Introduction to Empirical Methods: RBSI

Lecture 1, Part 2: Measurement and Research Design

The procedures we use to draw inferences (test hypotheses)

- What data are relevant?
- What comparisons do we need to make?
  - A good research design ensures that the data we examine will allow us to draw inferences and answer our research question.

Purpose of Research Design

- To ensure that the inferences we draw are valid.
  - To achieve this goal, good designs:
    - Account for threats to valid inference
    - Identify an appropriate unit of observation
    - Identify an appropriate temporal-spatial domain

Cornerstone

- Comparison is the cornerstone of research design
  - Compare values of Y across space (e.g., individuals, states, countries)
  - Compare values of Y over time (e.g., decades, years, months)
- The design identifies the comparisons we will make

Causal Process Must Be the Same Across Units

- We must compare two or more units to assess if we may infer that x causes y
  - Does rainfall (x) influence turnout (y)?
    - Is the turnout greater in a precinct that received no rain compared to a precinct that received rain?
- Answer the "How Else?" question.
  - Were those two precincts different in ways besides rainfall that might be important?

How Do We Compare?

- Maximize Comparability: the units should be identical except for the independent variable of interest
  1. Random assignment of units to value of x
  2. Statistical control
  3. Matching by selecting cases into dataset by finding cases that differ only on key x

Example: Comparing over Time

Figure: Traffic Deaths: 1955: 324; 1956: 284

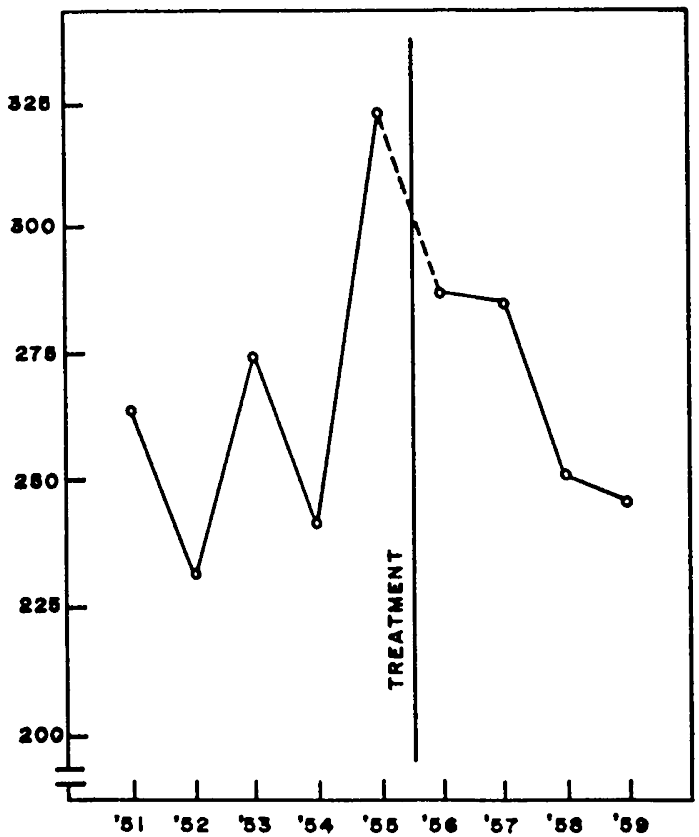


Figure 2. Connecticut Traffic Fatalities, 1951-1959

Did the Program *Cause* Fatality Decline?

- Internal Validity Defined:
  - Does the change in our independent variable really cause the change we observe in the values of our dependent variable?

Threats to Internal Validity

- History (concurrent events other than treatment)
- Maturation (passage of time)
- Testing ("pre-test"; events before experiment)
- Regression to the mean (free throw example)
- Selection (choosing y)

How do we Improve Internal Validity?

- That is, how can we isolate a causal effect?
  - Experimental designs seek to ensure that only X changes
  - Manipulate X while holding all other possibly relevant variables constant.
- The goal is to rule out all rival explanations for change in Y, except the change in X.

Standard Experimental Design

- Randomly divide subjects into two groups:
  - A treatment group and a control group.
- Do not present control group with the stimulus (i.e., they get a different value of X).
- Measure Y in each group afterwards.
- Any difference in Y across the two groups was caused by the treatment.
  - Why? Because the random assignment to the groups makes it very, very likely that the groups are, on average, the same in every way except for the stimulus, assuming large enough groups.

Key Elements of Experimental Design

- Random Selection
  - Each case has the same chance of being in the experiment.
- Random Assignment
  - Each case has the same chance of being assigned to control or treatment.
- Used together, random selection and assignment ensure that the groups are equivalent in all ways except the value of X, again assuming large enough groups.

The Beauty of Randomization

- Randomization makes the two groups identical (on average) in all ways *except for the treatment*.
  - Eliminates the threat of a spurious relationship.
  - Controls for observables:
    - Things we could measure, but have not!
  - Controls for unobservables:
    - Differences between units we cannot measure!

Challenges that limit use of Experimental Design

- The value of many independent variables cannot be randomly assigned.
  - Examples: Gender, religion, war
- Differences between lab and actual world.
  - External Validity
- Convenience samples (i.e., the sample we can get)

Experimenting in Connecticut

- Connecticut's crackdown not a true experimental design.
- Treatment not randomly assigned
  - Perhaps high fatalities led to program
- How could we design a standard experiment to test the effect of speeding on fatalities?
  - Are there ethical problems?

Additional Experimental Designs

- Field Experiment:
  - Randomly assign individuals into groups, but perform the manipulation in the real world.
- Natural Experiment:
  - An event outside the social scientist's control separates people into "control" and "treatment" groups.

Field Experiment Example

- What is the best method of getting out the vote?
  - Door to door canvassing
  - Telephone calls
  - Direct Mail
- Random assignment of X, but...
  - Because people's homes are not randomly distributed across neighborhoods, assignment not truly random.

Natural Experiment Example

- Newspapers and Political Information
  - Pittsburgh newspaper strike
- Random assignment of X, but...
  - Because people's homes are not randomly distributed across cities, assignment not truly random.

Must Cross the 4 Hurdles

- When we cannot conduct experiments, we collect data as they occur and study them, but the logic of inference is precisely the same.
- There are, however, different challenges to valid causal inference.

What is an Observational Study?

