Political Science 209 - Fall 2018

Measurement

Florian Hollenbach

27th September 2018

- A sample is a small share of the population in that we are interested in

Survey Sampling

- A sample is a small share of the population in that we are interested in

- How do we draw samples in such a way that polls accurately reflect what is going to happen?

- How to construct samples that will represent the population?

- Example: We want to know the voting intentions of Texans (or Americans)

- We can hardly ask all eligible voters about their intention

- Example: We want to know the voting intentions of Texans (or Americans)

- We can hardly ask all eligible voters about their intention

- We take a *sample*

- The size of the sample is less important than its composition

DRUDGE POLL WHO WON THE FIRST PRESIDENTIAL DEBATE?

DONALD TRUMP	90.37% (17,424 votes)
HILLARY CLINTON	9.63% (1,857 votes)
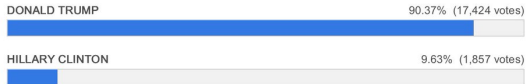
Total Votes: 19,281

Suka Kongsi 607

Tweet

Literary Digest Sample

- Mail questionnaire to 10 million people
- Addresses came from phone books and club memberships
- Problems?

- Mail questionnaire to 10 million people
- Addresses came from phone books and club memberships
- Problems?

- Biased *sample*

Quota Samping

- Sample certain groups until quota is filled
- Does not mean unobservables are representative

Simple Random Sampling

- Think of all voters sitting in a box, survey firm randomly draws voters
- Random draws without replacement give us an unbiased estimate of the population
- Everybody has the same chance of being in the sample

Simple Random Sampling

- Pre-determined number of units are randomly selected from population
- Sample will be representative of population on observed and unobserved characteristics

Simple Random Sampling

- Not every single sample will be exactly representative
- If we were to take a lot of random samples (say 1000 samples of 1000 respondents), on average the samples would be representative

Simple Random Sampling

- Each single sample can be off and different
- Polls are associated with uncertainty

Ted Cruz leads Beto O'Rourke 54 to 45, new poll says

The new Quinnipiac University poll surveyed likely voters instead of registered voters like it did in past iterations.

BY **PATRICK SVITEK**   SEPT. 18, 2018   11 AM

Simple Random Sampling

- Each single sample can be off and different
- Polls are associated with uncertainty

Ted Cruz leads Beto O'Rourke 54 to 45, new poll says

The new Quinnipiac University poll surveyed likely voters instead of registered voters like it did in past iterations.

BY **PATRICK SVITEK**   SEPT. 18, 2018   11 AM

Beto O'Rourke leads Ted Cruz by 2 among likely voters in U.S. Senate race, new poll finds

O'Rourke has been closing the gap over the last several months, but this is the first poll that puts him ahead of Cruz.

BY **KATHRYN LUNDSTROM**   SEPT. 19, 2018   8 AM

Random Sampling is hard

- How to create sampling frame?
- Random digit dialing? Walking to random houses?
- Multi-stage cluster sampling

- Unit non-response bias:

Non-reponse bias

- Item non-response bias: *What was the last crime you committed?*

- Sensitive questions: non-response, social desirability bias
  Turnout, racial prejudice, corruption

- unit non-response bias:

- unit non-response bias:

some citizens might not want to anser the door

Examples of Problems of Running Surveys in Afghanistan

- unit non-response bias:

some citizens might not want to anser the door

- item non-response bias:

Examples of Problems of Running Surveys in Afghanistan

- unit non-response bias:

some citizens might not want to anser the door

- item non-response bias:

Taliban supporters may be less likely to answer questions about Taliban

Examples of Problems of Running Surveys in Afghanistan

- unit non-response bias:

some citizens might not want to anser the door

- item non-response bias:

Taliban supporters may be less likely to answer questions about Taliban

- social desirability bias:

Examples of Problems of Running Surveys in Afghanistan

- unit non-response bias:

some citizens might not want to anser the door

- item non-response bias:

Taliban supporters may be less likely to answer questions about Taliban

- social desirability bias:

Taliban supporters may not want to admit to supporting Taliban

List Experiments

- list of groups respondent might support
- Asked to name number of groups they support
- Treated subject with controversial group, control group without controversial group

List Experiments - Control

I'm going to read you a list with the names of different groups and individuals on it. After I read the entire list, I'd like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don't tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.

Groups: Karzai Government; National Solidarity Program; Local Farmers

List Experiments - Treated

I'm going to read you a list with the names of different groups and individuals on it. After I read the entire list, I'd like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don't tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.

