Lecture 1

Populations, Samples, and Processes

Today's Updates / Reminders

- Every lecture's 2nd slide will show important information concerning upcoming due dates or testing information. Be sure to review this. I start each class with a brief mention of these items.
- WebAssign Orientation assignment – "Getting Started with WebAssign – Statistics with SALT" is open now, due a week from today.
- Assignments are due at the "end of the day". Homework is always due on a "class day." But you have until 4am in case you're a night owl. This means the due date will show in the **homework software** as a Wed, Fri, or Sat at 4am. Know this means "end of the day" on a Tuesday or a Thursday. Moodle has artificial due dates that show 11:59pm. The homework software will show 4am.

Instructor Information

- Spencer Hamrick
  - Website: https://hamrick.wordpress.ncsu.edu/
  - YouTube Channel: https://www.youtube.com/spencerhamrick
  - Read more about me: https://hamrick.wordpress.ncsu.edu/about-me/
- Office: SAS 5268
- Email: schamri2@ncsu.edu
- Telephone: 919-515-1939
- I have an open door policy. I want to get to know you.

Course Information

- Office Hours
  - M-Th, 11:45am-1:00pm, in my office
  - Zoom by appointment
- Communication
  - Email is best. Identify your section and be specific.
- Moodle – look for here for common questions/solutions.

Syllabus

- Find all these details and more inside Moodle
- You must buy access to WebAssign. Text is optional (and electronic version is included with WA).
- Access to R and Excel (R is free/on campus computers)
- Calendar of dates is a separate attachment. Make note of test dates and let me know ASAP of any conflicts.
- Grades are in Moodle. WA will never be fully accurate.

Grades

- Grades are calculated as follows:
  - Labs (5%) – long-form problems in WA using software
  - PRWA (10%) – peer-reviewed writing assignments
  - Homework (15%) – short-form problems in WA
  - Midterms (40%, 20% each) – two midterms
  - Final Exam (30%) – cumulative, see exam schedule

Midterm Exams

- Given during class time, lasting 75 minutes. Graded using Gradescope
- Exams are closed book, closed notes
- 1 letter size page formula sheet allowed – must be handwritten – can use both sides
- Cheating = F
- Makeup exams must be taken within 5 business days
- Failure to show up or give notice of absence prior to the start of the exam results in a letter grade (-10 point) deduction

Labs

- Homework and labs are completed in WebAssign.
- Labs are a series of questions around one theme. You will use either R or Excel to complete the work. Tutorials are included with the assignment.
- No late labs accepted. Your lowest lab will be dropped.
- Typically due on Fridays of each week. None due the first week, Friday closest to exams, or finals week.

Homework

- Homework and labs are completed in WebAssign.
- Homework are short-form questions. You may use SALT (a software inside WA) or any software you choose.
- No late homework accepted. Two lowest HWs dropped.
- Five tries for each homework, the higher score counts.
- Due dates will vary. Typically due a week after we finish learning the topics in that homework set.
- Don’t wait to start. The longer between due dates, the larger the homework assignment. Putting it off until the day it is due will be overwhelming.

PRWA

- PRWA = peer-reviewed writing assignment (in Moodle)
- There are two assignments, each having 2 due dates.
- The first due date, you must make your own submission.
- The second due date (a week later), you must grade 3 classmates.
- Work is completed anonymously. No names on papers.
- You are graded, both on your work and your grading.
- If you miss the first due date, you get a zero for all of it!

How Class Runs

- I post notes to my website as a PDF – print and bring with you. Or download and write using stylus.
- Notes, as posted, will be incomplete. If you miss class, watch the recording to get the info.
- Ask questions. I also do my best to bring “entertainment value” to class. Don’t be surprised if I know your name.
- Come see me. Talk to me. I want to get to know you.
- Questions?

Chapter 1.1

Populations, Samples, and Processes

What is Statistics?

- What is the science of Statistics about?
  - Statistics is the branch of mathematics that deals with the collection, organization, analysis and interpretation of data.
- Statistics is specifically used in drawing general conclusions about a group known as the population, from a sub-group of it known as the sample.

Population vs. Sample

- **Population:** entire group of "entities" that we want to study and get information about.
- **Sample**: part of the population that we actually examine in order to gather information.

Important: Population and Sample are made up of the same objects!

