1 Collecting Data

1.1 Planning a Study

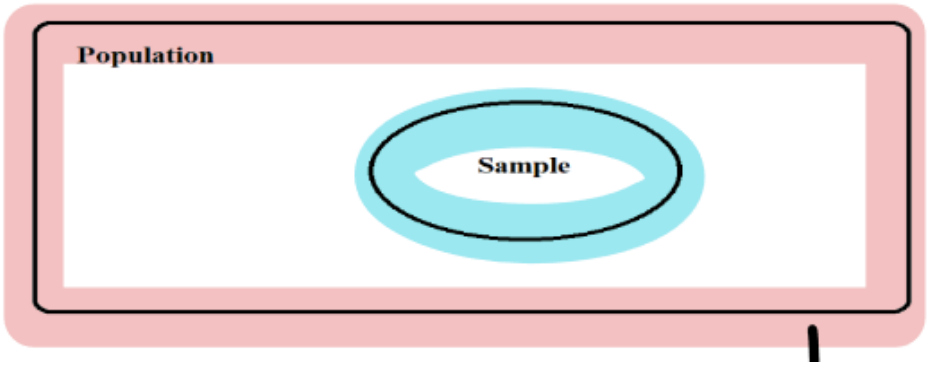
In order to better understand the characteristics of a population, statisticians and researchers often use a sample from that population and make inferences based on the summary results from the sample.

Population - the entire group we want information about.

A population can be huge like "all women" or small like "top 200 grossing movies in 2023."

Sample - a part of the population we actually examine.

The size of the sample can vary and depends on several factors we will examine through the course.



Census - collects data from every individual in the population.

The way we collect data influences what we can and cannot say about a population.

Observation Study - observes individuals and measures variables of interest but does not attempt to influence the responses.

- In an observational study, treatments are not imposed
- Investigators examine data for a sample of individuals (retrospective) or follow a sample of individuals into the future collecting data (prospective) in order to investigate a topic of interest about the population.
- A sample survey is a type of observational study that collects data from a sample in an attempt to learn about the population which the sample was taken

Experiment - deliberately imposes some treatment on individuals to measure their responses.

- We will discuss experiments in a later topic in this unit.

For now, let's discuss the various ways we plan an observational study.

It is only appropriate to make generalizations about a population based on samples that are randomly selected or otherwise representative of that population.

- A sample is only generalizable to the population from which the sample was selected.
  - For example, if you poll a sample of women asking about their shopping habits, you cannot take that result and apply it to men, since they were not represented in the sample that was taken.
- It is not possible to determine casual relationships between variables using data collected in an observational study.

  - While we observe variables and gather data in an observational study, we cannot make inferences between the variables. As long as it is a well-designed observational study, we can only apply the findings from the sample to the population.

Because we make inferences about a population from the sample, it is very important that the sample is collected appropriately and that is it representative of the population being studied.

Convenience Sample:

- Definition: Uses subjects that are readily avaliable.
- Advantage: Easy and less costly to collect
- Disadvantage: Not representative of the population
- Example: In order to get an idea of how students think of the new school policy, the principal stands outisde the library and asks a few students their opinions.

Voluntary Response Sample:

- Definition: A sample obtained by allowing subjects to decide whether or not to respond.
- Advantage: Easy to collect
- Disadvantage: Overrepresents people with strong opinions
- Example: After the State of the Union speech, ABC tells its audience to call a 1-900-555-1234 if they thought the speech was good and 1-900-555-7890 if they thought the speech was bad (there is a $0.50 charge for the call).

Simple Random Sample (SRS)

- Definition: Consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance in the sample selected.
- Advantage: Easy to accomplish using a table of random digits; likely to produce samples that are good representatives of the population
- Disadvantage: Cost or time could be an issue
- Example: In order to determine how happy students are with their education at a high school, the principal assigns each student a number from 1 to 1230 and then uses a random number generator to choose 50 numbers between 1 and 1230. He then surveys all the students with the chosen numbers.

Stratified Random Sampling

- Definition: Divide the population into groups of similar individual (strata) then select an SRS within each strata. Combine the SRSs from each strata to form your full sample.
- Advantage: Can produce more exact information (especially in large populations) by taking advantage of the fact that individuals in the same strata are similar to one another
- Disadvantage: Not appropriate unless strata are easily defined
- Example: In order to get a better idea of what a high school's athletes thought about homecoming last year, the director divides all the athletes into the teams they play for, and then selects a random sample from each sports team. His full sample consists of aggregating the random samples from each team.