- Take world "as it is" and study naturally occurring differences between units
- Two types of Observational Studies:
  - Cross-Sectional
    - Many units sampled over a single time period
  - Time-Series
    - Single unit sampled over many time periods (Connecticut speeding)

Data Set Structure

- Cross-Sectional Studies
  - Variation is across space
- Time-Series Studies
  - Variation is over time

Cross-Sectional Data

Figure: Debt and Unemployment

Nation	Debt as % of GDP	Unemployment
Finland	6.6	2.6
Denmark	5.7	1.6
United States	27.5	5.6
Spain	13.9	3.2
Sweden	15.9	2.7
Belgium	34.0	2.4
Japan	11.2	1.4
New Zealand	44.6	0.5
Ireland	63.8	5.9
Italy	42.5	4.7
Portugal	6.6	2.1

Time-Series Data

Figure: Approval and Inflation

Month	Presidential Approval	Inflation
January 2002	83.7%	1.14
February 2002	82.0%	1.14
March 2002	79.8%	1.48
April 2002	76.2%	1.64
May 2002	76.3%	1.18
June 2002	73.4%	1.07
July 2002	71.6%	1.46
August 2002	66.5%	1.80
September 2002	67.2%	1.51
October 2002	65.3%	2.03
November 2002	65.5%	2.20

Observational Example

- What is the effect of incumbency on Democratic vote share in US House elections?
- Ideally:
  - Examine the vote share in a district with a Democrat seeking re-election.
  - Go back in time to the same district and now make the Democrat not an incumbent.
  - The difference between the two vote totals would be the effect of incumbency.
- Collect time-series data on elections:
  - % of Democratic candidates who were incumbents
  - % of Democratic seats after the election
- Compare across the elections

Comparing Vote Share

- 2008
  - With Democratic Incumbent: 69.85%.
  - No Democratic Incumbent: 40.97%.
- 2006
  - With Democratic Incumbent: 71.89%.
  - No Democratic Incumbent: 41.80%.
- 2004
  - With Democratic Incumbent: 69.20%.
  - No Democratic Incumbent: 37.15%.

The Problem

- The previous results would suggest incumbency is worth about 30%.
- But incumbency is not randomly assigned! Districts with Democratic incumbents are different.
- What can we do?
  - Control for other factors: This would suggest incumbency is worth closer to 9%.
  - Rematches: Sophomore Surge.

Control Variables

- In experiments we control through random assignment and selection, and by holding specific X values constant.
- In non-experimental studies we cannot do this.
- Instead, we measure the Xs we want to "hold constant."
  - Statistics permit us to estimate the impact of a given X upon Y, *as if* the other Xs had been held constant.
- Problem: we can only "control" for other Xs that we measure and include in our study.

Research Design Summary

- Theory drives design
- Good research design...
  - Helps establish validity of causal inferences
  - Considers other factors that may be moving the dependent variable:
    - Spuriousness
    - Controlling for other factors to allow comparison across "like" units

Measurement (aka Operationalization)

- The hypothesis is a testable statement, derived from theory, that indicates a cause & effect between two concepts.
  - Theory is at the abstract level.
  - Hypothesis is at the empirical level.
- Must be able to measure theoretical concepts of interest (DV, IV, controls) in order to test for suspected cause & effect.
  - Without good measurement, inference is suspect (i.e., theory testing suffers).

Steps for measuring social & political phenomena:

- Begin with good theoretical understanding of phenomenon of interest.
- Construct good theoretical definition (our "concept").
- Use that theoretical definition to develop the operational definition.
  - Operational definition explains what the concept will look like and how it can be measured in the empirical world.
  - Consider how others have measured this concept of interest.
- We want:
  - valid measures
  - reliable measures
  - unbiased measures

Example: Environmental Protection

- We must have a thorough understanding of our concept in order to properly measure it.
  - What precisely are we talking about?
- How can we measure the concept of environmental protection?
  - First, what do we mean by environmental protection on a conceptual level?
  - What measure best captures this concept?

How Do We Measure Environmental Protection?

- What do we mean by "environmental protection"?
  - Legislation (policy)
    - Restrictions on leaded gasoline
    - Signatory to international environmental treaties
    - Restrictions on CO2 emissions
  - Performance (outcomes)
    - Level of air pollution
    - Level of CO2 emissions
    - Area of preserved land
    - Number of endangered species

Measurement Matters

- Hypothesis: Democracies do a better job of environmental protection
- Midlarsky (1998) finds:
  - Democracies have **more** protected land area.
  - Democracies have **worse** deforestation, carbon dioxide emissions, and soil erosion by water.
  - Democracy has **no significant effect** on freshwater availability and soil erosion by chemicals.
- Our evaluation of our theory depends on how we measure our concept.

Conceptual Clarity

- Define the characteristics and boundaries of concept or construct of interest.
- Know your unit of interest (e.g., Individuals? States? countries?).
- Know your variation of interest (Over time? Between units?).
- Be precise.

Example: Participation

- Interested in explaining public participation in politics.
  - "Participation" is our construct of interest.
- Construct good theoretical definition. Define the characteristics and boundaries of concept or construct of interest.
  - The concept of "participation" may be defined as the extent to which individuals within an electorate participate directly in the election of their representatives.
- Operationalize theoretical constructs.
  - We can operationalize "participation" as the % of a state's voting age population that turned out to vote for the highest office in the most recent statewide election.

Definition: Reliability

- Extent to which re-application of a measurement method produces identical values for a variable.
- If you cannot generate same values for DV or IV successively, your confidence in your results is diminished.
- Example: Individual coding newspaper articles on the president as having either a "positive" or "negative" tone.
  - Assessment: inter-rater reliability
- Weighing oneself on a scale.
  - Assessment: Test-retest

Reliability Assessment

- Test-retest (same measurement to observations at different points in time)
- Alternative Form (two different measures of same concept at two different times; e.g., partisanship and typical vote choice)
- Split-halves (split cases and use two different measures of same concept at same time)
- Inter-rater Reliability (multiple coders of same case)

Bias (aka Systematic Measurement Error)

- Definition: measurement is reliable but is consistently 'off the mark' (i.e., low or high).
  - Consistently records values for your variable of interest (DV, IV) that are either too low or too high.
  - Can still uncover associations between DV & IV
  - But must be skeptical of the size of the descriptives (e.g., mean) and estimated relationships (e.g., regression coefficients).

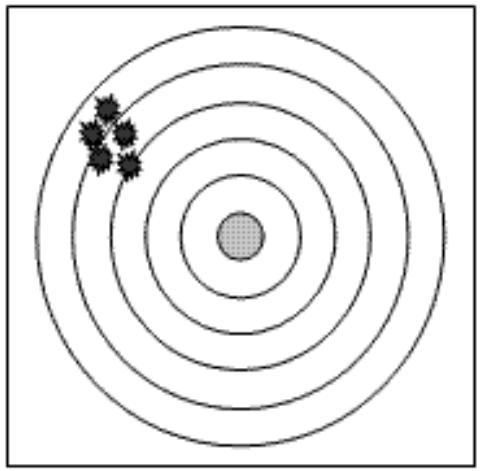
Definition: Validity

- Extent to which your instrument measures the concept of interest.
- Does what you are measuring map onto the theoretical concept you intended to measure?
- For example, how does one measure racial attitudes?
- What about asking how prejudiced someone is?
  - Really measures the willingness to reveal prejudice and not prejudice.

