PS0700
Basic Statistical Methods: Introduction to Descriptive Statistics

Political Science Research Methods

Professor Steven Finkel

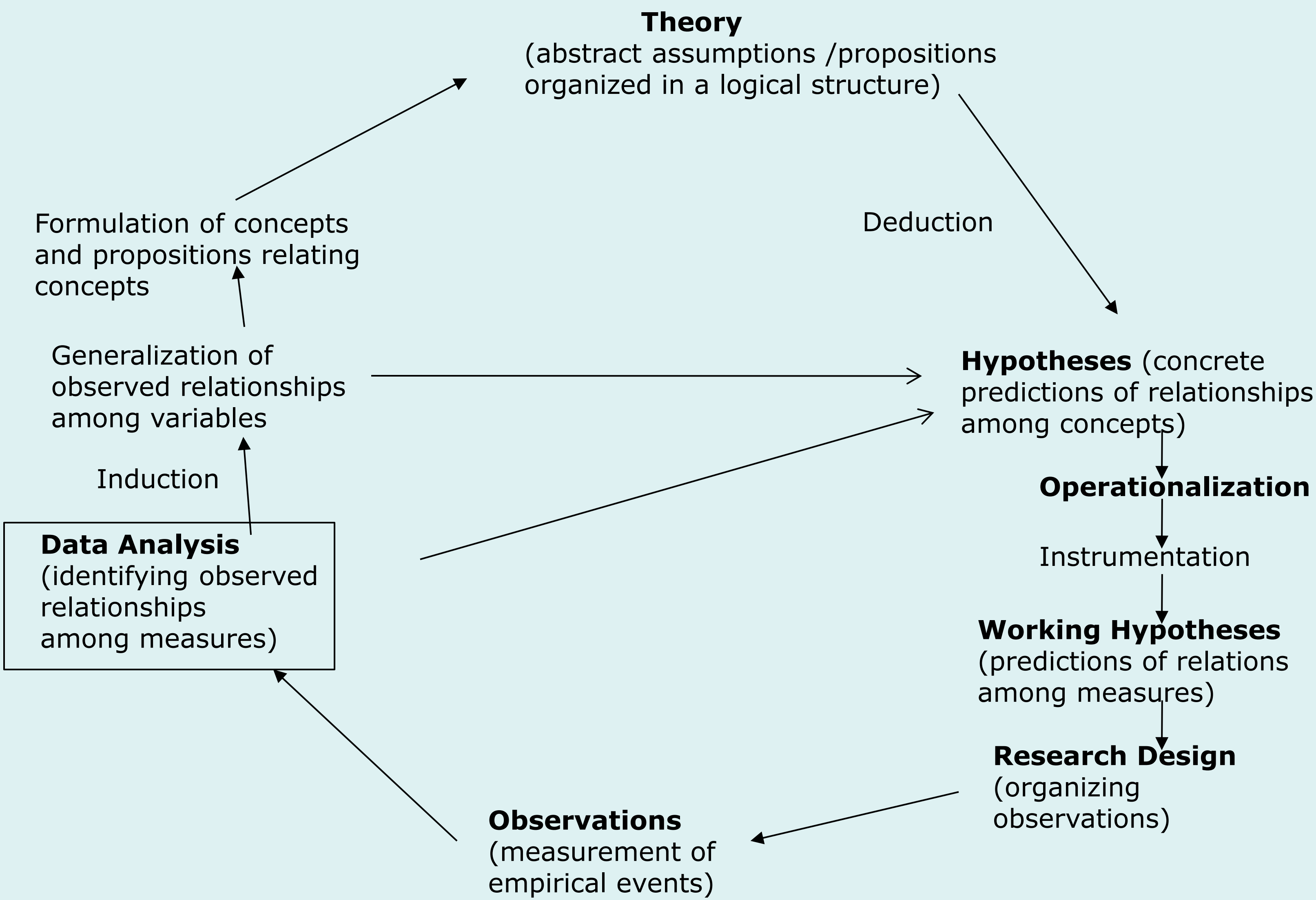
Fall Semester 2022

Week 9

Goals for the Sessions

- Discuss different general kinds of statistical analysis
- Provide introduction to descriptive statistics, including presenting and summarizing distributions
- Discuss measures of central tendency and measures of dispersion
- Discuss the "normal distribution" and its properties

