Stat 201: Statistics I
Chapter 1

Chapter 1
Introduction to Statistics

Section 1.1
Statistical and Critical Thinking

What is statistics?

"Statistics is the language of science."

What is statistics?

"Statistics is the language of science."

Statistics is also the language of...

- Politics (both campaigns and public policy)
- Economics
- Business
- Psychology and social sciences
- . . .

What is statisitcs?

Statistics is the science of using data to learn about the world.

What is statisitcs?

Statistics is the science of using data to learn about the world.

Data are collections of observations, such as measurements, biographical information or survey responses.

What is statisitcs?

Statistics is the science of using data to learn about the world.

Data are collections of observations, such as measurements, biographical information or survey responses.

Statistics is involved in...

- Designing studies and experiments
- Collecting data
- Producing informative summaries of data
- Analyzing data
- Interpreting results (answering questions)

Populations and samples

A **population** is any group that we are interested in knowing something about.

Populations and samples

A **population** is any group that we are interested in knowing something about.

A **census** is when data is collected from *every* member of a population.

Populations and samples

A **population** is any group that we are interested in knowing something about.

A **census** is when data is collected from *every* member of a population.

A **sample** is a subset of a population used to represent the whole population.

Population and sample examples

Example

Population	Sample
The entire population of the United States	Respondents to an internet survey

Population and sample examples

Example

Population	Sample
The entire population of the United States	Respondents to an internet survey
Males over 40 who have high blood pressure	High blood pressure patients in a clinical trial

Population and sample examples

Example

Population	Sample
The entire population of the United States	Respondents to an internet survey
Males over 40 who have high blood pressure	High blood pressure patients in a clinical trial
Students enrolled at Metro State in 2017	You (the students in this class)

Population and sample examples

Example

Population	Sample
The entire population of the United States	Respondents to an internet survey
Males over 40 who have high blood pressure	High blood pressure patients in a clinical trial
Students enrolled at Metro State in 2017	You (the students in this class)
Statistics classes in Minnesota	The summer semester statistics classes at Metro State

Statistical thinking

Prepare

1. Context
2. Source of the data
3. Sampling method

Analyze

1. Graph the data
2. Explore the data
3. Apply statistical method

Conclude

1. Significance

Prepare: Context

- What do the data mean?
- What is the goal of the study?
- Can the data answer the question of interest?

Prepare: Context

- What do the data mean?
- What is the goal of the study?
- Can the data answer the question of interest?

Example

Suppose a group of researchers wants to study the association between intelligence and grades. So, they collect the GPAs of a random sample of students and measure their skull circumference...

Prepare: Context

- What do the data mean?
- What is the goal of the study?
- Can the data answer the question of interest?

Example

Suppose a group of researchers wants to study the association between intelligence and grades. So, they collect the GPAs of a random sample of students and measure their skull circumference...

Note

This is not a completely made up example. Phrenology was the study of skull sizes and shapes, and was used as recently as the early 20th century to "prove" that non-white races were inferior and to diagnose mental illness.

Prepare: Source of the data

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

Prepare: Source of the data

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

Example

According to an article in the NY Daily News from June, 2014, titled, “Strip down: Sleeping naked is good for your relationship, survey says” (link). . .

From a survey of 1000 British couples, “57% of those who reported sleeping in the buff said they felt happy, compared with 48% of pajama wearers and 43% of nightie wearers.”

Prepare: Source of the data

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

Example

According to an article in the NY Daily News from June, 2014, titled, "Strip down: Sleeping naked is good for your relationship, survey says" (link)...

From a survey of 1000 British couples, "57% of those who reported sleeping in the buff said they felt happy, compared with 48% of pajama wearers and 43% of nightie wearers."

- The survey was conducted by Cotton USA.

Prepare: Sampling method

- Were the data collected in a way that is biased?

Prepare: Sampling method

- Were the data collected in a way that is biased?

A **voluntary response sample** (or **self-selected sample**) is one in which the respondents themselves decide whether to be included.

Prepare: Sampling method

- Were the data collected in a way that is biased?

A **voluntary response sample** (or **self-selected sample**) is one in which the respondents themselves decide whether to be included.

Example

- Call-in polls to radio or tv stations. . .

Prepare: Sampling method

- Were the data collected in a way that is biased?

A **voluntary response sample** (or **self-selected sample**) is one in which the respondents themselves decide whether to be included.

Example

- Call-in polls to radio or tv stations...
- Online surveys...

