

Chapter 1: Introduction to Statistics

p. 1-39

Chapter objectives:

Overarching: Conceptualize the importance of collecting appropriate sample data!!!

Improper sampling can lead to wrong conclusions!

Section 1: Statistical and critical thinking

Objectives:

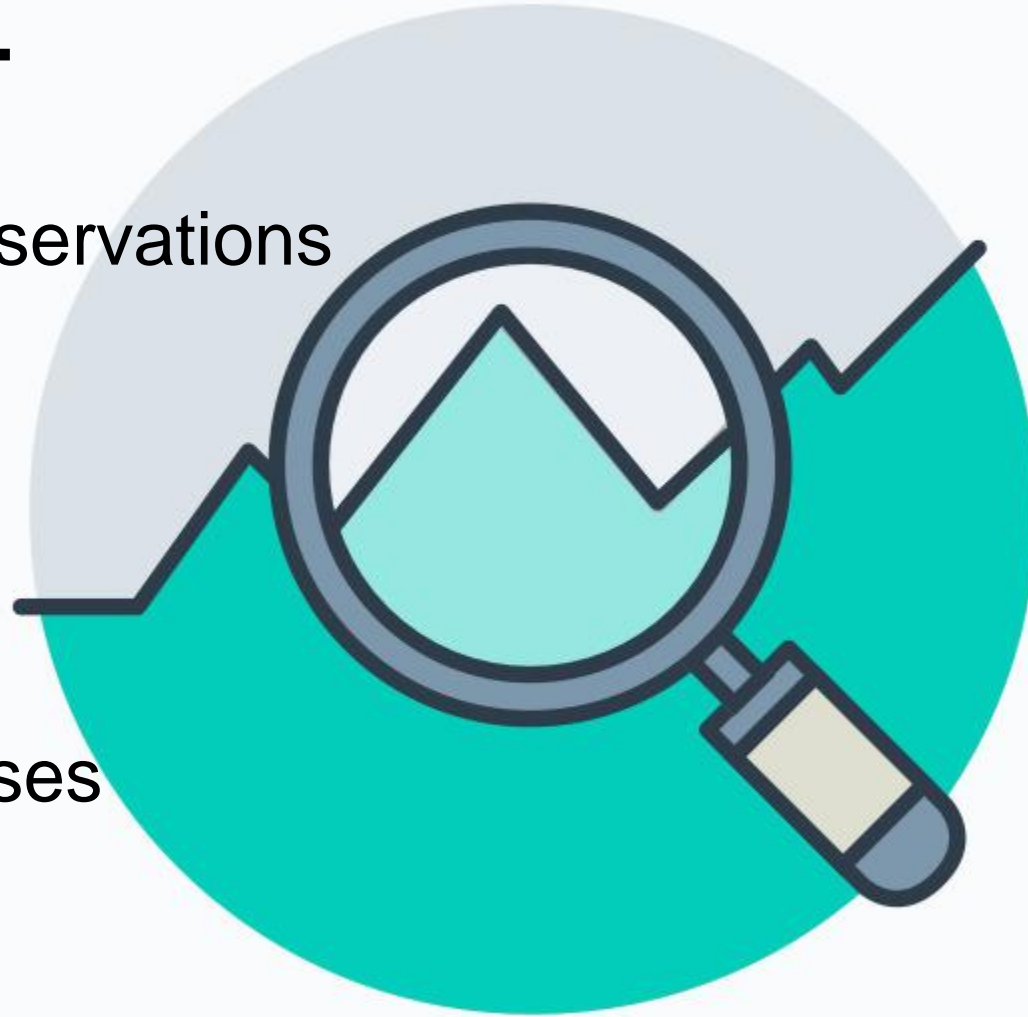
- Analyze sample data relative to context, source, and sampling method.
- Understand the difference between statistical significance and practical significance.
- Define and identify a voluntary response sample and know that statistical conclusions based on data from such a sample are generally not valid.

Data are...

collections of observations

Examples:

- Measurements
- Survey responses



Statistics is...

- The science of planning studies and experiments
- Obtaining data
- Organizing, summarizing, presenting, analyzing, and interpreting those data and then drawing conclusions based on them.





A population is

The complete collection of all measurements or data that are being considered.

Typically, the population is the complete collection of data that we would like to make inferences about.

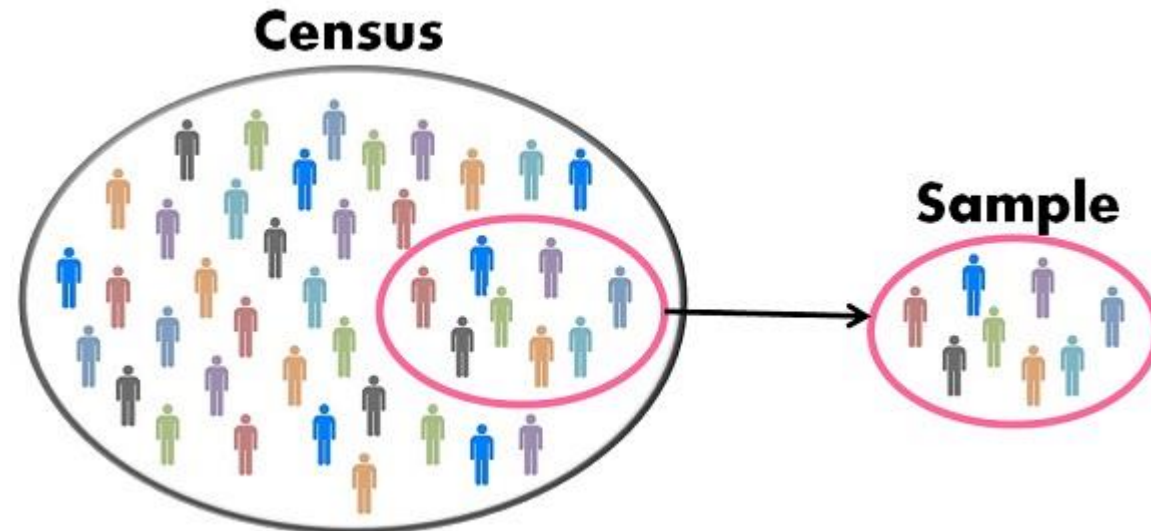
Data collected from the population:

A census is

The collection of data from every member of the population

A sample is

A subcollection of members selected from a population.



Examples

NIEHS' Sister Study recruited women that had at least one sister with a breast cancer diagnosis.

What population are they attempting to target?

What would be a census of this population?

Is this population a sample of a larger population?



The process involved in a statistical study

Prepare

Analyze

Conclude



Prepare

1. Context

What do the data represent?

What is the goal of the study?

Prepare

2. Source of the Data

Are the data from a source with a special interest so that there is a pressure to obtain results that are favorable to the source?

3. Sampling Method

Were the data collected in a way that is unbiased?

Analyze

1. Graph the Data
2. Explore the Data
 - Are there any outliers?
 - What important statistics summarize the data?
 - How are the data distributed?
 - Are there missing data?
 - Did many selected subjects refuse to respond?
3. Apply Statistical Methods
 - Use technology to obtain results

Conclude

1. Significance

- Do the results have statistical significance?
- Do the results have practical significance?

Example

1/8 women are diagnosed with breast cancer.

If we see 70/80 women in Sister Study with breast cancer, this is unlikely to occur by chance.

Alternatively, seeing 11/80 women with breast cancer is not statistically significant because this could occur by chance.

Would it be practical for all women to have a mastectomy to minimize the risk of breast cancer?

Section 1 Homework

1-4

5,11,15

17-20

25, 26, 28, 29

This is not to be turned in,
but beneficial for your
understanding.



Section 2: Types of Data

Objectives:

- Distinguish between a parameter and a statistic
- Distinguish between quantitative data and categorical data.
- Distinguish between discrete and continuous data.
- Determine whether basic statistical calculations are appropriate for a particular data set.

The background of the slide is a dark blue gradient with a pattern of binary code (0s and 1s) and various numbers (0-9) in a light blue, monospace font. The numbers are scattered across the background, some appearing larger and more prominent than others. A large, semi-transparent white circle is positioned on the left side of the slide, containing the text.