Cluster Sampling:

- Definition: Divide the population into sectors (clusters) then randomly choose a few of those clusters. Each member of the cluster becomes your sample.
- Advantage: Don't need a list of entire population
- Disadvantage: More variability between samples depending on how clusters are determined.
- Example: A psychologist at the University of Texas collects a sample by first dividing up the students into their respective schools (engineering, nursing, arts and sciences) then by the departments that their major is in, and then she selects a few departments that their major is in, and then she selects a few departments at random and surveys every student within those chosen departments.

Systematic Random Sampling

- Definition: Randomly select an arbitrary starting point and then select every kth member of the population.
- Advantage: Every member has an equal probability of being selected.
- Disadvantage: Not every sample of size n has an equal chance of being selected.
- Example: HP selects every 200th computer off the assembly line and inspects it for quality control.

When an item from a population can be selected only once, this is called without replacement. When an items from the population can be selected more than once, this is called with replacement.

Samples are biased if they are systematically not representative of the desired population.

- Bias occurs when certain responses are systematically favored over others.

Voluntary Response: When a sample is comprised entirely of volunteers or poeple who choose to participate, the sample will typically not be representative of the population (voluntary response bias).

- Example: A radio talk show host asks listeners to call in with their opinions of making wearing masks in public space mandatory.

Undercoverage: Occurs when some groups in the population are left out of the process of choosing a sample.

- Because they are generally fearful of government intrusion, many immigrants from Latin America did not return their census questionaire during the 1990 census.

Non-response: Occurs when an individual chosen for a sample can't be contacted or refuses to respond. Non-response is a big problem in mail surveys.

- Example: Our administration sends out 100 survey questions to a sample of parents in order to gage their attitudes towards returning to school in 2020. Only 23 respond.

Response Bias - bias caused by the behavior of the respondent or interviewer.

Untruthful Answers: people give untruthful answers for several reasons..

1. Sensitive Questions: How often do you drink alcohol?
2. Socially Acceptable Answers: Do you use corporal punishment with your children?
3. Interview Bias: One year after the Detroit race riots of 1967, interviewers asked a sample of black residents in Detroit if they felt they could trust most white people, some white people, or none at all. When the interviewer was white, 35% answered "most", when the interviewer was black, 7% answered "most".

Ignorance: people will give silly answered just so they appear to know something about the subject.

- Example: In a study, educators were asked how they would rank Princeton's undergraduate business program. In every case, it was rated among the top 10 departments in the country, even though Princeton doesn't offer an undergraduate business major.

Lack of Memory: giving a wrong answer simply because respondent doesn't remember the correct answer.

- Example: Students were asked to report their grade point averages. Researches then determined the actual GPA's. Over 17% of the students reported a GPA that was .4 or more above their actual average, and about 2% reported a GPA more than .4 below their actual GPA. (most inflated their GPA's!)

Timing: When a survey is taken can have an impact on the answers

- Example: In January, the National Football League reported a poll that revealed football as the nation's favorite sport (this is at the time of the Super Bow;)

Phrasing: Subtle differences in phrasing make large differences in the results.

Example:

- Should the president hae the line-item veto to eliminate waste? 97% said "yes"
- Should the president have the line-item veto? 57% said "yes"

When drawing a sample, two types of errors may occur:

Sampling Error: The difference between a sample result and the true population result. This error results from chance variation.

Example: Place 50 red and 50 green balls in a bag. Mix the balls thoroughly and randomly sample 30 balls. In your sample you find that 12 balls are red and 18 are green. Your sample result is different than the true population ratio of 1 to 1. The difference is due to sampling error. Virtually any experiment involving a sample will have sampling error. We can minimize sampling error through various statistical techniques, the most obvious is to increase sample size.

Non-sampling error: Occurs when the sample data is incorrectly collected, recorded, or analyzed. Usually occurs when the sample is selected in a non-random fashion.

Example: In order to gage student opinion on a new grading policy, an administrator stands outside the library during common time and asks a sample of 50 students if they agree with the new policy. The administrator finds that 25 out of the 50 students sampled agree with the new policy. When the entire student body is asked for their opinion, however, the results were 30% in factor 70% against. The difference between the sample percentage and the true population percentage is due to non-sampling error, because the sample was collected in such a way that a lot of bias was involved (convenience sampling).