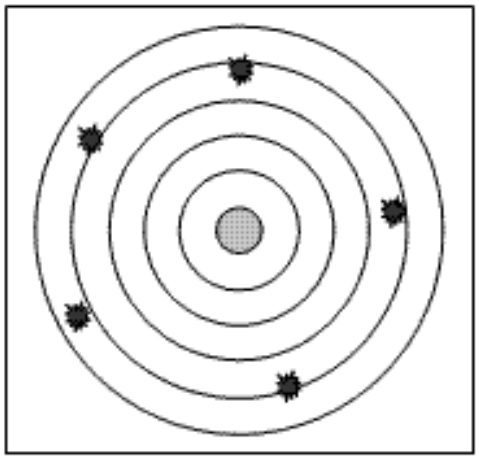
Types of Validity

- Face Validity (reasonable people agree?)
- Content Validity (has all the necessary elements; e.g., what makes a Democracy?)
- Construct Validity (related to other measures of other variables of interest as predicted by theory?)
  - Ex: If measuring individual political knowledge by a battery of factual questions about government & politics, this measure should also be associated with the individual's income, education level, as well as the likelihood of voting.

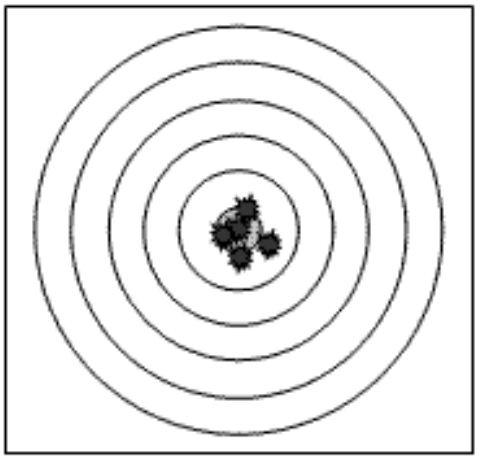
Reliability and Validity, Depicted



1. Good reliability, poor validity.



2. Poor reliability, good validity (on average).



3. Good reliability, good validity.

Source: `http://www4.state.nj.us/dhss-shad/image/reliability_validity.gif`.

Types of Variables

- Discrete (e.g., number of children)
- Continuous (e.g., age, income)

Level of Measurement

- The mathematical qualities of the values assigned
  - Nominal: Categories only—cannot rank them.
  - Ordinal: Values—can be ranked, but distance between categories unknown.
  - Interval-Ratio: Numbers—the distance between values has the same value across all values.

Nominal

- Discrete.
- They cannot be ranked or "operated" on by any mathematical function.
- Categories must be mutually exclusive and collectively exhaustive.
- Examples: gender, location
- Assign "Dummy variables" that can be 1 or 0 to turn into numbers.
- These numbers mean *nothing* beyond identifying category.

More Complicated Example: Religion

- Christian, Muslim, Jewish, Other, Atheist
- 5 categories - 4 dummy variables
  - Leaving one reference category.
- Are you a Christian or not?
- Are you a Muslim or not?
- Are you Jewish or not?
- Are you another religion or not?

Ordinal

- Discrete.
- Observations are in categories that can be ranked.
- The distance between those ranks is undefined.
- Categories must be mutually exclusive and collectively exhaustive.
- Examples: Highest schooling completed; "none, little, often, always" questions

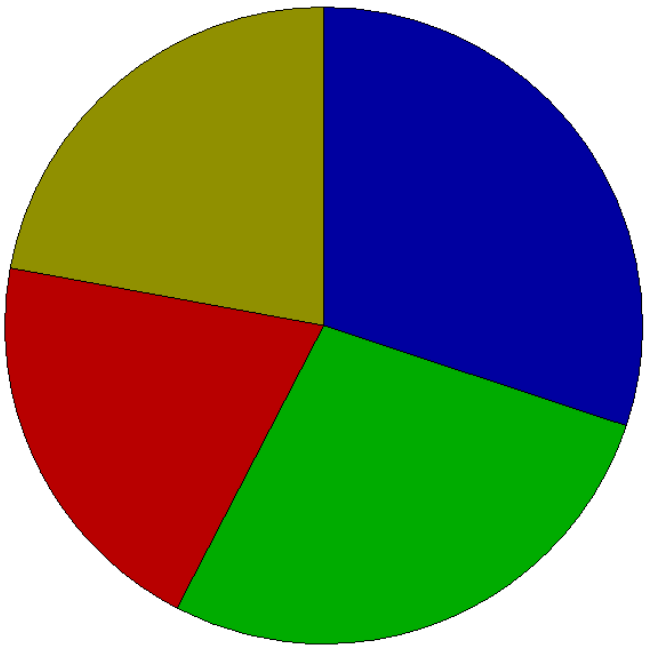
Interval-Ratio

- Might be discrete or might be continuous.
- Constant distance between values.
- Interval: arbitrary zero point.
- Ratio: Zero is meaningful.
- Categories must be mutually exclusive and exhaustive.
- Examples: years, income, temperature

Level of Measurement

FAM INCOME -- Respondent's reported family income.

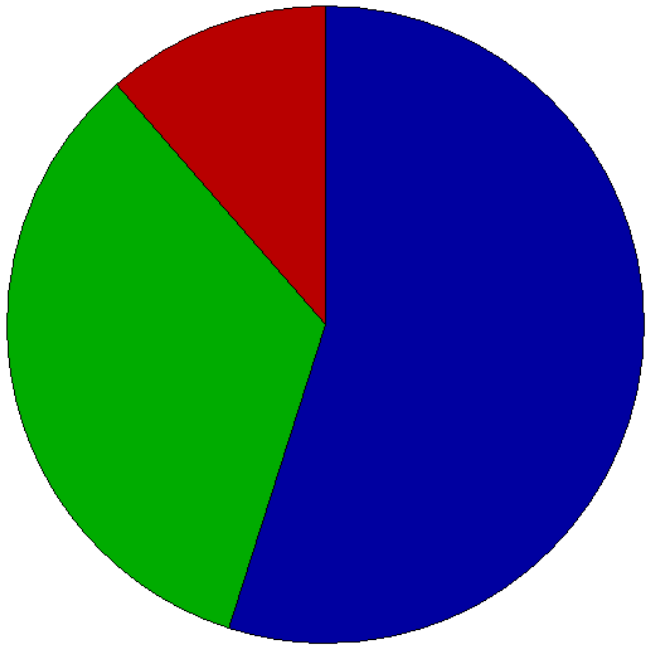
		Freq.	%
0)	< $25,000	478	30.1
1)	25k to 50k	437	27.5
2)	50k to 75k	322	20.3
3)	>$75,000	351	22.1
	TOTAL (N)	1588	100.0
	Missing	219	



Level of Measurement

WHO IN 96? -- IF VOTED IN 1996 which candidate did you vote for?

