Quantitative Data and Measurement

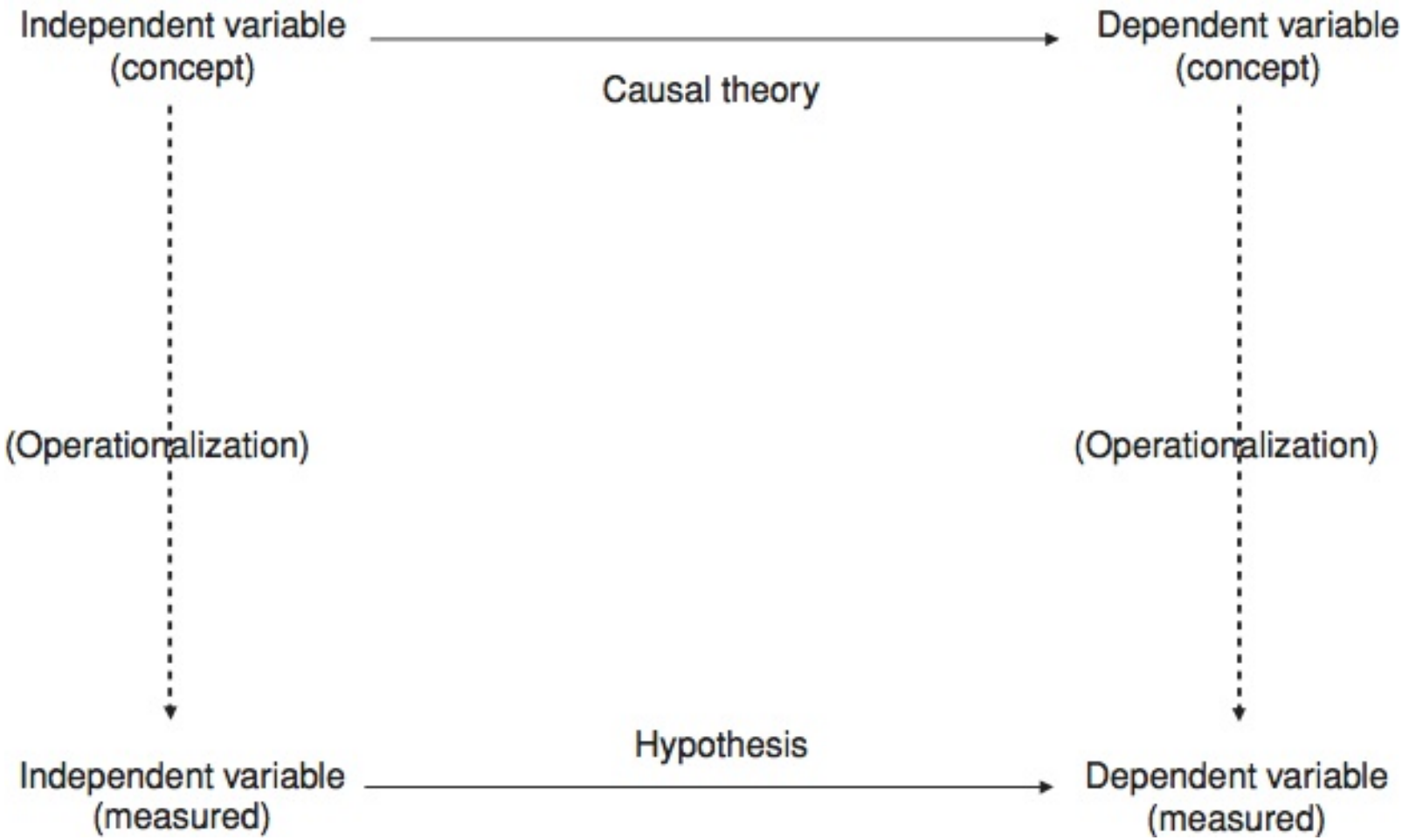
POLI 205 Doing Research in Politics

Fall 2015

Theory and Measurement

- We need to test our theories with empirical data
  - Inference
- **Measurement**: Systematic observation and representation of concepts
  - *Quantitative*: measures are numeric
  - *Qualitative*: measures based on the *qualities* that something possess
- **Problem of Measurement**: The need to be as confident as possible that our concepts in our theory correspond as closely as possible to our empirical observations (variables)
- Why is measurement important?