A Model of the Research Process



Two General Branches of Statistics

Descriptive Statistics: Provides information about variables, and relationships between variables, in a given sample of observations

Inferential Statistics: Use information about sample *statistics* to make generalizations about the likely values of the *parameters* in the overall population.

(NOTE: correct inferences are only possible with ***random sampling***)

Descriptive Statistics

- Provide information on the distribution of responses, or “frequencies” for the various categories of a single variable
- Provide a sense of the overall shape of a distribution
- Provide measures of the “average” value of a variable
- Provide measures of the amount of spread or “dispersion” in a variable
- Provide information on the relationship between variables in a given sample

Remember the "Levels of Measurement" of Variables?

- Four Types of Measures:
  - **Nominal**: categories that have no intrinsic ordering or ranking
  - **Ordinal**: ordered or ranking of units is possible, but there is no fixed or meaningful distance between categories
  - **Interval**: ordered with equal distance between categories
  - **Ratio**: ordered with equal distance between categories *and* a meaningful zero point

  (Practically the distinction between Interval and Ratio is not that important)
- Examples:
  - Nominal
    - Religion (Protestant, Catholic, Jewish, Muslim, Other)
  - Ordinal Measures
    - Social Class (lower, middle, upper)
  - Interval-Ratio Measures
    - Age
    - Revenues

Basic Statistics for Frequency Distributions

- Frequencies
- Proportions
- Percentages
- Percentiles

See File: "PS0700.Descriptive Statistics Output.PDF"

Measures of Central Tendency: The "Average" Value of a Distribution

1. Mode: The Most Frequent Value in a Distribution
   - Suitable for Nominal Variables
   - May be calculated for ordinal, interval-ratio variables, but often not particularly meaningful or informative about "average" value
2. Median: The 50th Percentile in a Distribution, or the place where half the cases are above and half are below
   - *Not* suitable for Nominal Variables
   - May be calculated for ordinal, interval-ratio variables, usually highly informative about "average" value

Measures of Central Tendency (Continued)

- Mean: The Arithmetic Average of a Distribution, or ("the sum of Y sub i, over N")

$$\frac{\sum_{i=1}^{n} Y_i}{N}$$

where Y_i is the value of Y for a particular case *i*, and N is the total number of cases

  - Denoted as $\bar{Y}$ ("Y-bar")
  - Used mainly with Interval-Ratio Variables
  - *Not* suitable for Ordinal Variables

Measures of Central Tendency (Continued)

- Which Measure to Use?
  - Nominal Variables: only the mode
  - Ordinal Variables: mode or (more commonly) the median
  - Interval-Ratio: depends on "skewness" of distribution
    - Skewed positive: mean will be "unnaturally" large compared to median, gives distorted picture of "average" value
    - Skewed negative: mean will be "unnaturally" small compared to median, gives distorted picture of "average" value
    - Mode is generally unhelpful for interval-ratio variables
  - For inferential statistics, the mean is **almost always used** because of its attractive statistical properties

Population Versus Sample Means

- The mean of the overall *population* is denoted as **μ** (the Greek letter "mu").
- The mean of a *sample* is denoted as $\bar{Y}$

- The population mean is the arithmetic average of **all units in the overall population**; the sample mean is the arithmetic average **of all units in the selected sample**

Measures of Dispersion: The Amount of Spread or "Variance" in a Distribution

- How tightly packed are cases around the mean (or median)?
- How typical is the "typical" or "average" value of the distribution as indicated by the mean (or median)?
- More commonly used for interval-ratio variables, but some do exist for nominal and ordinal variables (which we will not cover!)

