Economics 103 – Statistics for Economists

Rossa O'Keeffe-O'Donovan

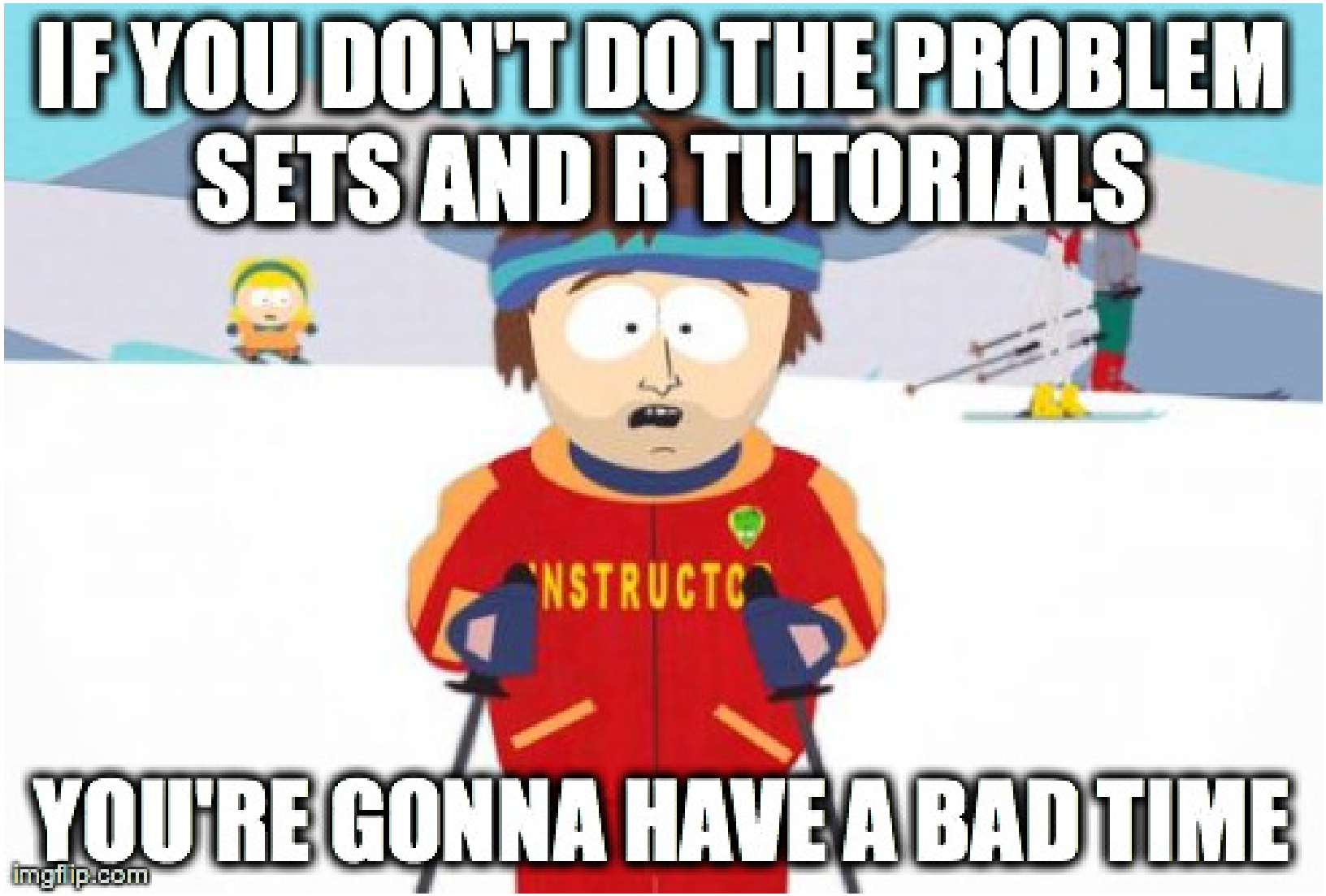
(heavily indebted to Francis J. DiTraglia)

University of Pennsylvania

Lecture # 1

Welcome to 103

- Introductions
- Syllabus
- Office hours: Days vary (see course calendar in syllabus or on website), 10:45am-11:45am, McNeil 430
- Piazza, Canvas
- Homeworks, quizzes
- Midterm (Mon Jun 6), Final (Wed Jun 29)


IF YOU DON'T DO THE PROBLEM SETS AND R TUTORIALS
YOU'RE GONNA HAVE A BAD TIME
imgflip.com

Today: What we will cover

Overview of Econ 103

Goals:

- Get a taster for the material ahead
- Understand the 'big picture'

Descriptive Statistics I

Goals:

- Understand the different types of variables
- Understand histograms and how they are constructed

This Course: Use Sample to Learn About Population

Population

Complete set of all items that interest investigator

Sample

Observed subset, or portion, of a population

Sample Size

of items in the sample, typically denoted n

Examples...