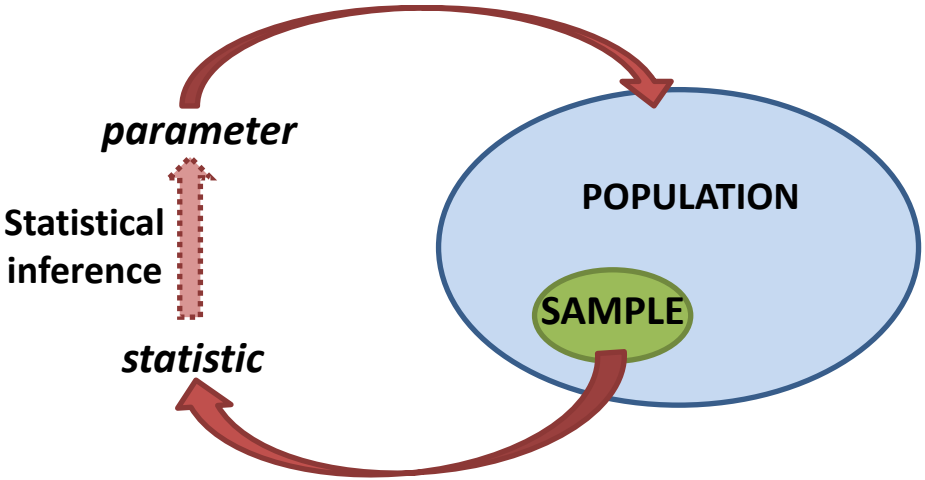
We could try to gather information from every single individual in the population; that is called a **Census.**

- **Census:** attempt to collect information from the entire population

Problem with census? Time consuming, of course. Not practical.
That is why we study a **sample;** but we have to be careful on how we choose that sample.

Parameter vs. Statistic

- **Parameter:** It is a fixed number that describes the population, but in practice we do not know its value.
- **Statistic:** it is a number that describes the sample. Its value is known when we have chosen the sample, and it can change from sample to sample. We use a statistic to estimate an unknown parameter or draw a conclusion about the parameter.



Example *: We design a study to estimate the average lifetime of a certain brand of battery "X" produced at a given factory during a month. Suppose that the factory produces 1 million of those batteries per month. We measure the average lifetime of 200 batteries.*

- What is the **population**? 1 million batteries
- Does it make sense to measure the lifetime of all of them? What that would be? No, they are 1 million! It would be a **census**.
- What is the **sample**? 200 batteries we choose to measure the lifetime.
- What is the **parameter** (what we want to know about the population)? The average lifetime of battery "X" (which is a fixed number).
- What do we collect? **Data**; we measure the lifetime of 200 batteries.
- What is the **statistic** (or what we calculate based on the data)? The average lifetime of the sample (which changes if we change the sample).

Conclusion: Based on the value of the statistic (taken from the sample) we will infer the value of the parameter (that describes the population).

Types of Samples

How do we choose the sample from the population?

1. **Voluntary response sample:** consists of people who choose themselves by responding to a general appeal. It is usually biased* because people with strong opinions are more likely to respond.

2. **Simple random sample (SRS):** consists of N individuals from the population chosen in such a way that every possible set of N individuals has an equal chance to be the sample selected*. It is not biased. It gives every individual an equal chance to be chosen.

***Bias:** a leaning in favor of or against something or someone; partiality or prejudice.
*The sample in a SRS is a **probability sample** or a sample chosen by chance.

The next two types of samples restrict the random selection.

Sometimes it is better to sample individuals with similar characteristics in separate groups. Each group is referred as a **strata.**

3. **Stratified random sample:** Is the result of dividing the population into different strata with similar characteristics, and then choosing separate simple random sample (SRS) in each of them to form a full sample.

4. **Multistage random sample**: Selects consecutive smaller groups within the population in stages, resulting in a sample consisting of clusters of individuals. Each stage may employ a SRS, a stratified sample, or another type of sampling method.

Examples: What kind of sample?

a) At a party there are 32 students over age of 21 and 16 students under age 21. You choose at random 4 of those over 21 and separately choose at random 2 of those under 21 to interview about attitudes toward alcohol.

Stratified random sample

b) You are planning a report on apartment living in a college town. You decide to select 5 apartment complexes at random for interviews with residents.

Simple random sample (SRS)

c) An online poll asks people who visit the website to choose their favorite TV show.