From Concepts to Variables



Problem of Measurement

- The relationship that we care about most is one we cannot directly observe. We therefore have to rely on potentially imperfect measures of the concepts we care about
- That means that measuring our concepts with care is one of the most important parts of social science

Conceptual Clarity

- What is the exact nature of the concept we're trying to measure?
- Example: How should a survey question measure "income"?
  - "What is your income?"
  - "What is the total amount of income earned in the most recently completed tax year by you and any other adults in your household, including all sources of income?"

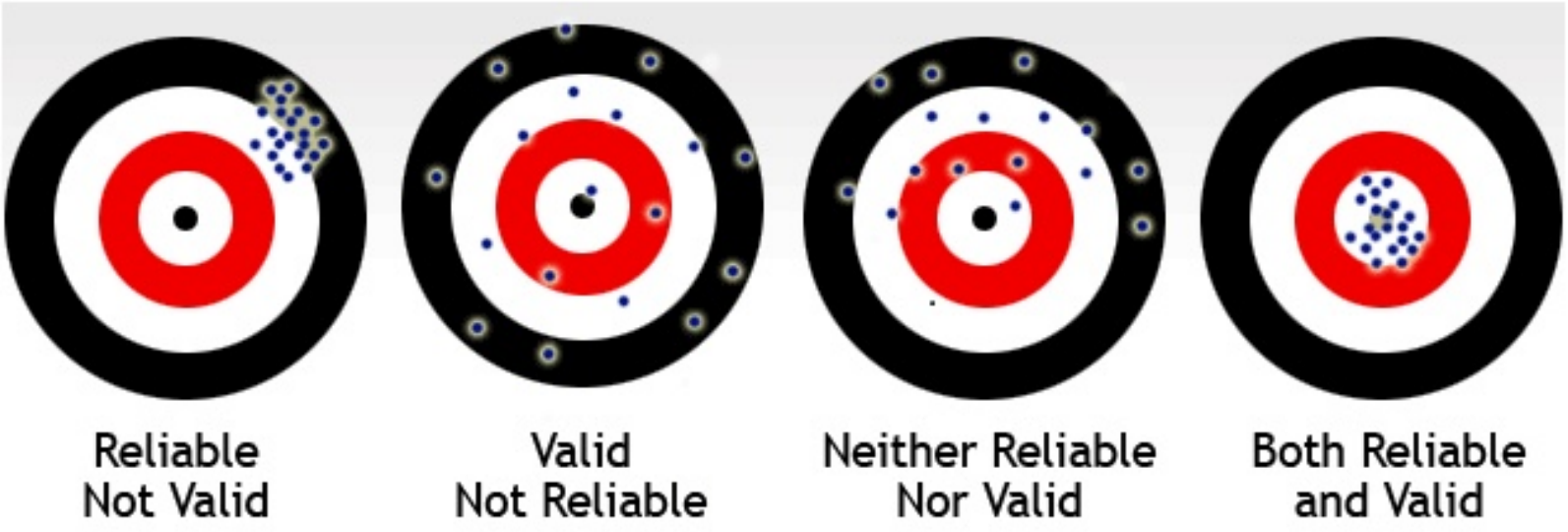
Reliability

- Applying the same measurement rules to the same case or observation will produce identical results
  - Example: The bathroom scale
- *Measurement bias*: Systematic over-reporting or under-reporting of values for a variable

Validity

- A *valid* measure accurately represents the concept that it is supposed to measure, while an invalid measure measures something other than what was originally intended
- Three types of validity
  - *Face validity*: measure appears valid on its face
  - *Content validity*: contains the essential elements of the concept
  - *Construct validity*: the measure is related to other measures in expected ways

Reliability and Validity



Types of Values

- **Variable**: A definable quantity that can take on two or more values
  - Labels: description of the variable
  - Values: denominations in which the variable occurs
- **Measurement Metric**: the *type of values* the variable takes on
- Three types of variables
  - Categorical
  - Ordinal
  - Continuous

