# Introduction to Social Statistics

Alex Shpenev

May 31, 2017

# Statistics and Data

- Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data.
- Data are a set of values of qualitative or quantitative variables (usually presented in numerical form)

# Steps of Research Process

- Asking the Research Question
- Formulating Research Hypotheses
- Collecting Data
- Analyzing Data
- Evaluating the hypotheses

It is crucial to furmulate your hypotheses before data collection. Otherwise, our results might be not accurate (later about it in the following classes)

# Empirical Research

- Our hypotheses should be tested using direct experience
- Subjective judgment should be avoided

# Research Hypotheses

- Tentative answers to research questions
- Frequently postulated in the form of a hypothized relationship between two variables

# Variable

- A logical set of properties our observations have

Two properties:

- Exhaustiveness
- Mutual Exclusiveness

# Example of a bad variable

| Age | Population size |
|---|---|
| 0 - 10 | 356246 |
| 20 - 30 | 45634 |
| 40 - 50 | 43653 |
| 60 - 70 | 34554 |

Here the variable is not exhaustive.

# Example of a bad variable

Religion:

- Buddhist
- Muslim
- Sunni
- Shia
- Christian
- Catholic
- Protestant
- Methodist
- Other

Here the variable categories are not mutually exclusive

# Units of Analysis

Variables can be referring to different units

- Number of years of Schooling (individual)
- Family size (Family. Not a property of any individual)
- Fertility rate among African Americans (The property of a racial group, not any particular individual)
- A country's GDP in 2010 (country, not a property of any individual in that country)

# Dependent and Independent variables

- Dependent variables are variables that we are trying to explain. We are interested in what causes variation
- Independent variables are variables that we investigate to be associated (or sometimes even cause) the variation in the dependent variable

# Causality

- Causes precede effects
- There should be a plausible empirical link between causes and effects
- No other explanations should seem to be plausible

Nicholas Cage and Drowning in Pools
`http://tylervigen.com/spurious-correlations`

# Levels of measurement

- Nominal (Race, Religion, Gender). No intrinsic order (If Susan is Christian and Madhu is Hindu, we can't say that Madhu or Susan scores more in religion (religiousity would be a different variable though))
- Ordinal (Social Class, Self-Reported Health). Intrinsic order, but no way to quantitatively compare values (Ex: Susan is healthier than Eric. We don't know by how many times )
- Interval-Ratio (Income, Age). Intrinsic Order. Can say by how many units and by how many times observations differ (Ex: John is 5 years older than Mary. Mary is 2 times younger than John)

# Cumulative Property

- Interval-Ratio variables can be treated as ordinal or even nominal
- Ordinal variables can be treated as Nominal.

But the opposite is not true. Nominal variables can't be treated as ordinal or interval-ratio

# Exception

Dichotomous variables can be analyzed on any of the scale. Gender is a nominal variable, but treating it as ordinal or interval-ratio will not affect any results.

# Discrete and Continuous Variables

- Discrete variables can only take certain values. E.g. The number of kids has to be a non-negative integer. You can have 0, 1, 2... children, but you can't have 3.5 or $\pi$ children
- Contious variables can theoretically take any value within a given interval E.g. Age, weight, height. Note, however, that very frequently these variables are rounded to integers and might seem discrete.

# Validity, Reliability and Measurement Error

- Validity: does the measure correspond accurately to the real world? Asking "What caste are you?" in the US might be not valid because castes are not a valid category for the majority of the US population.
- Reliability: does the measure give consistent results each time? Asking "What caste are you?" in India might be not consistent because people are likely to interpret the question differently depending on the circumstances.

A nice illustration of reporting heterogeneity
`https://gking.harvard.edu/files/anhal.mpg`

# Population and Sample

- Population is the total set of all individuals. It might be not always feasible to collect information on everyone.
- Sample is a set of individuals that we collect information on to make inferences about the population.