Voluntary response sample

d) There are seven sections of an statistics course. A random sample of three sections is chosen, and then a random sample of 8 students from each of these sections are chosen.

Multistage random sample

Problems with Samples

Random selection eliminates bias in the choice of a sample, however, bias can still be present for other reasons.

- **Undercoverage:** occurs when some groups in the population are left out of the process of choosing the sample. Example: an opinion poll aimed at all American households; if the poll is conducted by phone, it will leave out all the American households without residential phones.
- **Nonresponse:** occurs when an individual chosen for the sample cannot be contacted or doesn't cooperate.
- **Response Bias:** misleading or unreliable response due to the behavior of the interviewer or the respondent. Example: a survey asks the respondent about an illegal behavior.
- **Poorly worded questions:** occurs mainly in surveys when questions are confusing, or when they try to influence the answer.

Example: The NYC Marathon

The New York city marathon is a renowned annual event organized since 1970. It is the largest marathon in the world with 50,304 finishers in 2013. A group of students is interested in analyzing how the "average age" of the finishers has changed throughout the years. To do that, they randomly select 3 different years from each decade (70', 80', 90', 00'), then they separate the finisher runners from the selected years in men and women, and finally they choose randomly 100 men and 100 women from each year for the study.

a) What kind of sampling method are the students using? Multistage random sample
b) What is the population for this study? All the NYC marathon finishers from 1970, up to 2010
c) What is the sample? 2,400 finishers: 1,200 men & 1,200 women
d) What is the parameter? Average age of all the NYC marathon finishers from 1970 to 2010
e) What is the statistic? Average age of the 2,400 finishers in the sample.

Suppose that after concluding the study, the students decide that their results are valid for all the US marathons, not just the NYC one. What kind of problem could they have with the sample they originally chose?

 Undercoverage

To perform a statistical study we need to collect data.
Besides the data we collect, data could be:

- **Anecdotal data:** come from stories about cases that do not necessary represent a larger group. Example: *A friend tells you that she received a flu shot and then got the flu. Can you conclude that flu shots don't work?*

- **Available data:** were produced for some other purpose but may help answer the question of interest. Example: *If you want to know the average salary per hour for workers in the US you can use available data from U.S. Bureau of Labor Statistics website.*

Note: producing new data is expensive, so it is good to search the Internet or a library to use available data when possible.

Enumerative vs. Analytic

- Enumerative studies are based on an existing population. We can make inferences. There is a sampling frame from which to choose.
  - I need to test and compare the heat capacity of these 5 alloys.
- Analytic studies are based on a population yet to be created. There is no existing population so inferences can only be made by someone with expertise. A prototype may not be anything like the final product.
  - I want to create an alloy that has a specific heat capacity. I will try varying levels of alloying elements to see the effect.

How do we get data?

Two main sources (or ways) to obtain data:

- **Observational study** (example: sample survey): we observe individuals and measure variables of interest but do not attempt to influence the responses.

- **Designed experiment**: we impose some treatment on individuals and we observe their responses.

Observational Study, Experiment, or Anecdote?

- A group of researchers want to estimate the average concentration of lead in Indiana's bodies of water. To do that, they randomly select 50 bodies of water across the state and measure the concentration of lead in each of them.
  Observational Study

- It is suspected that learning to play the piano could improve the attention span in young children. To investigate that, a group of psychologists decide to conduct a study. From all the elementary schools in the US, they first select 20 schools randomly, and from each of those schools, they select 10 children randomly. The attention span of each kid is measured before the study begins. After that, these kids attend piano lessons for 6 months. At the end, the attention span for each of them is measured again.
  Designed Experiment

- A friend tells you that reading before bedtime makes her sleep better. You assume that must be true and start doing the same.
  Anecdote

Design of Experiments

Terms associated with experiments:

1. **Experimental unit**: units on which the experiment is done (if the units are human beings, they are called **subjects**).
2. **Treatment**: experimental condition applied to the units.
3. **Factors**: explanatory variables that we control (variables the outcome of the experiment might depend on). Also known as independent variables.
4. **Levels**: Specific values of factors. Each level (or combination of levels) is associated with a specific treatment.
5. **Response variable or outcome**: what is measured for each unit. Also known as dependent variable.