1.2 Selecting a Random Sample

The Hat Method

- This is a classic description of an SRS that can be used on the AP exam to describe selecting a random sample but it is rarely done in practice due to it being being so time consuming.
- Script: "Write down all the names or numbers on their own slip of paper. Then put all the pieces of paper into a hat, mix well in-between selections, and pull out the desired amount of slips (mention with or without replacement)"

Calculator - Random Number Generator

- MATH - PROB - 5:randInt(lower, upper, n)
- Can be used to describe on the AP Exam but not really used for "doing" random samples on the AP Exam because there is no way to "check" for it.
- In practice, computer generated numbers are the least time consuming.
- If using this method, make sure you "seed" your calculator, otherwise everyone will get the same "random numbers".
  - MMDDYYYY - STO→ - MATH - PROB - 1:rand-ENTER

Random Digit Table

- Given as a long string of digits 0-9 within the question. The digits are grouped in 5s to make it easier to read but it has no significant meaning.

TABLE B Random digits

Line								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335

Choosing SRS with a Random Digit Table

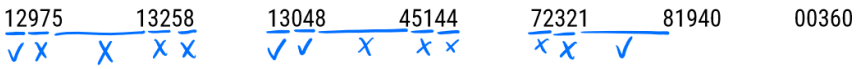
1. Label: Assign a number label to every individual in the population.
2. Random Digits: Start at the very first number and identify how many digits you will take at a time.

3. Stop: Indicate when you should stop sampling (toss out repeated numbers or numbers out of your range).
4. Identify Sample: Use the random numbers to identify subjects to be selected from your population. This is your sample!

Example

The school newspaper is planning an article on family-friendly places to stay over spring break at a nearby beach town. The editors intend to call 4 randomly chosen hotels to ask about their amenities for families with children. They have an alphabetized list of all 28 hotels in the town. Starting at Line 140 (given below), find an SRS of 4 hotels. Describe how you would select your SRS and then collect your sample.

01 Aloha Kai	08 Captiva	15 Palm Tree	22 Sea Shell
02 Anchor Down	09 Casa del Mar	16 Radisson	23 Silver Beach
03 Banana Bay	10 Coconuts	17 Ramada	24 Sunset Beach
04 Banyan Tree	11 Diplomat	18 Sandpiper	25 Tradewinds
05 Beach Castle	12 Holiday Inn	19 Sea Castle	26 Tropical Breeze
06 Best Western	13 Lime Tree	20 Sea Club	27 Tropical Shores
07 Cabana	14 Outrigger	21 Sea Grape	28 Veranda



1. Each hotel is given a number 01 to 28.
2. Choose 2 digits at a time. If the number is 01-28, it will represent a hotel in the sample.
3. Ignore numbers 29-99 and repeats. Repeat process until 4 hotels are chosen.
4. The checkmarks in the image are those to include in the sample, and the x's are skips.

The four hotels are 04, 12, 13, and 18.

Observational Study: Observes individuals and measures variables of interest but does not attempt to influence the response.

Experiment: Deliberately imposes some treatment on individuals to measure their responese.

Experimental Unit: the things on which the experiment is done

Subjects: experimental units that are human beings

Treatment: a specific experimental condition applied to the units

Example

Are the following scenarios an experiment or an observational study? Explain your answer.

(a) A medical team examines the records of 5 large hospitals and compares the survival times of those cancer patients who had surgery versus those who had chemotherapy.

Observational Study - medical team did not impose any treatments

(b) In a gym class, the effect of exercise on blood pressure is studied by requiring that half of the students walk a mile each day while the other students run a mile each day.

Experiment - treatments are running or walking a mile.

(c) The relationship between weights of bears and their lengths is studied by measuring bears that have been anesthetized.

Observational Study - weights and lengths are recorded, no treatments

(d) People who smoke are asked to halve the number of cigarettes consumed each day so that any effect on pulse rate can be measured.

Experiment - treatment is halving cigarettes

Example

Determine if it is an observational study or an experiment, and then identify the explanatory and response variables in each situation.

(a) One effect of alcohol is a drop in body temperature. To study this effect, researchers give several amounts of alcohol to mice, and then measured the change in each mouse's body temperature.

Experiment, explanatory: amount of alcohol, response: body temperature

(b) A study is done to try and find the correlation between verbal and math SAT scores. The scientist wants to use the verbal score to predict the math score.

Observational Study, explanatory: verbal SAT score, response: math SAT score

(c) Some breast cancer patients were given each a new treatment. The patients were closely followed to see how long they lived following surgery.