Interval-Ratio Variables: Two Less-Widely Used (but Important) Measures of Dispersion

- "Range": The difference between the largest and smallest values in the distribution
  - Gives some sense of spread of distribution but:
  - Highly sensitive to an extreme and possibly atypical value at high or low end of distribution
- "Interquartile Range": The difference between the value represented by the 75th percentile and the value represented by the 25th percentile
  - Eliminates extreme value distortion
  - Gives range within which 50% of the distribution lies
  - But loses exact information on where the outer 50% lies
  - Is used in "boxplots" to give a sense of the overall distribution, its range and general dispersion [We won't cover this further].

The Usual Measures: Variance and Standard Deviation

- Based on deviations or distances from individual cases from the mean
  - $Y_i - \bar{Y}$ ("Y sub i minus Y-bar")
  - The sum of these deviations is **always** 0
- *Variance* is based on *squared deviations* from the mean. It is the average squared deviation of a given case from the mean
  - Formula: $\sigma^2 = \dfrac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{N}$
  - Denoted by the lower case Greek letter "sigma" squared
  - Large Values means the average case is far away from the mean, small values means the average case is close to the mean

- The ***standard deviation*** (denoted as σ) is the square root of the Variance
  - Formula: $\sigma = \sqrt{\dfrac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{N}}$
  - Denoted by the lower case Greek letter "sigma"
  - In words:

    "the square root of the sum of Y_i minus $\bar{Y}$ squared, all divided by N"
  - What it means: **the square root of the average squared deviation of a given case from the mean, or, informally, it is the "average deviation" from the mean**
  - Advantage over the variance: It is expressed in the "raw units" of the given variable, not "squared units"
  - Otherwise same interpretation as variance

Issues Related to S.D. and Variance

- Difference between formulas for *population* and *sample* standard deviations and variances that we will discuss in the next weeks:

Population:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \mu)^2}{N}}$$

Sample:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{N-1}}$$

- SD Interpretations are best when made in relation to the mean, i.e. larger means usually have larger expected SDs and variances
- Can express how many SDs a given case is from the mean as what is known as a "z-score":

$$z_i = \frac{Y - \mu}{\sigma} \text{ for a unit in a population}$$

$$z_i = \frac{Y - \bar{Y}}{\hat{\sigma}} \text{ for a unit in a sample}$$

So a case with a z score of +1 is one SD *above* the mean, a z score of -1 is one SD *below* the mean

Can use Z scores to compare location of cases *within* distributions and even *across* different distributions

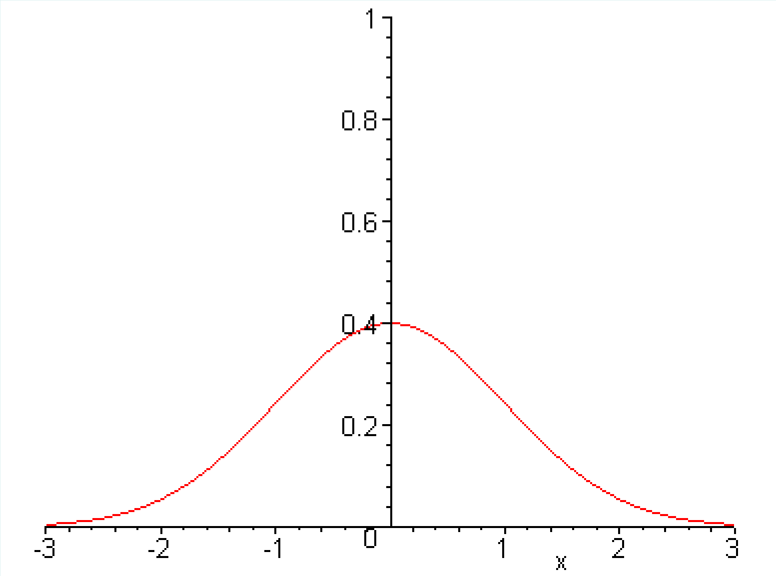
The Normal Distribution

- Most important distribution in statistics
- A Theoretical Distribution (with infinite number of cases) in which:
  - The single peak is the mean, median and mode
  - There is perfect symmetry as go toward either “tail” of the distribution
  - Is “Bell-shaped” according to the formula (which you don’t have to know but is given here in case you are interested):