Example (Introduction to the Practice of Statistics, by Moore, McCabe and Craig, page 174)

The Tennessee STAR program was an experiment on the effects of class size. The subjects were 6385 students who were beginning kindergarten. Each student was randomly assigned to one of three treatments: regular class (22 to 25 students) with one teacher, regular class (22 to 25 students) with one teacher and one teacher aide, and small class (13 to 17 students) with one teacher. The students stayed in the same type of class for four years, and then all returned to regular classes. In later years, students from the small classes had higher scores on standard tests.

a) What or who are the **experimental units**? The **subjects** are the 6385 students.

b) What is the **treatment** applied to the units? To place them in different types of classes.

c) What is the **factor** in this experiment? The type of class.

d) What are the **levels** of the factor? We have 3 levels: regular size + 1 teacher, regular size + 1 teacher + 1 aide, small size + one teacher.

e) What is the **response variable** or **outcome**? The scores in the standard tests.

When performing an experiment or observational study we have to be careful with different types of problems that can arise if the study is not well designed or thought.

- **Lurking variables:** hidden variables, variables that are not controlled and that can influence the final result (lurk = to stay hidden)

Example: if in the Tennessee STAR program instead of choosing students randomly to place them in smaller classes, we just did an observational study comparing the standard test performances of students in smaller classes with those in bigger classes →lurking variables: students in smaller classes came from communities with more resources, or from more educated parents.

- **Placebo effect:** In medicine, a placebo is a dummy treatment (like a sugar pill). The placebo effect happens when people have a favorable response (due to personal attention) to a placebo that they hope will help them.

- **Bias:** results are skewed in a certain direction (usually due to failure in random allocation).

  Example: An endurance study applies two different levels of exercise (easy and difficult) and sends mostly well fit subjects to the higher level of exercise.

- **Lack of realism:** Happens when the experimental units, or the treatment, or the setting of an experiment doesn't duplicate the conditions we want to study.

  Example: Psychologists wanted to study how layoffs at a workplace affect the workers that remain on the job. To do that, they asked college students to proofread a text for extra credit and then they "let go" some of them who were accomplices with the experimenters. Lack of realism: we cannot be sure the reactions of the students are the same as those of workers who survive a real layoff.

How to avoid/reduce some of the above?
An experiment should be able to show causation.

Causation: the changes in the factors are the only cause of the changes in the response variables (or outcomes). A good experiment should be able to minimize the effect of lurking variables.

There are **three principles of experimental design** (to show causation)

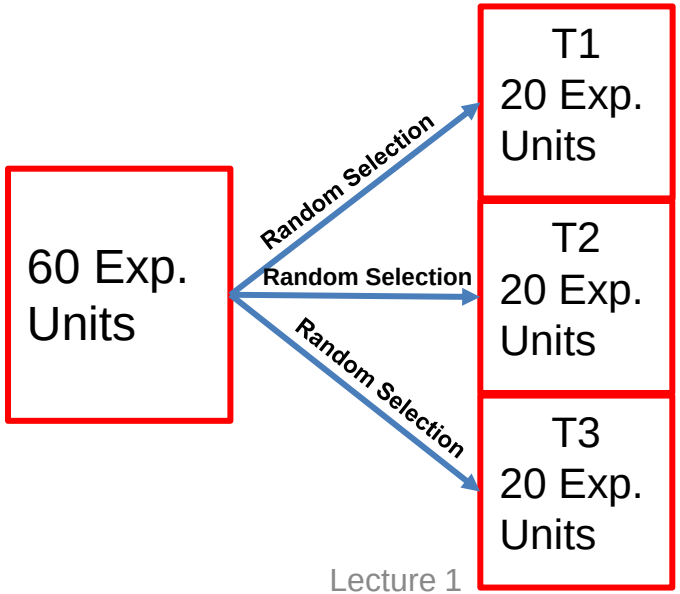
1. **Compare:** An experiment should have two or more treatments to control the effects of lurking variables on the response. One group should receive no treatment; that is the **control group**.

2. **Randomize:** Use impersonal chance to assign experimental units to the different treatments. This avoids bias. Also, ideally, an experiment should be **double-blind** (neither the subjects nor the experimenters know which treatment the subject received).