Categorical

- Categorical variables are variables for which cases have values that are either different or the same as the values for other cases, but about which we cannot make any universally-holding ranking distinctions
- Example: "Religious Identification." Some values for this variable are "Catholic," "Muslim," "non-religious," and so on.
- *Qualitative* data

Ordinal

- Ordinal (*ranking*) variables have values that are either different or the same as the values for other cases, but we *can* make universally-holding ranking distinctions across the variable values for ordinal variables
- Ordinal variables do not have *equal unit differences*
- Example: Respondent's evaluation of Bush's handling of the War on Terror
  - -2 disapprove strongly
  - -1 disapprove not strongly
  - 0 don't know
  - 1 approve not strongly
  - 2 approve strongly

Continuous

- Continuous variables are variables that have equal unit differences
- Example(s): Age, Bush feeling thermometer
- *In analyses, we often treat ordinal variables as if they were continuous*

Descriptive Inference

- **Descriptive inference**: using observations about the world to learn about other unobserved facts
- **Descriptive statistics**: provide summaries of the *data*
- Types of descriptive statistics that are most relevant in the social sciences:
  - **Central tendency**: tell us about typical values for a particular variable.
  - **Variation or (dispersion)**: tell us the distribution (or spread, or range) of values that it takes across the cases for which we measure it.
  - **Rank/Order statistics**: summaries of values based on position in an ordered list of all values
  - **Moments**: provides information on the *shape* of a distribution

Central Tendency

- **The Mode**: the most frequently occurring value
- **The Median**: the value of the case that sits at the exact center of our cases when we rank them from the smallest to the largest observed values
- **The Mean**: the "average" value for the variable
  - $\bar{Y} = \frac{\sum Y_i}{n}$

Variance

- **Variance**: illustrates how a variable is spread or distributed around its mean
  - *Population* variance: σ_Y^2
  - *Sample* variance: $var(Y) = s_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$

Standard Deviation

- **Standard deviation (sd)**: average difference between values of Y (Y_i) and the mean of Y ($\bar{Y}$)
  - *Population* sd: σ_Y
  - *Sample* sd:

$$sd(Y) = sd_Y = s_Y = \sqrt{var(Y)} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

Rank/Order Statistics

- *Minimum Value*: The lowest value of a distribution
- *Maximum Value*: The highest value of a distribution
- *Median*: The value at the center of a distribution
- *Quartiles*: Divides the values into quarters
- *Percentiles*: Divides the values into hundredths

Moments

- Mean (Expected Value)
- Variance
- *Skewness*: A measure of the *asymmetry* of a distribution. When the mean and median of a variable are roughly equal, $\bar{Y} \approx Md_Y$, then the distribution is considered approximately symmetrical, $S = 0$
- *Kurtosis*: The kurtosis of a distribution refers to the the peak of a variable (i.e., the mode) and the number of observations in the tails. Higher kurtosis is indicative of a distribution where the variance is a result of low frequency yet more extreme observed values.

Categorical Data

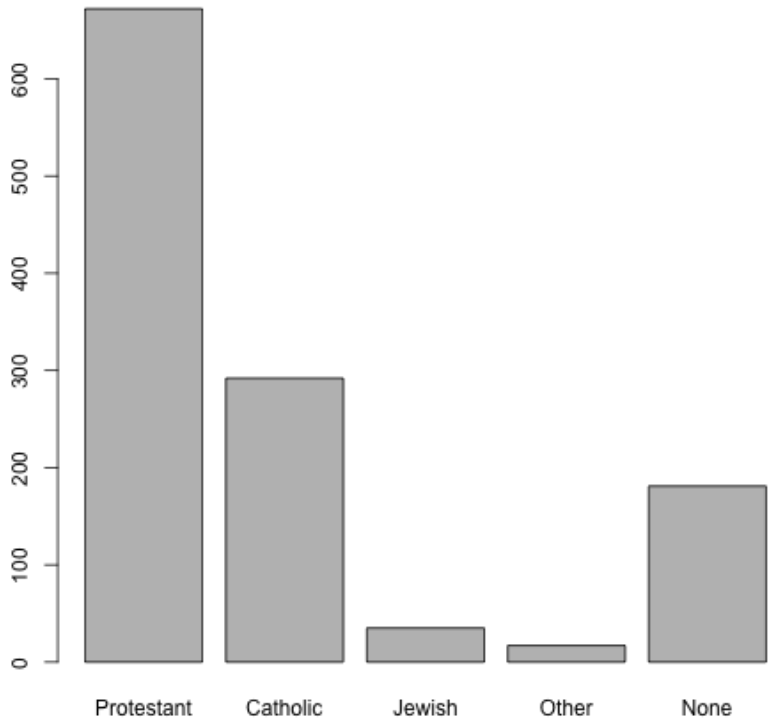
- The only measure of central tendency that is appropriate for a categorical variable is the *mode*
- Why couldn't we compute the median or mean?