Groups: Karzai Government; National Solidarity Program; Local Farmers; ISAF (Taliban)

List Experiments

- Average difference between Treated and Control group is the estimated percentage of people who support controversial group

Summarizing Bivariate Relationships

- Bivariate relationships are associations between two variables
- Example: treatment of Spanish confederates (X or T) and exclusionary attitudes (Y)

Simple Summaries of Bivariate Relationships

If X (independent variable) is categorical:

- Comparison of means
- boxplots

If both X (independent variable) and Y (dependent variable) are continuous:

Simple Summaries of Bivariate Relationships

If both X (independent variable) and Y (dependent variable) are continuous:

- Scatterplots

Simple Summaries of Bivariate Relationships

If both X (independent variable) and Y (dependent variable) are continuous:

- Scatterplots

- Correlation

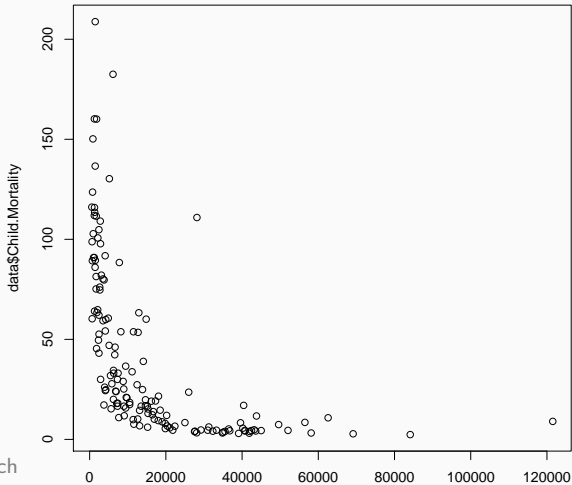## Simple Summaries of Bivariate Relationships: Scatterplot

- Direct graphical comparison of two variables
- Use plot(y,x) in *R*

Scatterplot

```
data <- read.csv("bivariate_data.csv")
data <- subset(data, year == 2010)
plot(data$GDP,data$Child.Mortality)
```

Scatterplot

That looks weird, no? What do we do with skewed variables?

What do we do with skewed variables?

What do we do with skewed variables?

- When variable have a small number of observations with extremely large or small positive values, we often take the natural log
- The natural logarithm is the logarithm with base e, which is a mathematical constant approximately equal to 2.7182 (inverse of e^y)

Scatterplot

```
plot(log(data$GDP),data$Child.Mortality)
```
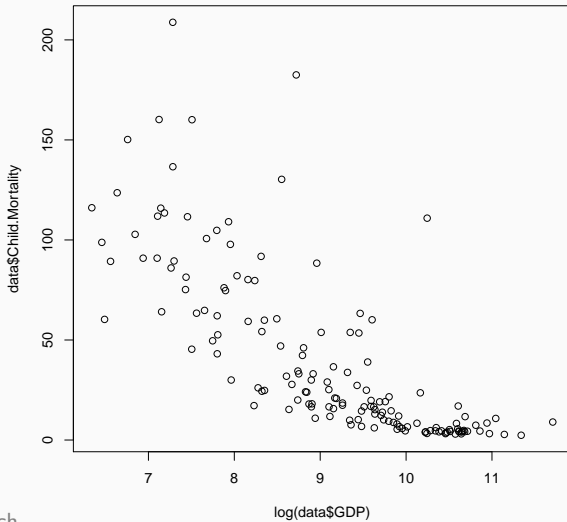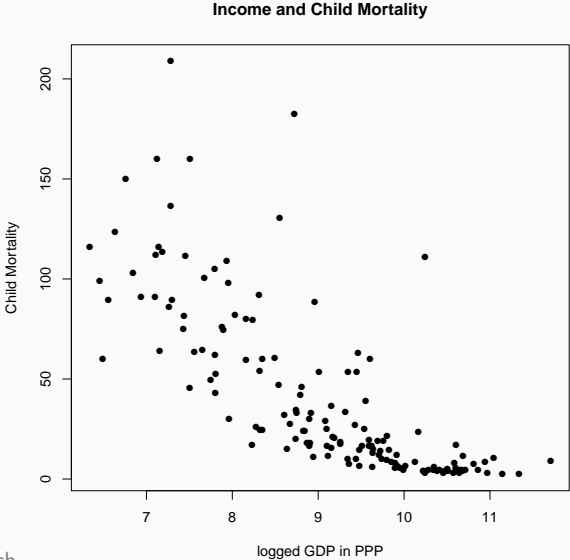
Scatterplot

```
pdf("~/Documents/GitHub/Polisci209_2018/slides/week5/scatter.pdf")
plot(log(data$GDP),data$Child.Mortality, pch = 16, col = "black",
xlab = "logged GDP in PPP", ylab = "Child Mortality", main = "Income and Child Mortality")
dev.off()
```
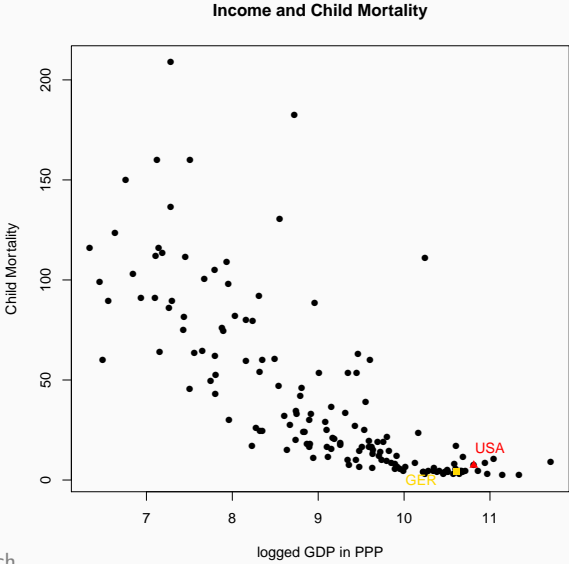
Income and Child Mortality

