1. Introduction to concepts of probability and statistics

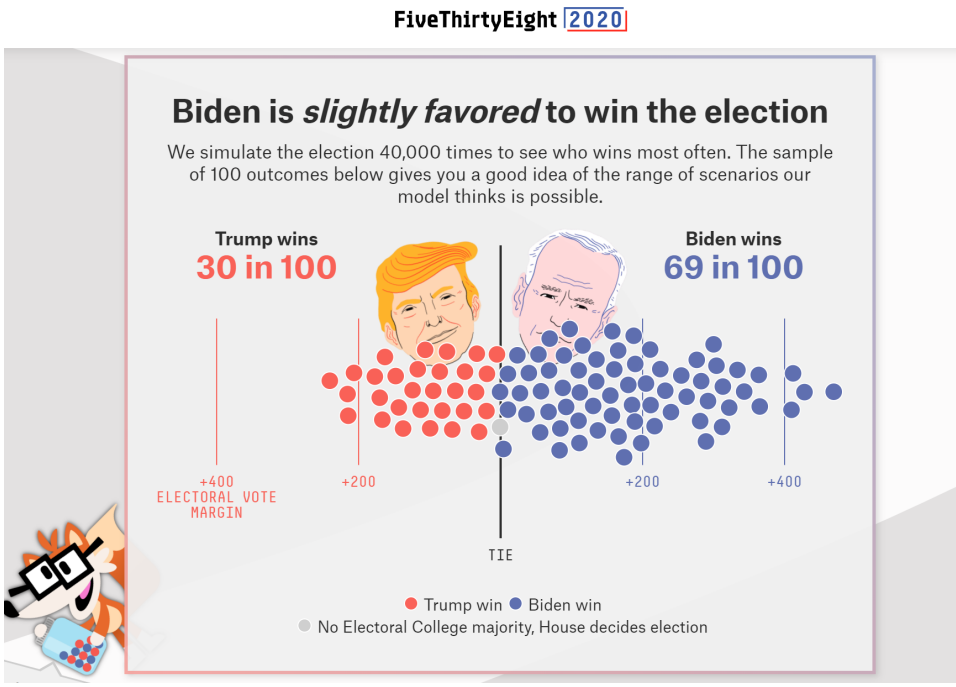
LPO.8800: Statistical Methods for Education Research

Sean P. Corcoran

Why study statistics?

- Data, statistics, and visualizations of data are *everywhere*—and increasingly so
- Technology has vastly increased the availability of raw data and raised the sophistication of everyday statistical reporting

Why study statistics? (2020)



Why study statistics? (2016)

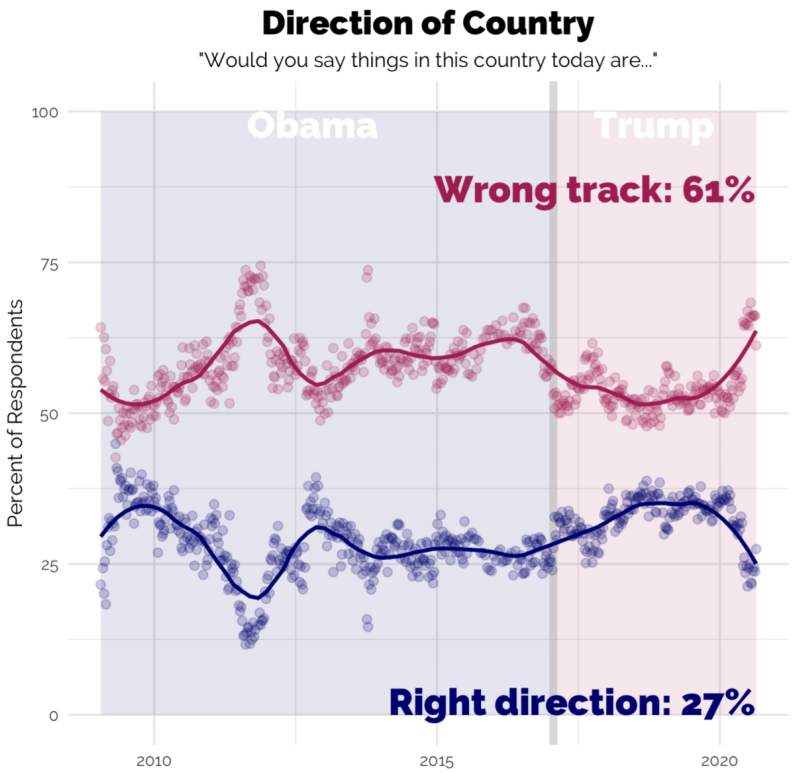
Who's ahead in each state and by how much

Our win probabilities come from simulating the election 10,000 times, which produces a distribution of possible outcomes for each state. Here are the expected margins of victory. The closer the dot is to the center line, the tighter the race. And the wider the bar, the less certain the model is about the outcome.