In Particular: Use Statistic to Learn about Parameter

Parameter

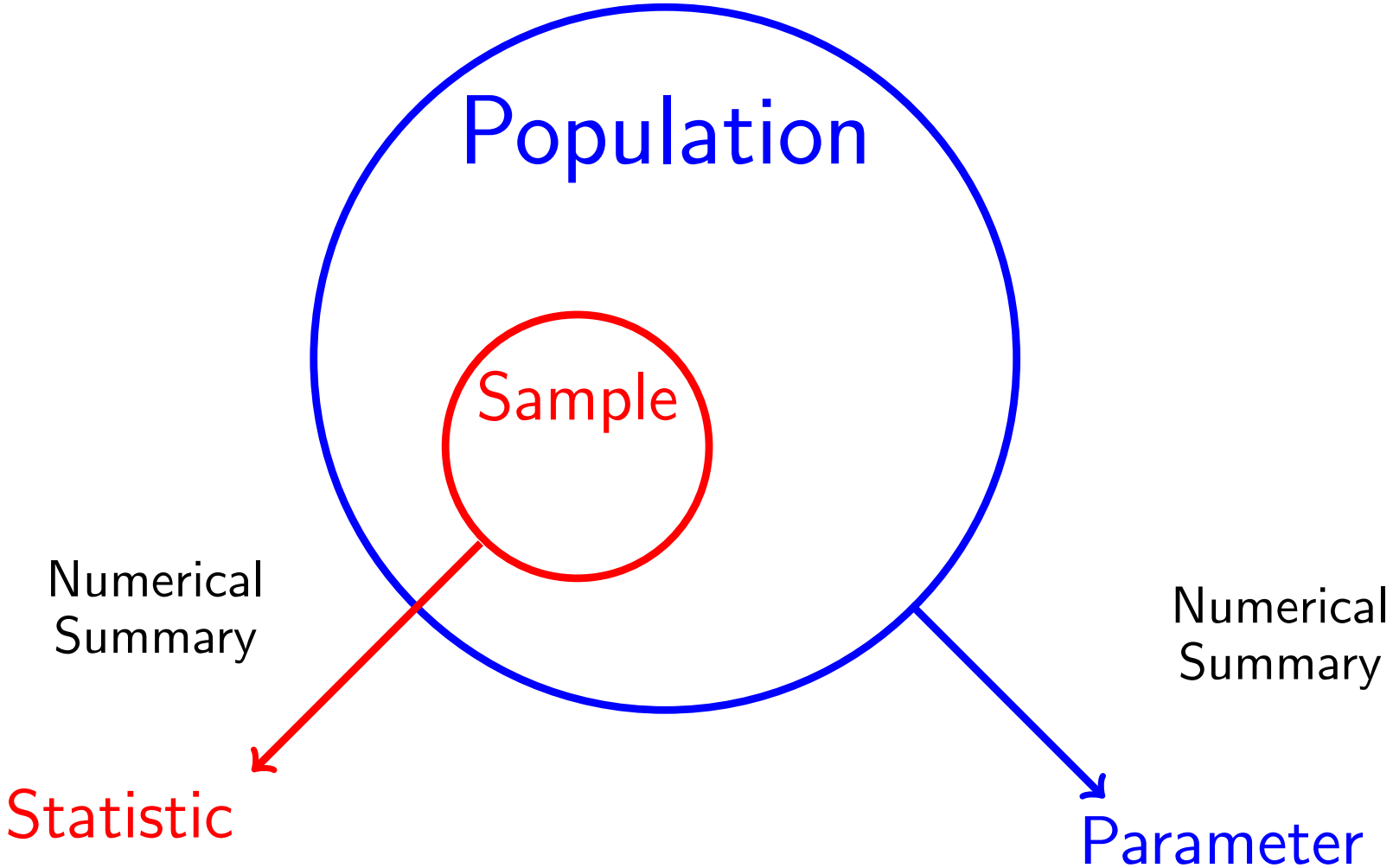
Numerical measure that describes specific characteristic of a *population*.

