POL3371

Quantitative Analysis for Political Science
Lecture 02: Descriptive Statistics

Today's Lecture

1. Preliminaries
2. Justification
3. Terminology
4. Variables
5. Central Tendency
6. Variance and Standard Deviation

Preliminaries

- If you have not done so, please sign into Lore.com
  - It will be hard to retrieve materials and submit your assignments if you don't have an account.
- Course code = 66XKTY

Preliminaries

- Newly arrived students:
  - Please collect a syllabus from me.
  - See me after class if you have questions.

Justification

- Why quantitative analysis?
  - Allows us to compare variables in the political and social world.
  - Allows us to measure important features in the political and social world.
  - Allows us to test hypotheses about the political and social world
  - Allows us to draw inferences about the political and social world.

Justification

- Quantitative versus Qualitative?
  - Quantitative analysis is often used to supplement qualitative analysis in the study of politics (and vice versa)
  - Same fundamental approach to knowledge.
  - Anything comparable can be quantified
  - Quantity makes analysis reproducable

Terminology

- Population: everything/everybody of interest in a study
- Sample: a subset of the population that we actually look at.
    - May be a random or non-random sample.

Terminology

- Descriptive Statistics
  - Quantitatively describes the main features of a collection of data.
    - Not based on probability theory.
- Inferential Statistics
  - Used to draw conclusions or generalizations about the characteristics about a population from a sample.
  - Based on probability theory.

Terminology

- Parameters: the characteristics of the population
    - eg. average age of the entire student body at University of Ottawa
- Statistics: the characteristics of the sample.

Terminology

- Variable:
  - Measurement of a characteristic of a subject that can vary across subjects in a population of subjects.
    - eg. age; gender; opinion about same sex marriage;
- Levels of Measurement
  - The different nature of variables means that we have to examine different types of data in different ways

Let's Play

- Form groups of two or three students with your immediate neighbours.
- Take five minutes to talk about these questions

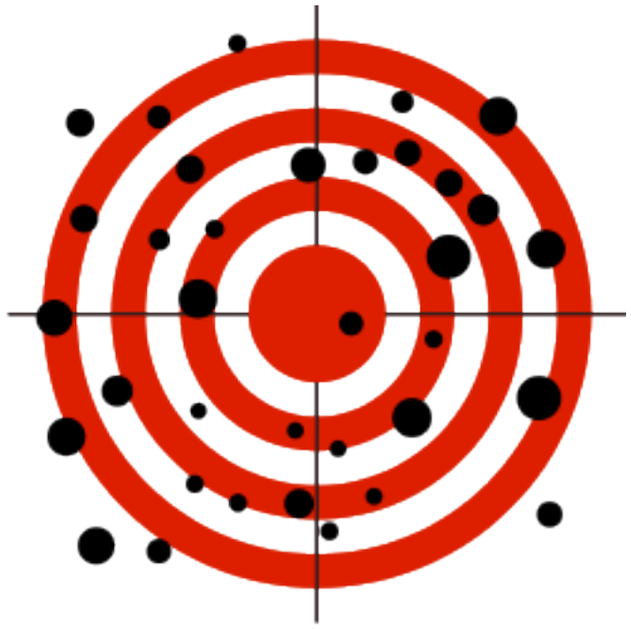
Let's Play

- Measurement
  - How would you measure the ideological position of a Supreme Court of Canada judge?
  - How would you measure the ethnicity of the Canadian people?
  - How would you measure prime ministerial power?
  - How would you measure inequality?

Let's Play

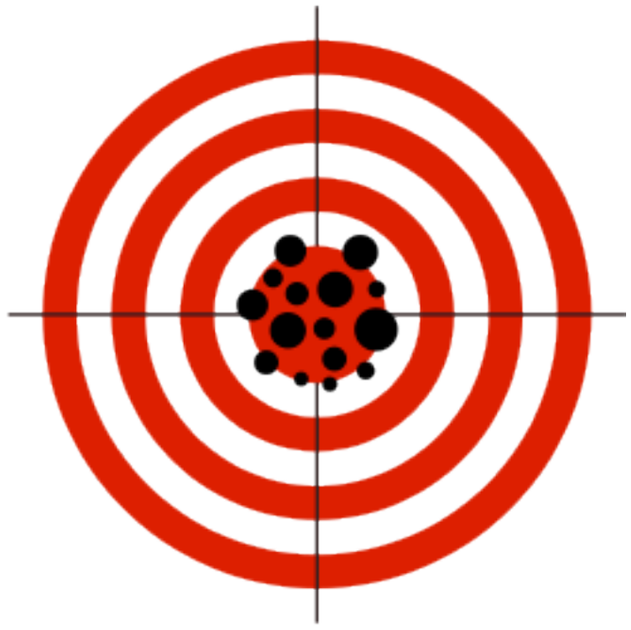
- Important considerations re: measurement
  - Reliability:
    - Do I get the same result if I do the test/analysis over and over again.
  - Validity:
    - Is the measure and the concept connected