KEY AVERAGE — 80% CHANCE OUTCOME FALLS IN THIS RANGE

	Expected margin of victory (+75 +50 +25 +25 +50 +75)		Chance of tipping election
Florida		D+1.4	16.4%
Ohio		D+1.4	12.0%
Pennsylvania		D+4.8	10.2%
North Carolina		D+0.6	8.1%
Michigan		D+5.8	7.6%
Wisconsin		D+5.3	6.4%
Virginia		D+6.1	6.1%
Colorado		D+5.3	3.7%
Minnesota		D+7.1	3.3%
New Jersey		D+9.5	3.2%
Arizona		R+2.8	2.7%
Iowa		D+0.3	2.5%
New Hampshire		D+4.7	2.5%
Georgia		R+4.0	2.4%

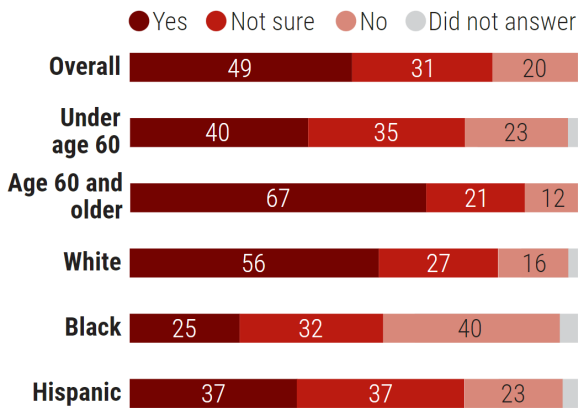
Why study statistics?



Why study statistics?

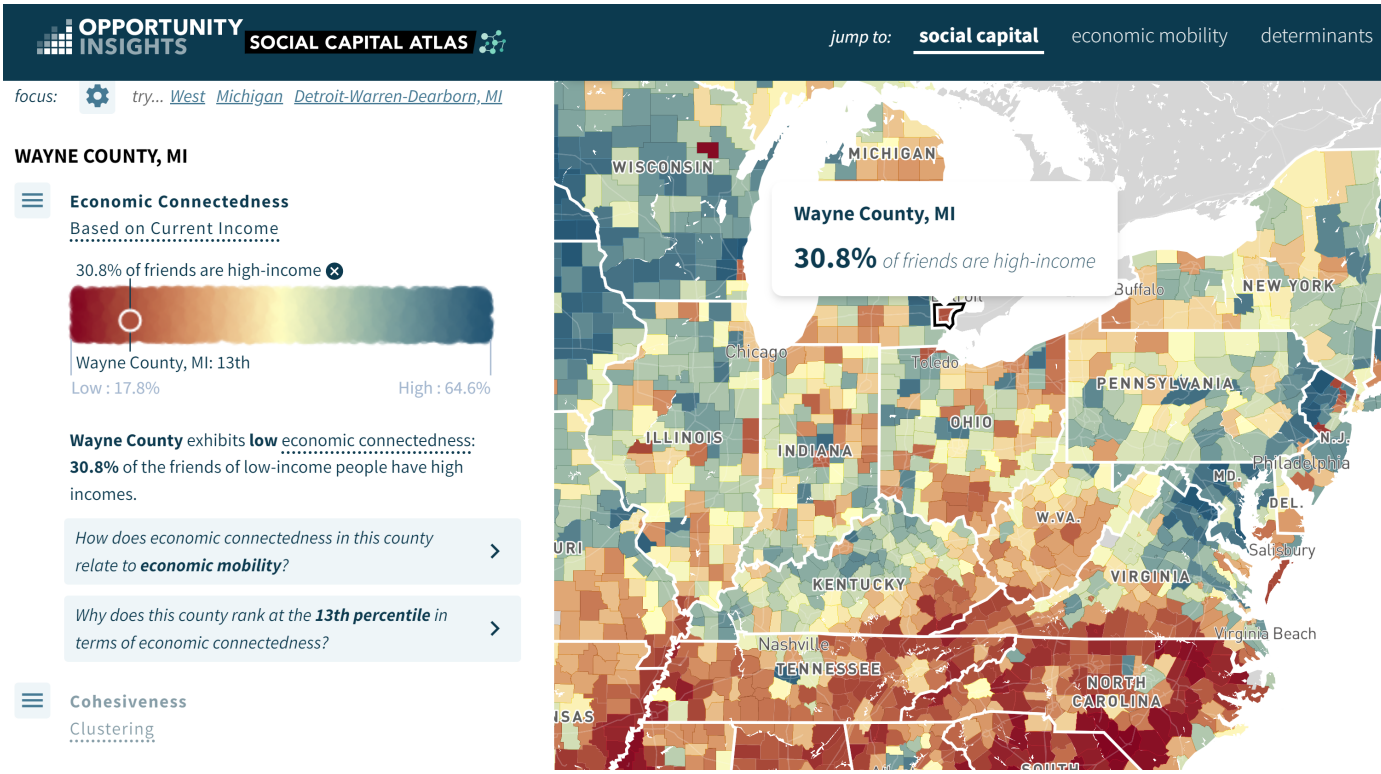
Do you plan to get a coronavirus vaccine when one is available?

For some in the United States, the answer is no, according to a survey of 1056 people in mid-May.

