

Introduction to Statistics

Michael C. Hackett
Assistant Professor, Computer Science

Community
College
of Philadelphia

A First Case Study: Treating Chronic Fatigue Syndrome

- To introduce some basic ideas, we'll consider an experiment that evaluates the effectiveness of cognitive behavior therapy for chronic fatigue syndrome.
- Participant pool
 - 142 patients who were recruited from referrals by primary care physicians and consultants to a hospital clinic specializing in chronic fatigue syndrome.
- Actual participants
 - Only 60 of the 142 referred patients entered the study. Some were excluded because they didn't meet the diagnostic criteria, some had other health issues, and some refused to be a part of the study.

Deale, et. al. 1997. *Cognitive behavior therapy for chronic fatigue syndrome: A randomized controlled trial*. The American Journal of Psychiatry 154:3.

A First Case Study: Treating Chronic Fatigue Syndrome

- Patients were randomly assigned to treatment and control groups, 30 patients in each group.
 - Treatment Group: Patients in this group received cognitive behavior therapy.
 - Control Group: Patients in this group did not receive cognitive behavior therapy.

A First Case Study: Treating Chronic Fatigue Syndrome

- The table below shows the distribution of patients with good outcomes at 6-month follow-up.
 - Note that 7 patients dropped out of the study: 3 from the treatment and 4 from the control group.

		Good Outcome		Total
		Yes	No	
Group	Treatment	19	8	27
	Control	5	21	26
	Total	24	29	53

A First Case Study: Treating Chronic Fatigue Syndrome

- A **summary statistic** is a single figure that summarizes a large amount of data.
- Proportions of good outcomes:
 - Treatment Group: $19/27 \sim .70 \sim 70\%$
 - Control Group: $5/26 \sim .19 \sim 19\%$

		Good Outcome		Total
		Yes	No	
Group	Treatment	19	8	27
	Control	5	21	26
	Total	24	29	53

A First Case Study: Treating Chronic Fatigue Syndrome

- Do the data show a "real" difference between the groups?
 - If you flip a coin 100 times, and while the chance a coin lands heads in any given coin flip is 50%, we probably won't observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process.
 - The observed difference between the two groups ($70 - 19 = 51\%$) may be real or may be due to natural variation.
- The bigger the difference, the more believable that the difference is real.
 - The difference between 70% for the treatment group and 19% for the control group is quite large, so it is more believable that the difference between the groups is real.

A First Case Study: Treating Chronic Fatigue Syndrome

- But how big of a difference is big enough to determine if the difference is real?
 - In other words, how do we determine if the difference observed was by chance (like the coin flip example) or real?
 - These are the type of questions we'll answer as we progress in the course

A First Case Study: Treating Chronic Fatigue Syndrome

- Can the results of this study be generalized to all patients with chronic fatigue syndrome?
 - No.
 - These patients had specific characteristics and volunteered to be a part of this study; therefore they may not be representative of all patients with chronic fatigue syndrome.
- While we cannot immediately generalize the results to all patients, this first study is encouraging.
 - The method works for patients with some narrow set of characteristics, and that gives hope that it will work, at least to some degree, with other patients

Data Matrices

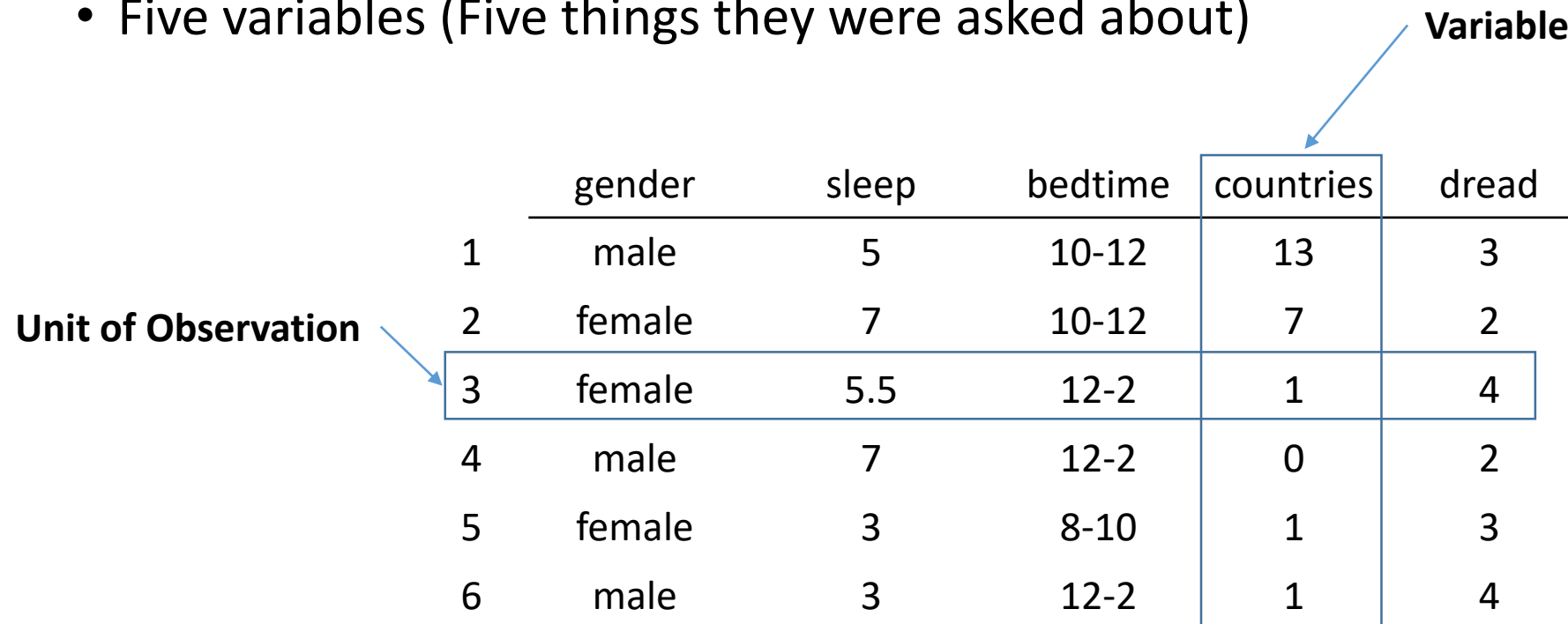
- A data matrix is a two-dimensional table used to organize data.
- Each row in the data matrix is a unit of observation (also called a case).
- Each column in the data matrix is a variable: a characteristic of the observation.
- The general organization of a data matrix is like a table in a database

