Probability and Inference

POLI 205 Doing Research in Politics

Fall 2015

Population versus Sample

- **Population**: data for every possible relevant case
- **Sample**: a *subset* of cases that is drawn from an underlying population
- Inference

Parameters and Statistics

- A parameter is a value, usually unknown (and which therefore has to be estimated), used to represent a certain population characteristic.
- Within a population, a parameter is a fixed value which does not vary. Each *sample* drawn from the population has its own value of any statistic that is used to estimate this parameter.

Parameters and Statistics

Concept	Sample Statistic	Population Parameter
Mean	$\bar{X} = \frac{\sum X_i}{n}$	$\mu_X = E(X)$
Variance	$s_x^2 = \frac{\sum (X - \bar{X})^2}{(n-1)}$	$\sigma_x^2 = Var(X)$
Standard Deviation	$s_x = \sqrt{\frac{\sum (X - \bar{X})^2}{(n-1)}}$	$\sigma_x = \sqrt{Var(X)}$

Figure: Sample and Population Notation

Samples

Probability Samples

- **Probability sample**: A sample for which each element in the total population has a *known* probability of being included in the sample
  - *Random* sample: each member of the population has an *equal probability of being selected*
  - *Systematic* sample: the K th element is selected
  - *Stratified* sample: elements sharing one (or more) characteristics are grouped and selected by proportion to the population
  - *Cluster* sample: initially samples based on clusters (generally geo- graphic units, such as census tracts) and then samples participants within those units

Samples

Nonprobability Samples

- **Nonprobability samples**: A sample for which each element in the total population has a *unknown* probability of being selected
  - *Purposive* sample: researcher exercises considerable discretion over what observations to study
  - *Convenience* sample: elements are included because they are convenient for a researcher to select
  - *Snowball* sample: respondents are used to identify other persons who might qualify for inclusion in the sample

Defining Probability

- **Probability**: tells us how likely something is to occur
  - All outcomes have some probability ranging from 0 to 1
  - The sum of all possible outcomes must be exactly 1
- **Outcome**: the result of a random observation
  - *Independent* outcomes: the realization of one of the outcomes does not affect the realization of the other outcomes. The probability of those events both occurring is equal to the *product* of them individually.
    - Example: probability of three tails in a row, $1/2$ X $1/2$ X $1/2 = 1/8$

Types of Probabilities

- *Simple* probability: number of ways your outcome can be achieved over all possible outcomes
  - Example: Rolling a 2 on a six-sided die, 1 over $6 = .167$
- *Conditional* probability: the probability of some event A, given the occurrence of some other event B
- *Joint* probability: tells the likelihood of two (or more) events both occurring

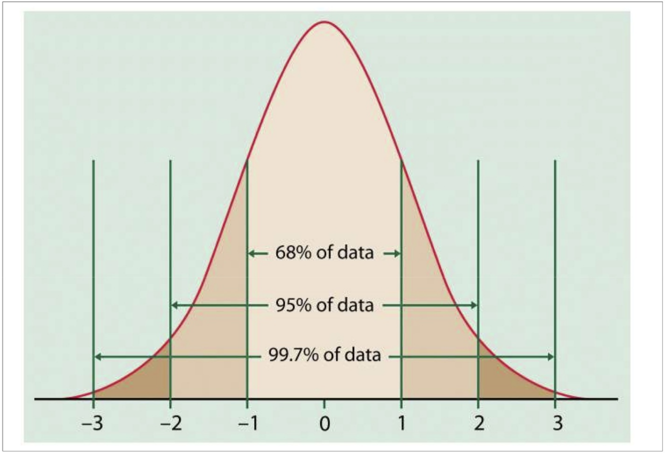
Normal Distribution

- There are some things (like the mean) that we can know (with certainty) about a sample. But we care about the population. How can we learn about the population from a sample?
- The Central Limit Theorem will invoke a particular kind of distribution called the *normal distribution*

Properties of the Normal Distribution

- It is symmetrical around its mean and median, μ
- The highest probability (aka "the mode") occurs at its mean value
- Extreme values occur in the tails
- It is fully described by its two parameters, μ and σ
- If a distribution is normally shaped, we know a certain % of cases fall within a certain distance of the mean
- The standard normal distribution has a $\mu = 0$ and $\sigma = 1$

Normal Distribution Plot



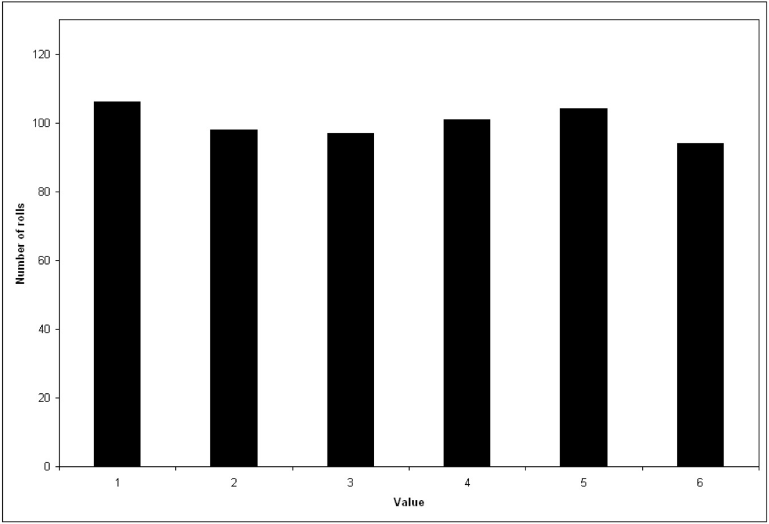
Frequency Distribution

- **Frequency distribution**: The distribution of actual scores in a sample
- Most frequency distributions are not normally shaped
- Even if a frequency distribution is not normally shaped, if we imagine a (hypothetical) world in which we took an infinite number of samples, and took the mean of each sample, and then plotted those means, then how would those plotted means be distributed?

Example: Dice

- Imagine that we rolled a six-sided dice, it can come out as a 1, 2, 3, 4, 5 or 6 with equal probability
- Let's say you rolled that dice 600 times. What would that distribution look like?

Example: Dice



Example: Dice

- Let's say we rolled that dice 600 times. What do you think the mean would be (about)?
- Would it be exactly 3.5? Every time?
- But what would happen if we rolled it a billion times, then plotted the means?

Example: Dice

- **It would be normal**:
  - In our frequency distribution, we could get a score of 1 to 6 with equal likelihood. But in our sample means, we would never get means of 1 or 6. All of our means would be somewhere around 3.5. Moreover, they would be distributed around that mean (3.5) normally

Central Limit Theorem

- The *Central Limit Theorem* says that, no matter what the underlying shape of the frequency distribution (whether it's uniform, normal, or whatever), the *sampling distribution*– the hypothetical distribution of sample means – will be normal, with mean equal to the true population mean, and standard deviation equal to the *standard error of the mean*
- $$\sigma_{\bar{Y}} = \frac{S_Y}{\sqrt{n}}$$