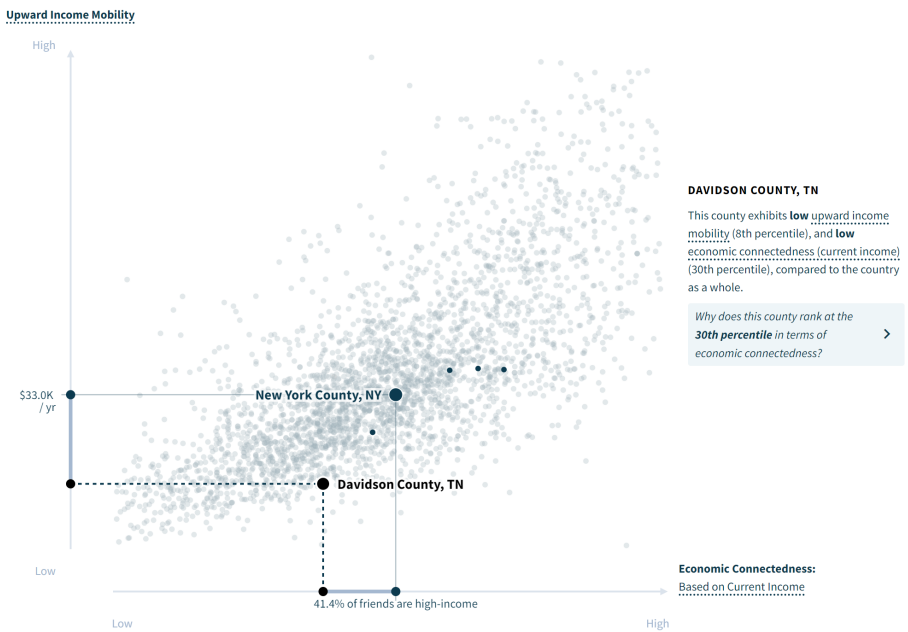


(GRAPHIC) V. ALTOUNIAN/*SCIENCE*; (DATA) ASSOCIATED PRESS–NORC CENTER FOR PUBLIC AFFAIRS RESEARCH AT THE UNIVERSITY OF CHICAGO

Why study statistics?

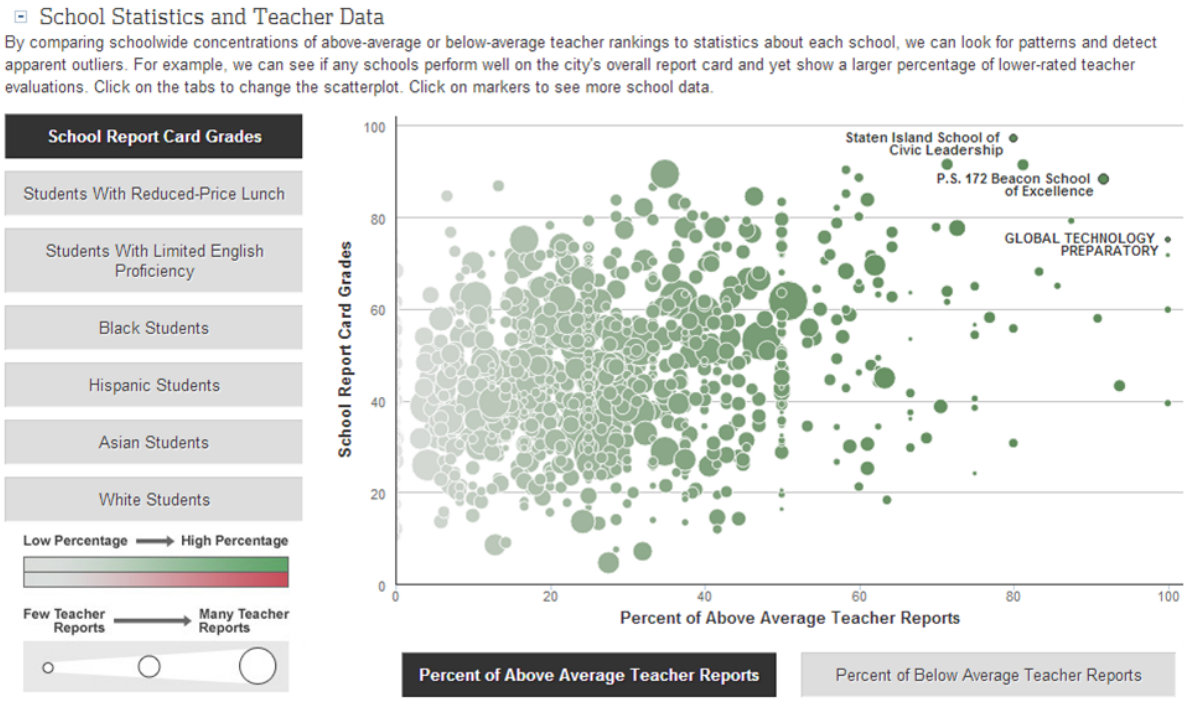


Why study statistics?