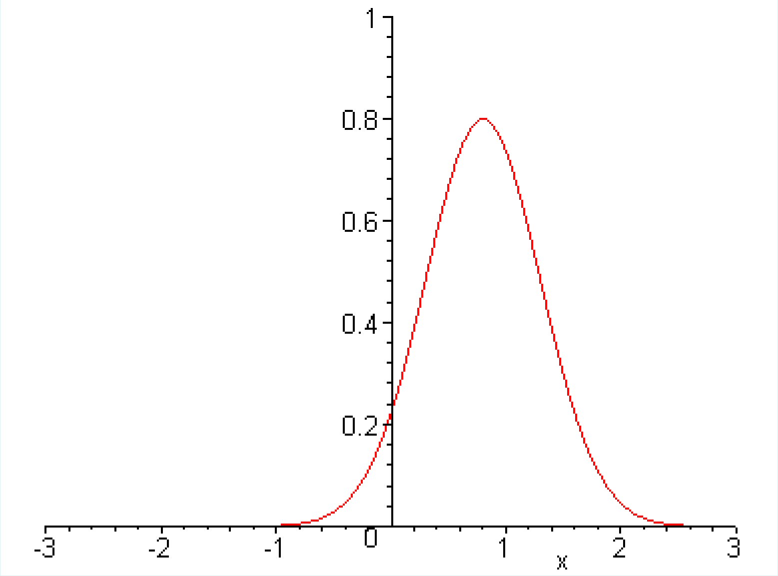
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2}$$

$$\text{for} \quad -\infty < x < \infty$$

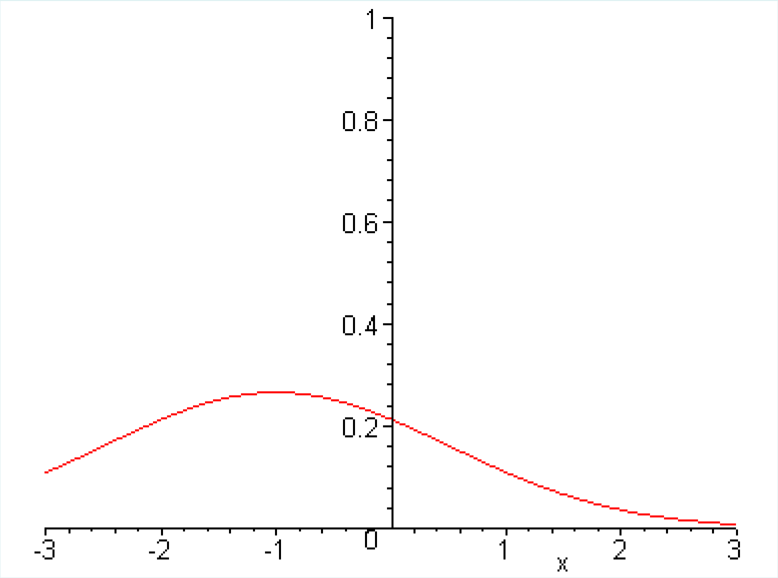
- The Normal Distribution is:
  - Perfectly defined by the mean and SD
  - Approximated by many empirical distributions
  - Used extensively in inferential statistics, as will be seen in next few sessions
- Examples of Graphs (Histogram) of Normal Distributions
  - Mean of 0, SD of 1: **"standard normal curve"**