Experiment, explanatory: "new treatment", response: length of life

(d) To find out how well a child's height predicts their age a study was done where they measured the heights of a group of children at age 6, wait until they are 16 and then measure their heights again.

Observational study, explanatory: height, response: age

1.3 Experimental Design

Two advantages of an experiment over an observational study:

1. We can study the specific factors we are interested in while controlling the effects of lurking variables.
2. Experiments also allow us to study the combined effects of several factors.

How to design an experiment

- Factor: The explanatory variables in an experiment
- Level: the various groups the factors take

For example, if an experiment compared the drug doses 50 mg, 100 mg, and 150 mg, then the factor "drug dosage" would have three levels: 50 mg, 100 mg, and 150 mg

Prinicples of Experimental Design

1. Comparison
    - We want to make sure we are using a design that compares two or more treatments
    - We need to make sure the groups we are comparing don't differ greatly before our experiment begins or bias can result.
2. Randomization
    - The most important element of any experiment. It must be incorporated either in the selection process of experimental units and/or the distribution of experimental units into treatment and control groups.
    - You can use your calculator (random number generator), random digit table, the hat method, or flipping a coin to randomize an experiment.
    - Randomization produces groups of experimental units we expect to be similar in all respects before treatment is applied. Therefore, measured differences must be due either to treatment or by change of random assignment.

3. Control
    - Control group is treated identically in all respects to the group receiving the treatment except that the members of the control group do not receive the treatment.
    - The control group is our baseline and our experimental group has only one thing changed - the explanatory variable.
    - This reduces variability in the reponse variable. If one group is controlled, we would expect their responses to be controlled as well.
4. Replication
    - Use enough experimental units in each group so that any difference in the effects of the treatments can be distinguished from chance differences between groups.
    - Even with control, natural variability occurs among experimental units.
    - We would like to see units within a treatment group responding similarly to one another, but differently from units in other treatment groups (then we can be sure that the treatment is responsible for the differences).
    - If we assign many individuals to each treatment group, the effects of chance (and individual differences) will average out.

Experimental Terms

- Placebo: treatment designed to have no therapeutic value
- Placebo Effect: subjects receiving the placebo have a response that is similar to what we would expect from the treatment
- Single Blind: subject does not know what treatment they are receiving
- Double Blind: subject and administrator do not know who receives the treatment

The design of an experiment is crucial. Experiments are designed with the purpose of isolating the effect of the treatment on the response variable and removing any confounding effects.

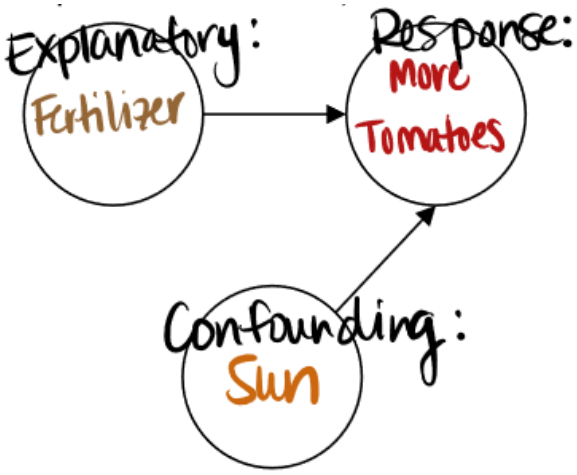
- In a poorly designed experiment, it might be difficult to tell if the explanatory variable causes a change or if it was another variable that wasn't measured
- Confounding variables are variables that might affect the outcome, but we did not control or account for them in our experiment.

One way to remove the effect of any confounding variable is to randomly assign subjects to the treatment or control group. This allows for any possible bias in the population to be evenly spread among the treatment and control groups. Sometimes instead of relying on randomization to make groups as even as possible we actually force the groups to be similar.

An extraneous variable is one that is not an explanatory variable in the study but is thought to affect the response variable. There are two types of extraneous variables present in studies:

Confounding variable refers to another variable that may affect the response and is in some way tied together with the factor under investigation. It leaves us unable to tell which of the two variables (or perhaps some interaction) caused the observed response.