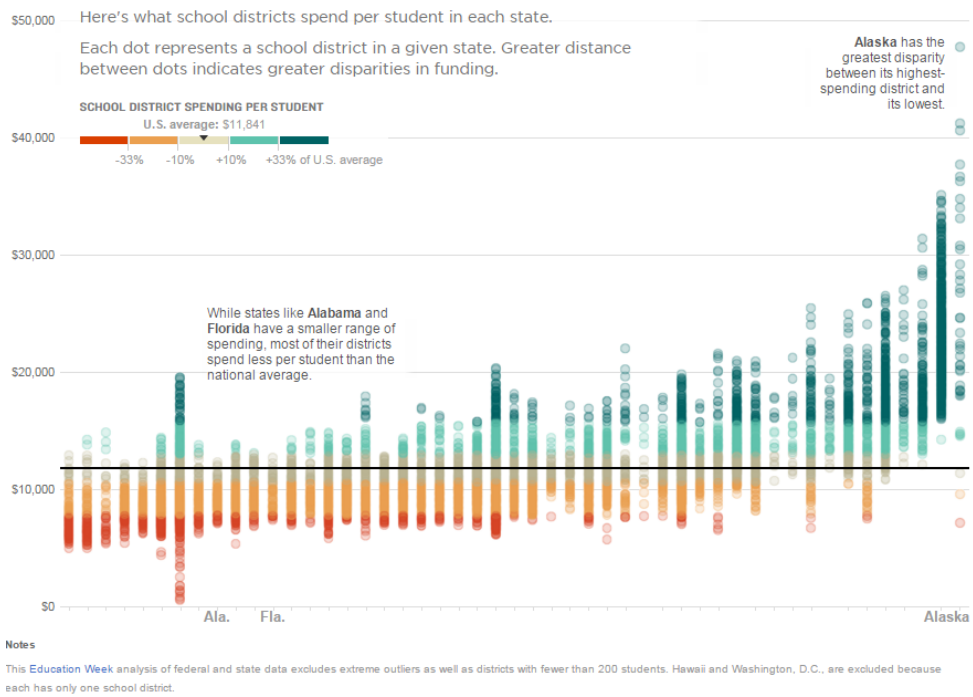


Source: Social Capital Atlas `https://socialcapital.org/`

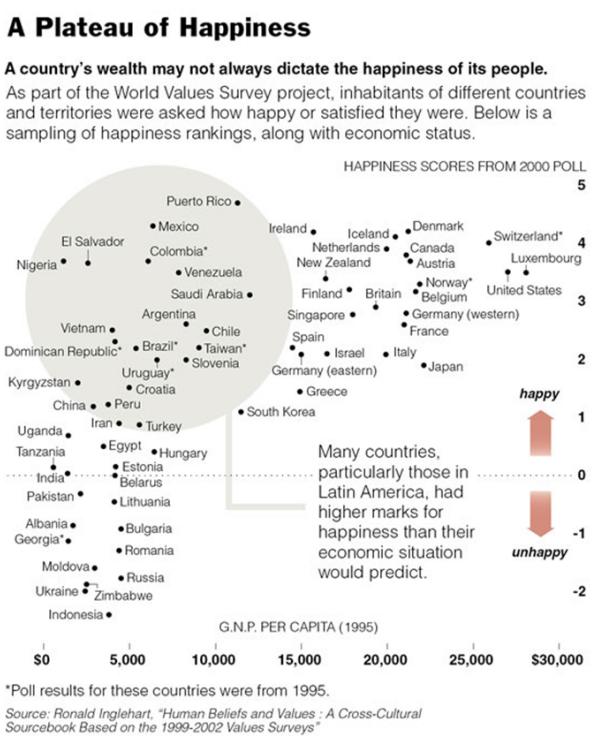
Why study statistics?

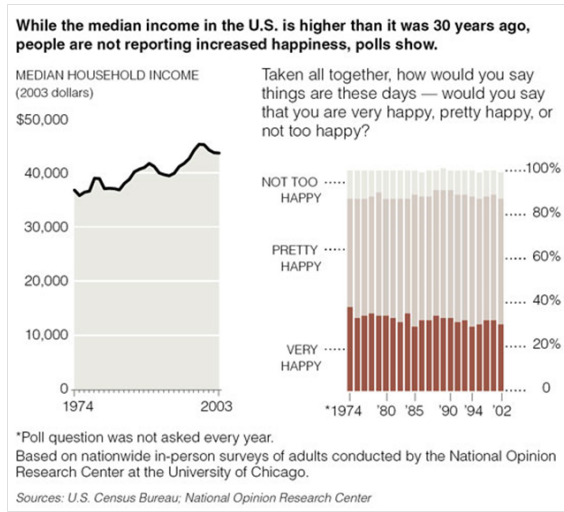


Why study statistics?



Why study statistics?





Why study statistics?

- This is especially true in education, as student- and school-level data, accountability, and public reporting have all grown in use (NCLB)
- The use (and misuse) of data is ubiquitous in education policy
- A strong foundation in statistics will help you to be an intelligent producer and consumer of data analysis

Descriptive vs. inferential statistics

- Statistics deals primarily with *variation* and/or *uncertainty* in an *outcome, characteristic*, or phenomenon of interest.
- A collection of observations on these outcomes is referred to as **data** or a **data set**.
- The **unit of observation** in one's data set may depend on the research question of interest (and/or data availability).