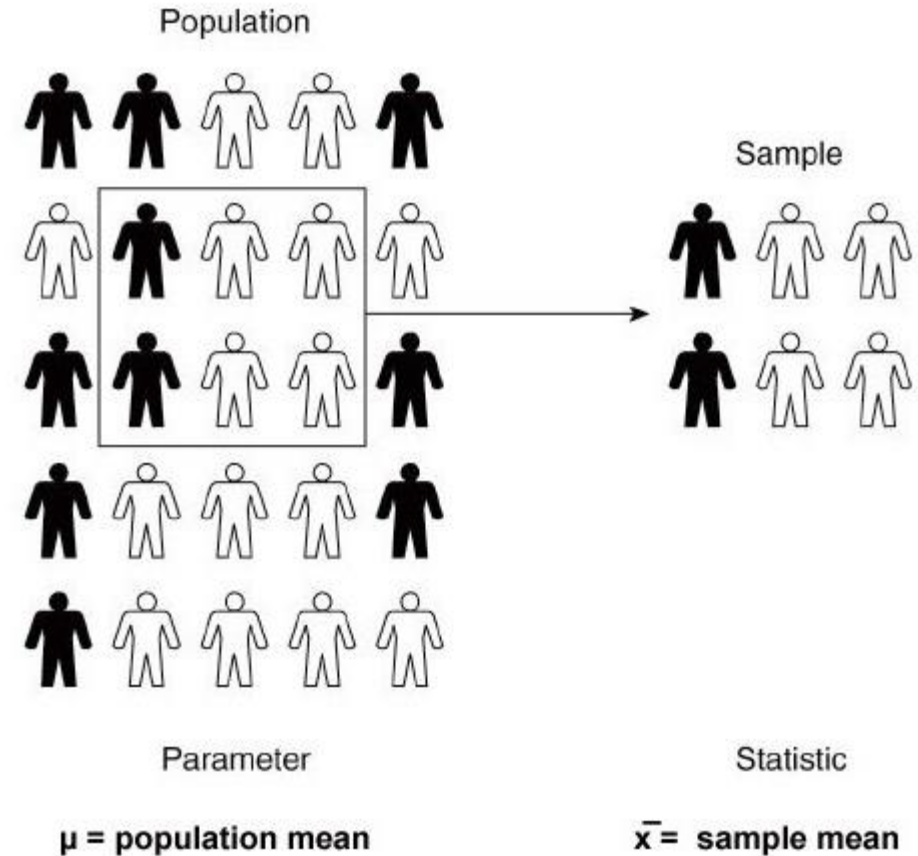
Section 2, Part 1

Basic types of Data

Parameters vs. Statistics

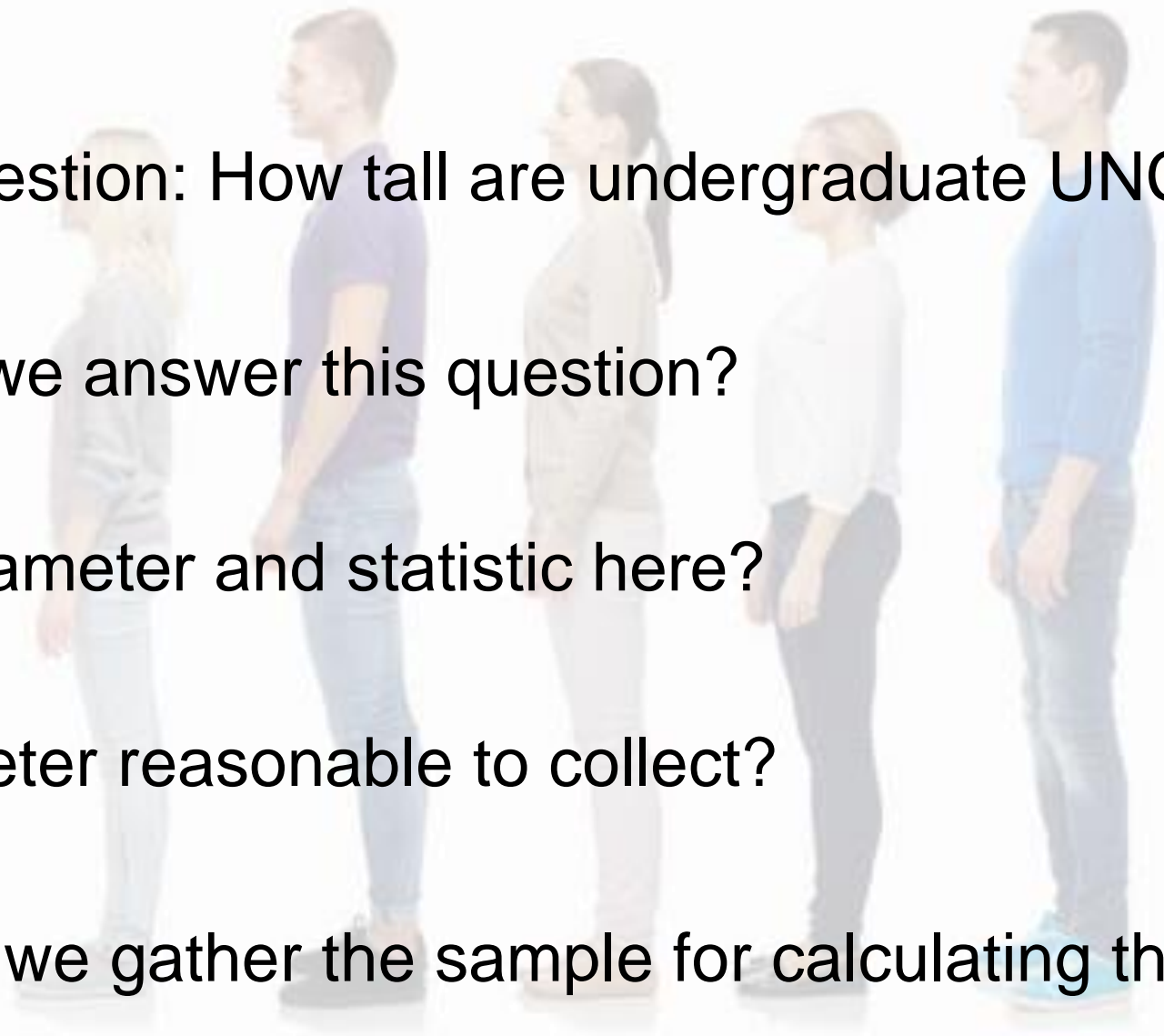
A **parameter** is a numerical measurement describing some characteristic of a *population*.

A **statistic** is a numerical measurement describing some characteristic of a *sample*.



Example!

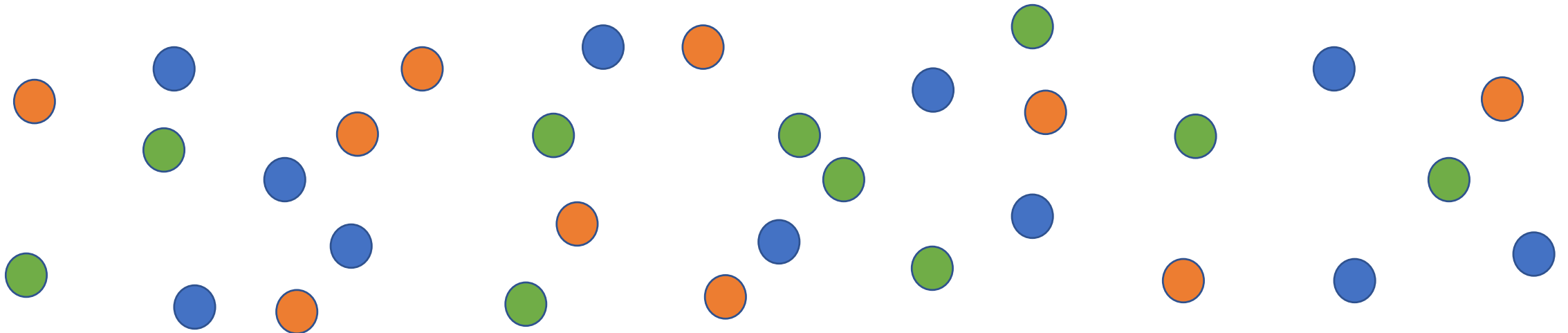
- Research question: How tall are undergraduate UNCW students?
- How should we answer this question?
- What's a parameter and statistic here?
- Is the parameter reasonable to collect?
- Where could we gather the sample for calculating this statistic?



Quantitative vs. categorical data

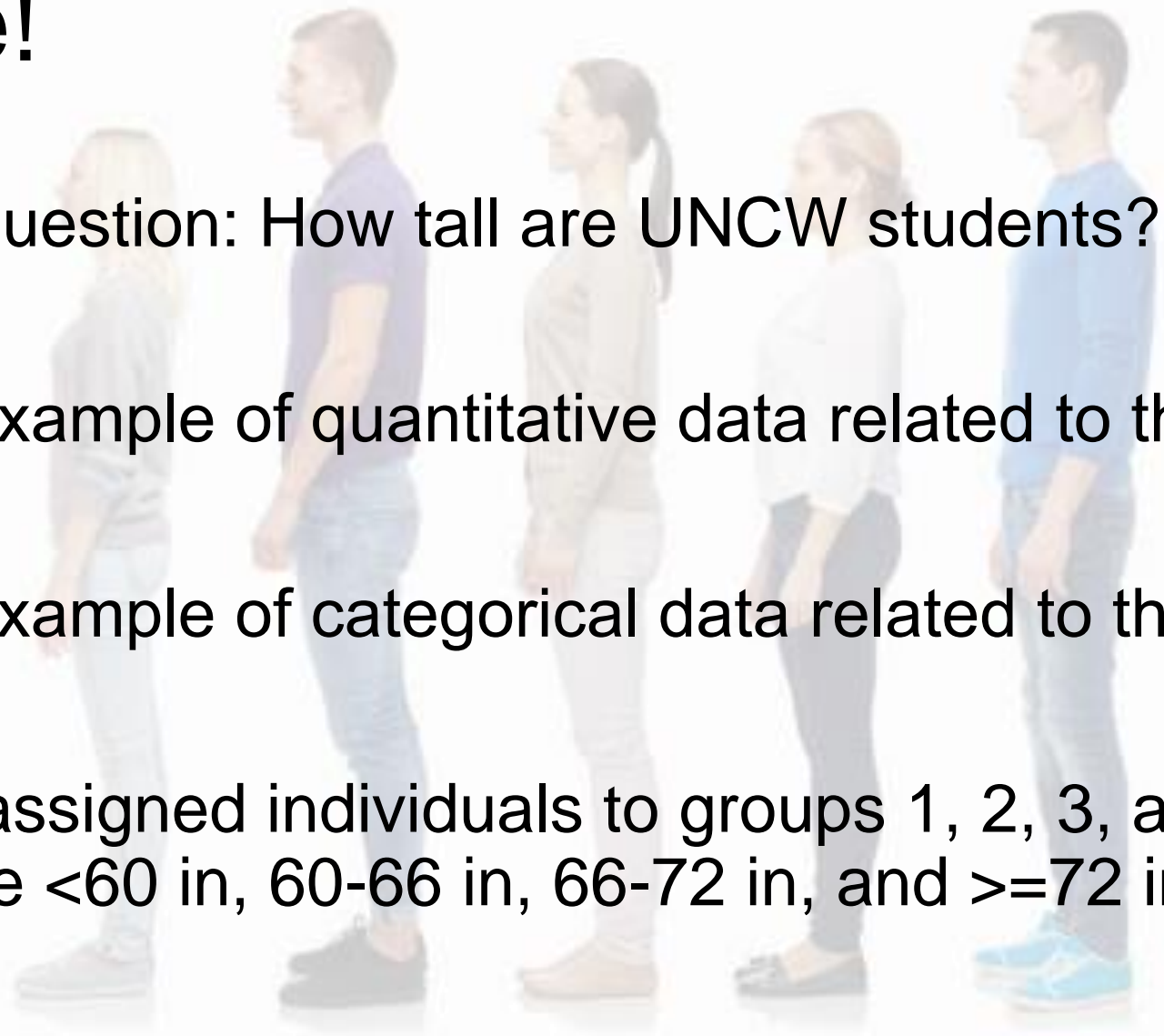
Quantitative (or **numerical**) **data** consist of *numbers* representing counts or measurements.