Prepare: Sampling method

- Were the data collected in a way that is biased?

A **voluntary response sample** (or **self-selected sample**) is one in which the respondents themselves decide whether to be included.

Example

- Call-in polls to radio or tv stations...
- Online surveys...
- Trending on twitter...

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?
3. Apply statistical method

Most of the course concerns the analyze step.

Conclude: Significance

- Do the results have statistical significance?
    - Statistical significance is a measure of how unlikely observed results are given certain assumptions.
    - Statistical significance is determined by many factors, including study design.
- Do the results have practical significance?
    - Do the results matter?

Conclude: Significance

- Do the results have statistical significance?
  - Statistical significance is a measure of how unlikely observed results are given certain assumptions.
  - Statistical significance is determined by many factors, including study design.
- Do the results have practical significance?
  - Do the results matter?

Example

A clinical trial shows a new drug lowers systolic blood pressure by an average of 3 mmHg. Results might be statistically significant, but are probably not practically significant.

Potential pitfalls: Misleading conclusions

Mistaking an association or relationship between two variables or factors for one factor causing the other.

Potential pitfalls: Misleading conclusions

Mistaking an association or relationship between two variables or factors for one factor causing the other.

!!!

Correlation does not imply causation.

Potential pitfalls: Misleading conclusions

Mistaking an association or relationship between two variables or factors for one factor causing the other.

!!!

Correlation does not imply causation.

Example

Recall the sleeping naked study.

Though the article made the claim that sleeping naked **caused** happier relationships, the study merely pointed to an association.

There are many other possible explanations for that association. This study alone does not provide evidence for which explanation is “true”.

Other potential pitfalls

- **Reported results** are data provided by the subjects of a study, rather than measured directly.

Other potential pitfalls

- **Reported results** are data provided by the subjects of a study, rather than measured directly.
- **Sample size** is important. Be wary of results drawn from very small samples.

Other potential pitfalls

- **Reported results** are data provided by the subjects of a study, rather than measured directly.
- **Sample size** is important. Be wary of results drawn from very small samples.
- **Loaded questions** are those designed to elicit a particular response or to influence the subject.
  - Also known as: push polls

Other potential pitfalls

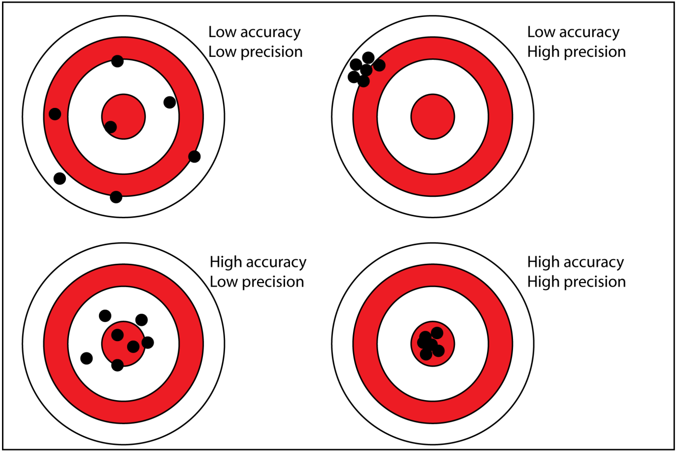
- **Reported results** are data provided by the subjects of a study, rather than measured directly.
- **Sample size** is important. Be wary of results drawn from very small samples.
- **Loaded questions** are those designed to elicit a particular response or to influence the subject.
  - Also known as: push polls
- The **order of questions** can influence responses.

Other potential pitfalls

- **Reported results** are data provided by the subjects of a study, rather than measured directly.
- **Sample size** is important. Be wary of results drawn from very small samples.
- **Loaded questions** are those designed to elicit a particular response or to influence the subject.
  - Also known as: push polls
- The **order of questions** can influence responses.
- **Missing data** can introduce bias if there are characteristics shared by subjects who have missing data or those who do not.

Potential pitfalls: Precise numbers

Precision is not the same thing as accuracy.



Potential pitfalls: Percentages

Sometimes percentages are used in confusing ways. Remember, 100% of a thing is all of it. Percentages above 100, or phrases like "a reduction of 100%", do not always have clear meanings.

Percentages: Review

- A **percentage** is number describing a proportion as an amount out of 100 (per cent).