Descriptive vs. inferential statistics

- Statistical methods can be classified as **descriptive** or **inferential**.
- **Descriptive statistics** are used to *describe* outcomes in a population or sample (e.g., central tendency, variation, distribution shape; overall or by subgroup; correlation).
- **Inferential statistics** are used to make *inferences* or *predictions* about a population larger than that observed in the data.

Descriptive vs. inferential statistics

- **Population**: the universe of outcomes of interest
  - GPAs of all Vanderbilt students
  - Commuting time for all Vanderbilt graduate students
  - Incomes of U.S. households
  - Math ability of 4th grade students
  - Favored candidate of registered voters in a political race
- Notice each of these examples includes an *outcome* and a *unit of observation* (and often a time/place)
- The researcher may or may not observe (or be able to observe) the population of interest.

Descriptive vs. inferential statistics

- **Sample**: a subset of the population, chosen at random or by some other method.
- Descriptive statistics can be conducted on either a population or a sample.
- The key difference is how these statistics are used / interpreted.

Descriptive vs. inferential statistics

- It may be impossible, or cost-prohibitive, to observe the full population. In these cases a sample can be used to make inferences about the population.
- An important step in inferential statistics is the *quantification of uncertainty* (e.g., standard errors or a "margin of error"). Covered in Lecture 5 and following.

Examples

- What proportion of Nashville Labor Day weekends saw rain in the 1990s? (descriptive statistics)
- What is the likelihood that it will rain next Labor Day weekend in Nashville? (inferential statistics)
- What fraction of the incoming freshman class in NYC in 2004 graduated in four years? (descriptive statistics)
- What is the four-year high school graduation rate in the United States? (inferential statistics)
- What is the probability that a randomly drawn 18-year-old will be a high school graduate? (inferential statistics)

Examples

- What is the average annual household income in this class? (descriptive statistics)
- What is the average annual household income in the United States? (inferential statistics)
- What is the probability that a randomly drawn U.S. household will earn $250,000/year or more? (inferential statistics)

Descriptive vs. inferential statistics

In data analysis it is important to be precise in defining:

- The unit of observation (level at which the data are observed)
- The outcome of interest (and how it is measured)
- The population of interest (who, what, when, where)
- How the sample is drawn (when applicable)

Descriptive vs. inferential statistics

What is the (1) unit of observation; (2) outcome variable of interest; (3) population of interest?

- What proportion of Nashville Labor Day weekends saw rain in the 1990s?
- What is the likelihood that it will rain *next* Labor Day weekend in Nashville?
- What is the average annual household income in the U.S.? What is the probability that a randomly drawn household will earn $250,000/year or more?
- Math ability of 4th grade students:
  - Unit of observation: a 4th grade student
  - Population: your classroom? Tennessee? U.S.?
  - Outcome of interest: "math ability" - how measured?
  - Sample: NAEP, for example

Descriptive vs. inferential statistics

In this course:

- Lectures 1 - 3 and some of 10-12: descriptive statistics
- Lectures 4 - 9: inferential statistics

Variables and measurement

The "outcomes" described above are also called **variables** (or *random variables*) because they can vary from one unit of observation to another.

- Many variables are inherently numeric; others are not.
- **Categorical** or **qualitative** variables can be assigned numeric values for convenience.
- Examples: male / female, employed / unemployed, strongly disagree ... strongly agree, marital status

Variables and measurement

The possible values a variable can take on (and their meaning) form its **measurement scale**.

- The measurement scale is important, as it dictates which kinds of statistical analysis are appropriate.
- The scale can be characterized in a number of ways.
  - Quantitative or categorical ("qualitative")
  - Nominal, ordinal, interval or ratio
  - Discrete or continuous

Variables and measurement

One dimension is whether the variable is quantitative or categorical:

- A **quantitative** variable is on a numeric scale, where the numeric values express the magnitude of some property or characteristic.
- A **categorical** ("qualitative") variable consists of a number of distinct categories that may or may not have a natural ordering.

Examples

Quantitative or categorical?

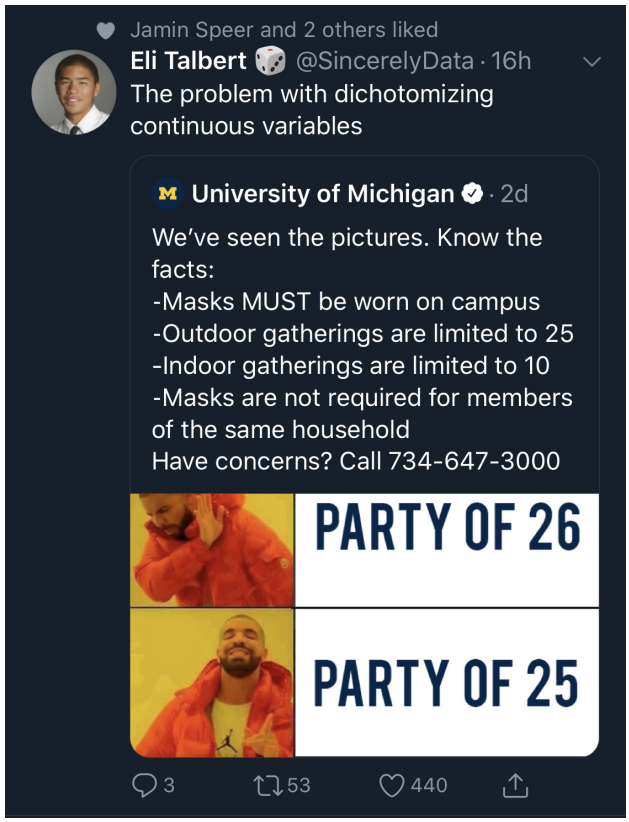
- Height
- Profession
- Percentage of HS seniors each year who report they have never smoked cigarettes
- Annual rainfall in Savannah, GA to the nearest inch
- Number of points scored in a football game
- Weight
- Verbal aptitude as measured by SAT verbal
- Salary of various government officials
- Age
- Eye color