Categorical Data: R example

- R Code:
  - `relig <- table(nes2004$religion)`
  - `relig`
  - `1  2  4  6  7`
  - `672 292  35  17 181`

Categorical Data: R example

- R Code:
  - `barplot(relig, names.arg=c("Protestant","Catholic","Jewish","Othe`

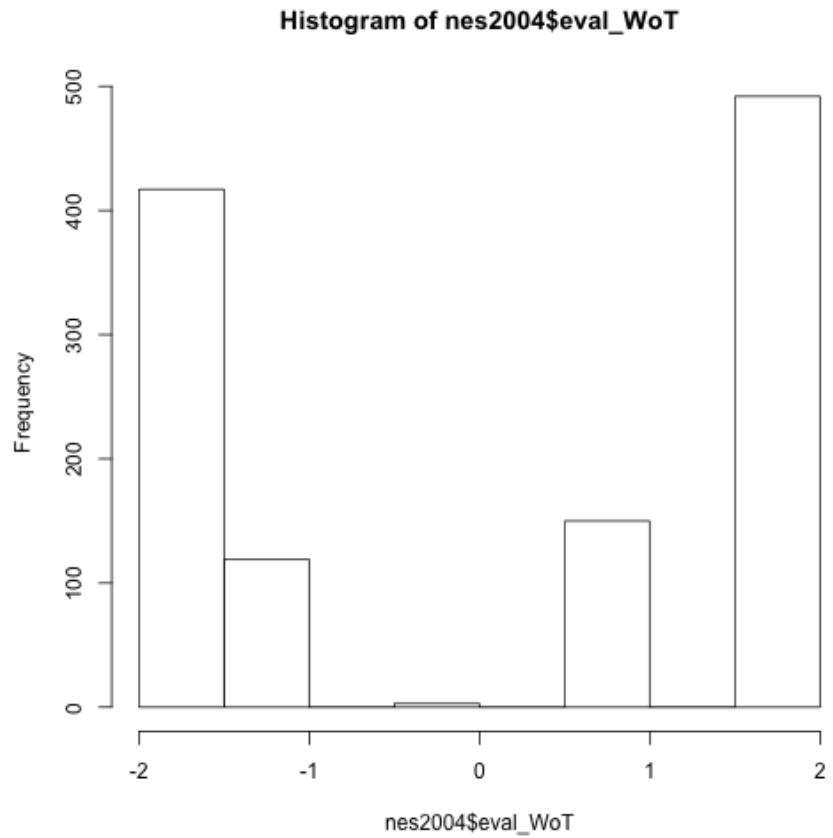


Continuous data

- With continuous variables, we want to know about the central tendency and the spread or variation of the values around the central tendency
- With continuous variables we also want to be on the lookout for *outliers*.
    - **Outliers** are cases for which the value of the variable is extremely high or low relative to the rest of the values for that variable

Continuous (Ordinal) data: R example

- R Code:
  - `hist(nes2004$eval_WoT)`



Continuous data: R example

- R Code:
  - summary(nes2004$bush_therm)
  - Min. 1st Qu. Median Mean 3rd Qu. Max. NAs
  - 0.0 30.0 60.0 54.9 85.0 100.0 5
  - mean(nes2004$bush_therm, na.rm=TRUE)
  - [1] 54.94
  - median(nes2004$bush_therm, na.rm=TRUE)
  - [1] 60
  - sd(nes2004$bush_therm, na.rm=TRUE)
  - [1] 33.55

Continuous data: R example

- R Code:
  - `hist(nes2004$bush_therm)`