Percentages: Review

- A **percentage** is number describing a proportion as an amount out of 100 (per cent).
- We can also describe a **proportion** as a fraction of 1.

$$\frac{50}{100} = \frac{1}{2} \quad \Rightarrow \quad 50\% = .50$$

Percentages: Review

- A **percentage** is number describing a proportion as an amount out of 100 (per cent).
- We can also describe a **proportion** as a fraction of 1.

$$\frac{50}{100} = \frac{1}{2} \quad \Rightarrow \quad 50\% = .50$$

- 100% represents a whole, just as for proportions 1 represents a whole.

Percentages: Review

- A **percentage** is number describing a proportion as an amount out of 100 (per cent).
- We can also describe a **proportion** as a fraction of 1.

$$\frac{50}{100} = \frac{1}{2} \quad \Rightarrow \quad 50\% = .50$$

- 100% represents a whole, just as for proportions 1 represents a whole.
- It often doesn't make sense to talk about percentages greater than 100%.

Percentages: Calculations

To convert from percentage to proportion, divide by 100:

$$56\% \quad \Rightarrow \quad \frac{56}{100} = 0.56$$

To convert from proportion to percentage, multiply by 100:

$$\frac{5}{8} = 0.625 \quad \Rightarrow \quad 0.625 \times 100 = 62.5\%$$

To find the quantity a percentage represents:

$$13\% \text{ of } 264 \quad \Rightarrow \quad \frac{13}{100} \times 264 = 34.32$$

To find the percentage a quantity represents:

$$135 \text{ out of } 475 \quad \Rightarrow \quad \frac{135}{475} \times 100 = 28.42\ldots\%$$

Group work

- Get into groups of 2 - 4
- Introduce yourself. Tell your groups your major and why you are taking this class.
- For each question, read the scenario. Discuss and note some answers to part (a).
- Don't worry about getting exactly the right answers. Thinking about and discussing the question is more important.
- After about 15 minutes we will discuss as a class.

Section 1.2
Types of Data

Parameters and statistics

A **parameter** is a value describing an aspect of a population.

Parameters and statistics

A **parameter** is a value describing an aspect of a population.

A **statistic** is a value describing an aspect of a sample.

Parameters and statistics

A **parameter** is a value describing an aspect of a population.

A **statistic** is a value describing an aspect of a sample.

Example

- The average height of adult men in the U.S. is 72 inches: **Parameter**
- The average height of 30 randomly selected male Metro State students is 68.5 inches: **Statistic**

Types of data

Quantitative data are numbers representing amounts, sizes, time or other measurements.
Also known as: Numeric

Example

Class size, height, age, systolic blood pressure, temperature

Types of data

Quantitative data are numbers representing amounts, sizes, time or other measurements.
Also known as: Numeric

Example

Class size, height, age, systolic blood pressure, temperature

Categorical data are values representing groups or categories.

- Also known as: qualitative, attribute

Example

Gender, state of residence, football player's numbers, pain scale

Types of data: Quantitative

Discrete data have a finite, or countably infinite, number of possible values. There are gaps in the possible values.

Example

Class size: can't have a class size of 22.5

Types of data: Quantitative

Discrete data have a finite, or countably infinite, number of possible values. There are gaps in the possible values.

Example

Class size: can't have a class size of 22.5

Continuous data have an infinite number of possible values. There are no gaps in possible values.

Example

Height: a height of 70.2641... inches is possible (not necessarily useful, but possible)

Levels of measurement

- Nominal
- Ordinal
- Interval
- Ratio

Levels of measurement: Nominal

The **nominal** level of measurement is categorical data that are names or labels for groups or categories. There is no reasonable order or ranking to the categories.

Levels of measurement: Nominal

The **nominal** level of measurement is categorical data that are names or labels for groups or categories. There is no reasonable order or ranking to the categories.

Example

- Gender: *male* or *female*
- State of residence: *Minnesota*, *Wisconsin*, etc.

Levels of measurement: Nominal

The **nominal** level of measurement is categorical data that are names or labels for groups or categories. There is no reasonable order or ranking to the categories.

Example

- Gender: *male* or *female*
- State of residence: *Minnesota*, *Wisconsin*, etc.

Hint

The root word *nom* means "name".

Levels of measurement: Ordinal

The **ordinal** level of measure is categorical data that are naturally ordered or ranked.

Levels of measurement: Ordinal

The **ordinal** level of measure is categorical data that are naturally ordered or ranked.

Example

- Pain scale: *No pain* $<$ *Moderate pain* $<$ *Heavy pain*
- Grades: $A > B > C > D > F$

Levels of measurement: Interval

The **interval** level of measurement is quantitative data where the difference between values has meaning but where there is no natural "zero".