Nominal scales

Categorical variables that lack a natural ordering are on a **nominal scale**.

- Categories simply recognize *differences*
- Numeric values can be assigned to categories; the values themselves are arbitrary
- Order is not meaningful; values cannot be compared
- A **dichotomous** variable has *two* categories (special case of a nominal measure). In practice called a "dummy" or "indicator" variable (0-1).

Dichotomous danger



Nominal scales

Variables with a **nominal scale**:

- gender 1 = Male, 2 = Female
- state of residence 1 = NY, 2 = NJ, 3 = CT
- high school graduate 1 = Yes, 2 = No
- employment status: 0 = unemployed, 1 = employed
- favorite color: 10 = blue, 20 = green, 30 = red, 40 = yellow
- favorite type of music: 1 = Rock, 2 = Classical, 3 = Jazz, 4 = Country, 5 = Other

Ordinal scales

Categorical variables with a natural ordering are on an **ordinal scale**.

- Numeric values are assigned to categories, *and* their order is meaningful
- The values themselves are *arbitrary*, but higher value implies *more* of some property
- Numeric values can be compared, but only to determine which has more or less of some property
- The *interval* between numeric values is not meaningful, and thus their difference and ratio are not meaningful

Ordinal scales

Variables with an **ordinal scale**:

- NAEP math skills 1 = basic, 2 = proficient, 3 = advanced
- job satisfaction 1 = very dissatisfied, 2 = somewhat dissatisfied, 3 = neutral, 4 = somewhat satisfied, 5 = very satisfied (a **Likert**-type item)
- education completed 1 = less than HS, 2 = HS, 3 = some college, 4 = college or higher

Ordinal scales

In education, state assessments commonly use an ordinal scale to indicate performance levels. Ex:

- New York State: Levels 1-4 (3 or 4 is "proficent")
- Tennessee: Level 1 = below grade level, 2 = approaching, 3 = on-track, 4 = mastered

These are often aggregated to "percent proficient". See the excellent article from Andrew Ho on the problems with proficiency when comparing results across locations or over time: `https://www.gse.harvard.edu/news/uk/15/12/when-proficient-isnt-good`

Interval scales

Quantitative variables may have an **interval** or **ratio** scale:

- Numeric values have meaning, and can be used to compare magnitude of some property
- The intervals between numeric values are informative: equal increments on the scale implies equal intervals of some property
- The *ratio* of two values *may or may not* be meaningful (if they are, the measure can also be said to have a **ratio** scale)
- Most quantitative variables have a ratio scale, but some are interval but not ratio.

Interval scales

Variables with an **interval** (and **ratio**) scale:

- days of instruction (150 - 205)
- annual income (\$0 to ∞?)
- calories consumed in a day (e.g., 1000 - 10000)
- Can make ratio comparisons: Person A consumed twice as many calories (3000) as Person B (1500)
- Ratio scales have a clear "zero value"

Interval scales

Some scales are **interval** but not ratio:

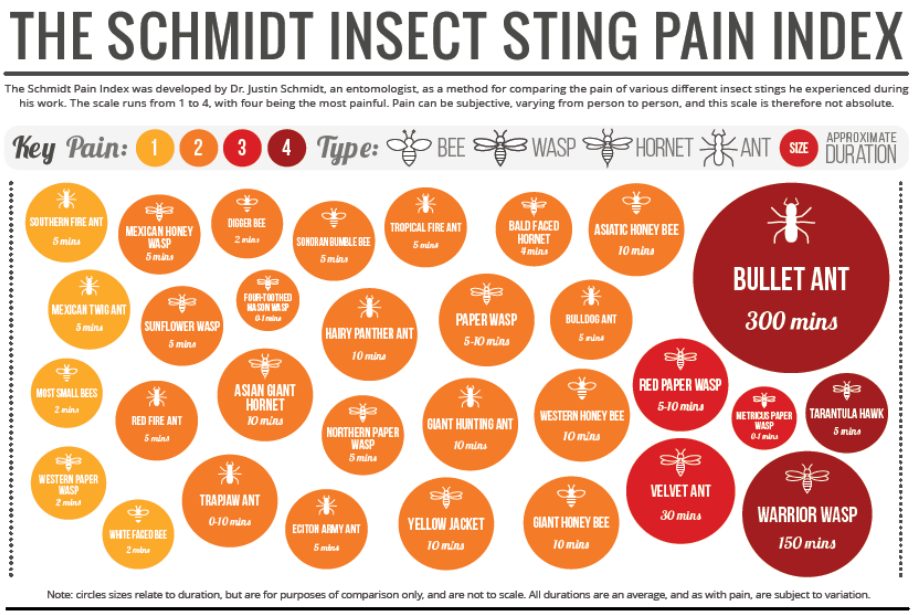
- Temperature is the most common example:
  - 60°F is 20°F more than 40°F
  - 80°F is 20°F more than 60°F
  - But 80°F is not "four times as warm" as 20°F.
  - Arbitrary zero point: "temperature" does not begin at a specific floor (e.g. zero) that would enable ratio comparisons.
- Date, measured from an arbitrary starting point
- Opinion and attitude scales (maybe, if not just ordinal)
- IQ, math achievement (maybe, if not just ordinal)