Scatterplot – more fun

```
## add special points for USA and Germany
pdf("~/Documents/GitHub/Polisci209_2018/slides/week5/scatter_points.pdf")
plot(log(data$GDP),data$Child.Mortality, pch = 16, col = "black",
xlab = "logged GDP in PPP", ylab = "Child Mortality", main = "Income and Child Mortality")
points(log(data$GDP[data$Country.code == "USA"]), data$Child.Mortality[data$Country.code == "USA"], pch = 17, col = "red") ##USA
text(11, 16, "USA", col = "red")
points(log(data$GDP[data$Country.code == "DEU"]), data$Child.Mortality[data$Country.code == "DEU"], pch = 15, col = "gold") ##USA
text(10.2, 0, "GER", col = "gold")
dev.off()
```

Income and Child Mortality

- How can we quantify the relationship between two continuous variables?

Bivariate Relationships in Numbers

- How can we quantify the relationship between two continuous variables?

- Correlation: the most used measure of bivariate relationships
- Correlation measures how two variables move together relative to their respective means.

- By standardizing we bring all variables on the same scale
- The resulting mean will be zero, the standard deviation will be one
- We standardize by subtracting a variables mean and dividing by the standard deviation

Calculating correlations – Step 1: Standardizing a variable

- Standardized variables are also called z-scores:

$$z_i = \frac{x_i - \bar{x}(\text{mean of x})}{sd_x(\text{standard deviation of x})}$$

- The z-score is independent of the scale of the variable or shifts in the variable

Calculating correlations – Step 1: Standardizing a variable

- Standardized variables are also called z-scores:

$$z_i = \frac{x_i - \bar{x}(\text{mean of } x)}{sd_x(\text{standard deviation of } x)}$$

- The z-score is independent of the scale of the variable or shifts in the variable

- This means GDP and (GDP*100 + 10000) will have the exact same z-scores

Correlation $(x,y) = \frac{1}{N} \sum_{i=1}^{N}$ z-score of $x_i \times$ z-score of y_i

Correlation $(x,y) = \frac{1}{N} \sum_{i=1}^{N}$ z-score of $x_i \times$ z-score of y_i

Correlation $(x,y) = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i - \bar{x}}{sd_x} \times \frac{y_i - \bar{y}}{sd_y}$

- Correlation measures *linear* association
- Correlations are between -1 and 1

```
cor(log(data$GDP),data$Child.Mortality, use = "pairwise")
```

= -0.7684907

cor(data[, c("GDP", "Child.Mortality", "PolityIV")], use = "pairwise.complete.obs")

```
### by hand
z_gdp <- (log(data$GDP) - mean(log(data$GDP), na.rm = T))/sd(log(data$GDP), na.rm =T)
z_CM <- (data$Child.Mortality - mean(data$Child.Mortality, na.rm = T))/sd(data$Child.Mortality, na.rm =T)
cor <- sum(z_gdp*z_CM)/(length(z_gdp)-1)
 -0.7684907
```

First draft: Due on October 5th Final Version: Due on October 15th

Writing Assignment

Structure of the Assignment

- Short Introcution
    - What is this paper about?
    - Why does this matter?
- Main body of text
- Short Conclusion

How to write well

- Cite every single outside source!
- See TAMU Library: Citing your Sources
- Preferred: APA style with parenthetical citation

How to write well

- Be clear and concise in your statements

- Short sentences

- Do not use words you do not understand yourself

How to write well

- Limited (or no) passive voice:
  - things don't just happen
- Be concrete, use active verbs

- Each paragraph should contain one important point, or should be "a complete discussion of one topic"

- Punctuation should clarify

How to write well

- Use correct grammar, limit spelling errors
- Revise, proofread, revise!

Some more tips are available here:
http://fhollenbach.org/WritingAcademic/