Unreliable & Unvalid



Unreliable, But Valid

Reliable, Not Valid



Both Reliable & Valid

Variables and Measurement

- Types of Variables
- Categorical Variables
- Numerical Variables

Variables and Measurement

- Nominal Level Measures
- Ordinal Level Measures
- Interval Level Measures

Variables and Measurement

- Nominal Level Measures
  - Represents a category
    - Gender
      - Male
      - Female
    - Party Affiliation
      - Conservative Party of Canada
      - Liberal Party of Canada
      - New Democratic Party
      - Green Party

Variables and Measurement

- Nominal Level Measures
  - Nominal variables can be quantified by tabulating them:
  - Normally represent nominal data in a simple table with percentages.

2012 House of Commons

Party	#	%
PCP	163	52.92
NDP	100	32.47
LPC	35	11.36
BQ	4	1.30
GP	1	0.32
Ind. Cons.	1	0.32
Ind.	1	0.32
Vacant	3	0.97
Total	308	100

Variables and Measurement

- Ordinal Level Measures
  - Like nominal variables, dealing with categories.
  - But this time the categories are ordered.
    - eg. Likert scale
      - Strongly agree; somewhat agree; somewhat disagree; strongly disagree
        - Usually includes a “don’t know” category.

Variables and Measurement

- From the 2008 Canadian Election Study

The Liberal Party's Green Shift will really hurt the Canadian economy.

Strongly agree: 16.8%
Somewhat agree: 22.9%
Somewhat disagree: 26.0%
Strongly disagree: 13.4%
Don't know: 20.0%

Variables and Measurement

- Ordinal Level Measures
  - N.B. The distance between each catagory is unknown
    - Why is this important?

Variables and Measurement

- Interval Level Measures
  - Numbers represent a quantitative variable
    - Age, Income, Votes, ministerial resignations
    - There is a specific distance between each level
      - Can say that Harper has sacked fewer ministers in his term than Martin in his first term and we can say that he sacked ?? fewer ministers than Martin.
        - This is an example of a continuous variable

Variables and Measurement

- Interval Level Measures
  - Can also say that Harper had ?? ministers in his first cabinet
  - This is an example of a discrete variable
    - You can't have 15.25 minister! Has to have 1, 2, 3 ... ministers.
- Most of what we talk about today (and on Thursday) will focus on interval level measures.

Central Tendency

- We are dealing with descriptive statistics.
  - We want to describe a large amount of data/information in as parsimonous a manner as possible.
- Why bother?

Central Tendency

- We want to reduce a lot of interval level measurements to a few numbers

- Here are the salaries of my best friends (population = Kerby's best friends)

Name	Salary
Stephen	315462
Thomas	233247
Nycole	186151
Bob	211425
David	157731

Central Tendency

- The Mean
  - The most common way of measuring the central tendency is to use the mean (or average)
  - Sum of the measurements devided by the number of observations
  - So, for my best friends:

$$\frac{315426 + 233247 + 186151 + 211425 + 157731}{5}$$

Mean = 220803.2

Central Tendency

- Put another way:
  - Suppose we have *n* observations, with each value denoted by X_1, X_2 and so on until X_n. Then the mean is described as follows:

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

or

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Central Tendency

- Good things about the mean
  - Shift of origin of measurement
    - If my friends all get a $2000 bonus or a $2000 deduction then the new mean salary is just old mean salary plus or minus $2000
  - Change of scale
    - If Canada suddenly adopts the US dollar as the national currency then the new mean salary is simply the old mean salary multiplied by 1.0288.
- Sum of two variables
  - Imagine that income = salary + bribes
  - Mean income = mean salary + mean bribes

Central Tendency

- The median
  - If we rank all the observations in ascending order the median is just the observation in the middle (1/2 observations above and 1/2 below)

157731; 186151; 211425; 233247; 315426

Median =211425

Central Tendency

- Good things about the median
  - Shift of origin of measurement: YES!
  - Change of Scale: YES!
  - Sum of two variables: NO!
    - This is why we usually use the mean and not the median in most statistical analysis.
- But, it's not as sensitive to outliers as the mean!

