We ♥ Statistics!

Or, wrap-up of class

Why did we have to endure this class?

- Quantitative analysis skills are important in any professional job
- These skills are highly marketable
- Statistical literacy, or, knowing when you're being bamboozled with numbers

Exploring and Describing Data

- Simple descriptive statistics (percentages, rates)
- Central tendency (mean, median, mode)
- Dispersion (variance, standard deviation)
- Tables, charts, and graphs

Exploring and Describing Data

- Would you walk across a river that is, on average, three feet deep?
- Which river would you walk across? Why?

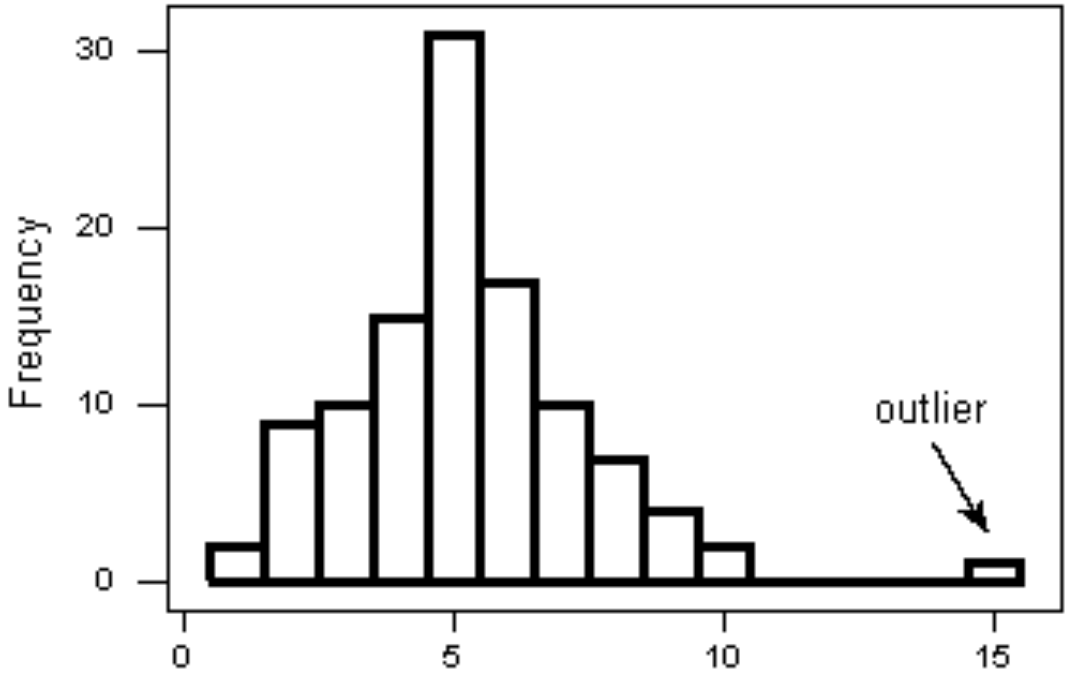
River 1	*River 2*
mean = 3	mean = 3
std = 0.5	std = 10

Exploring and Describing Data

- Ability to assess distribution is important to know what is *really* going on.
- Skew? Outliers? Range?
- The mean is the most used and most comprehensive statistic of central tendency, but it doesn't tell the whole story

Exploring and Describing Data

- Graphs help you see things that you might have otherwise missed.