Categorical (or **qualitative** or **attribute**) **data** consist of names or labels (not numbers that represent counts or measurements).



Example!

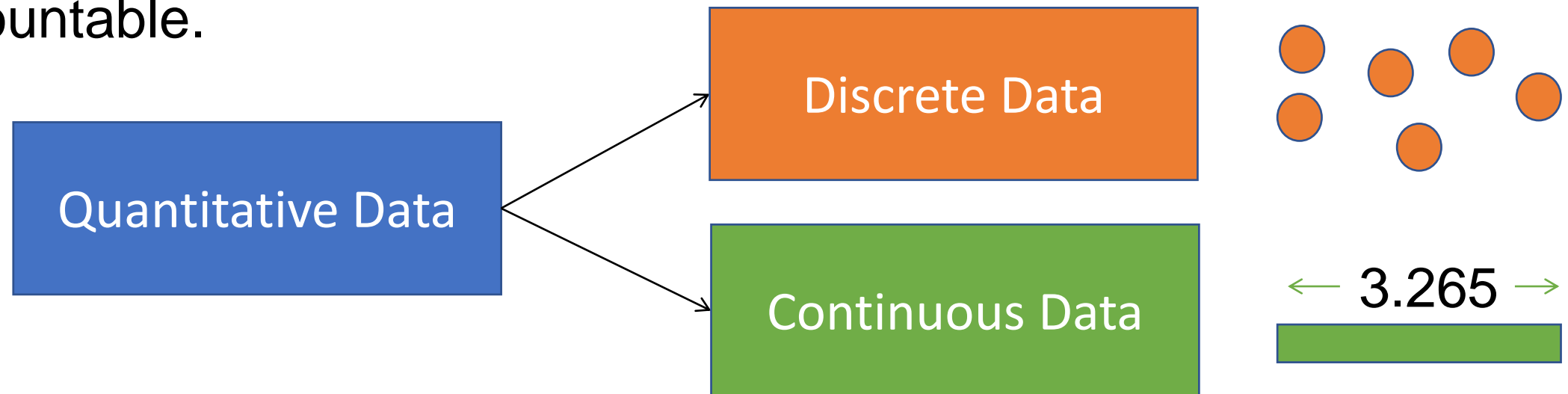
- Research Question: How tall are UNCW students?
- What's an example of quantitative data related to this question?
- What's an example of categorical data related to this question?
- What if we assigned individuals to groups 1, 2, 3, and 4 if their heights were <60 in, 60-66 in, 66-72 in, and ≥ 72 in?



Discrete vs. continuous data

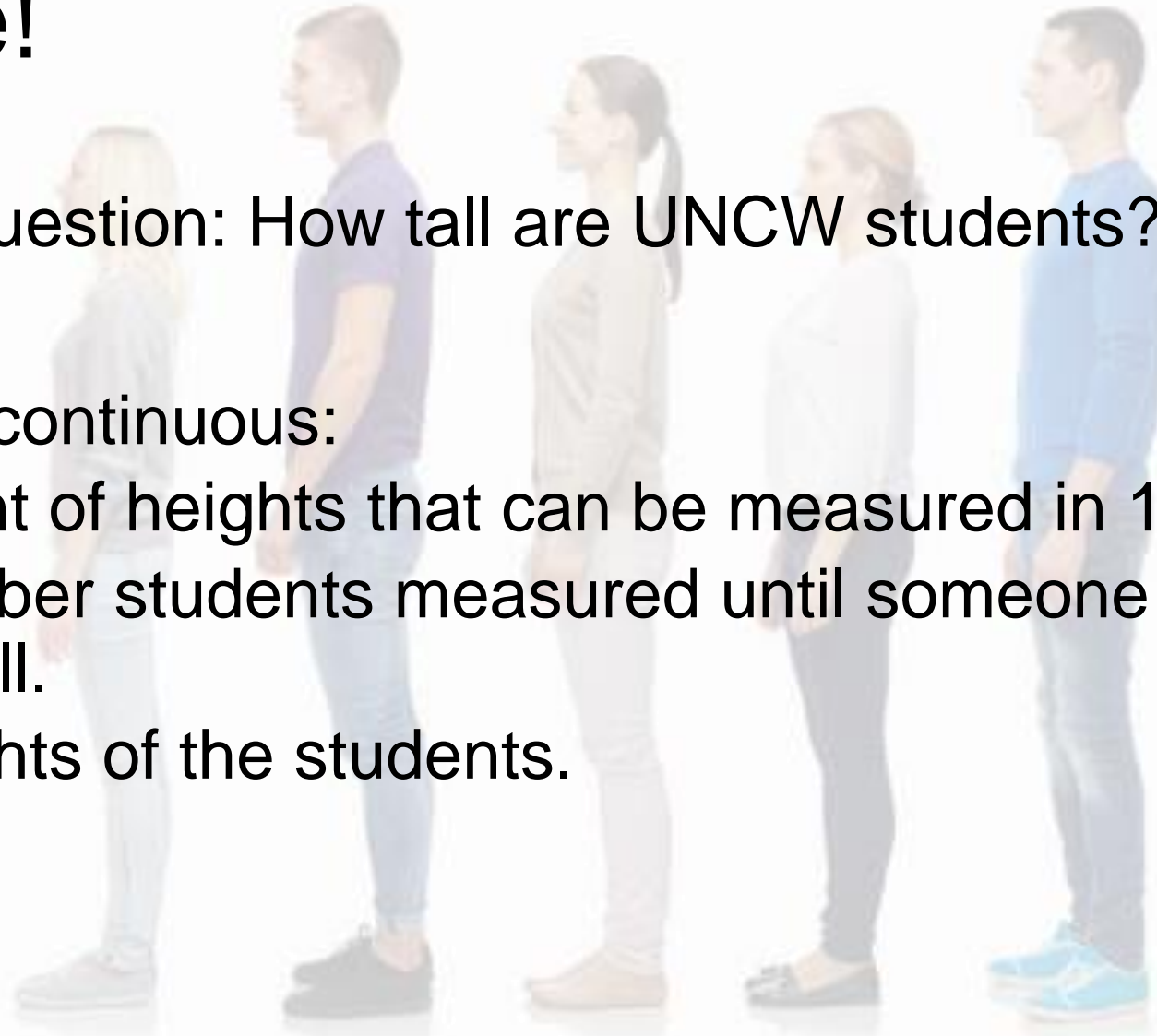
Discrete data result when the data values are quantitative and the number of values is finite or “countable.”

Continuous (numerical) data result from infinitely many possible quantitative values, where the collection of values is not countable.



Example!

- Research question: How tall are UNCW students?
- Discrete or continuous:
 - The count of heights that can be measured in 15 minutes.
 - The number students measured until someone is under 60 inches tall.
 - The heights of the students.



Four levels of measurement

Examples

Nominal	Categories (no ordering or direction)	Marital status, car make/model
Ordinal	Ordered categories (rankings, order, or scaling)	Service quality rating, letter grades
Interval	Differences between measurements but no true zero	Temperatures in Fahrenheit,
Ratio	Differences between measurements, true zero exists	Height, age, weekly food spending

Interval vs. ratio levels of measurement

- **Ratio test**

Ask yourself the question: “Does the word twice make sense?”

- **True zero**

Ask: “Is there a meaning to the value of 0 or does it even exist?”

Hot topics:

Big data refers to data sets so large and so complex that their analysis is beyond the capabilities of traditional software tools. Analysis of big data may require software simultaneously running in parallel on many different computers.

Data science involves applications of statistics, computer science, and software engineering, along with some other relevant fields (such as biology and epidemiology)

Missing data

A data value is **missing completely at random** if the likelihood of its being missing is independent of its value or any of the other values in the data set.

A data value is **missing not at random** if the missing value is related to the reason that it is missing.