Central Tendency

- Good things about the median
  - Imagine what would happen if Stephen made $2.5 million rather than $315,462

$$\frac{2500000 + 233247 + 186151 + 211425 + 157731}{5}$$

Mean = $657710.8 vs. $220803

Median = $211425

Central Tendency

- Median vs. Mean?
  - When the distribution is highly 'skewed', the median works best.
    - eg. 2, 3, 5, 7, 9, 10, 125
    - Mean is 23; Median is 7
    - Which better describes central tendency in the data?
    - Take out the extreme value (125) and the mean is 6 and the median 6

Central Tendency

- The median and ordinal level data
  - The median can be used for ordinal level data too
  - Ask five imaginary friends about position on Canada pulling the embassy out of Iran.
    - 2 strongly agree, one agreed, one disagreed and one strongly disagreed
  - We can rank these answers and find the median

Strongly agree; strongly agree; agree; disagree; strongly disagree

Median = Agree

Central Tendency

- Mode
  - Can't use median or mean for nominal data
  - Instead, use the mode.
    - This is the most commonly occurring value
    - Looking back at the parties in the House of Commons...

2012 House of Commons

Party	#	%
PCP	163	52.92
NDP	100	32.47
LPC	35	11.36
BQ	4	1.30
GP	1	0.32
Ind. Cons.	1	0.32
Ind.	1	0.32
Vacant	3	0.97
Total	308	100

2012 House of Commons

Party	#	%
PCP	163	52.92
NDP	100	32.47
LPC	35	11.36
BQ	4	1.30
GP	1	0.32
Ind. Cons.	1	0.32
Ind.	1	0.32
Vacant	3	0.97
Total	308	100

Central Tendency

- One special case where we can use the mean for nominal data
  - Binary data, which is coded as '0' or '1'
  - eg. Sex: Male = 0; Female = 1
  - Mean score for 0s and 1s is the proportion of women.

$$\text{Mean} = \frac{0+0+1+1+1}{5} = 0.6 = 60\%$$

Team up with your group again...