Data Matrices

- To illustrate, let's consider a survey of students that asks:
 - The student's gender
 - Their average hours of sleep each night
 - Time frame that they usually go to bed
 - The number of countries they have visited
 - On a scale of 1 to 5, how much they dread going to class

Data Matrices

- The data matrix below is a hypothetical result of the survey
 - Six observations (Six students surveyed)
 - Five variables (Five things they were asked about)



	gender	sleep	bedtime	countries	dread
1	male	5	10-12	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	male	7	12-2	0	2
5	female	3	8-10	1	3
6	male	3	12-2	1	4

Variables

- All variables in a data matrix are described as either *numerical* or *categorical*.
- **Numerical variables** contain numerical data, specifically where it would make sense to perform calculations with those values.
 - Finding averages, adding, subtracting, etc.
- **Categorical variables** contain data from a selection of possible values.
 - Each category is called a **level**

Variables

- Gender: Categorical

- Two levels seen
 - male and female

- Sleep: Numerical

- Bedtime: Categorical

- Three levels seen
 - 8-10, 10-12, 12-2

- Countries: Numerical

- Dread: Numerical *or* Categorical

- We could treat it as a numerical variable to find the average rating
- We could treat it as a categorical variable with five levels, say 1 through 5

	gender	sleep	bedtime	countries	dread
1	male	5	10-12	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	male	7	12-2	0	2
5	female	3	8-10	1	3
6	male	3	12-2	1	4

Variables

- Numerical variables are either *discrete* or *continuous*
- Numerical variables are **discrete** when the possible values are finite.
 - A good rule of thumb is discrete variables have values that can be counted
 - For example, the number of pages in a book.
- Numerical variables are **continuous** when the possible values are infinite.
 - A good rule of thumb is continuous variables have values from measurements
 - For example, temperature.
 - There are infinite values between 75.1 and 75.2 (75.1, 75.11, 75.111, 75.1111...)

Variables

- Gender: Categorical

- Two levels seen
 - male and female

- Sleep: Numerical, **Continuous**

- Bedtime: Categorical

- Three levels seen
 - 8-10, 10-12, 12-2

- Countries: Numerical, **Discrete**

- Dread: Numerical *or* Categorical

- We could treat it as a **discrete** numerical variable to find the average rating
- We could treat it as a categorical variable with five levels, 1 through 5

	gender	sleep	bedtime	countries	dread
1	male	5	10-12	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	male	7	12-2	0	2
5	female	3	8-10	1	3
6	male	3	12-2	1	4

Variables

- Categorical variables can be either *ordinal* or *nominal*
- Categorical variables are **ordinal** when there exists a natural ordering of the values.
- Categorical variables are **nominal** when there is not a natural ordering of the values.