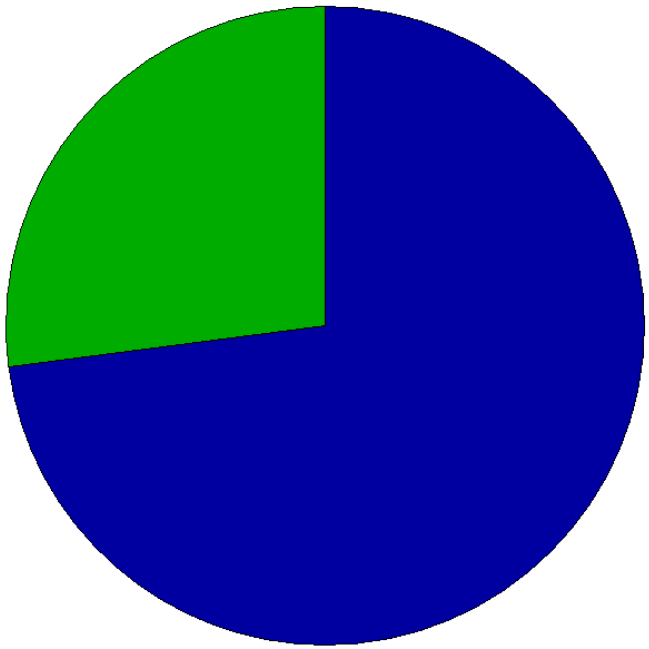
	Freq.	%
1) Clinton	665	54.9
2) Dole	408	33.7
3) Perot	138	11.4
TOTAL (N)	1211	100.0
Missing	596	



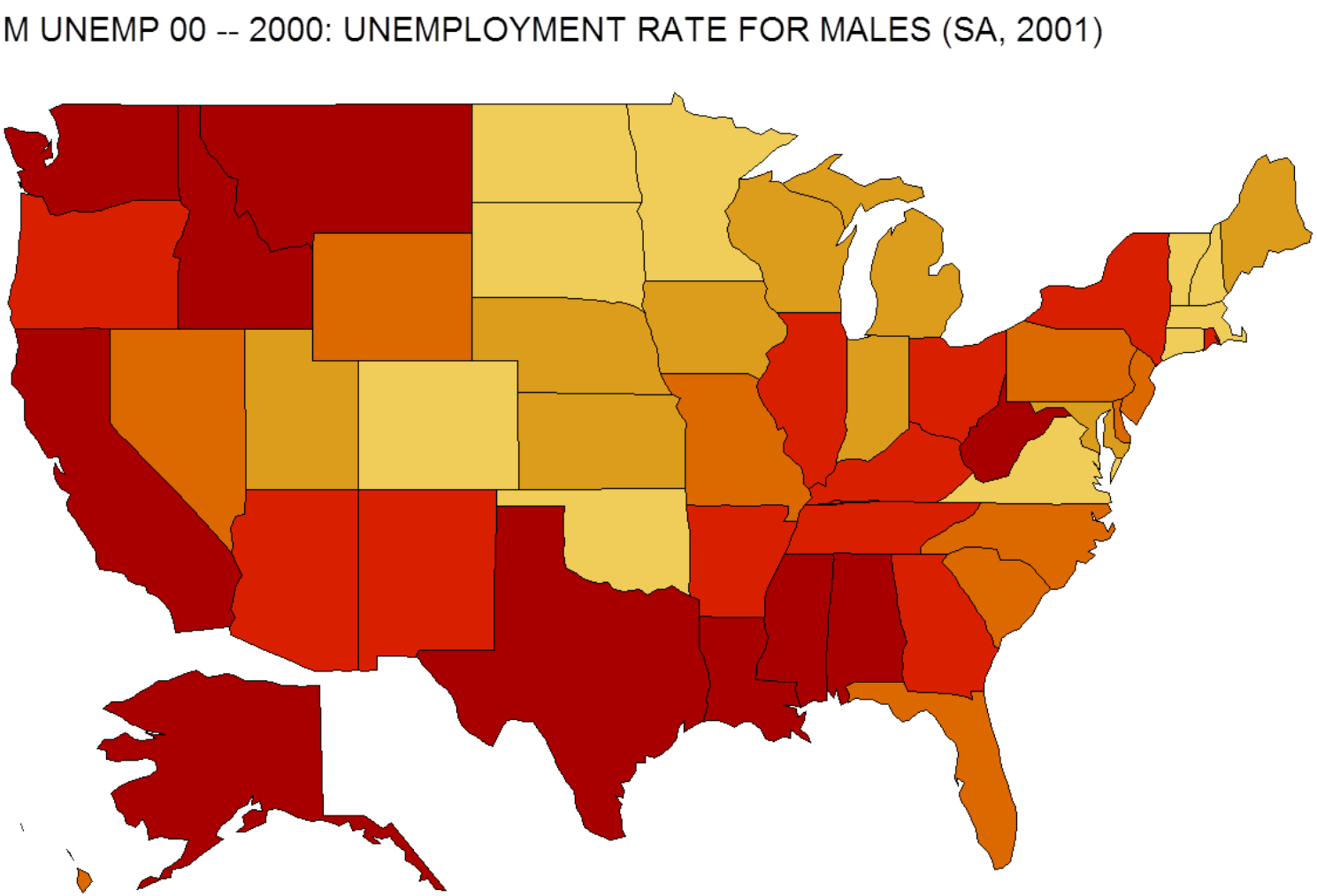
Level of Measurement

ENGL OFFIC -- Do you favor a law making English the official language of the United States, mea government business would be conducted in English only, or do you oppose such a law?

	Freq.	%
1) FAVOR	1202	72.9
2) OPPOSE	446	27.1
TOTAL (N)	1648	100.0
Missing	159	



Level of Measurement



What Level?

- You may need to decide at what level of measurement you will measure your variable.
- For example, measuring education:
  - Ratio: Years in school?
  - Ordinal: highest level of schooling completed
- This decision should be based on theory.
  - Basic rule of thumb: dont throw away information, unless there is a theoretical reason to do so (e.g., drinking age).

Population and Samples

- "Any well-defined set of units of analysis."
- Samples are drawn from a theoretically constituted population.
- Sample parameters are estimates of population parameters.
  - Our real goal
- The larger the sample, the smaller the sampling error
  - This means our estimates of population parameters are more precise.

Samples as Estimates of Population Parameters

- Who was leading in the race for governor of Florida in 2010?
  - The only way to know for sure is to ask everyone in Florida.
- The typical way we answer this question is by using a survey of a sample of voters.
  - Reuters/Ipsos poll from September 15, 2010

Candidate	Supporters
Scott	47%
Sink	45%

So, How Many People Support Scott & Sink?

