpha + \beta_1 + \beta_2 + \beta_3
\end{aligned}$$

- This gives us the following interpretations of the coefficients:
    - α: the intercept when both d_1 and d_2 are 0.
    - β_1: the effect of d_1 when $d_2 = 0$.
    - β_2: the effect of d_2 when $d_1 = 0$.
    - β_3: the change in effect for d_1 when $d_2 = 1$ OR the change in effect of d_2 when $d_1 = 1$.

Varying slopes across a continuous variable

- What if the effect of one variable depends not on a dummy variable, but instead on a continuous variable?

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3(x_1 * x_2)$$

- We can rewrite this to see how the interaction works:

$$\begin{aligned} y &= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2) \\ &= \alpha + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2 \end{aligned}$$

- Look carefully at how we wrote this, now the effect of a one-unit change in x_1 depends is $(\beta_1 + \beta_3 x_2)$. You can see that the effect of x_1 depends on the value of x_2. Thus, a one-unit change in x_2 leads to a β_2 unit change in y, but also a β_3 unit change in the **effect** of x_1.
- If x_1 is the main variable we care about, we can interpret the coefficients in the following way:
  - α: the intercept when $x_2 = 0$.
  - β_1: a one-unit change in x_1 leads to a β_1 unit change in y when $x_2 = 0$.
  - β_2: a one-unit change in x_2 leads to a β_2 unit change in y when $x_1 = 0$.
  - β_3: a one-unit change in x_2 leads to a β_3 unit change in the effect of x_1.
- Note that we could interpret β_3 in the opposite way:
  - β_3: a one-unit change in x_1 leads to a β_3 unit change in the effect of x_2.
- Again, we have to be careful in how we interpret our coefficients. The effects of x_1 and x_2 are no longer "holding all other variables constant" because they depend on each other.

Interactions in R

- When we go to add our interactions to our regression model in R only requires one change: that we add another variable which is just the two variable names connected by a `:`. That is, we should use the following syntax:

```
mymod <- lm(depvar ~ indvar1 + indvar2 + indvar1:indvar2, data = mydata)
```

- Here we just added one more term which is the interaction term.
- We can interact two dummy variables, a dummy variable and a continuous variable, or two continuous variables.

PSC200: Lecture 24

Matthew Blackwell

12/10/2012

What if an effect is non-linear?

- When we look at a regression, the effect of an independent variable is linear. That is, if x changes by 1, then the effect is β and if x changes by 100 then the effect is $100 \times \beta$.
- Does this always make sense? Probably not, think of the effect of education on income: do we think that 1 year of eduation in high school has the same effect as one year of education in college or beyond?
- Sometimes economists call this diminishing marginal returns: the effects of something tend to get smaller as we consume more of it. Think of how good the first piece of cake is compared to the the 10th piece of cake. It's probably the case the effect of cake on your happiness changes over time.
- Why do we care about this? Because sometimes trying to fit a line will hide a relationship that is strong, but non-linear.

Adding a non-linear term to a regression

- How can we allow the effect of a variable to change at different values of that variable? It might seem strange, but we have already learned how to let the effect of a variable change in response to another variable: we interact those two variables.
- When we want to see how the effect of a variable changes across itself, we just interact that variable with itself:

$$y = \alpha + \beta_1 x + \beta_2 (x * x)$$

- Let's just rewrite this formula to see what is going on:

$$y = \alpha + (\beta_1 + \beta_2 x)x$$

- What does this tell us. It says that α is the y-intercept and that the effect of x ($\beta_1 + \beta_2 x$) depends on the value of x. So that if $x = 0$, then the effect will be β_1. And for a one-unit increase in x the effect of x changes by β_2.

- So if $\beta_2 < 0$ and $\beta_1 > 0$, then the effect get smaller and smaller as x goes up, until at some point, the effect becomes negative and then gets more and more negative. This might sound familiar because it looks like an inverted-U shape or a parabola.
- In fact, we can think of this as modeling the effect of x as a parabola, instead of as a line:

$$y = \alpha + \beta_1 x + \beta_2 x^2$$

- This is exactly the same as the equation for a parabola from any old algebra class, where we have replaced α, β_1, and β_2, with c, b, and a:

$$y = ax^2 + bx + c$$

- This gives use some leverage on how to quickly interpret the parameters. β_2 (the coefficient on the squared term) tell us if the parabola is facing up (U-shaped, $\beta_2 > 0$) or down (inverse-U shaped, $\beta_2 < 0$). α tells us what the y-intercept is and β_1 is a little more complicated. It's the slope/effect for a small change around $x = 0$. In general, it tells us how steep the parabola is around $x = 0$.
- One last point: we can use the regression to tell us if there really is a non-linear component or if the effect looks roughly linear. We do this by checking the statistical signficance of β_2. This is because if $\beta_2 = 0$, then we just get back the original regression, with a linear effect. Thus, we can use this to check to see if a variable has a non-linear effect on y.
- Checking substantive significance is a little different here too. Since the effect depends on the value of x, you have to pick a value of x and a change in x and evaluate it. For instance, you might pick the mean of x and see how a one-SD change in x would increase y:

$$(\hat{\alpha} + \hat{\beta}_1(\bar{x} + SD) + \hat{\beta}_2(\bar{x} + SD)^2) - (\hat{\alpha} + \hat{\beta}_1\bar{x} + \hat{\beta}_2\bar{x}^2)$$

Non-linear effects in R

- In order to run a regression with a non-linear term in R, we have to add a slightly different variable to the model. We have to use this syntax:

```
mymod <- lm(depvar ~ indvar + I(indvar^2), data = mydata)
```

- What's going on here. First, note that the `I(indvar^2)` adds the non-linear/quadratic term to the regression. Unfortunately, we can't just add `indvar^2`, because R doesn't know what these mean. You have to wrap it in the `I()` function (the capital is important).
- Once you have the summary of the model, you can use all the usual tools to check significance, etc.