Variables

- Gender: Categorical, **Nominal**

- Two levels seen
 - male and female

- Sleep: Numerical, Continuous

- Bedtime: Categorical, **Ordinal**

- Three levels seen
 - 8-10, 10-12, 12-2

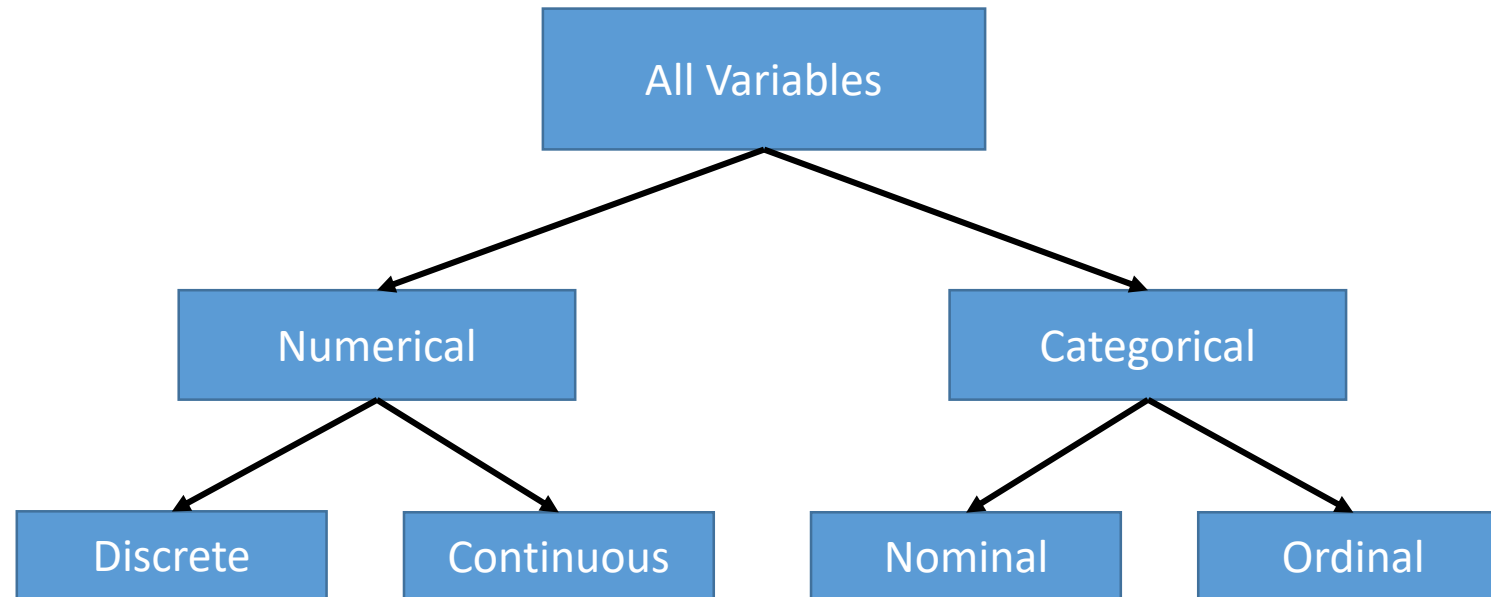
- Countries: Numerical, Discrete

- Dread: Numerical *or* Categorical

- We could treat it as a discrete numerical variable to find the average rating
- We could treat it as an **ordinal** categorical variable with five levels, 1 through 5

	gender	sleep	bedtime	countries	dread
1	male	5	10-12	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	male	7	12-2	0	2
5	female	3	8-10	1	3
6	male	3	12-2	1	4

Variables

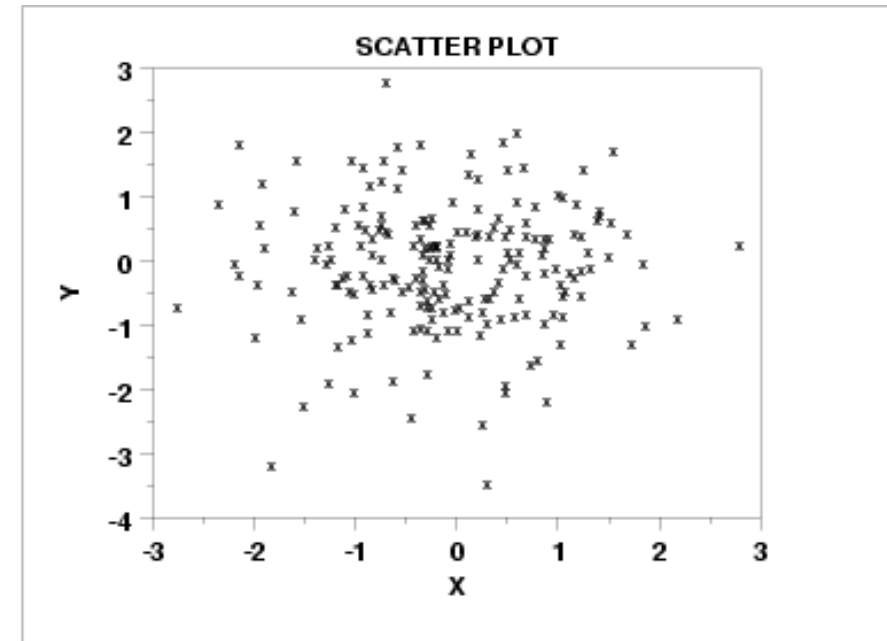


Relationships Among Variables

- **Associated variables** (also called a **dependent variables**) show some connection with one another.
- If two variables are not associated (no evident relationship between them) they are called **independent variables**
- Two variables are either dependent or independent, never both.

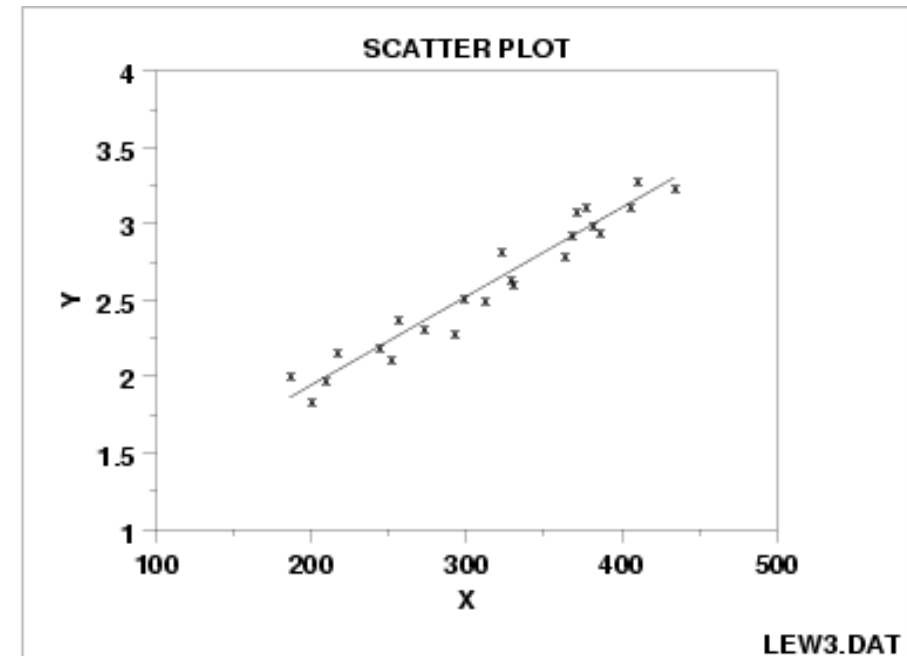
Relationships Among Variables

- In the scatterplot below, there is no discernable relationship between the variables X and Y
 - X and Y are **independent** variables