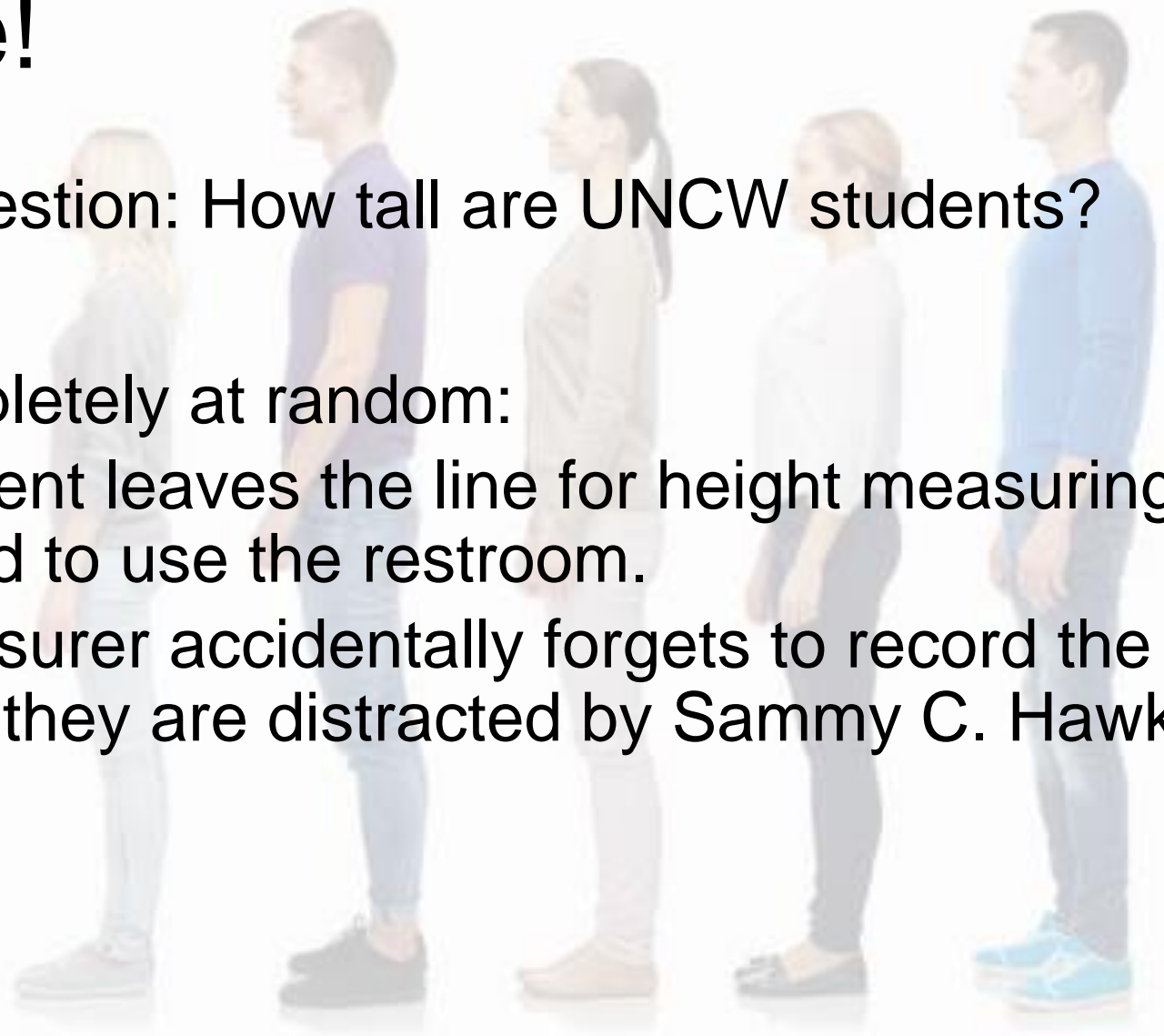


Example!

Research question: How tall are UNCW students?

Missing completely at random:

- The student leaves the line for height measuring because they need to use the restroom.
- The measurer accidentally forgets to record the height because they are distracted by Sammy C. Hawk.

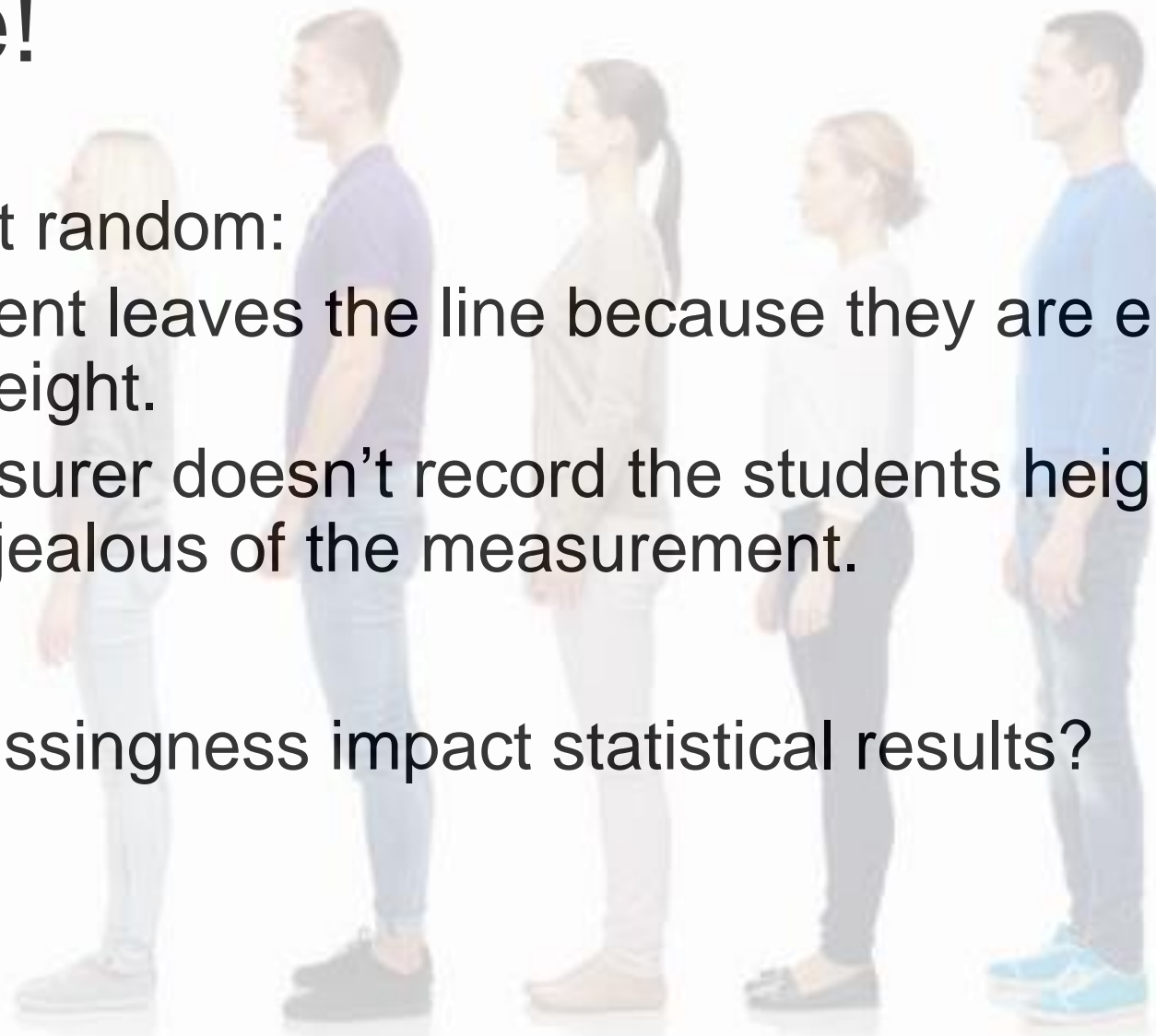


Example!

Missing not at random:

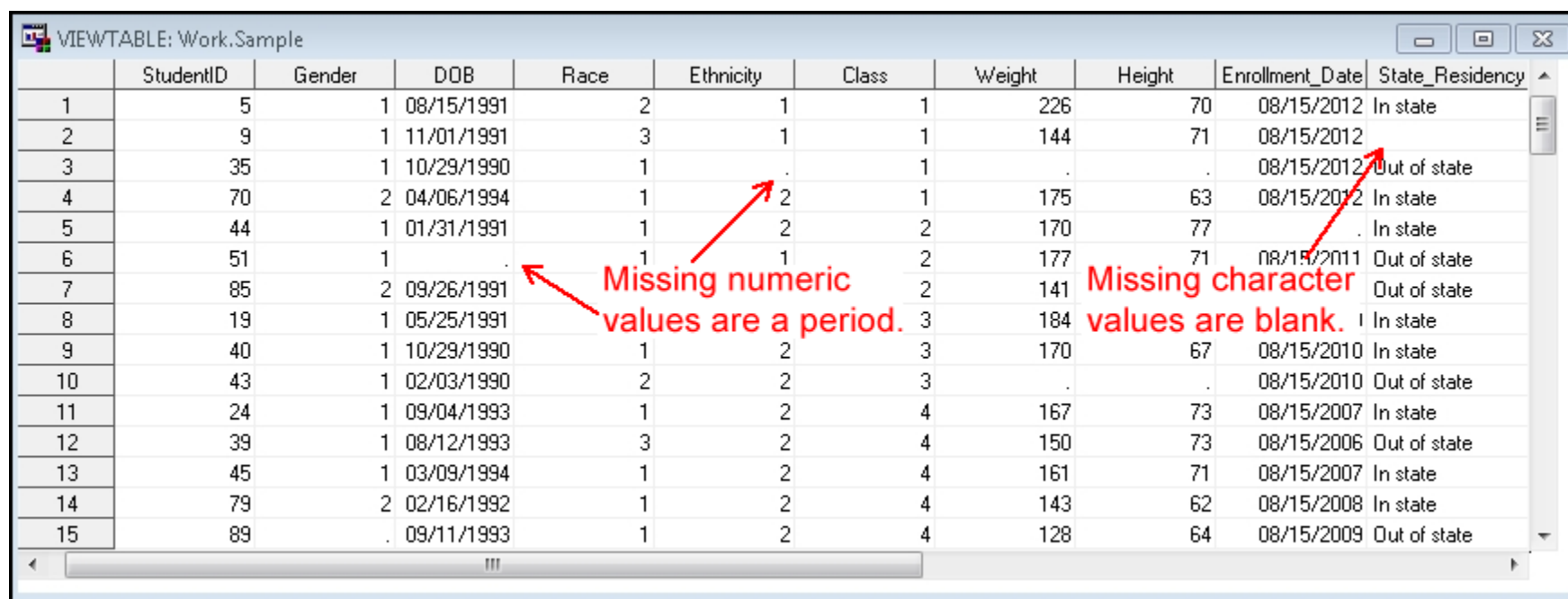
- The student leaves the line because they are embarrassed of their height.
- The measurer doesn't record the students height because they are jealous of the measurement.