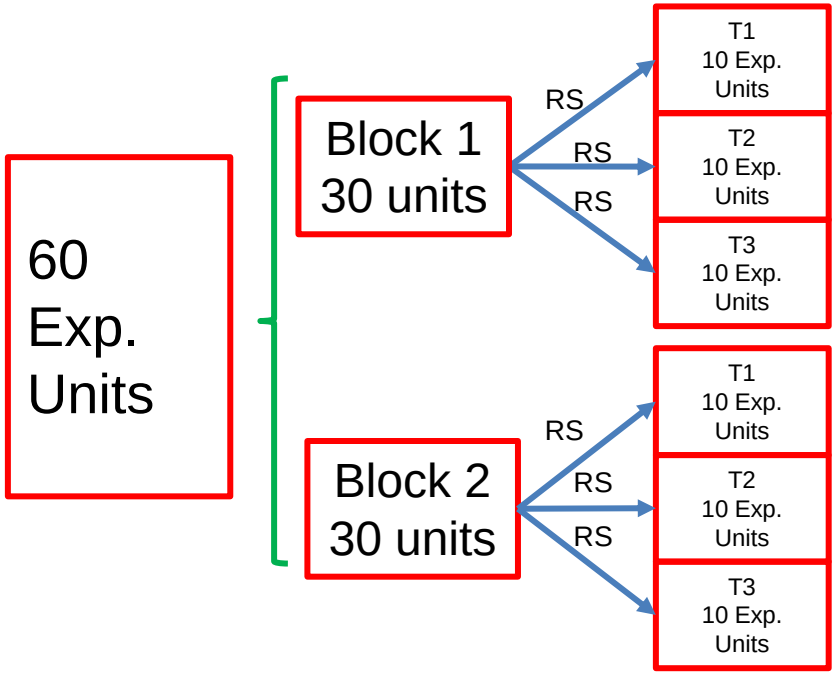
3. **Repeat:** Each treatment should be applied to many units to reduce chance of variation in the results.

Types of Experimental Design

1. **Completely randomized design (CRD):** all experimental units are allocated at random among the different treatments.



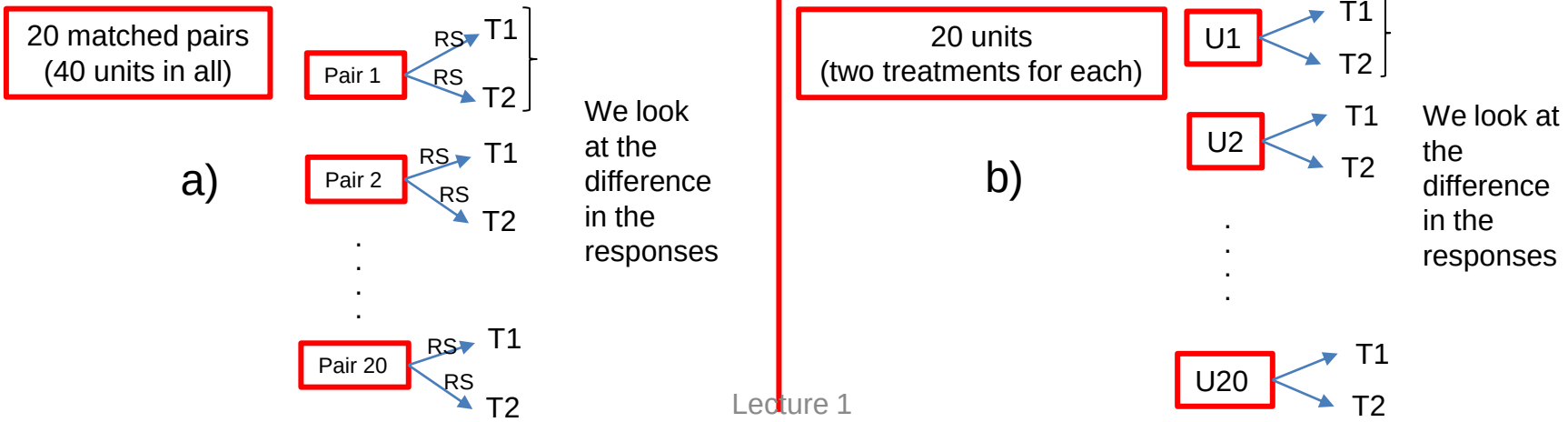
2. **Randomized block design:** the units are separated into blocks based on similarities before performing the random assignment to all the treatments.