- "The survey of 600 registered Florida voters... has a margin of error of plus or minus 4 percentage points.
  - Margin of error?
- The survey allows us to *estimate* the level of support for each candidate.
- There is still uncertainty about the actual level of support in the population.
  - That is, we don't know the exact level of support among all voters.
- If we're dealing with the population, we know the exact levels.
  - It's a survey of a sample; we cannot tell who is winning.
  - If these values were for the population, we would say that Scott is winning.

Scalars and Vectors

- Scalar
- Vector
  - Concept
  - Addition
  - Scalar Multiplication
  - Dot (Scalar) Product
- Utility in Statistics

Get to Know Your Data

- When dealing with a dataset your first step should be to summarize your data: before checking for *co-variation*, examine *variation*.
- These summaries are called *descriptive statistics*.
- The particular descriptive statistics you use depend on the level of measurement.

How do we summarize data?

- The Mode
  - The most common score.
- The Median
  - The middle score.
    - Or the mean of the two middles if our sample has an even number of elements.
- The Mean
  - The sum of all scores divided by the number of scores.

Describing Nominal (aka Categorical) Variables

- Nominal variables can be described by their frequency.
  - How many cases fall into a particular category?

Category	Number of cases	Percent
Protestant	672	56.14%
Catholic	292	24.39%
Jewish	35	2.92%
Other	17	1.42%
None	181	15.12%

- The most suitable descriptive statistic is the *mode*.
  - What's the most frequent category?
  - Rank and quantitative differences between values of the variable are meaningless with nominal data.

Describing Ordinal Level Variables

- With ordinal variables we can describe the data in other ways.
  - Mode: most frequent value (i.e., category) of a variable in a dataset.
  - Median: when data are arranged from lowest to highest, median is middle value (50th percentile).
    - Also useful to report the Interquartile Range (IQR): data value at the 75th percentile minus data value at the 25th percentile.

Describing Interval-Ratio Level Variables

- With interval-ratio level variables, we can describe the "moments" of the variable.
  - The moments describe the "central tendency" of a variable and the distribution of values around it.
- The first moment is the *mean*.
- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - The sum of all scores divided by the number of scores.
  - Sometimes we'll write $\sum X_i$ rather than $\sum_{i=1}^{n} X_i$ to save space.

    They mean the same thing.

Properties of the Mean

- Zero Sum Property
- $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$

Properties of the Mean

- Zero Sum Property
- $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$
- Least Squares Property
- $\sum_{i=1}^{n}(X_i - \bar{X})^2 < \sum_{i=1}^{n}(X_i - c)^2$
  - where c is any constant other than the mean.

Properties of the Mean

- Zero Sum Property
- $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$
- Least Squares Property
- $\sum_{i=1}^{n}(X_i - \bar{X})^2 < \sum_{i=1}^{n}(X_i - c)^2$
  - where c is any constant other than the mean.
- Because of these two properties, the mean is the "expected value" of the variable.

The Expected Value

- What does this mean?
- Imagine you want to predict how much a random person makes.
- The mean of the population is the best guess.
  - You'll probably be wrong, but if you did this an infinite number of times no other value would be closer more often.
- The mean is essentially our first model.
  - With more information, we'll be able to construct a better model and make better predictions.

The Effect of Outliers (2012)

- Imagine our research question is how much money does the average 25 year old make.
- Take a sample of ten 25 year-olds.

Name	Income
Ricardo	$21,256

The Effect of Outliers (2012)

- Imagine our research question is how much money does the average 25 year old make.
- Take a sample of ten 25 year-olds.

Name	Income
Ricardo	\$21,256
Thomas	\$10,134

The Effect of Outliers (2012)

- Imagine our research question is how much money does the average 25 year old make.
- Take a sample of ten 25 year-olds.

Name	Income
Ricardo	$21,256
Thomas	$10,134
Jeanette	$42,113

The Effect of Outliers (2012)

- Imagine our research question is how much money does the average 25 year old make.
- Take a sample of ten 25 year-olds.

Name	Income
Ricardo	$21,256
Thomas	$10,134
Jeanette	$42,113
David	$32,243

The Effect of Outliers (2012)

- Imagine our research question is how much money does the average 25 year old make.
- Take a sample of ten 25 year-olds.

Name	Income
Ricardo	$21,256
Thomas	$10,134
Jeanette	$42,113
David	$32,243
Sara	$5,984

The Effect of Outliers (2012)

- Imagine our research question is how much money does the average 25 year old make.
- Take a sample of ten 25 year-olds.

Name	Income
Ricardo	$21,256
Thomas	$10,134
Jeanette	$42,113
David	$32,243
Sara	$5,984
Clyde	$12,204

The Effect of Outliers (2012)

- Imagine our research question is how much money does the average 25 year old make.
- Take a sample of ten 25 year-olds.

Name	Income
Ricardo	$21,256
Thomas	$10,134
Jeanette	$42,113
David	$32,243
Sara	$5,984
Clyde	$12,204
Randi	$15,928

The Effect of Outliers (2012)

- Imagine our research question is how much money does the average 25 year old make.
- Take a sample of ten 25 year-olds.

Name	Income
Ricardo	$21,256
Thomas	$10,134
Jeanette	$42,113
David	$32,243
Sara	$5,984
Clyde	$12,204
Randi	$15,928
Monique	$25,706

The Effect of Outliers (2012)

- Imagine our research question is how much money does the average 25 year old make.
- Take a sample of ten 25 year-olds.

Name	Income
Ricardo	$21,256
Thomas	$10,134
Jeanette	$42,113
David	$32,243
Sara	$5,984
Clyde	$12,204
Randi	$15,928
Monique	$25,706
Ted	$26,003

The Effect of Outliers (2012)

- Imagine our research question is how much money does the average 25 year old make.
- Take a sample of ten 25 year-olds.

Name	Income
Ricardo	$21,256
Thomas	$10,134
Jeanette	$42,113
David	$32,243
Sara	$5,984
Clyde	$12,204
Randi	$15,928
Monique	$25,706
Ted	$26,003
LeBron	$14,410,581

The Effect of Outliers (2012)

- LeBron is going to have a large effect on the mean, but not the median.
  - With LeBron the mean is $1,460,215.
  - Without LeBron the mean is $21,286.

The Effect of Outliers (2012)

- LeBron is going to have a large effect on the mean, but not the median.
  - With LeBron the mean is $1,460,215.
  - Without LeBron the mean is $21,286.
  - With LeBron the median is $23,481.
  - Without LeBron the median is $21,256.

The Effect of Outliers (2012)

- LeBron is going to have a large effect on the mean, but not the median.
  - With LeBron the mean is $1,460,215.
  - Without LeBron the mean is $21,286.
  - With LeBron the median is $23,481.
  - Without LeBron the median is $21,256.
- The lesson: always look at your data
- Look at the range and check out those outliers.

Deviations from the Mean

- The mean does not perfectly represent all of the data points.
- The difference between an observed value and the mean is called a *deviation.*
- The deviations tell us how spread out the data are.
  - This is also called the *dispersion*.
  - If the deviations are small, then the data are clustered around the mean.
  - If the deviations are large, then the data are spread out.