- Mean of .8, SD of .5



- Mean of -1, SD of 1.5



Some Characteristics of the Normal Curve

- If you know the value of Y_i, then you also know:
  - Proportion of all cases *above* Y_i,
  - Proportion of all cases *below* Y_i (this is the "percentile" if expressed as percentage, as we discussed earlier)
  - Proportion of all cases *between* Y_i and the mean
  - Proportion of all cases *between* Y_i and another Y_i

- The proportion of all cases between Y_i and any other point is equal to the "area" under the normal curve that is set off by those two points [derived via integral calculus which we (thankfully!) will not concern ourselves with]

- The proportion of all cases between Y_i and any other point is equal to the *probability* of observing a case in that portion of the normal curve

Some Empirical Results

- 68% of all cases lie between $\bar{Y}$ $\pm$ 1 S.D.
- 95% of all cases lie between $\bar{Y}$ $\pm$ 1.96 S.D.
- 99% of all cases lie between $\bar{Y}$ $\pm$ 2.58 S.D.
- So in a normal distribution with Mean of 50 and S.D. of 10:
  - 68% of all cases between 40 and 60
  - 95% of all cases between 30.4 and 69.6
  - 99% of all cases between 24.2 and 75.8

- Could also do this via ***z-scores*** *(see slide 17)*:

  $Z_i = \mathbf{(Y_i - \mu) / \sigma}$, or, with sample notation:

  $Z_i = (Y_i - \bar{Y}) / \hat{\sigma}$

  So a z-score of 1 on this distribution is:

  $1 = (Y_i - 50) / 10$, and $Y_i = 60$

  And a z-score of -1 on this distribution is:

  $-1 = (Y_i - 50) / 10$, and $Y_i = 40$

  Therefore, $\pm$ 1 S.D on this distribution is between 40 and 60, encompassing 68% of all cases

- These calculations have been worked out for ALL possible z-scores in any normal distribution!!!!!

Problem Solving with Normal Distribution

- What proportion of cases lies between a z-score of 2.25 and the mean?

  - Look up Z of 2.25 in Normal Curve Table and see the value .012. What does this mean? That 1.2% of all the cases are *above* a z-score of 2.25.

  - Since 50% of the cases in general are above $\bar{Y}$, this means that 48.8% of all cases are between a z of 2.25 and $\bar{Y}$ (50%-1.2%). In proportion terms, it is .488.

  - This also means that 98.8% of all cases in a normal distribution are *below* a z-score of 2.25, or that a z of 2.25 is associated with the 98.8th percentile

- On our distribution with $\bar{Y}$ of 50 and SD of 10, this means that the value associated with a z of 2.25 is

  $2.25 = (Y_i - 50) / 10$, and $Y_i = 72.5$

  So 48.8% of all cases lie between 50 and 72.5

- Negative Z scores are exactly the same because the normal distribution is perfectly symmetric

  $-2.25 = (Y_i - 50) / 10$, and $Y_i = 27.5$

  So 48.8% of all cases lie between 50 and 27.5, and 1.2% lie below 27.5

Exercise

- Assume the population of student incomes is normally distributed with a mean of 10000 and a standard deviation of 2000
  - What proportion of this population has an income above 11000?
    - Solution:
      1. Find out what the z-score is that corresponds to 11000 on this normal distribution
      2. Look on z table for the proportion of cases above that z-score
  - What proportion has an income below 9500?
      1. Find out what the z-score is that corresponds to 9500 on this normal distribution. This will be a negative number.
      2. Look on z table for the proportion of cases above the *positive equivalent* of that number. The same proportion are *below* a negative z-score as are *above* a positive z-score of the same magnitude!

- Assume a university only accepts those in the upper 15% on their SAT quantitative scores, which are normally distributed with a mean of 500 and a SD of 100. You scored 600. Is that good enough to get in?