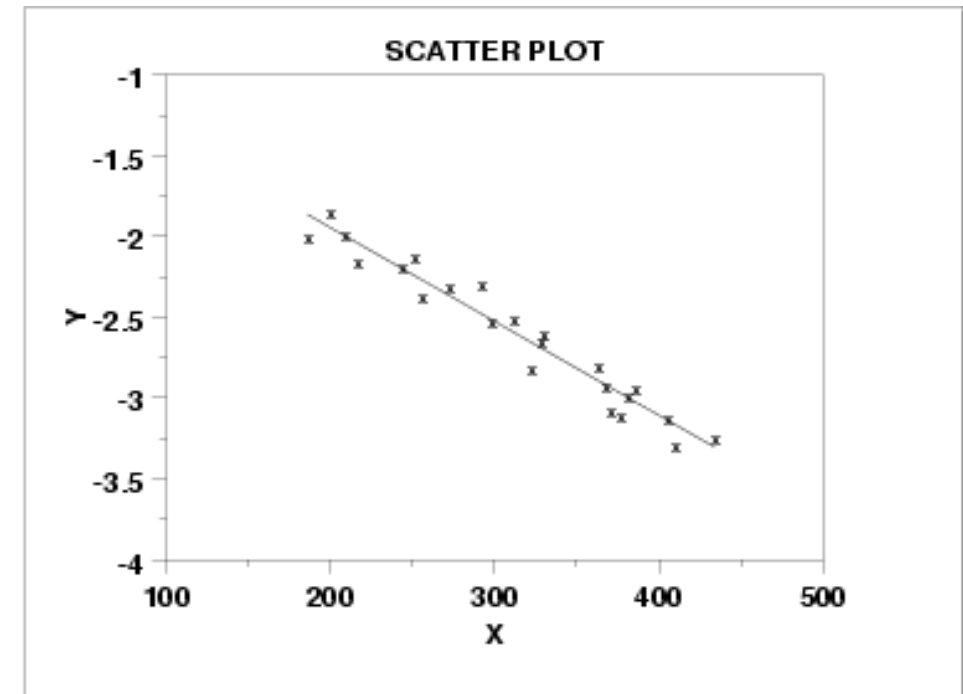
Relationships Among Variables

- In the scatterplot below, there is a clear relationship between the variables X and Y
 - X and Y are **dependent** variables
- Large values of X are associated with large Y values
 - This indicates a **positive** association
 - An upward slope



Relationships Among Variables

- In the scatterplot below, there is again a clear relationship between the variables X and Y
 - X and Y are **dependent** variables
- Large values of X are associated with small Y values
 - This indicates a **negative** association
 - A downward slope



Observational Studies

- Observational studies show possible associations between variables to form hypotheses that can be tested in an experiment.
- Here is a classic example of why observational studies alone are not sufficient for determining causality:
 - In 2009, a study showed that violent crimes were higher when ice cream sales were up and decreased when sales were down.
 - In 2006, a similar study showed that drownings increased when ice cream sales were up and decreased when sales were down.
 - One might conclude that ice cream sales are linked to death rates.

Observational Studies

- This example illustrates the maxim that *correlation does not imply causation*.
 - There may be a correlation between violent crimes, drownings, and ice cream but observational studies rarely account for every possible variable to support causality.
- In this example, the two observational studies fail to account for the weather.
 - In hot weather:
 - Ice cream sales are up
 - Violent crimes increase due to stress from the heat
 - Drowning increase because more people are swimming in hot weather

Observational Studies

- The weather was an example of a **confounding variable**, which is a variable that correlates with the explanatory (ice cream sales) and response (violent crime rate) variables.
- Because it is difficult (if not impossible) to eliminate all confounding variables, there is no guarantee that the results of an observational study will justify causal conclusions.

Observational Studies

- Observational studies are either prospective or retrospective.
- A **prospective study** collects information as events unfold.
 - Surveying students' opinions as they progress through their degree programs
- A **retrospective study** collects information about past events.
 - Surveying students' opinions about a course at the end of the semester.

Anecdotal Evidence

- **Anecdotal evidence** is data that may be true and verifiable but only represent extreme or one-time observations.
- For example, someone claiming, “My father smoked two packs a day and never had lung cancer, so smoking doesn’t cause lung cancer” is referencing anecdotal evidence.
 - It may be true, but it ignores the mountains of data to the contrary.
- Another example, a professor says, “A student got a 100% on the final exam, so my course adequately prepares students for the exam.”
 - This may be true, but is it representative of the class if no other student scored above a 70% on the same exam?