How could missingness impact statistical results?



Correcting for missingness

- Delete cases, perform a “complete case analysis”
- Impute (substitute for) missing values



VIEWTABLE: Work.Sample

	StudentID	Gender	DOB	Race	Ethnicity	Class	Weight	Height	Enrollment_Date	State_Residency
1	5	1	08/15/1991	2	1	1	226	70	08/15/2012	In state
2	9	1	11/01/1991	3	1	1	144	71	08/15/2012	In state
3	35	1	10/29/1990	1	.	1	.	.	08/15/2012	Out of state
4	70	2	04/06/1994	1	2	1	175	63	08/15/2012	In state
5	44	1	01/31/1991	1	2	2	170	77	.	In state
6	51	1	.	1	1	2	177	71	08/15/2011	Out of state
7	85	2	09/26/1991	.	2	2	141	.	.	Out of state
8	19	1	05/25/1991	.	2	3	184	.	.	In state
9	40	1	10/29/1990	1	2	3	170	67	08/15/2010	In state
10	43	1	02/03/1990	2	2	3	.	.	08/15/2010	Out of state
11	24	1	09/04/1993	1	2	4	167	73	08/15/2007	In state
12	39	1	08/12/1993	3	2	4	150	73	08/15/2006	Out of state
13	45	1	03/09/1994	1	2	4	161	71	08/15/2007	In state
14	79	2	02/16/1992	1	2	4	143	62	08/15/2008	In state
15	89	.	09/11/1993	1	2	4	128	64	08/15/2009	Out of state

Missing numeric values are a period.

Missing character values are blank.

Section 2 Homework

1-4,

3 of 5-12

3 of 13-20

21-32

This is not to be turned in,
but beneficial for your
understanding.



Section 3: Collecting sample data

Objectives:

- Define and identify a simple random sample.
- Understand the importance of sound sampling and the importance of good design of experiments.



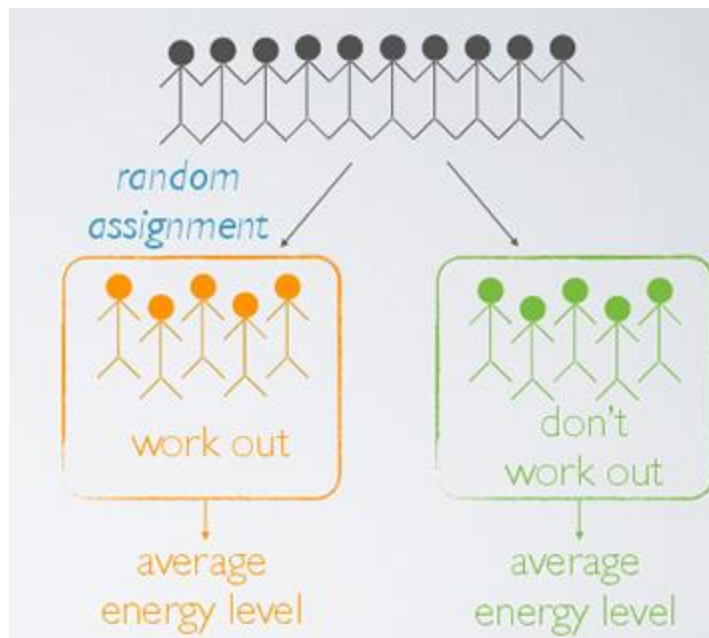
Randomization of individuals to placebo and treatment groups is considered **The Gold Standard** of experimental design.

This design for collecting data has proven to be ideal and been used time and time again.

Refer to your text to learn about the largest health experiment ever conducted.

Definitions

In an **experiment**, we apply some *treatment* and then proceed to observe its effects on the individuals.

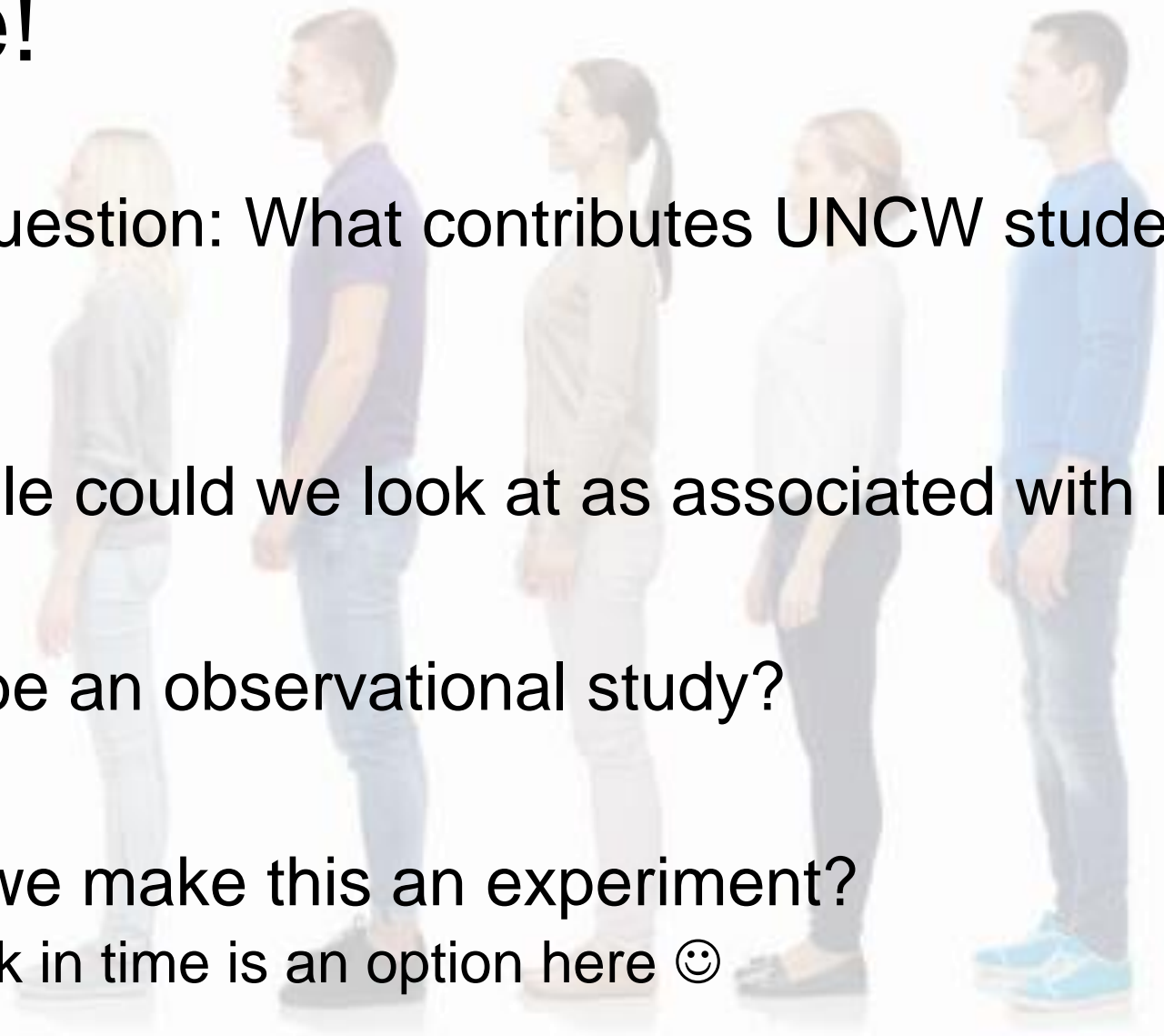


In an **observational study**, we observe and measure specific characteristics, but we don't attempt to *modify* the individuals being studied.



Example!

- Research question: What contributes UNCW students being taller?
- What variable could we look at as associated with height?
- Would this be an observational study?
- How could we make this an experiment?
---going back in time is an option here 😊



Design of Experiments

- **Replication:** Repetition of an experiment on more than one individual. Experiments need adequate sample size to detect a treatment effect.