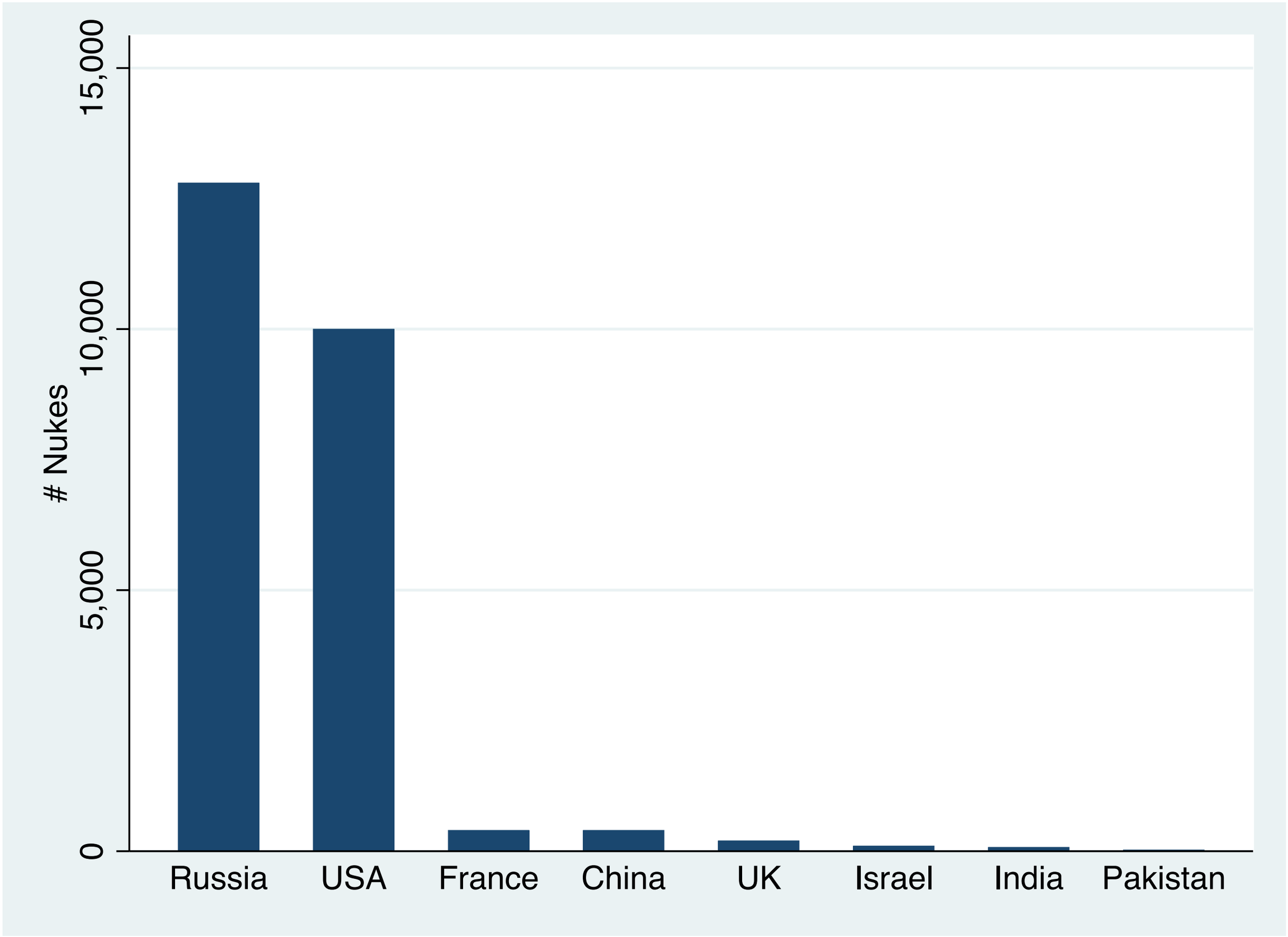
Exercise

- Population is all countries with nuclear capability; variable is approximate number of nuclear weapons.
- Calculate the mean, mode and median for the number of nuclear weapons
- How good is each of these at summarizing the data? Do we need more information than just a measure of central tendency?

Country	# Nukes
USA	10,000
India	75
China	400
France	400
UK	200
Russia	12,800
Israel	100
Pakistan	25