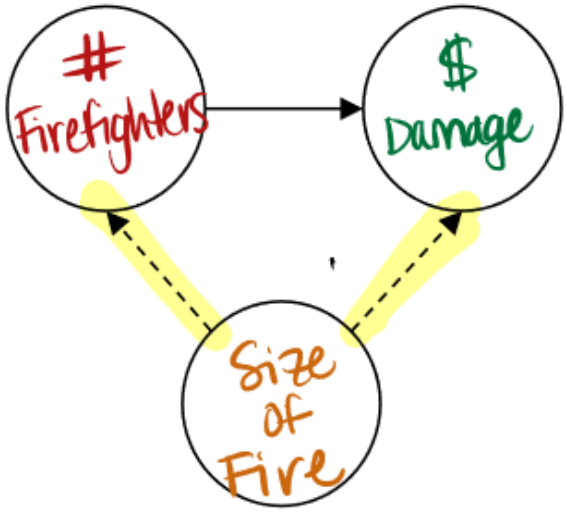
For example, we plant tomatoes in a garden that's half-shaded. We test a fertilizer by putting it on the plants in the sun and apply none to the shaded plants. Months later the fertilized plants grow more and better tomatoes. Why? Well, maybe it's the fertilizer, maybe it's the sun, maybe we need both. We're unable to conclude that the fertilizer works because any effect of fertilizer is confounded with any effect of the extra sunshine.



Lurking variable refers to a variable that drives each of the two variables under investigation, making it appear that there's some association between them.

For example, there is a strong association between the number of firefighters who respond to a fire and the amount of damage done. One shouldn't conclude that the firefighters may be responsible for the damage; the lurking variable is the size of the fire.

Lurking variables are the risk we face in sampling and observational studies. In an experiment, though, the factor under consideration isn't being driven by some lurking variable, because we are the ones in control there.



Example

State whether the relationship between the two variables involves a lurking or confounding variable.

(a) Does watching TV make you live longer? Measure the number of television sets per person, x, and the average expectancy, y, for the world's nation. There is a high positive correlation: nations with many TV sets have high life expectancies. Could we lengthen the lives of people in Rwanda by shipping them TV sets? Justify your answer.

Lurking variable is money.

Money pays for food and healthcare (increasing life expectancy). More TVs generally mean more money.

(b) Do artificial sweeteners cause weight gain? People who use artificial sweeteners in place of sugar tend to be heavier than people who use sugar. Does this mean that artificial sweeteners cause weight gain? Give a more plausible explanation for this association.

The confounding variable is diet.

People who use artificial sweeteners could be trying to lose weight so they may be heavier to begin with.

Inference is drawing conclusions beyond the data at hand.

Let's take a look at two different studies:

Study 1: The U.S. Census Bureau carries out a monthly Current Population Survey of about 60,000 households. Their goal is to use data from these randomly selected households to estimate the percent of unemployed individuals in the population.

Study 2: Scientists performed an experiment that randomly assigned 21 volunteer subjects to one of two treatments: sleep deprivation for one night or unrestricted sleep. The experimenters hoped to show that sleep deprivation causes a decrease in performance two days later.

- Random selection of individuals allows inference for the population
- Random assignment in an experiment allows inference for cause and effect

For the U.S. Census Bureau, individuals were randomly chosen to participate in the survey. The Bureau would be safe in making an inference about the population.

In the sleep deprivation experiment, subjects were randomly assigned to their treatments. If there is a large difference in the results, then we can assume it is not due to chance variation between the groups alone and must be due to sleep deprivation.

Well-designed experiments randomly assign individuals to treatment groups, but most don't select experimental units at random from the larger population, so their findings are limited to just cause and effect.

		Were individuals randomly assigned to groups?	
		YES	**NO**
Were individuals randomly selected?	**YES**	Inference about the Population: **YES** Inference about Cause and Effect: **YES**	Inference about the Population: **YES** Inference about Cause and Effect: **NO**
	NO	Inference about the Population: **NO** Inference about Cause and Effect: **YES**	Inference about the Population: **NO** Inference about Cause and Effect: **NO**

Both random sampling and random assignment introduce chance variation into a statistical study. When performing inference, statisticians use the laws of probability to describe this chance variation.

In some cases, it is not practical or ethical to do an experiment to establish a cause and effect relationship. Consider the following examples:

- Does texting while driving increase the risk of having an accident?
- Does going to church regularly help people live longer?
- Does smoking cause lung cancer?

There are laws now that must be followed when dealing with human subjects:

- Reviewed by an institutional review board - the board's purpose is "to protect the right and welfare of human subjects recruited to participate in research activities."
- Informed consent - subjects must be aware of the harm and danger a study could inflict and must be given written permission to participate.
- Confidentiality - protect individuals' privacy by keeping their identity separate from their results.