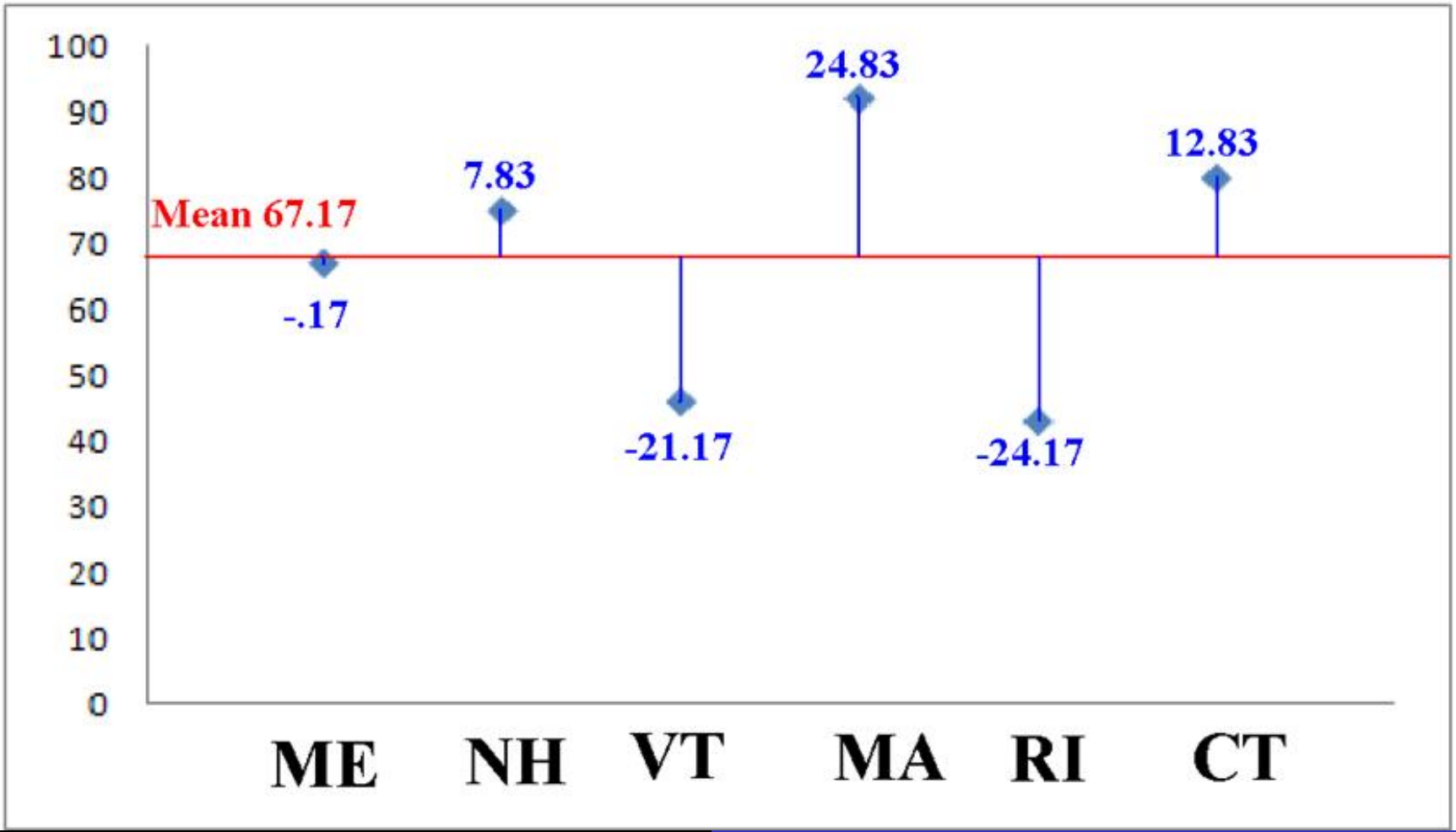
An Example with Fake Data

- Data on states' per month spending on students

State	ME	NH	VT	MA	RI	CT
Spending	67	75	46	92	43	80

- Median=71
- Mean=67.17

Deviations from the Mean



We Need a Measure of the Typical Deviation

- We could take the mean of the deviations

State	Spending	$\bar{X}$	$X_i - \bar{X}$
ME	67	67.17	-0.17
NH	75	67.17	7.83
VT	46	67.17	-21.17
MA	92	67.17	24.83
RI	43	67.17	-24.17
CT	80	67.17	12.83

We Need a Measure of the Typical Deviation

- We could take the mean of the deviations

State	Spending	$\bar{X}$	$X_i - \bar{X}$
ME	67	67.17	-0.17
NH	75	67.17	7.83
VT	46	67.17	-21.17
MA	92	67.17	24.83
RI	43	67.17	-24.17
CT	80	67.17	12.83

- The sum of the deviations is zero.
  - Some numbers are above the mean. Some below.
- So, the mean of the deviations is zero.

Squared Deviations

- We need to get rid of the negative signs.
  - So, we square the deviations.

Squared Deviations

- We need to get rid of the negative signs.
  - So, we square the deviations.

State	Spending	$\bar{X}$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
ME	67	67.17	-0.17	0.03
NH	75	67.17	7.83	61.36
VT	46	67.17	-21.17	448.03
MA	92	67.17	24.83	616.69
RI	43	67.17	-24.17	584.03
CT	80	67.17	12.83	164.69

Squared Deviations

- We need to get rid of the negative signs.
  - So, we square the deviations.

State	Spending	$\bar{X}$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
ME	67	67.17	-0.17	0.03
NH	75	67.17	7.83	61.36
VT	46	67.17	-21.17	448.03
MA	92	67.17	24.83	616.69
RI	43	67.17	-24.17	584.03
CT	80	67.17	12.83	164.69

- The sum of $(X_i - \bar{X})^2$ is 1874.83.
- This is known as the *sum of squared errors.*

The Sample Variance

- The sample **variance** is the sum of the squared errors divided by the number of data points, minus one.
- $s^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$

The Sample Variance

- The sample **variance** is the sum of the squared errors divided by the number of data points, minus one.
- $s^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$
- So, in this case 1874.83 divided by 5.
- The sample variance is 375.
  - The variance tells us typically how much a data point differs from the mean.
    - This is the second moment of the sample.

Sample Standard Deviation

- Of course, this number is quite large because we were looking at squared errors.
- We take the square root of the sample variance for the sample *standard deviation.*
  - $s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}$
- The variance is 375.
- The standard deviation is 19.4.

Back to Populations and Samples

- The calculations we just discussed apply if we are looking at a *sample* drawn from the *population*.
  - This is almost always the case. We are almost never looking at the population.
- *However*, if we were looking at the population, then we would divide the sum of the squared errors by the population size (i.e., without the minus one).
- So, for our first group of states: if that was the population, we would divide by six instead of five.
- We refer to the *population* mean as μ, and the *population* variance as σ^2. These are the first and second moments of the population. The population standard deviation is σ.