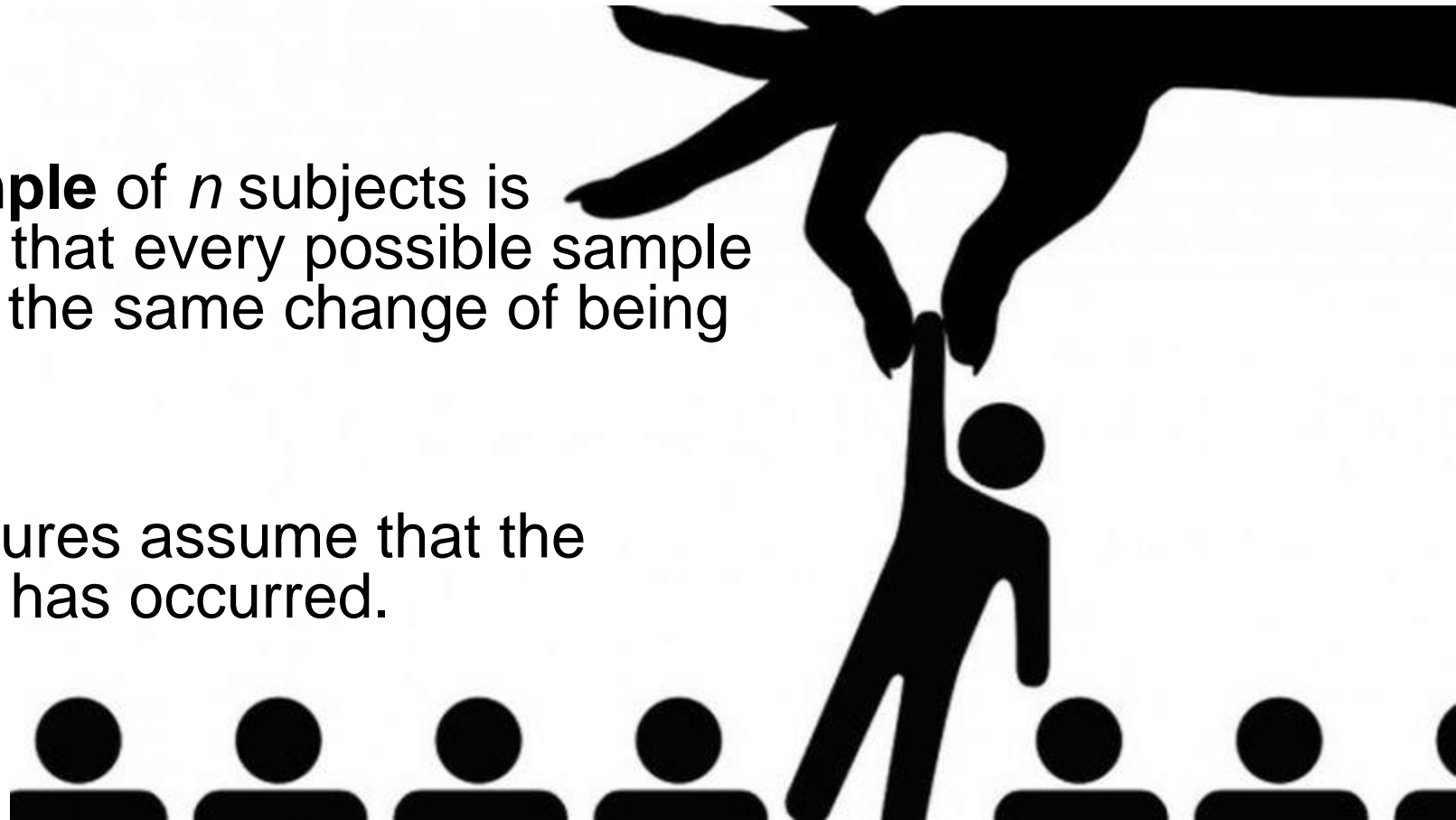
Design of Experiments

- **Blinding:** When someone (subject or researcher) involved in the experiment is unaware of who receives the treatment/placebo.
- Subject blinding is most common. This gets around the **placebo effect**.
- A **double blind** experiment means both the subject and researcher are blinded.



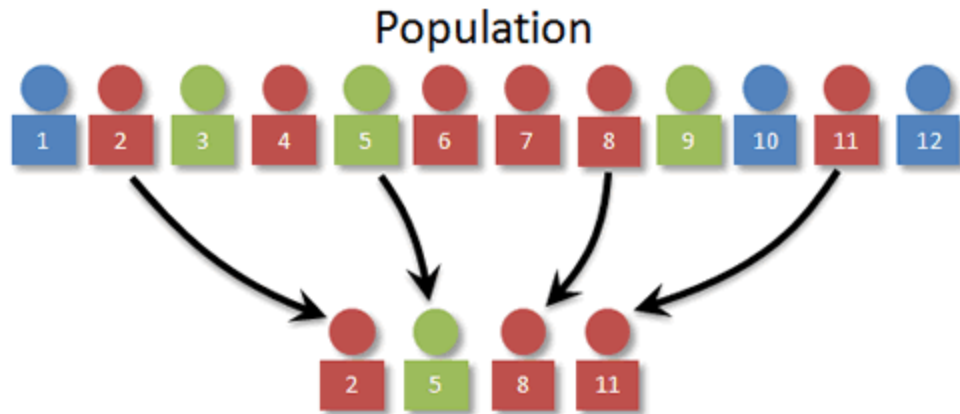
Design of Experiments

- **Randomization:** Random selection is used to assign individuals to different groups within the experiment.
- A **simple random sample** of n subjects is selected in such a way that every possible sample of the same size n has the same chance of being chosen.
- Many statistical procedures assume that the simple random sample has occurred.

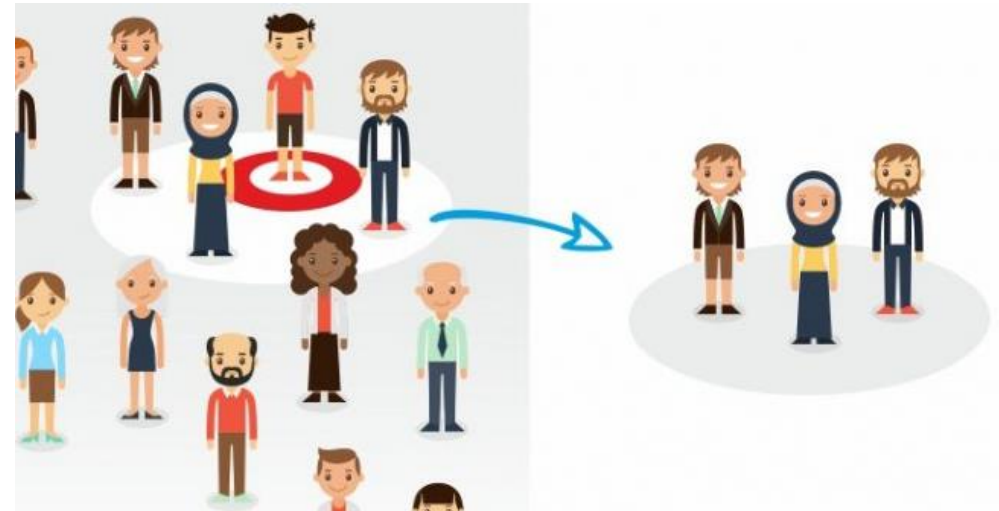


Other sampling methods with UNCW student height examples.

In **systematic sampling**, we select some starting point and then select every k^{th} (e.g. every 3rd) element in the population.

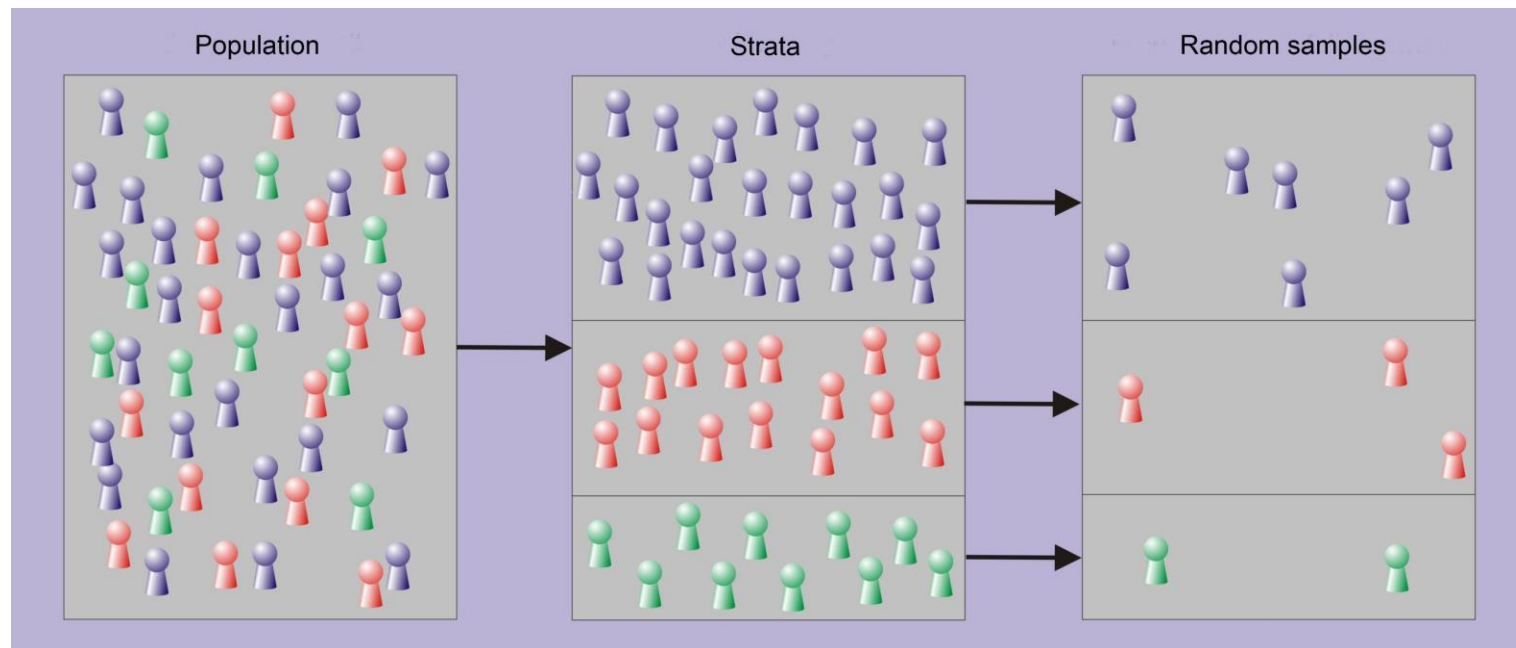


With **convenience sampling**, we simply use data that are very easy to get.