Writing up an experiment

1. Determine what type of design is best for your experiment. This will depend on the context of the question.
2. Diagram your experiment (time permitting).
3. Tell exactly how you will randomly assign your treatments.
    - Random Digit Table
    - The Hat Method

- Dice, Coin, or Playing Cards

4. Explain exactly what you are comparing once you gather the data

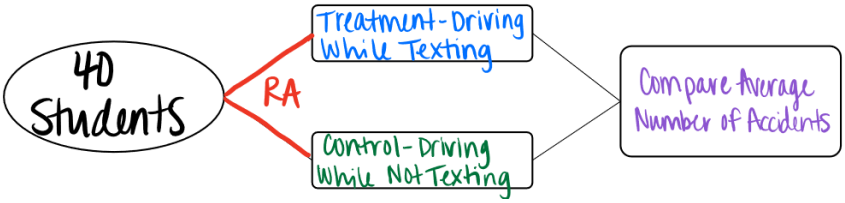
Completely Randomized Design

- The experimental units are assigned to treatments completely by chance.
- Treatment groups and control groups will be about equal in size in a completely randomized design.
- There are mathematical reasons for having groups of equal sizes, which we will discuss later.

Example

Is texting while driving causing accidents? There are 40 students that have volunteered for a study to determine if texting while driving causes more accidents. The county sheriff's department has given a driving simulator to use in the experiment. Design a completely randomized experiment.

- Write each student's name on identical slips of paper and place in a hat.
- Select 20 names from the hat without replacement and mixing well in between
- The 20 names selected will receive the treatment - driving while texting
- The remaining 20 students will receive the control - driving while not texting
- Compare the average number of accidents between the two groups



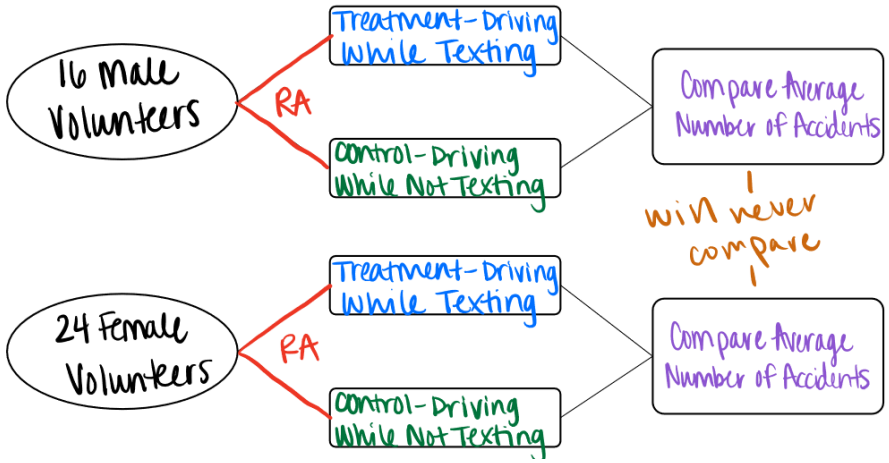
Randomized Block Design

- When groups of experimental units are similar, it's often a good idea to gather them together in blocks.
- Blocking isolates the variability due to the differences between the blocks so that we can see the differences due to the treatments more clearly.
- When randomization occurs only within the blocks, we call the design a randomized block design.
- Control what you can, block on what you can't control, and randomize to create compatible groups.

Example

It is brought to a teacher's attention that gender could be another contributing factor to the number of accidents people get into. Design an experiment to address this new information.

- Separate volunteers into two blocks based on gender.
- Number males 01-16. Use a random digit table to select 8 unique two digit numbers, ignoring 00 and 17-99.
- The 8 names selected will receive the treatment - driving while texting.
- The remaining 8 names will receive the control - driving while not texting.
- Repeat for females.
- Compare the average number of accidents between the two groups within each block.



Matched Pairs Design

- These are experimental designs in which either the same individual or two matched individuals are assigned to receive the treatment and the control.
- Often the "pair" in a matched pairs design is just one experimental unit which serves as its own control.
- In the case where an individual receives both the treatment and the control, the order in which this happens should be random.

Example

A student now brings up the fact that each student has different driving styles with many other variables that can influence the number of accidents. Design an experiment that would help addres the other variables present for individual drivers.

- Each volunteer will do both treatments.
- Randomly assign order of treatments by flipping a coin.
- Heads - driving while texting first
- Tails - driving while not texting first
- Perform remaining treatment the next day
- Repeat for other 39 volunteers
- Compare the difference in number of accidents for each of the volunteers