Statistic

Numerical measure that describes specific characteristic of *sample*.

Examples...

Essential Distinction You Must Remember!



The Field of Statistics

Descriptive Statistics (today and tomorrow)

Graphical and numerical procedures to summarize data.

Inferential Statistics (later in the course)

Using sample data to estimate things about the 'true' underlying population, and to quantify the uncertainty of these estimates (e.g. how accurate these estimates are on average)

This Course

1. Descriptive Statistics: summarize data
   - Summary Statistics
   - Graphics
2. Probability: Population $\rightarrow$ Sample
   - deductive: "safe" argument
     - All ravens are black. Mordecai is a raven, so Mordecai is black.
3. Inferential statistics: Sample $\rightarrow$ Population
   - inductive: "risky" argument
     - I've only every seen black ravens, so all ravens must be black.

Sampling and Nonsampling Error

In statistics we use samples to learn about populations, but samples are almost never *exactly* like the population they are drawn from.

1. Sampling Error
   - *Random* differences between sample and population
   - Cancel out on average
   - Decreases as sample size grows
2. Nonsampling Error
   - *Systematic* differences between sample and population
   - Does *not* cancel out on average
   - Does *not* decrease as sample size grows

Example: 'The poll that changed polling'

Literary Digest – 1936 Presidential Election Poll

FDR (Dem) versus Kansas Gov. Alf Landon (Rep)

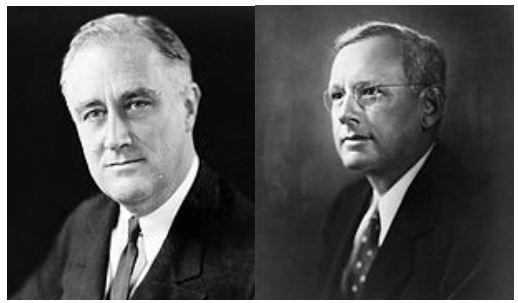
Huge Sample

Sent out over 10 million ballots; 2.4 million replies! (Compared to less than 45 million votes cast in actual election)

Prediction

Landslide for Landon

Spectacularly Mistaken!



FDR versus Kansas Gov. Alf Landon

	Roosevelt	Landon
Literary Digest Prediction:	41%	57%
Actual Result:	61%	37%

What Went Wrong? *Non-sampling Error (aka Bias)*

Source: Squire (1988)

Biased Sample

Some people more likely to be sampled than others.

- Literary Digest subscribers were on average much richer than most Americans
- Also used lists of automobile and telephone owners

Non-response Bias

Even if sample is unbiased, can't force people to reply.

- Among those who received a ballot, Landon supporters felt more strongly about the election and were more likely to reply.

Randomize to Get an Unbiased Sample

In the same election, George Gallup predicted the right result using a MUCH smaller *random* sample (50,000 vs 2.4m)

Simple Random Sample

Each member of population is chosen strictly by chance, so that: (1) selection of one individual doesn't influence selection of any other, (2) each individual is just as likely to be chosen, (3) every possible sample of size n has the same chance of selection.

What about non-response bias?

We can try to minimize this through sampling methods (e.g. telephone vs internet polling)

But we still get (random) sampling error - example

Source: Gallup

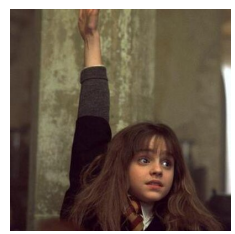
Results for this USA Today/Gallup poll are based on telephone interviews conducted Dec. 14-17, 2012, with a random sample of 1025 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, one can say with 95% confidence that the maximum margin of sampling error is 4 percentage points.

We can use this 'margin of sampling error' to construct a 'confidence interval' - more on this later in the course!

Does taking Econ 103 increase your earnings in the future?

Ask random sample of graduates who took Econ 103 what their earnings are. Also ask a random sample of graduates who didn't take Econ 103 what their earnings are, and then take the difference.

Would this procedure give a reliable *causal* estimate of the effect of taking Econ 103 on earnings?



Problem

Students who take Econ 103 may differ systematically from those who don't in *other ways* that impact that student's future income!

Students who take Econ 103 might be brighter, more motivated, or might be more likely to want to apply for a high-paying job in finace

Confounder

Think of taking Econ 103 as a 'treatment'.