Sampling

- A **population** is the set of all possible units of observation.
 - Every registered voter in the USA
 - Every person who lives Wyoming
 - Every trout in the Ramapo River
- It's difficult (if not impossible) to observe or survey large populations like the examples above.
- A **sample** is a smaller subset of the target population
 - Should be representative of the population as a whole
 - Sampling the *entire* population is called a **census**

Sampling

- A “soup analogy to sampling”
 - You are preparing a soup (**population**)
 - You take a spoonful to do a taste test (**sample**)
 - You tasted the spoonful and decided the soup isn’t salty enough (**exploratory analysis**)
 - You conclude the entire soup needs salt (**inference**)
- For the inference to be valid, your spoonful needed to be representative of the entire soup.
 - If the spoonful came from the surface of the soup and the salt has settled at the bottom, then what you tasted was probably not representative of the soup (a bad sample)
 - If you stirred the soup first, then what you tasted was probably more representative of the soup (a good sample)

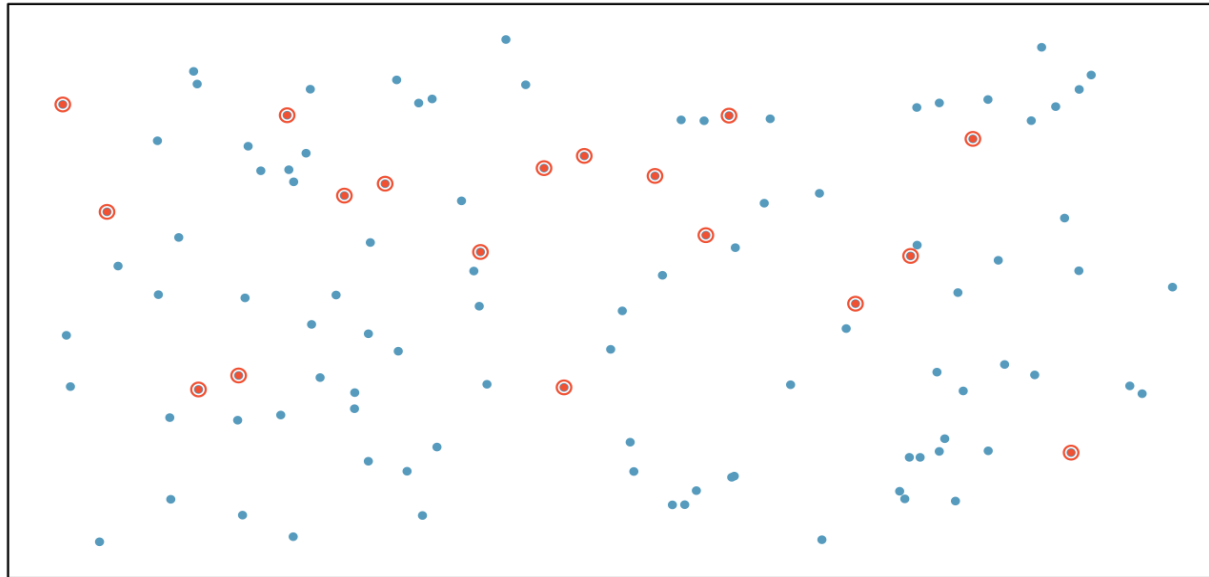
Sampling

- Nearly all statistical methods are based on the notion of implied randomness
- If observational data are not collected in a random framework from a population, these statistical methods (and the estimates and errors associated with the estimates) are not reliable.

Sampling

- **Simple Random Sampling**

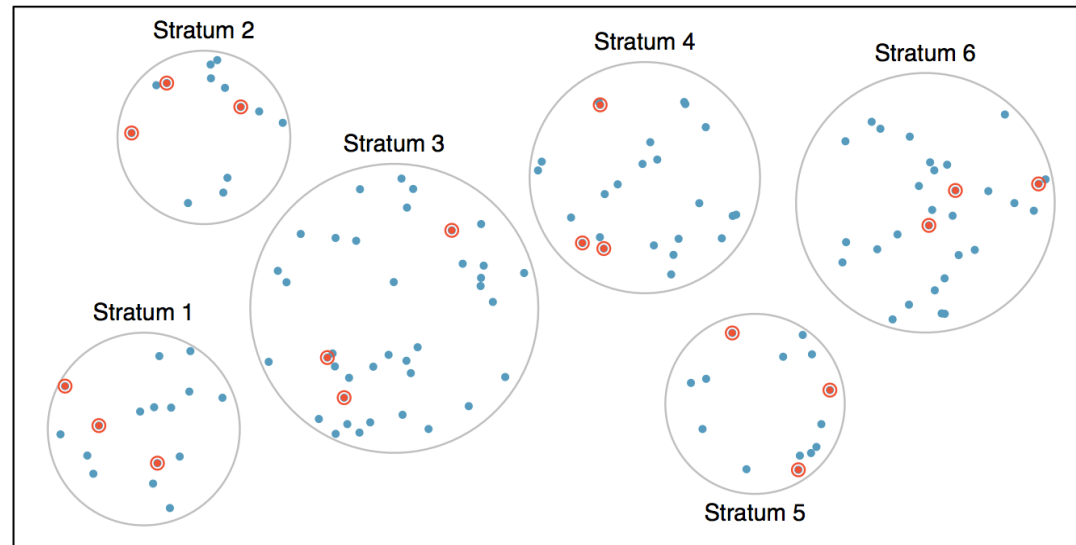
- Randomly choosing units of observation from a population
- Each unit has the same chance of being chosen