Why The Sample Size Minus One?

- It is because of the degrees of freedom.
  - Consider the following example.
- There are five people at a party and there are five sandwiches.

Why The Sample Size Minus One?

- It is because of the degrees of freedom.
  - Consider the following example.
- There are five people at a party and there are five sandwiches.
  - The person who arrives first can pick whatever sandwich she wants.
  - The second arrival can pick any sandwich except for the one the first person picked.
  - The final person who arrives must eat what is left.

Why The Sample Size Minus One?

- It is because of the degrees of freedom.
  - Consider the following example.
- There are five people at a party and there are five sandwiches.
  - The person who arrives first can pick whatever sandwich she wants.
  - The second arrival can pick any sandwich except for the one the first person picked.
  - The final person who arrives must eat what is left.
- Four out of the five people have a choice.
- The fifth person has no choice.
- We ran out of degrees of freedom.

How the Example Applies...

- We are using the variance and standard deviation of the sample to estimate the true variance and standard deviation of the population.
- In order to do so we are going to assume that the sample mean is the population mean.
- So, if we have a sample of 100, 99 of those values can be anything.
    - But one value has to be of a certain size to make the mean the value that we fixed it at.

Further Explanation

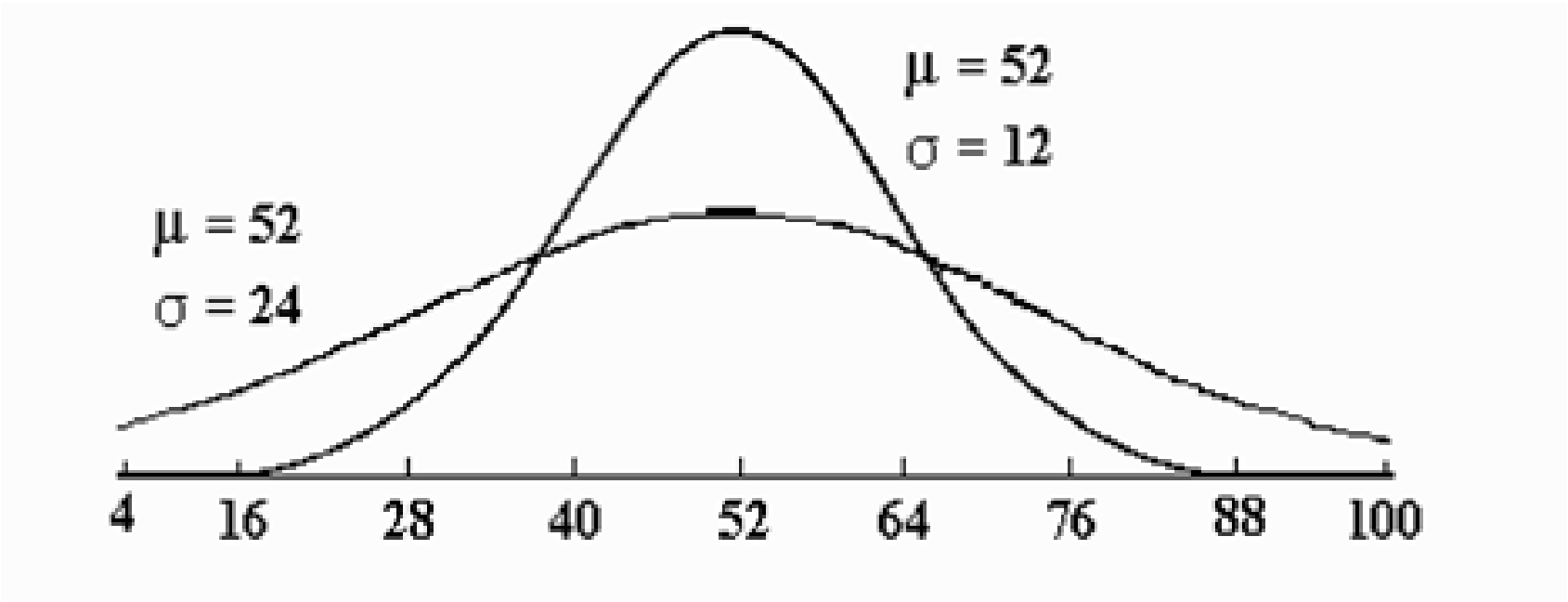
- Imagine we fix the mean at five.
- If we have four values that are 2,7,9,3, then we know the last value has to be four.
  - If it wasn't, then the mean wouldn't be five.
- The key point:
  - Because we hold the population mean to be the sample mean, we must exclude one value from the calculation.
  - So, we divide by the sample size minus one.
  - When considering the population as a whole we don't need to do this, because we aren't fixing anything.

The Formulas for the Sample

- $s^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$
- $s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}$
- Sometimes you will see $\hat{\sigma}^2$ and $\hat{\sigma}$ instead of s^2 and s. They mean the same thing.
- We need this to estimate the standard error of our sample mean without pulling hundreds of repeated samples.
- Standard error of the mean equals the standard deviation of the sample over the square root of the sample size.
  - $\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}}$
  - We'll get back to this more later.

Characteristics of Distributions I

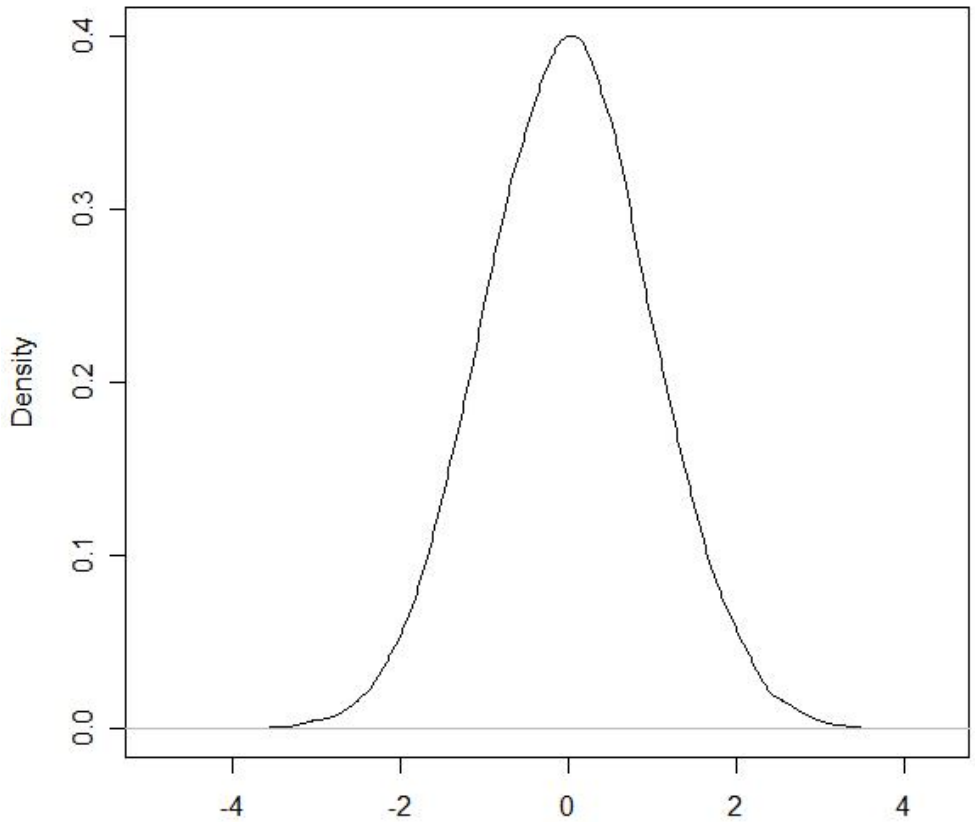
- First and Second Moments: Mean, Variance
- Taller curve has less dispersion than shorter curve (lower variance & standard deviation), but the same mean.