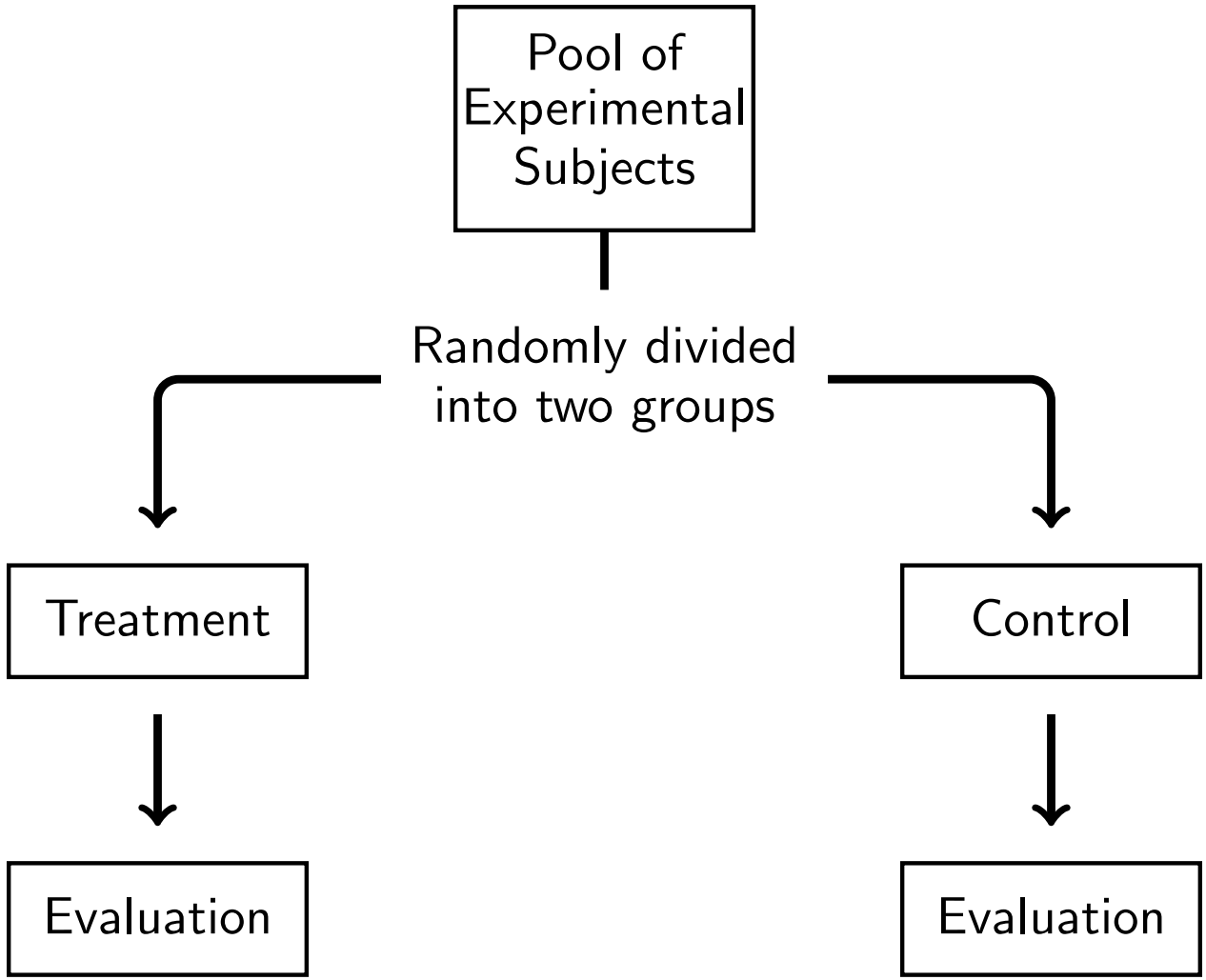
Confounder: A factor that influences outcomes and also whether subjects are 'treated' or not. Masks true effect of treatment.

Experiment Using Random Assignment: Randomized Experiment

Treatment Group Takes Econ 103, Control Group Doesn't

Essential Point!

Random assignment *neutralizes* effect of all confounding factors: since groups are initially equal, on average, any difference that emerges must be the treatment effect.


Pool of Experimental Subjects
Randomly divided into two groups
Treatment
Control
Evaluation
Evaluation

Gold Standard: Randomized Experiment

Randomized experiments ensure that on average the two groups are initially equal, and continue to be treated equally. Thus a fair comparison is possible.

Randomized experiments are generally the best way to untangle causation.