Sampling

- **Stratified Sampling**

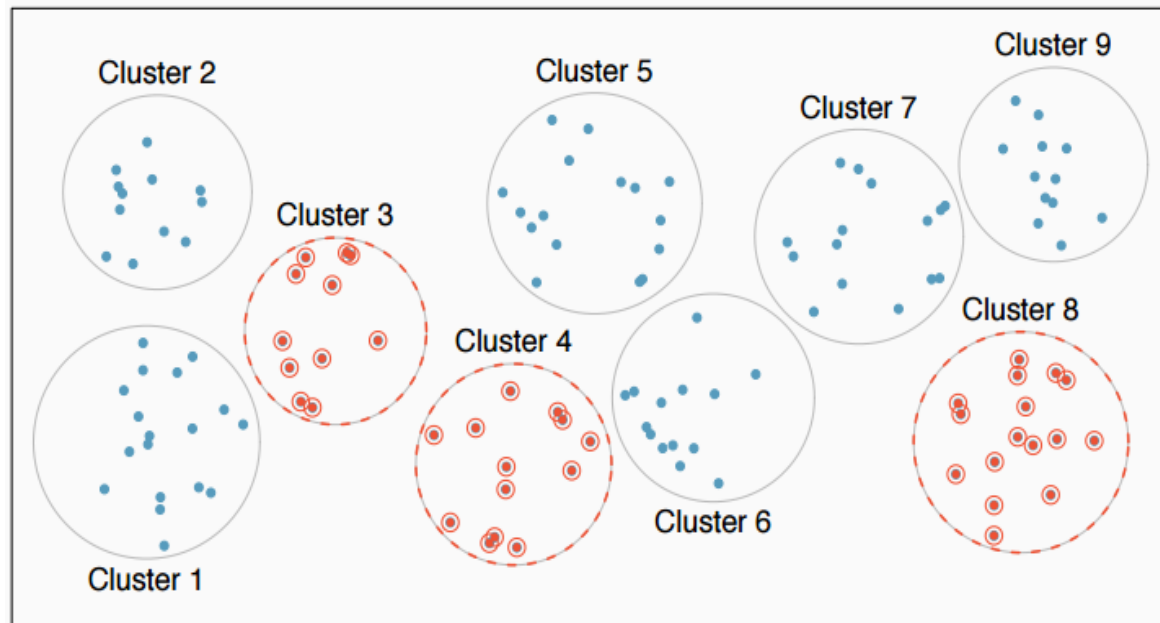
- Population is divided into groups containing similar observations (“strata”)
- A second sampling method (such as simple random) selects units from each stratum



Sampling

- **Cluster Sampling**

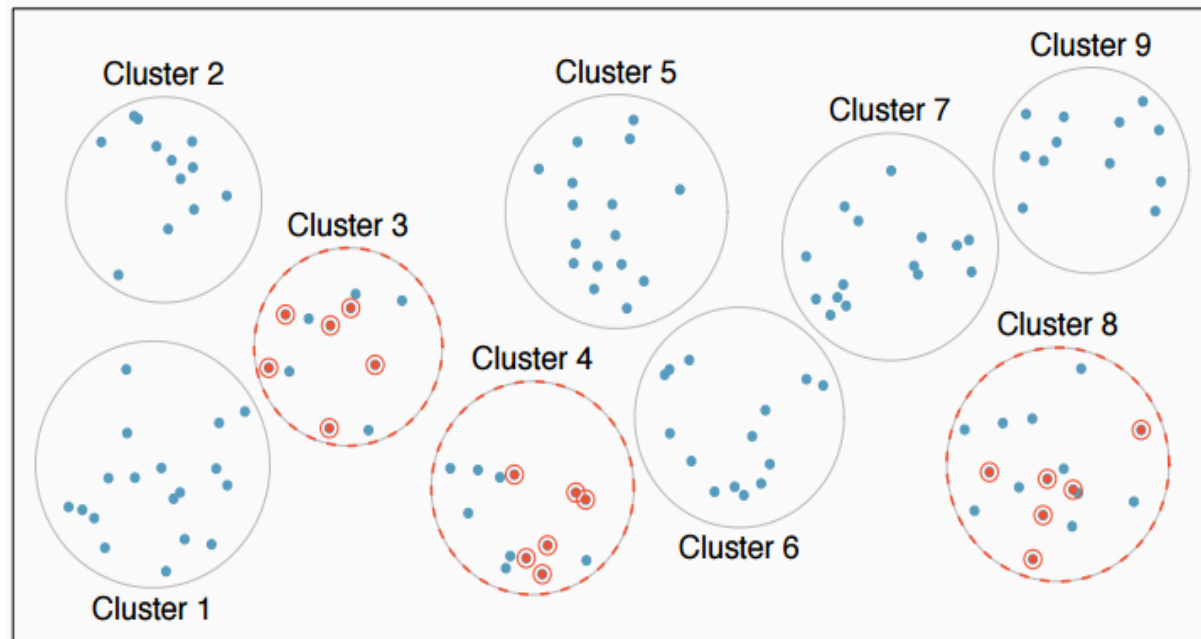
- Population is broken into groups (“clusters”)
- The samples consists of all observations in a certain number of clusters



Sampling

- **Multi-Stage Sampling**

- Like cluster sampling, but observations are selected using sampling method (such as simple random) to select observations from the chosen clusters



Sampling

- Let's say we want to poll every registered voter in Pennsylvania.
 - Target Population: Every registered voter in Pennsylvania
 - Unit of Observation: One registered voter in Pennsylvania
- We decide to use a sample size of 500 Pennsylvanians that are registered to vote.
- Think about the following:
 - Our entire sample of 500 voters all live in Philadelphia. Would our poll results accurately reflect all voters in Pennsylvania?
 - Our entire sample of 500 voters are all female. Would our poll results accurately reflect all voters in Pennsylvania?
 - Our entire sample of 500 voters are all White. Would our poll results accurately reflect all voters in Pennsylvania?