Examples

What level of measurement?

- Height
- Profession
- Percentage of HS seniors each year who report they have never smoked cigarettes
- Annual rainfall in Savannah, GA to the nearest inch
- Number of points scored in a football game
- Weight
- Verbal aptitude as measured by SAT verbal
- Salary of various government officials
- Age
- Eye color

(Not so) fun scales



Discrete vs. continuous

Quantitative variables can be discrete or continuous:

- **discrete**: the variable can take on a *countable* number of values (e.g., values can be represented by integers)

- **continuous**: the variable can take on a *continuum* of values (e.g., all values between some a and b)

Discrete vs. continuous

It is useful to separate in your mind features of the *underlying property* you are measuring and the scale of the *measurement tool* being used to measure it (and/or the observed data). Example: height

- The underlying property being measured is continuous (and on a ratio scale)
- The measurement (and data you use) may be discrete—e.g. rounded to the nearest inch
- The large number of possible values in the data lead us to treat the measure as continuous

Confusing, I know...

Examples

Discrete or continuous?

- Height
- Profession
- Percentage of HS seniors each year who report they have never smoked cigarettes
- Annual rainfall in Savannah, GA to the nearest inch
- Number of points scored in a football game
- Weight
- Verbal aptitude as measured by SAT verbal
- Salary of various government officials
- Age
- Eye color

Variables and measurement

When measuring quantitative variables in practice, it pays to capture more information rather than less, to the extent you can. (It's easy to go from a fine measure to something less fine, but hard to go the other direction).

- ex: if individuals' heights are known (6'1", 5'3", etc.—a ratio scale) one throws information away by simply classifying individuals as "short" or "tall" (an ordinal scale)

- Preserving the original variable in your data will permit you to create other versions of the variable on a different scale

Sampling

In inferential statistics, a sample is drawn and used to make inferences about a larger population. This sample can be taken in a number of different ways.

- In **simple random sampling** (SRS) every member of the population has an equal chance of being selected. (E.g., with N units in the population, n are drawn at random. n is the **sample size**).

- With SRS, each unit has an n/N probability of being drawn.

Sampling

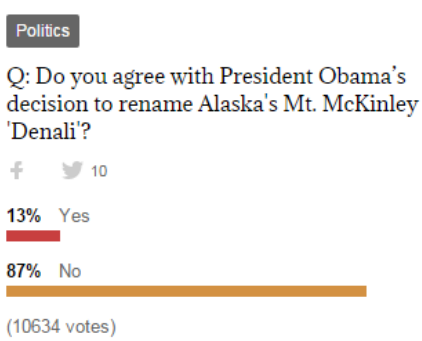
A random sample begins with a **sampling frame**, an enumeration of all units that could be sampled. This typically represents the population of interest.

- Ex: all registered students at Vanderbilt
- Ex: all 4th grade students in the U.S.
- all working telephone numbers in Iowa

Random samples are appealing because they are unlikely to be biased in some systematic way **(sampling bias)**

Sampling

- A complete sampling frame may not be available to the researcher; alternative sampling methods may be required
- **Nonprobability sampling**: when one cannot determine probabilities of drawing possible samples. Examples:
  - Volunteer sampling (e.g. web-based, cable news polls)
  - "Convenience sampling"



Sampling

Probability sampling: when the probabilities of drawing possible samples from the population can be determined:

- Simple random sampling
- **Stratified sampling**
  - Random samples are drawn within defined groups, or **strata**
  - May be **proportional or disproportional**, depending on whether the sample drawn from each stratum is proportional to its size in the population, or not. Sometimes groups are over-sampled to ensure representation
- **Cluster sampling**: sampled units are drawn in *groups*, such as schools, counties, or city blocks.

Example: stratified sample

Example: stratified sample

Record	Name	Group	
1	Bradburn Corp.	High	
2	Cochran Inc.	Highest	
3	Deming Design	High	One SRS of Size 4
4	Fuller & Fuller	Medium	
5	Habermann AG	Medium	
6	Hansen PLC	Low	
7	Hu Electronics	Highest	
8	HydeBev	High	
9	Kalton Group	Medium	→ Kalton Group
10	Kish Consulting	Low	
11	Madow USA	Highest	
12	M.P.H. Bank	Highest	
13	Norwood LC	Medium	→ Norwood LLC
14	Rubin Inc.	Low	→ Rubin Inc.
15	Sheatsley Co.	Low	
16	Steinberg Ltd.	Low	
17	Sudman Inc.	High	
18	Wallman AG	High	→ Wallman AG
19	Wolfe & Enix	Highest	
20	WXM Ventures	Medium	

Figure 4.5 Frame population of 20 establishments sorted alphabetically, with SRS sample realization of size $n = 4$.