Sugar Doesn't Make Kids Hyper

http://www.youtube.com/watch?v=mkr9YsmrPAI

Observational Data

Data that do not come from a randomized experiment.

(Assignment to treatment is not randomized).

It is very difficult to untangle cause and effect using observational data because of confounders.

Summary: Randomization

We have covered two cases:

Randomized sample

When interested in information about the 'true' overall population, we should take a random sample, to make sure people in our sample are not systematically different to overall population

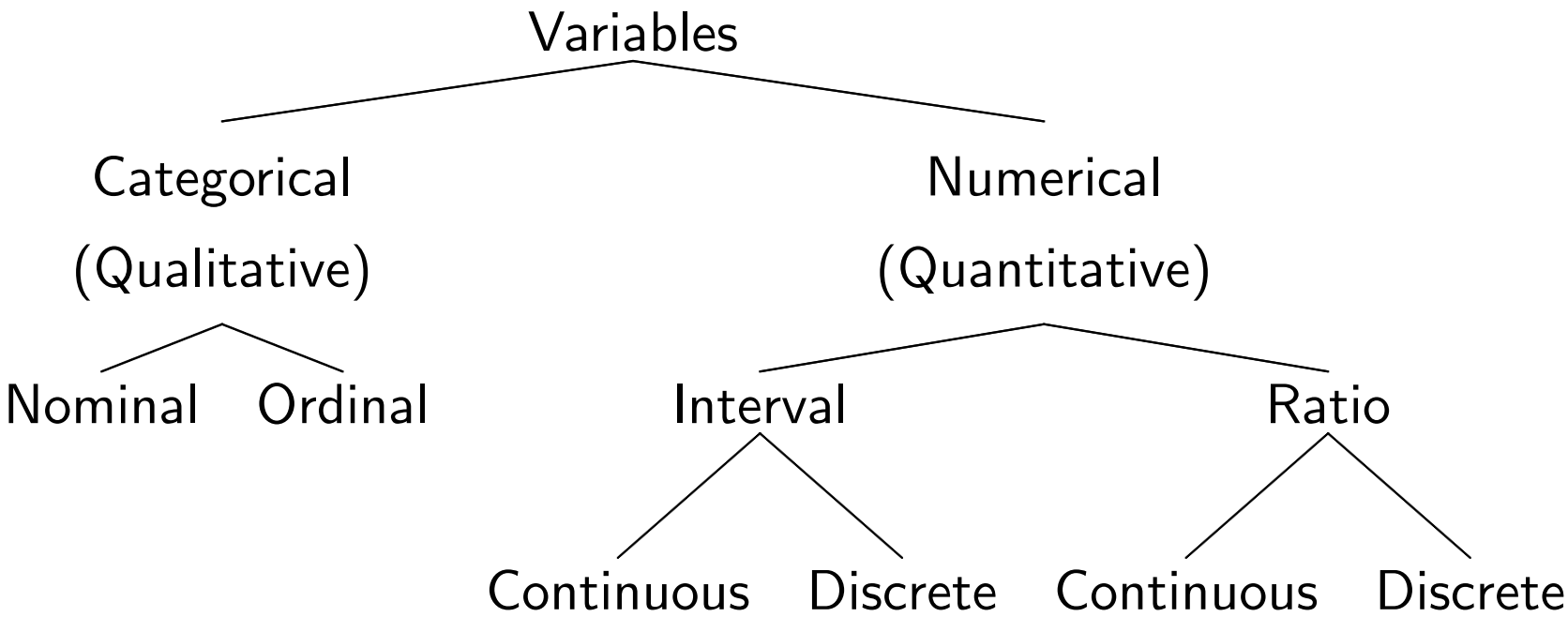
Randomized assignment in experiments

When trying to find the *causal* effect of a 'treatment' on outcomes, we should randomize assignment to treatment or control groups. On average, both groups are initially similar.

Randomization is not always possible, practical, or ethical.

Descriptive Statistics I: Types of Variables

A Taxonomy of Variables



From Weakest to Strongest

Categorical

Qualitative, assigns each unit to category, number either meaningless or indicates order only

- **Nominal** no order to the categories (e.g. mode of transport)
- **Ordinal** categories with natural order (poor, average, good)

Numerical

Quantitative, number meaningful

- **Interval** only differences meaningful, no natural zero (temperature)
- **Ratio** differences and ratios meaningful, natural zero (income in dollars)

And For Numerical Variables (interval or ratio)...