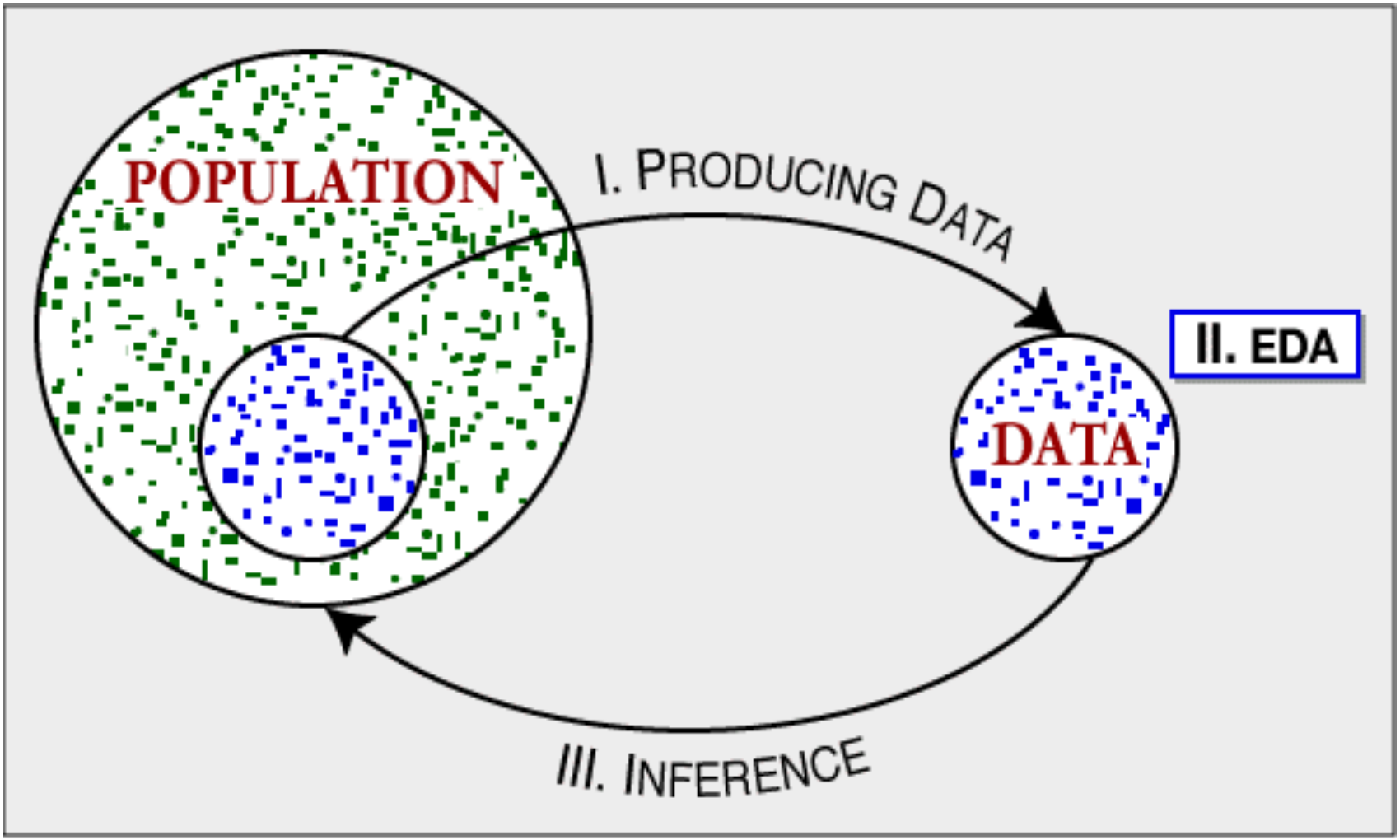


Managing Data

- Organized onto columns (variables) and rows (observations)
- Measurement: nominal, ordinal, interval-ratio

	Variables					
Individuals	Gender (M/F)	Age	Weight (lbs.)	Height (in.)	Smoking (0=No, 1=Yes)	Race
Patient #1	M	59	175	69	0	White
Patient #2	F	67	140	62	1	Black
Patient #3	F	73	155	59	0	Asian
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Patient #75	M	48	190	72	0	White

Inference



Inference

Why do we have data?