Characteristics of Distributions II

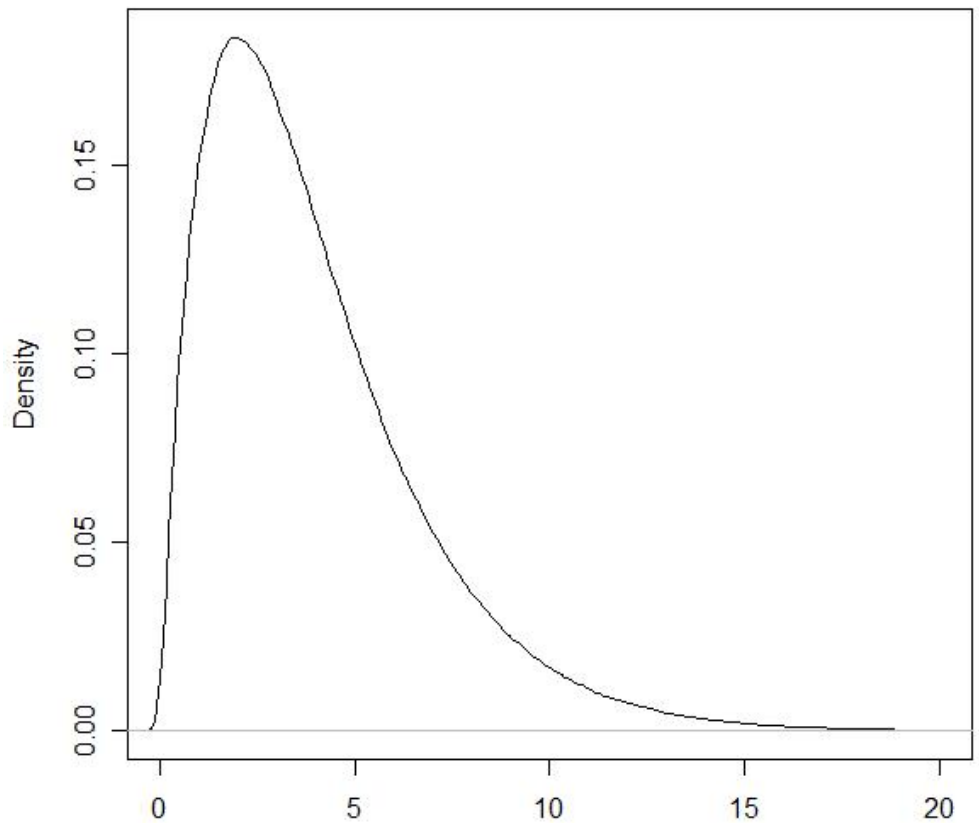
- Third Moment: Skew
  - A skewed distribution is not symmetric.
  - Most frequent scores more common at one end.
- Two types
  - Positive and negative skew

Normal Distribution



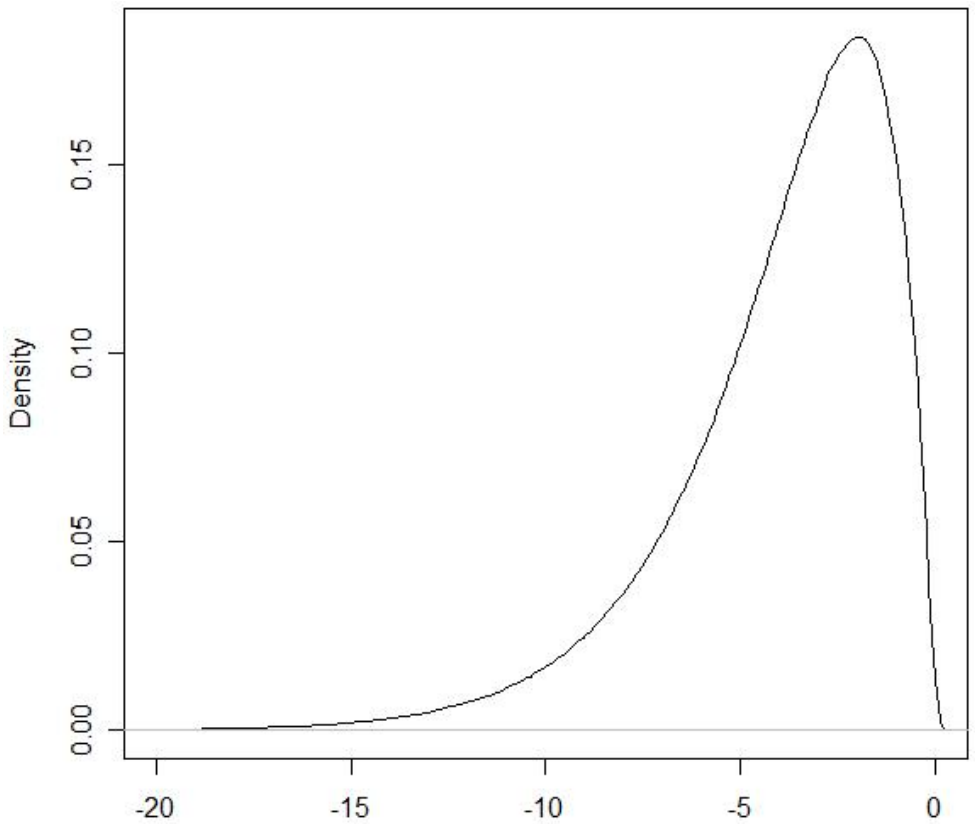
- Bell-shaped and symmetric
- Majority of scores lie around middle of distribution

Positive Skew



- Few scores at the upper end of the scale.

Negative Skew



- Few scores at the lower end of the scale.