3. Matched pairs design: Two possibilities:

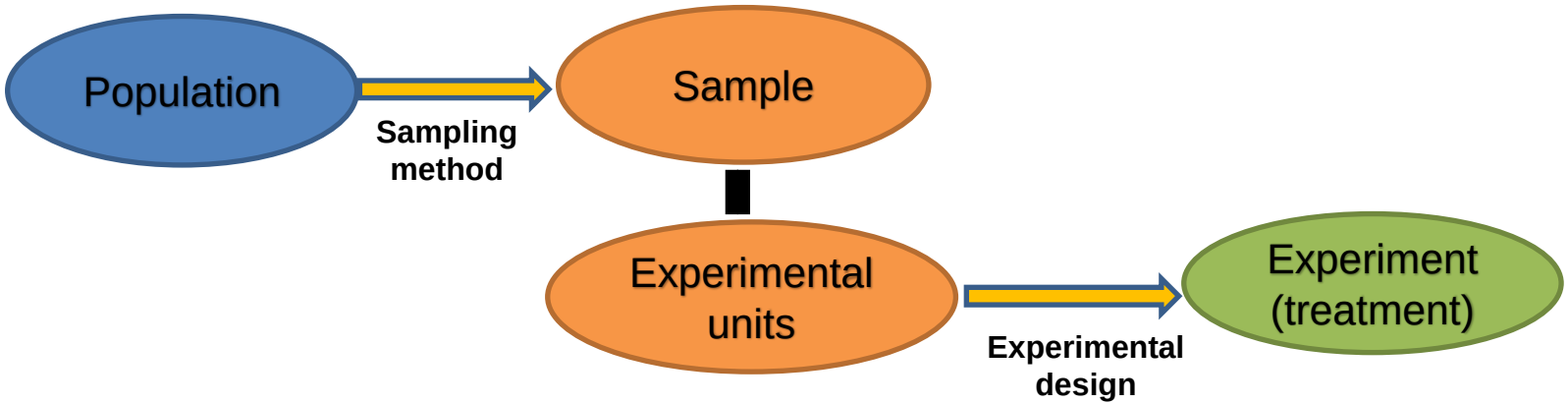
a) It compares 2 treatments; the experimental units are matched in pairs based on similar characteristics; the units in each pair are randomly allocated between the 2 treatments.

b) It compares 2 treatments; each experimental unit receives both treatments. Example: 2 results on each experimental unit, one result previous to a treatment and the other obtained after a treatment.



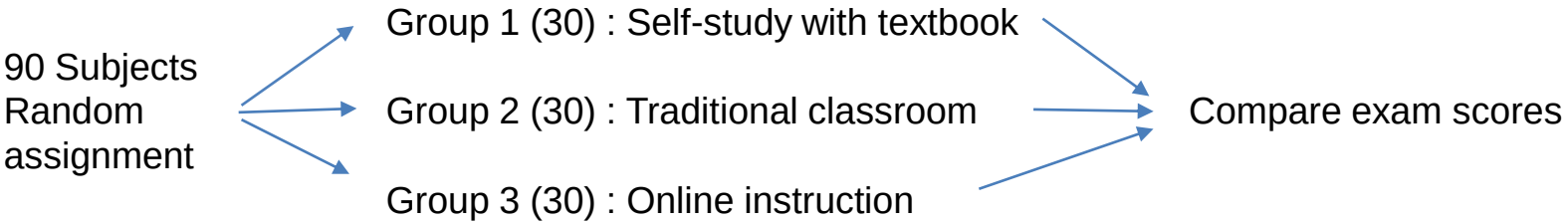
Do not confuse sampling method with experimental design!!

The sampling procedure comes before the assignment of the experimental units to the treatments. Then the sample will form or become all the experimental units to which the experiment is applied.



Example of Completely Randomized Design CRD:

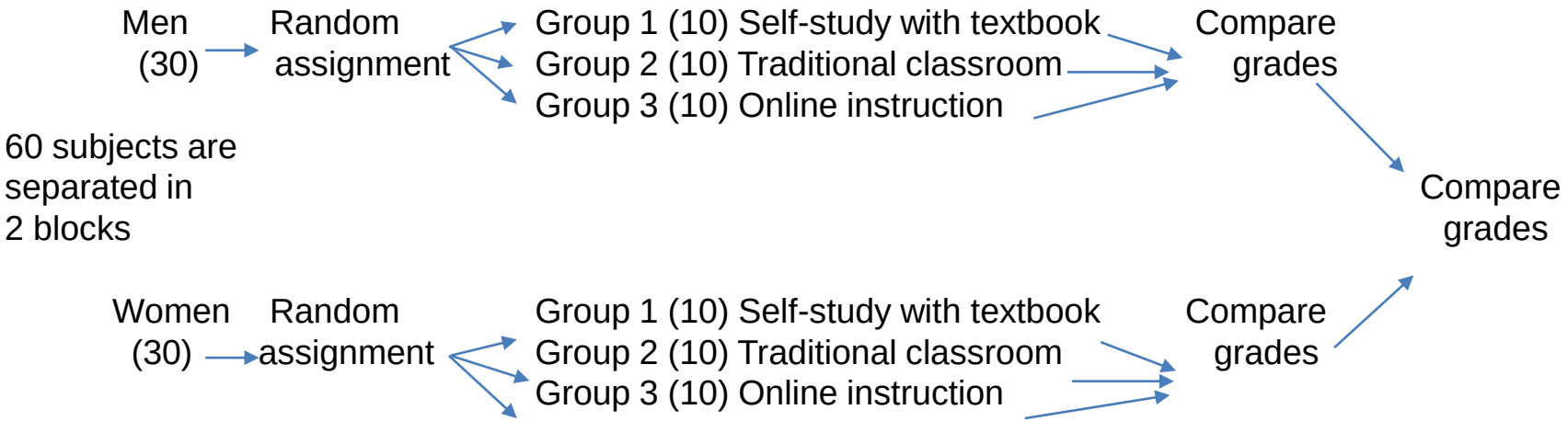
What instructional method helps students learn new material better: self-study from the textbook, traditional classroom, online instruction. A sample of 90 students are randomly assigned to one of these three groups (treatments). Scores on an exam are measured to compare these methods of teaching.



- What are the experimental units? Subjects→ The 90 students
- Treatment? The implementation of different instructional methods
- Factors? One factor: The type of instructional method
- Levels? How many? Three: Self-study, traditional, online.
- Response variable? Scores on an exam.

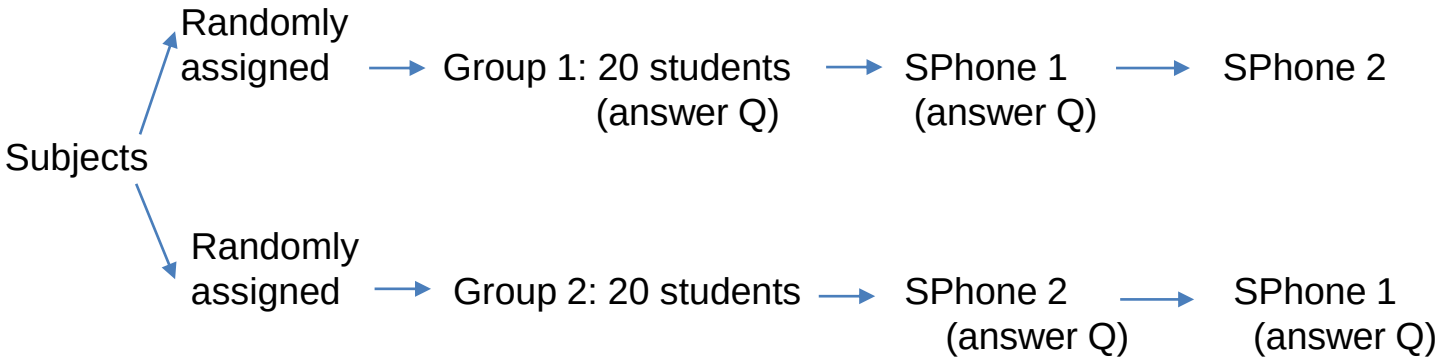
Example of Randomized Block Design:

Men and women may have different learning styles. Does this affect the instructional method that is most effective? A sample of 30 men and 30 women are assigned randomly to the three instructional methods from the previous example. An exam is given to determine the amount of material retained.



Example of Matched Pairs Design:

Two teams have each prepared a prototype for a new smartphone. Before deciding which one will be marketed , the smartphones will be evaluated by college students. Forty students will try the smartphones. Each student will use each phone for two weeks, and then answer some questions about how did they like the phone. Twenty students will be chosen randomly to evaluate Phone 1 first and Phone 2 second. The other 20 will evaluate Phone 2 first and Phone 1 second.



At the end we compared the answers to see which smartphone should be marketed.