- Sometimes we have data for an entire population (i.e. New School student database)
- But usually, we have data on a smaller group, which we intend to use to infer meaning about a larger group (a sample)

Inference

- The only samples that can be used for hypothesis testing are probability samples, which are achieved through EPSEM principles
- Hypothesis testing assumes randomness

Inference

- Why is randomness so important?
  - As it turns out, randomness is actually quite predictable (at least when it comes to means)
  - The p-value (Sig.) is the probability that the sample mean is different from the population mean because of sampling error.

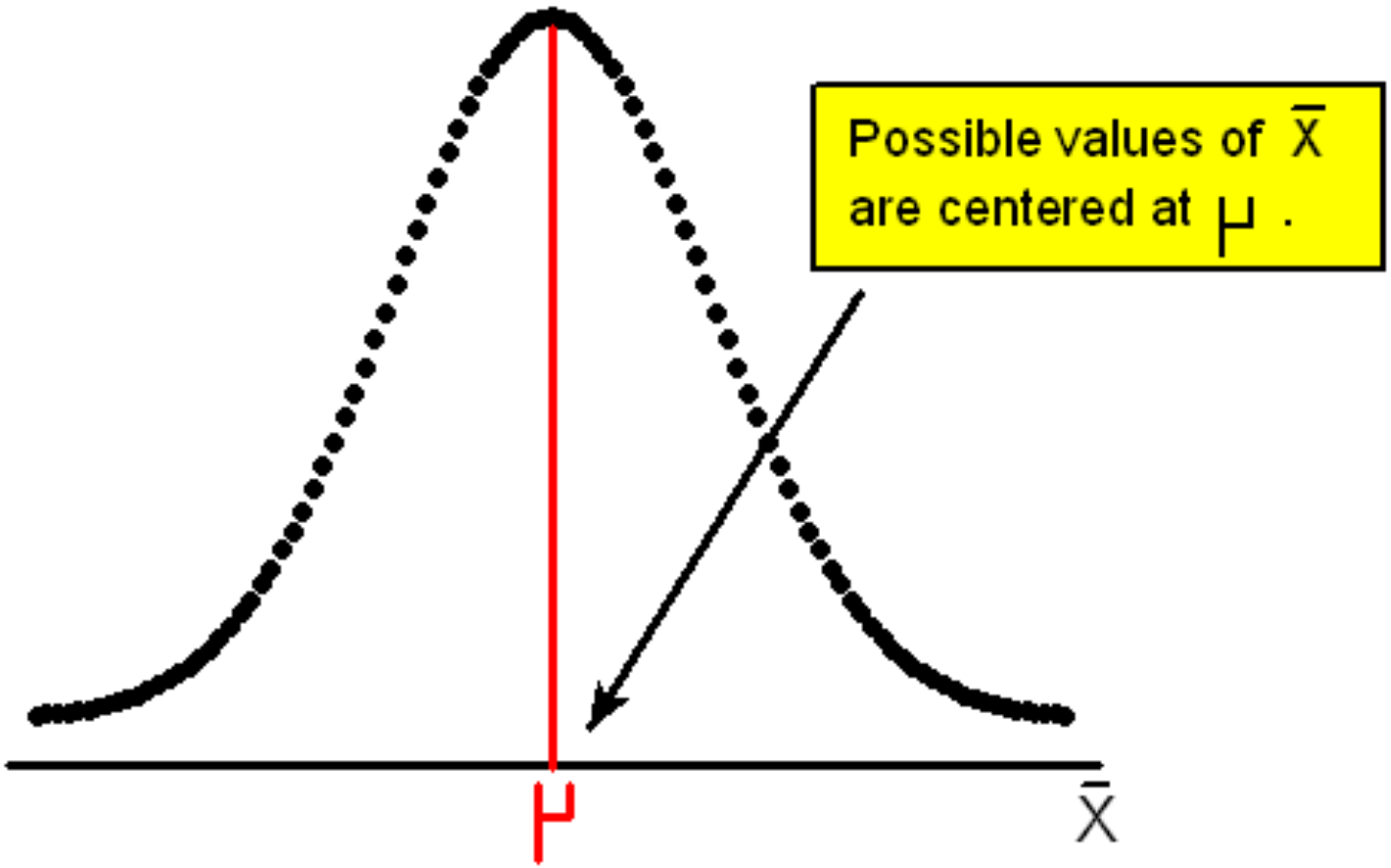
Inference

- Our work with the normal curve was so extensive because this is the most basic probability distribution (Z). We also learned other distributions: t, F, X^2