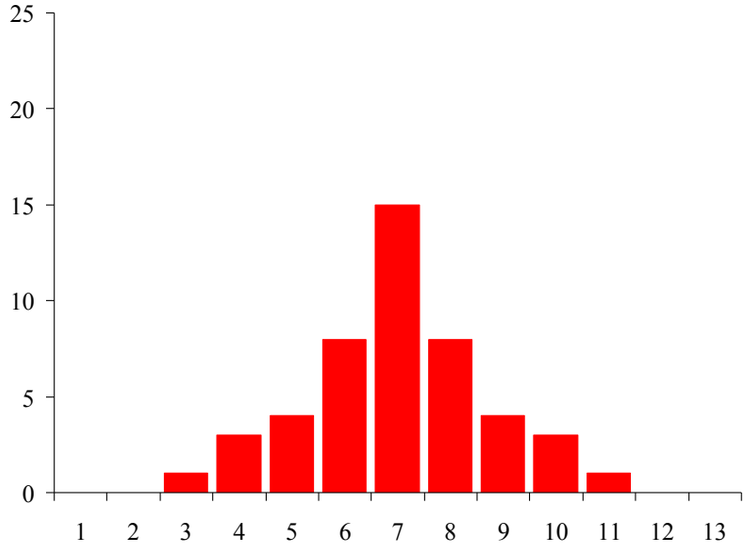
Answers

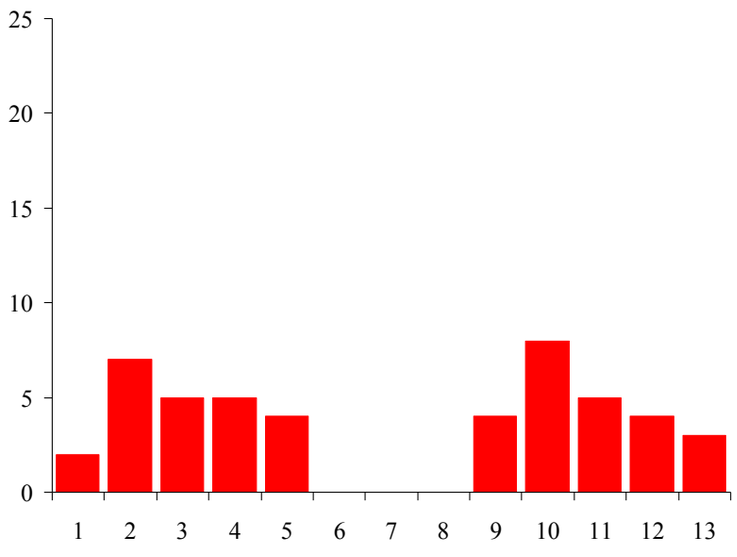
- Mean = 24000/8 = 3000
- Median = (400+200)/2 = 300
- Mode = 400
- These measures of central tendency are useful, but we need to know something about the distribution.
- Why? Because two countries account for almost all the nuclear weapons in the world.

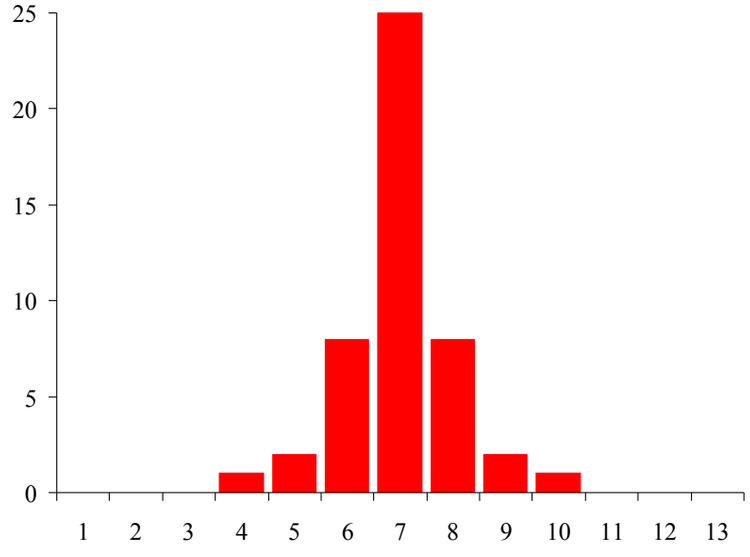

Nukes
15,000
10,000
5,000
0
Russia
USA
France
China
UK
Israel
India
Pakistan

Measures of Dispersion

- The mean (or median) tells us something about the centre of the distribution, but what about its dispersion
- The means/median of these distributions of children's scores on math tests in three different classes are all the same (48 obs, mean of 7, median of 7), but each tells a different story.







Measures of Dispersion

- Measures of dispersion tell us how widely, tightly and where the data is dispersed.
- There are a number of methods:
  - Range
  - Standard Deviation
  - Variance

Measures of Dispersion

1.3, 2.1, 9.6, 10.7, 11.2, 12.4, 13.2, 13.7

- The range of data is simply the distance between the lowest and the highest value

- You should sight those values when reporting the range

- eg. Range = 12.4 (1.3, 13.7)

- Problem: Ignores any variability in the middle range.

Measures of Dispersion

- Standard deviation
  - This is the most common and useful form of summarizing dispersion
  - It is also an integral part of many statisitical methods that we will move on to: remember it!
  - $\sigma = \sqrt{\sum(x- \mu)^2 / N}$ for the population
  - $s = \sqrt{\sum(x- \text{x-bar})^2 / n-1}$ for a sample
  - We know that 68 percent of the data is accounted for within ±1 SD of the mean, and 95 percent is within ±2 SDs

Measures of Dispersion

- Variance
  - Closely related to the Standard Deviation (SD)
    - It's the square of the Standard Deviation
    - $\sigma^2 = \sum(x- \mu)^2 / N$ for the population
    - $s^2 = \sum(x- \text{x-bar})^2 / n-1$ for a sample

Open Stata