Sampling Bias

- The three previous examples illustrate **sampling bias**
- When hand-picking units of observation, there is a risk of creating bias (either intentionally or unintentionally) in the sample
- Let's say I am asked to survey CCP graduates on their current salaries and I survey my former students.
 - Most of them will be CIS/Computer Science majors, which would not reflect the entire population of CCP graduates.
 - This would be a case of (most likely) unintentionally creating a biased sample.

Sampling Bias

- Let's say I am instead asked to survey CCP Computer Science graduates on their current salaries and I survey my former CS students.
 - If I only choose graduates who I know have a tech career with a high salary, would it represent the entire population of CCP Computer Science graduates?
 - This would be a case of intentionally creating a biased sample.
- Thus, when creating a sample it's best to apply one of the previously shown sampling methods, such as simple random sampling.

Sampling Bias

- Even with randomly chosen samples, other forms of sampling bias may manifest.
- **Non-response bias** is the skew of results based on a high rate of non-response to the survey.
 - If our randomly chosen sample selected 500 PA voters and only 15% responded to the survey, then the results are not likely to be representative of the entire population.

Sampling Bias

- **Convenience bias** occurs when easily accessible cases are more likely to be surveyed.
 - You conduct a survey of 100 shopping mall customers
 - You sit in front of a Hot Topic store and survey 100 customers as they leave the store.
 - Would the results of your survey be representative of all shopping mall customers?
- **Voluntary response bias** is the skew of results based on people who volunteer to respond because they have strong opinions on the topic.
 - A local newspaper puts a survey on their homepage about a new town ordinance. Website visitors are not required to give their opinion.
 - The website visitors who took the time to voice their opinion in the survey may not be reflective of the town's population.

Sampling Bias Example

- In 1936, the magazine *The Literary Digest* conducted a poll of its readers on the upcoming presidential election between Landon and Roosevelt.
- The magazine polled 10 million Americans and received 2.1 million responses.
- The result of *Digest's* poll showed only 43% support for Roosevelt, thus predicted a loss for the democrat president

Sampling Bias Example

- The actual election result? Roosevelt won with 62% of the vote.
- What happened? The magazine surveyed:
 - Its own readers who were also
 - Automobile owners
 - Telephone users
- These groups had incomes well over the national average at the time
 - This was during the Great Depression

Sampling Bias Example

- The sample *Digest* used was biased.
 - *Digest* had surveyed wealthier, Republican-leaning voters
 - Not the typical voter and not truly reflective of the American population at the time
- Completely discredited, *The Literary Digest* was soon discontinued.
- Going back to the “soup analogy”, this shows that even with a big spoon (a 2.1 million sample size) the soup still might not taste right if it isn’t properly stirred.

Designing Experiments

- A study where cases are assigned treatments is called an **experiment**.
- A **randomized experiment** randomizes which cases receive a treatment.
 - Randomized experiments are used to show a causal connection between two variables.

Designing Experiments

- To illustrate the main principles of designing experiments, we'll imagine being researchers that wish to test a new medication.
- To reduce bias in our sample of human volunteers, half are assigned to a treatment group and the other half are assigned to a control group.
 - The **treatment group** will receive the medication
 - The **control group** will not receive any treatment
 - The idea is that, at the end of the study, there will be enough evidence to show the medication's effect (if any) on the treatment group when compared to the control group

Designing Experiments

- **Control**

- This design principle is to ensure uniformity among cases.
- Example:
 - The medication we are testing is in pill form. Some treatment group participants swallow the pill with water while others swallow the pill without any water.
 - To account for any effects of water consumption, all participants may be required to take the pill with an 8 oz glass of water.
- (This is a different concept from the *control group* concept mentioned on the previous slide)

Designing Experiments

- **Randomization**

- This design principle helps to account for variables that cannot be controlled and helps to prevent introducing accidental bias.
- Example:
 - Diet may play a role in the effectiveness of the medication.
 - Unknown to us, there are patients in our sample that are vegan, vegetarian, or pescatarian, in addition to patients who consume red meat.
 - In other words, we aren't going to be monitoring what our volunteers eat.
 - Randomly assigning patients to the treatment and control groups would make it less likely to make a mistake of assigning the vegans and vegetarians to, say, the treatment group and the “meat-eaters” to the control group.