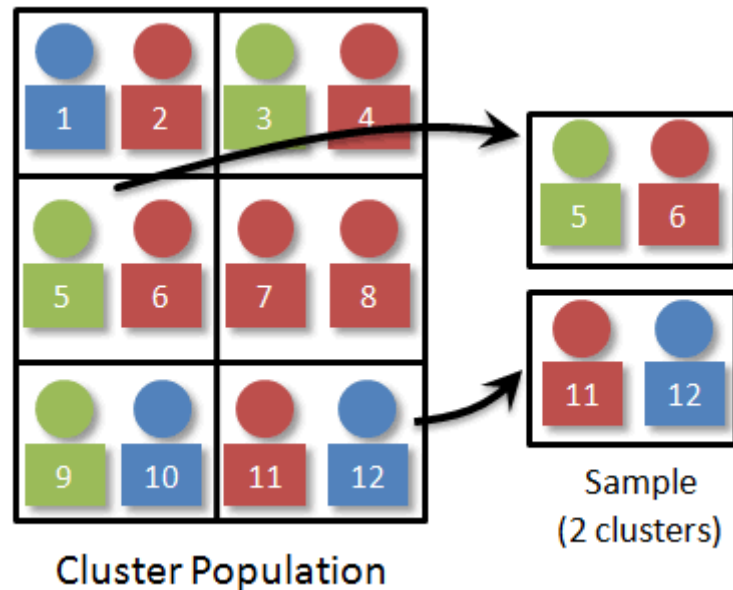
Other sampling methods with UNCW student height examples.

In **stratified sampling**, we subdivide the population into at least two different subgroups so that subjects within the sample subgroup share the same characteristics. Then we draw a sample from each subgroup.



Other sampling methods with UNCW student height examples.

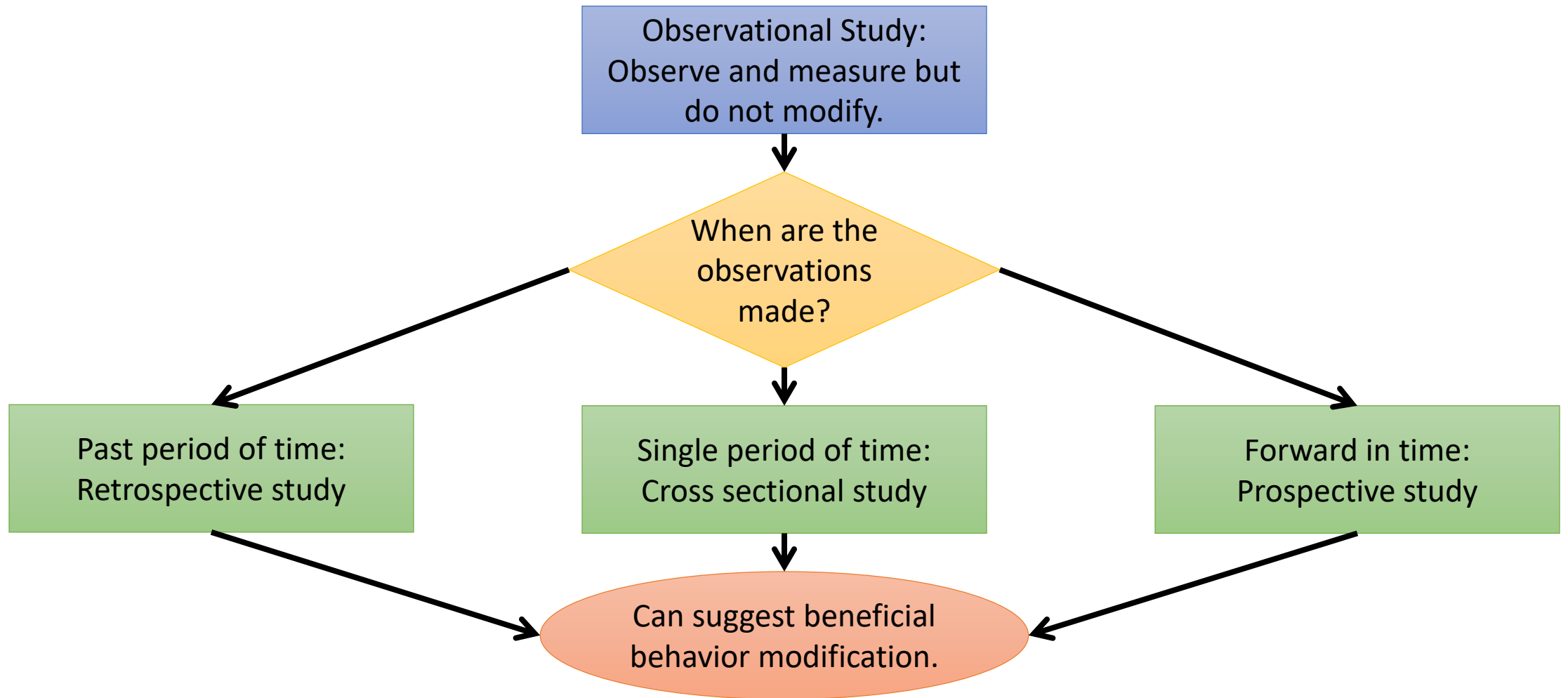
In **cluster sampling**, we first divide the population area into sections. Then we randomly select some of those clusters and choose all the members from those selected clusters.



Three types of observational studies

- In a **cross sectional study**, data are observed, measured, and collected at one point in time, not over a period of time.
- In a **retrospective** (or **case-control**) **study**, data are collected from a past time period by going back in time (through examination of records, interviews, and so on)
- In a **prospective** (or **longitudinal** or **cohort**) **study**, data are collected in the future from groups that share common factors (such groups are called cohorts)

Three types of observational studies



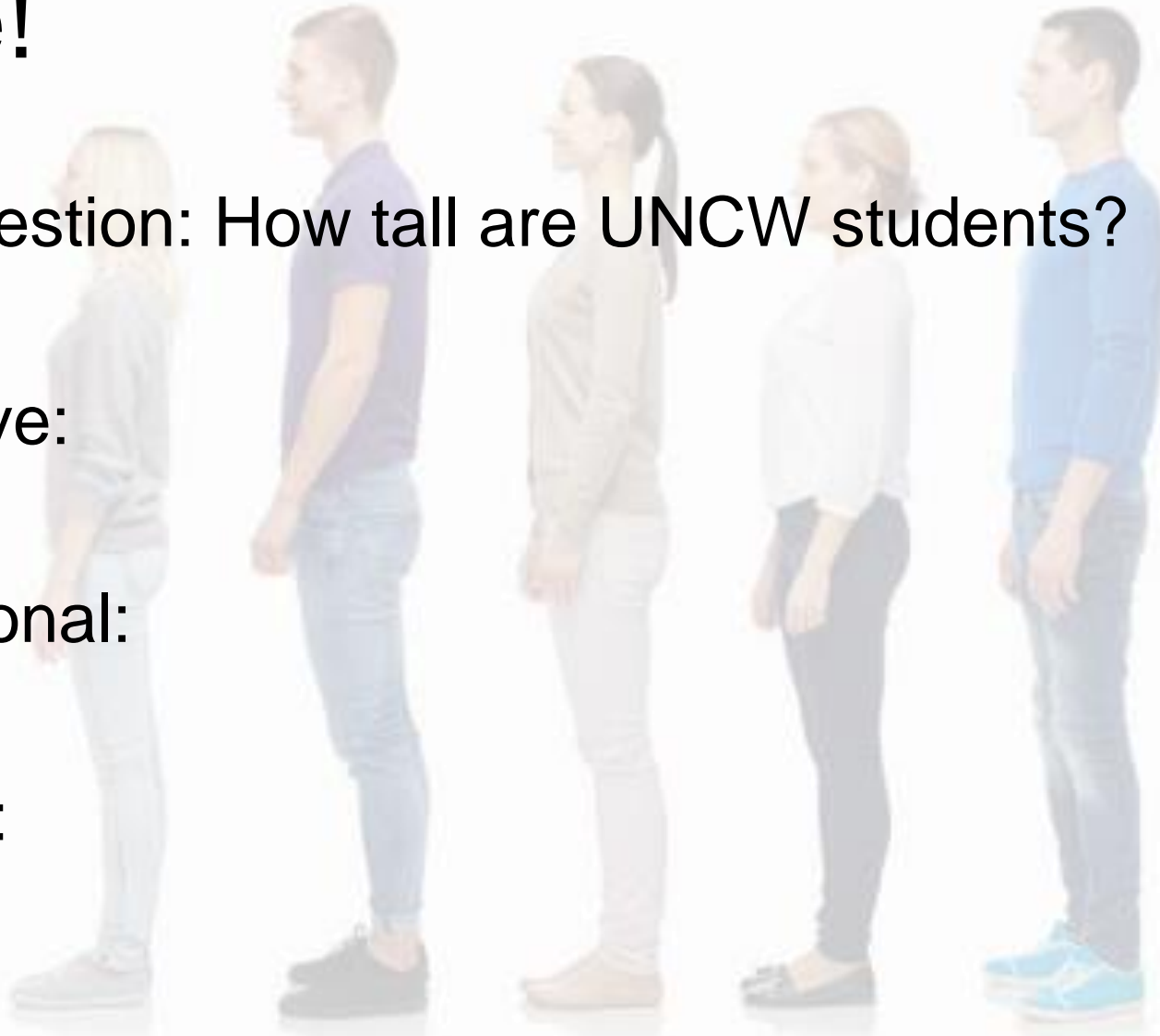
Example!