Discrete

Takes value from discrete set of numbers, typically count data

Continuous

Value could be any real number within some range (even though *measurements* are made with finite precision)

What kind of variable is...

...Handspan?

(a) Nominal

(b) Ordinal

(c) Interval

(d) Ratio

What kind of variable is...

...Temperature?

(a) Nominal

(b) Ordinal

(c) Interval

(d) Ratio

What kind of variable is...

...Eye Color?

(a) Nominal

(b) Ordinal

(c) Interval

(d) Ratio

What kind of variable?

On course evaluations you can rate your professor as follows:

0 = Poor, 1 = Fair, 2 = Good, 3 = Very Good, 4 = Excellent.

What kind of data is your rating?

(a) Nominal

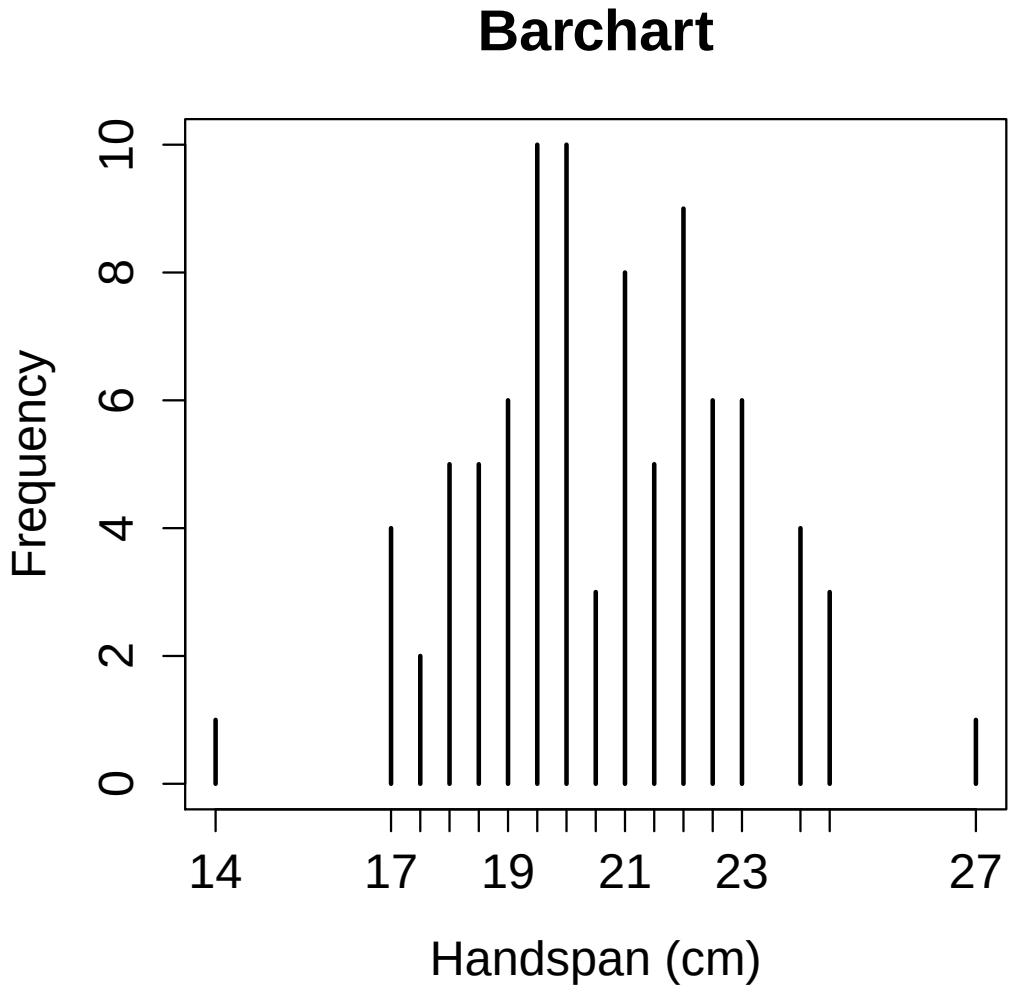
(b) Ordinal

(c) Interval

(d) Ratio

Handspan - Frequency and Relative Frequency

cm	Freq.	Rel. Freq.
14.0	1	0.01
17.0	4	0.05
17.5	2	0.02
18.0	5	0.06
18.5	5	0.06
19.0	6	0.07
19.5	10	0.11
20.0	10	0.11
20.5	3	0.03
21.0	8	0.09
21.5	5	0.06
22.0	9	0.10
22.5	6	0.07
23.0	6	0.07
24.0	4	0.05
24.5	3	0.03
27.0	1	0.01
	$n = 89$	1.00