Designing Experiments

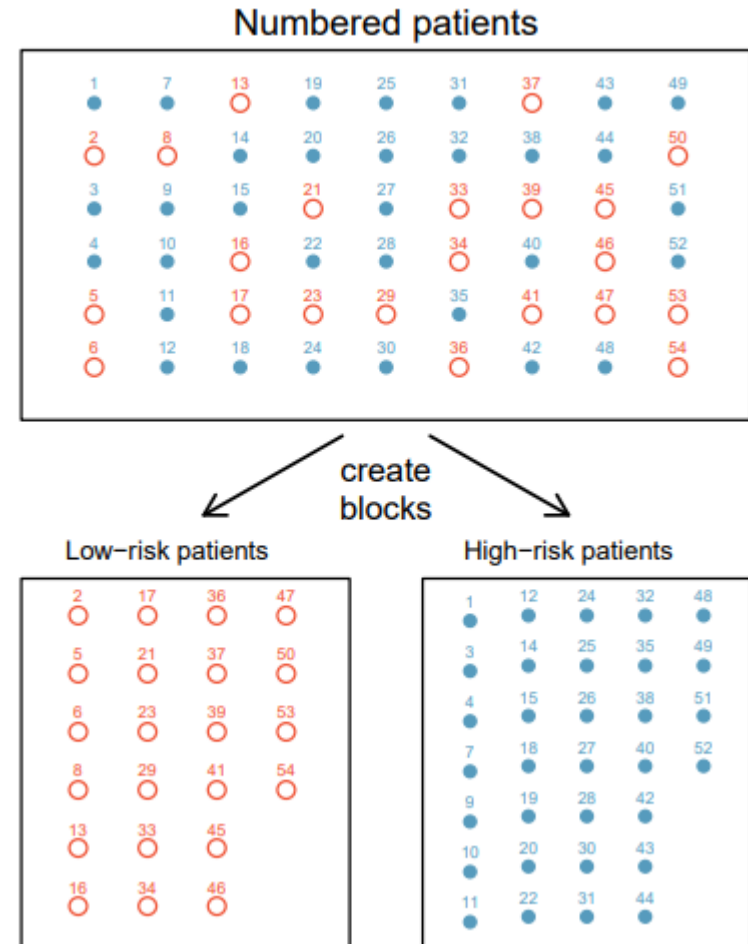
- **Replication**

- The more cases that are observed, the more accurate our conclusions
- The larger our sample (treatment and control group), the more observations we will make.
- Also, the entire experiment can be replicated to verify previous findings.
 - For example, repeating this medication study with entirely new treatment and control groups to see if the same results are achieved.

Designing Experiments

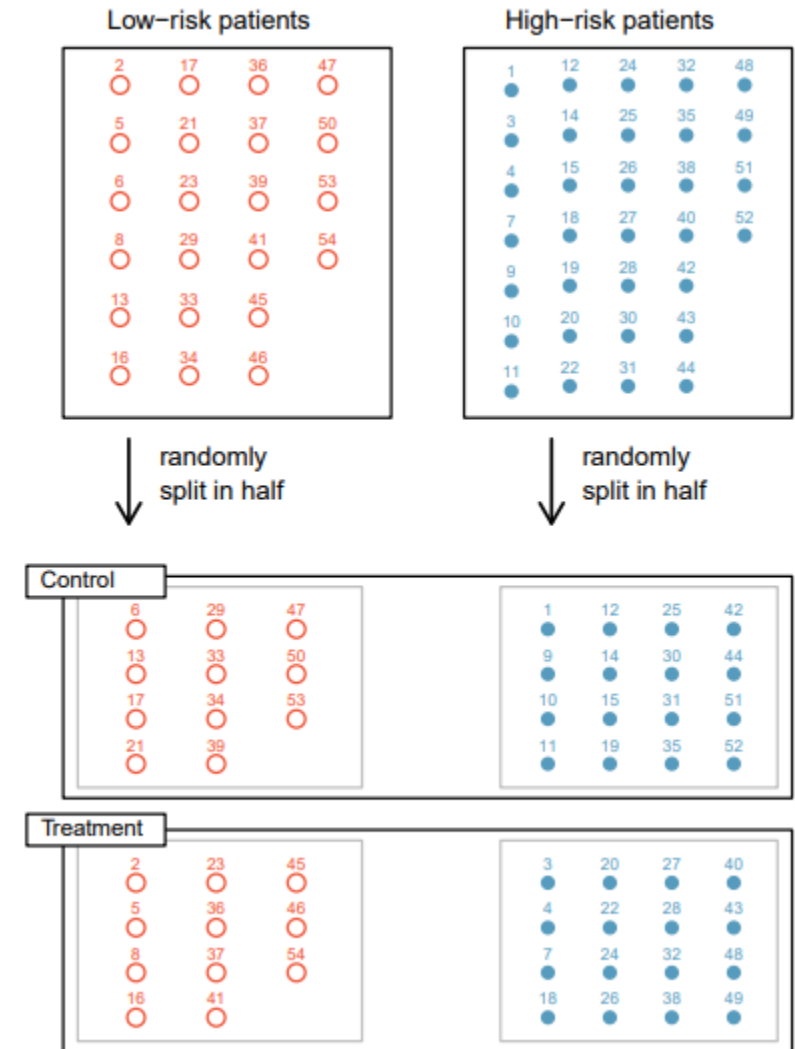
- **Blocking**

- Sometimes, other variables may be anticipated by the researchers to have influences on the response to a treatment
- For example, if our hypothetical medication is to prevent heart attacks, we might separate our participants into low-risk and high-risk groups.



Designing Experiments

- **Blocking**
 - Once separated, the blocks are split evenly into the treatment and control groups.
 - The assignment of participants to the treatment or control group is randomized
 - This will ensure both groups have an equal number of low-risk and high-risk patients



Designing Experiments

- When participants are not told whether they are a member of the treatment or control group, this is called a **blind study**
 - However, it becomes obvious to the participant that they are in our control group if we aren't giving them a pill to take
- To address this, the control group is given a **placebo**- a fake treatment (usually a sugar pill)
 - All participants will get a pill to take (Though the control group gets fake medicine)
 - Sometimes, taking a placebo results in improvement. This is called the **placebo effect**.
 - Example: Patient complains of pain and (unknown to the patient) their doctor prescribes a placebo. The patient says the pill makes them feel better even though the pill contained no medication.

Designing Experiments

- If they know which patients are in the treatment and control group, the researchers and doctors might give different attention to each group.
- To further eliminate possible bias, the researchers and doctors directly involved with the patients should not know which participants are in the treatment and control groups.
 - This is called a **double-blind study**- Neither the researchers and doctors nor the participants know who is in which group.