Research Question: How tall are UNCW students?

- Retrospective:

- Cross Sectional:

- Prospective:



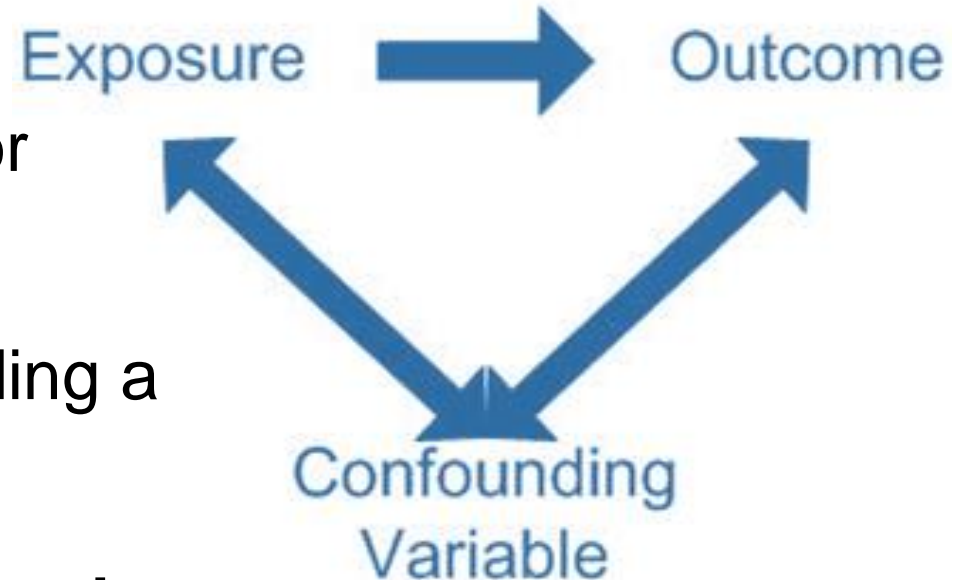
Confounding

Confounding occurs when we can see an effect but cannot determine the specific factors that caused it.

Confounding can occur in experiments or observational studies.

We call a variable that causing confounding a **confounder**.

In a experiment, a **completely randomized experimental design** can minimize confounding.

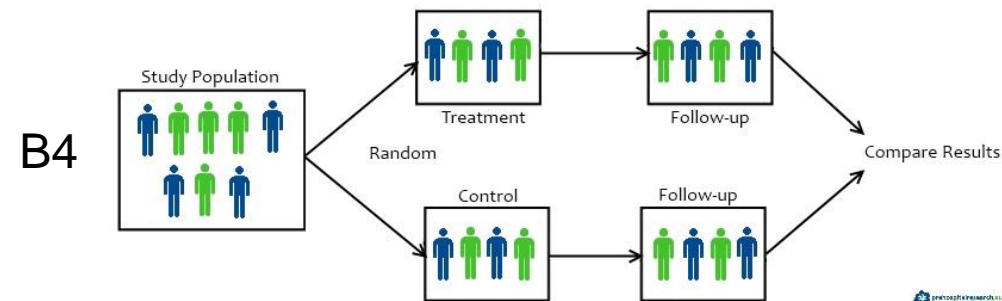
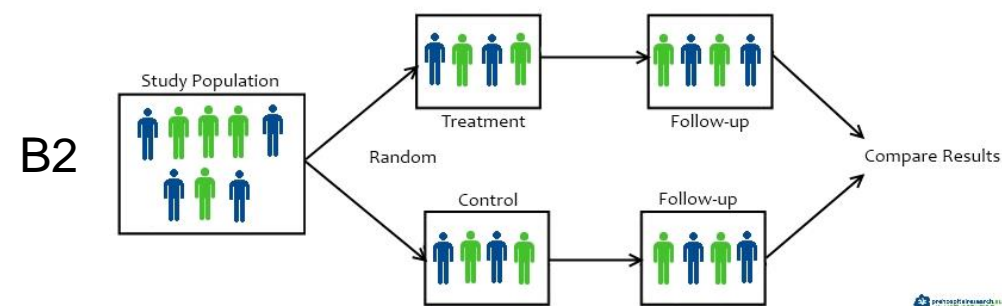
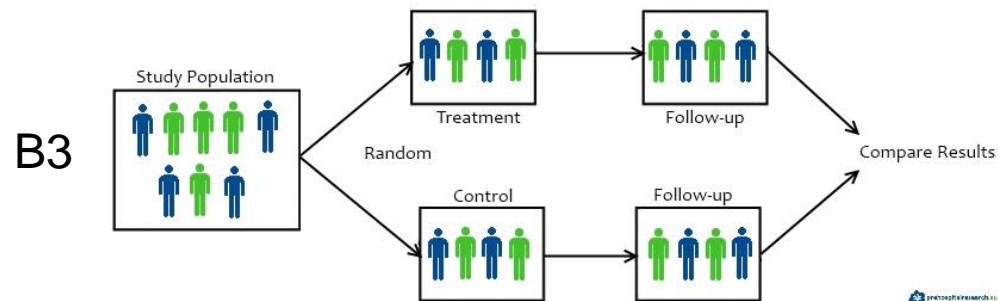
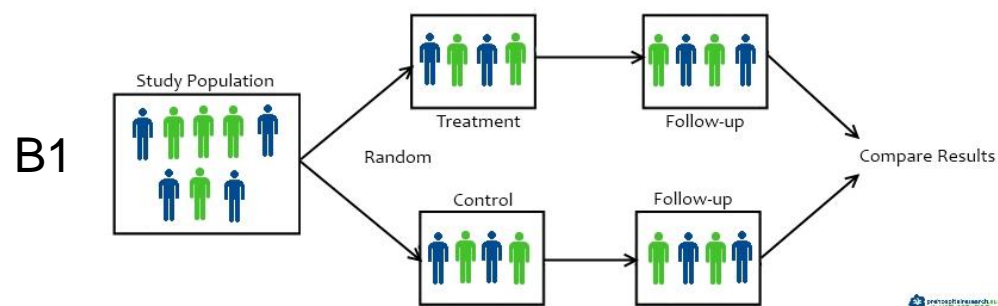


Other designs

- Randomized block design:

A block is formed of individuals with similar characteristics. Then, randomization occurs within each block.

e.g. blocks formed by race

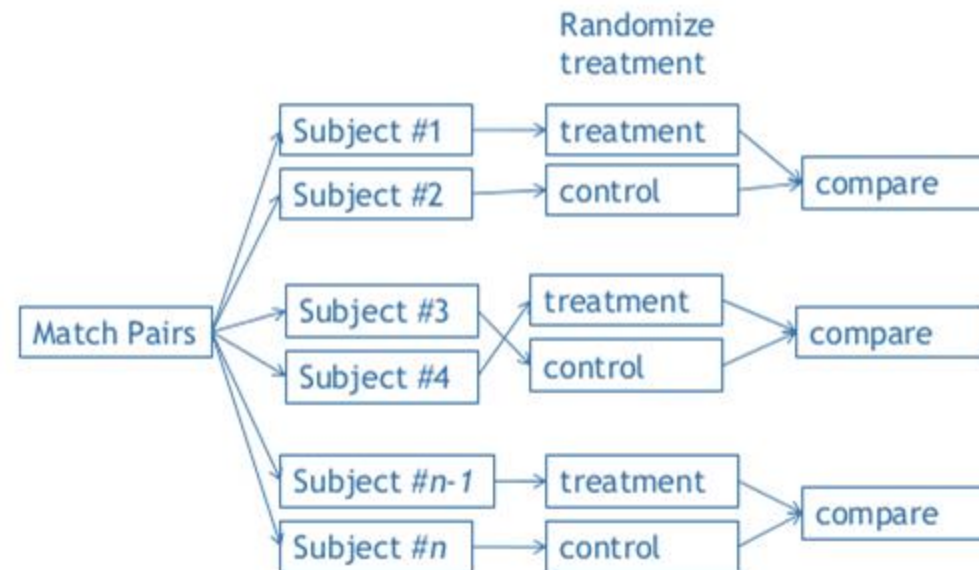


Other designs

- Matched pairs design:

Compare two treatment groups across matched subjects.

e.g. identical twins!



Other designs

- Rigorously controlled design:

Carefully assign subjects to different treatment groups, so that those given each treatment are similar in the ways that are important to the experiment.

Sampling errors

- **Sampling error**

- Sample selected with a random method
- Discrepancy between the sample result and the true population result
- Results from chance

- **Nonsampling error**

- Results from human error
- E.g. data entry error, wrong statistical method, etc

- **Nonrandom sampling error**

- Sampling method implemented was not random



ERROR

Section 3 Homework

1-8

5 of 9-20 (5,7,9,12,13)

21, 22, 24, 29

This is not to be turned in,
but beneficial for your
understanding.