Example: stratified sample

Record	Name	Group	
2	Cochran Inc.	Highest	
7	Hu Electronics	Highest	One Stratified Random Sample of Total Size 4
11	Madow USA	Highest	
12	M.P.H. Bank	Highest	
19	Wolfe & Enix	Highest	→ Wolfe & Enix
1	Bradburn Corp.	High	→ Bradburn Corp.
3	Deming Design	High	
8	HydeBev	High	
17	Sudman Inc.	High	
18	Wallman AG	High	
4	Fuller & Fuller	Medium	→ Fuller & Fuller
5	Habermann AG	Medium	
9	Kalton Group	Medium	
13	Norwood LC	Medium	
20	WXM Venture	Medium	
6	Hansen PLC	Low	
10	Kish Consulting	Low	
14	Rubin Inc.	Low	→ Rubin Inc.
15	Sheatsley Co.	Low	
16	Steinberg Ltd.	Low	

Figure 4.6 Frame population of 20 establishments sorted by group, with stratified element sample of size $n_h = 1$ from each stratum.

Political polling

As an example, visit `https://projects.fivethirtyeight.com/polls/` and select one of FiveThirtyEight's recent poll averages.

- Who is the population of interest? Who is the sampling frame?
- How many subjects were selected?
- How were the subjects selected? Was SRS used, or stratified sampling?
- Was a margin of error reported? (More on this later)

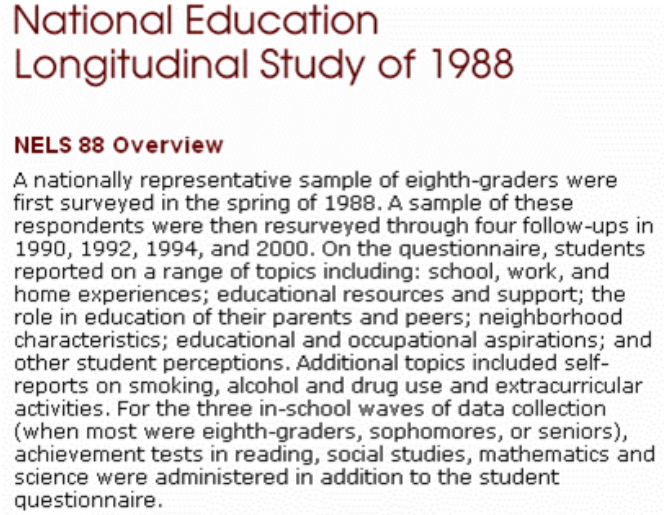
A crash course in Stata

You can view my introduction to Stata video on YouTube. See also the accompanying handout *Useful Stata commands*. These cover:

- Stata interface
- Syntax
- Working with .dta and .do files
- Reviewing contents of a dataset
- Simple descriptive statistics
- Dropping and keeping variables
- Using log files
- Creating new variables
- Variable manipulation
- ...and more

Introduction to NELS-88

A simple dataset we will use throughout the course is an extract from the National Education Longitudinal Study of 1988 (NELS-88) published by the National Center for Education Statistics.
`http://nces.ed.gov/surveys/nels88/`



National Education Longitudinal Study of 1988

NELS 88 Overview

A nationally representative sample of eighth-graders were first surveyed in the spring of 1988. A sample of these respondents were then resurveyed through four follow-ups in 1990, 1992, 1994, and 2000. On the questionnaire, students reported on a range of topics including: school, work, and home experiences; educational resources and support; the role in education of their parents and peers; neighborhood characteristics; educational and occupational aspirations; and other student perceptions. Additional topics included self-reports on smoking, alcohol and drug use and extracurricular activities. For the three in-school waves of data collection (when most were eighth-graders, sophomores, or seniors), achievement tests in reading, social studies, mathematics and science were administered in addition to the student questionnaire.

Example variables from NELS

Example variables from NELS—what level of measurement? Discrete or continuous?

- GENDER: 0 = male, 1 = female
- URBAN: 1 = urban, 2 = suburban, 3 = rural
- SCHTYP8: type of school attended, 8th grade 1 = public, 2 = private religious, 3 = private non-religious
- TCHERINT: agreement with "my teachers are interested in students," *Likert scale*
- NUMINST: number of post-secondary institutions the student attended

Example variables from NELS

Example variables from NELS—what level of measurement? Discrete or continuous?

- ACHRDG08: score on standardized test of reading achievement, 8th grade (ranges from 36.61 to 77.2)
- SES: socioeconomic status (ranges 0-35)
- SLFCNC12: self-concept score in 12th grade
- SCHATTRT: average daily attendance rate for the school student attended
- ABSENT12: number of times student missed school, categorical, 0 = never, 1 = 1-2 times 2= 3-6 times, etc.