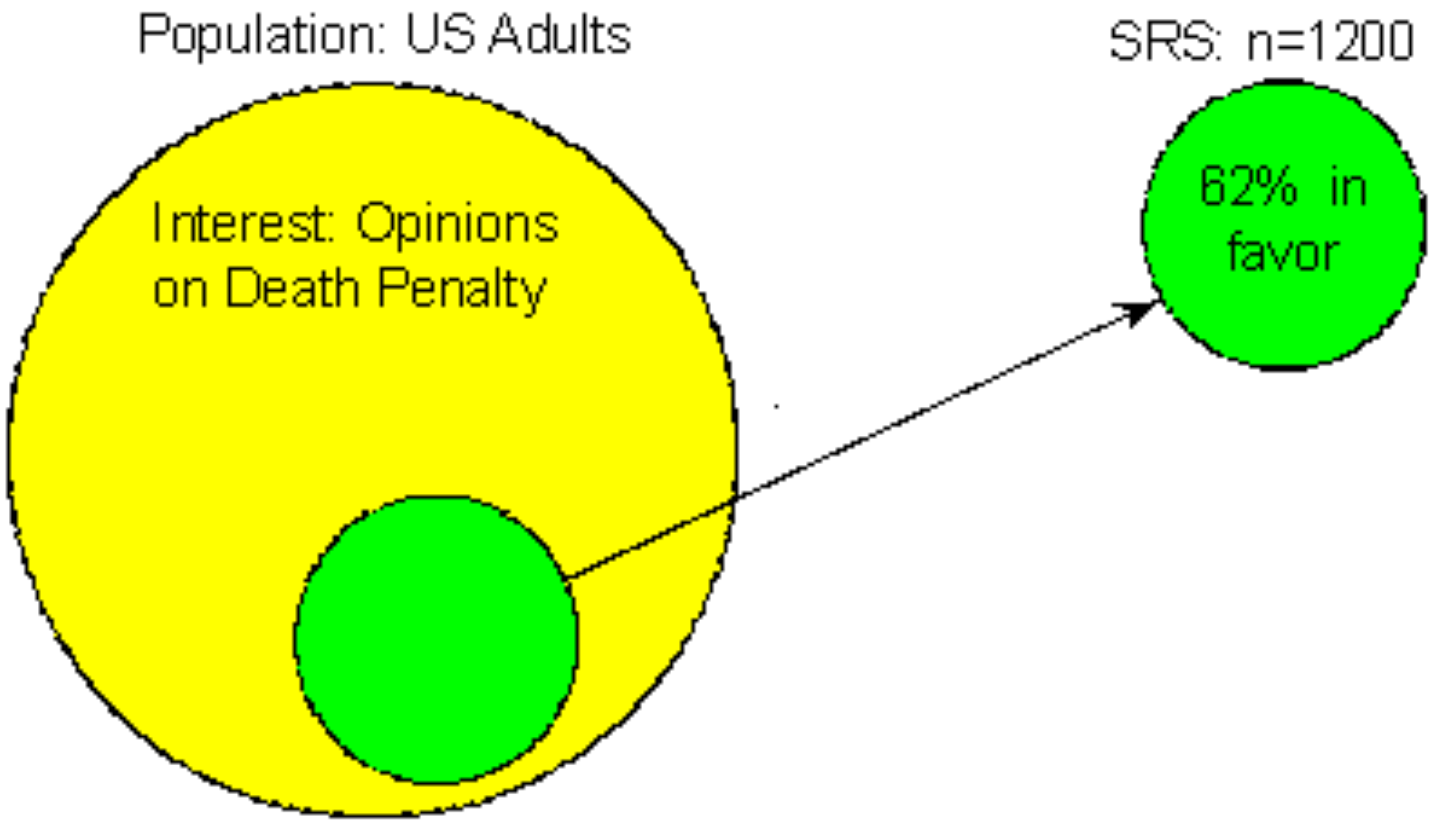
Inference

- The probability distribution allows us to understand the relationship between the sample mean (known) and the population mean (unknown)
- Thanks to: the Central Limit Theorum
- *Only when samples are random!*

Inference



Inference


Population: US Adults
Interest: Opinions on Death Penalty
SRS: n=1200
62% in favor

How is p-value determined?

- Observed means difference
- N
- Variance

Note of caution

- Even when EPSEM is strictly followed, there are other potential sources of bias:
  - Mode of interview (not everyone has a telephone)
  - Invasive or flawed questions (how would you respond if a survey asked intimate details about sex?)
  - Calculation/assumption/measurement errors

Hypothesis Testing

- Hypothesis tests, though quite different, are fundamentally doing the same thing:
  - Comparing means
  - Determining the probability that a mean is only different because of sampling error

Hypothesis Testing

- Because we want to be extremely cautious about claiming a relationship when none actually exists, we have a low tolerance for Type I error: Alpha is usually 0.05
- But don’t get too carried away: Type II error is bad too (claiming a relationship does not exist when one actually does)

Hypothesis Testing

- The type of test is usually chosen because of level of measurement
- But there are also other important assumptions that must be met

Two-Sample T-test

- Interval-Ratio dependent variable and Nominal or Ordinal independent variable (with only two categories)

ANOVA

- Interval-Ratio dependent variable and Nominal or Ordinal independent variable (with three or more categories (but generally no more than 5-7)
- Requires that the categories have roughly the same number of observations
- Doesn't say *which* mean is different

Chi-Square

- Nominal or ordinal variables for both independent and dependent variables
- Very flexible, thus very commonly used
- Careful not to have too many categories, especially with small sample size (doesn't do well with cells that have 5 or fewer observations)