Handspan - Summarize Barchart by "Smoothing"

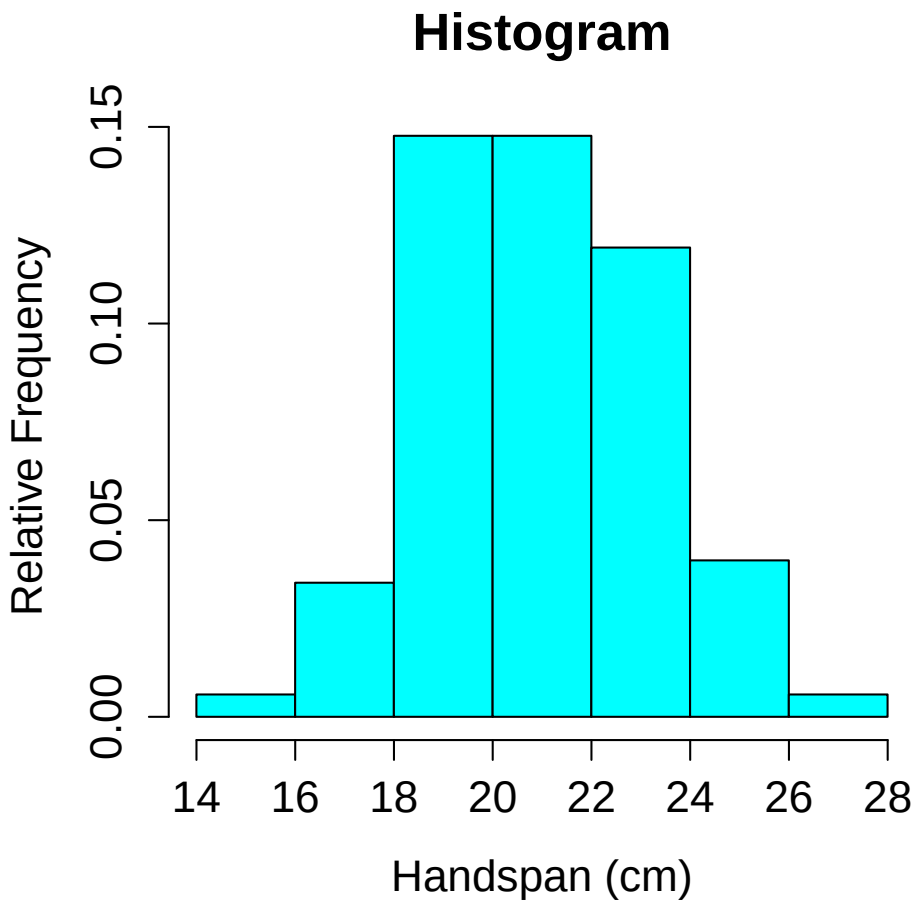
cm	Freq.	Rel. Freq.
14.0	1	0.01
17.0	4	0.05
17.5	2	0.02
18.0	5	0.06
18.5	5	0.06
19.0	6	0.07
19.5	10	0.11
20.0	10	0.11
20.5	3	0.03
21.0	8	0.09
21.5	5	0.06
22.0	9	0.10
22.5	6	0.07
23.0	6	0.07
24.0	4	0.05
24.5	3	0.03
27.0	1	0.01
	$n = 88$	1.00

Group data into non-overlapping bins of equal width:

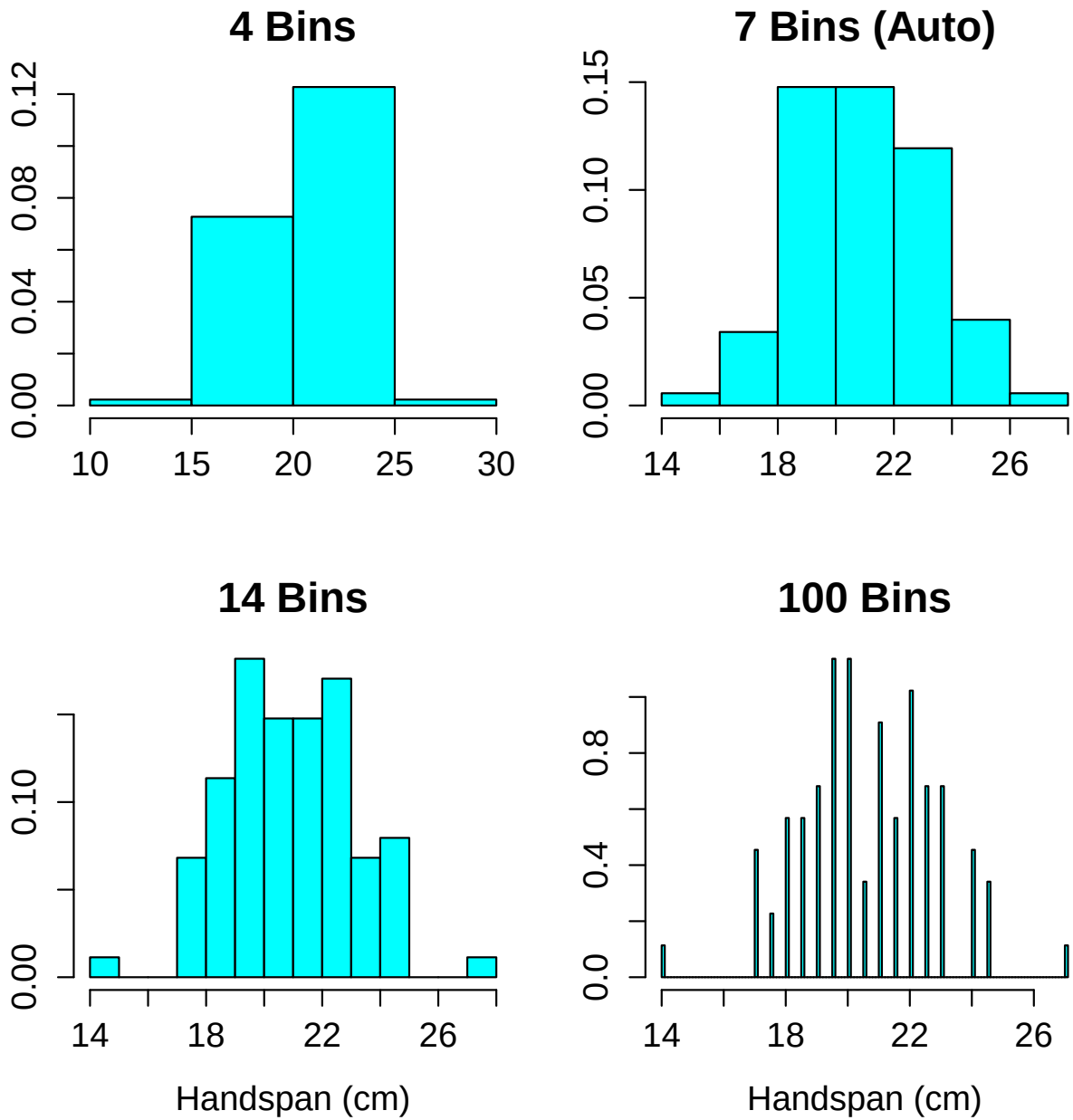
Bins	Freq.	Rel. Freq.
$[14, 16)$	1	0.01
$[16, 18)$	6	0.07
$[18, 20)$	26	0.30
$[20, 22)$	26	0.30
$[22, 24)$	21	0.24
$[24, 26)$	7	0.08
$[26, 28)$	1	0.01
	$n = 88$	1.00

Histogram – Density Estimate by Smoothing Barchart

Bins	Freq.	Rel. Freq.
$[14, 16)$	1	0.01
$[16, 18)$	6	0.07
$[18, 20)$	26	0.30
$[20, 22)$	26	0.30
$[22, 24)$	21	0.24
$[24, 26)$	7	0.08
$[26, 28)$	1	0.01
	$n = 88$	1.00



Number of Bins Controls Degree of Smoothing



Histograms are *Really* Important

Why Histogram?

Summarize numerical data, especially continuous (few repeats)

Too Many Bins – Undersmoothing

No longer a summary (lose the shape of distribution)

Too Few Bins – Oversmoothing

Miss important detail

Don't confuse with barchart!

Tomorrow: Summary Statistics

1. Measures of Central Tendency
   - Mean
   - Median
2. Measures of Spread
   - Variance
   - Standard Deviation
   - Range
   - Interquartile Range (IQR)
3. Measures of Symmetry
   - Skewness
4. Measures of relationship between variables
   - Covariance
   - Correlation
   - Regression