Regression

- All variables must be Interval-Ratio
- Independent variables that are nominal or ordinal can be used if they are converted to dummy variables
- Simple regression is rarely used; usually regression is used when there are multiple independent variables

Regression

- Multivariate analysis is much more revealing than bivariate
- Regression models should be parsimonious
- Potential problems: circularity and multicollinearity

Hypothesis Testing

- The process:
  - Formulate and express your theory and hypothesis
  - Based on your data, choose the appropriate test(s)
  - Run/Calculate Test and Report Results

Reporting Results

- Very important to report actual means differences by the independent variable
- Report the results technically (see lab handouts) and conversationally (what the results say about your theory and hypothesis)

Reporting Results (regression)

- For regression:
  - You must write out the equation
  - R^2 tells you the percent of variation in the dependent variable that is explained by the independent variable(s)
  - ANOVA Sig. tells you whether the model as a whole is significant

Reporting Results (regression)

- The coefficients tell us the rate of change in the dependent variable for each unit change in the independent variable
- Say whether each independent variable is a significant predictor
- How do you tell which is the strongest independent variable? Standardized coefficients.

Correlation/Association

- Correlation is NOT hypothesis testing, but it gives us more detail about the strength of the relationship
- There are measures of correlation for nominal and ordinal variables, and you know where to find them if you ever need them
- For interval-ratio variables, Pearson's R

The Final Word

Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: “There are three kinds of lies: lies, damned lies and statistics.”

- *Mark Twain's Own Autobiography: The Chapters from the North American Review*