Levels of measurement: Interval

The **interval** level of measurement is quantitative data where the difference between values has meaning but where there is no natural "zero".

Example

- Temperature: The difference between 101°F and 98.6°F is meaningful, but 0°F does not mean no temperature.
- Year: 2017 is four years after 2013, but year 0 does not mean no years.

Levels of measurement: Ratio

The **ratio** level of measurement is quantitative data where the difference between values and relative sizes of values have meaning. There is a natural "zero".

Levels of measurement: Ratio

The **ratio** level of measurement is quantitative data where the difference between values and relative sizes of values have meaning. There is a natural "zero".

Example

- Age: Someone who is 40 years old is *twice* as old as someone who is 20 years old. Zero does mean no age.
- Height: A tree that is 10 feet tall is *one third* as tall as a tree that is 30 feet tall. Zero does mean no height.

Group work

- Get back into your groups.
- Discuss and note some answers to part (b) for each question.
- Remember, getting exactly the right answer is less important than the discussion.
- After about 15 minutes we will discuss as a class.

Section 1.3
Collecting Sample Data

Samples

- Recall, when we want to know something about a population and we can't collect data from the entire population, we can collect data from a subset, or a **sample**, of the population instead.

Samples

- Recall, when we want to know something about a population and we can't collect data from the entire population, we can collect data from a subset, or a **sample**, of the population instead.
- We can then use statistics to learn something about the whole population.

Samples

- Recall, when we want to know something about a population and we can't collect data from the entire population, we can collect data from a subset, or a **sample**, of the population instead.
- We can then use statistics to learn something about the whole population.
- Therefore, how we pick our sample is very important in how valid the interpretation of our results are.

Samples

- Recall, when we want to know something about a population and we can't collect data from the entire population, we can collect data from a subset, or a **sample**, of the population instead.
- We can then use statistics to learn something about the whole population.
- Therefore, how we pick our sample is very important in how valid the interpretation of our results are.

Example

- Suppose an organization is interesting in the taco consumption by Metro State students. It would be difficult, if not impossible, to ask every student about their taco eating habits. A sample is needed.

Types of samples: Random sample

A **random sample** is a sample selected such that every individual member of a population has an equal chance of being included.

Types of samples: Random sample

A **random sample** is a sample selected such that every individual member of a population has an equal chance of being included.

A **simple random sample** is a sample selected such that every possible sample of a specific size has an equal chance of being selected.

- These are the "best" kind of samples for producing valid, unbiased results, but they are not always easy to get.

Types of samples: Random sample

A **random sample** is a sample selected such that every individual member of a population has an equal chance of being included.

A **simple random sample** is a sample selected such that every possible sample of a specific size has an equal chance of being selected.

- These are the "best" kind of samples for producing valid, unbiased results, but they are not always easy to get.

Example

- Given an alphabetical list of students, use a random number generator to select a sample.

Types of samples: Systematic sampling

Systematic sampling is a method where every kth member of of a population is selected.

- These samples are ofter easier to produce, but can lead to biased samples.

Types of samples: Systematic sampling

Systematic sampling is a method where every kth member of of a population is selected.

- These samples are ofter easier to produce, but can lead to biased samples.

Example

- Given an alphabetical list of students, select every fifth student until you have a sample of the desired size.

Types of samples: Convenience sampling

Convenience sampling is a method of choosing members of a population that are nearby or easy to access.

- The easiest of all methods, but by far the lowest quality data for producing results.

Types of samples: Convenience sampling

Convenience sampling is a method of choosing members of a population that are nearby or easy to access.

- The easiest of all methods, but by far the lowest quality data for producing results.

Example

- Wander the halls before class, asking students who happen to walk by.
- Put a poll on the Metro State website.

Types of samples: Stratified sampling

Stratified sampling is a method where the population is divided into groups and samples are selected from each group.

- Useful when you want to ensure that a factor of interest has enough representation, but it is not a random sample as we have defined it.

Types of samples: Stratified sampling

Stratified sampling is a method where the population is divided into groups and samples are selected from each group.

- Useful when you want to ensure that a factor of interest has enough representation, but it is not a random sample as we have defined it.

Example

- If we have particular interest in the taco consuming difference between graduate students and undergrads, select a sample from each group.

Types of samples: Cluster sampling

Cluster sampling is a method where the population is divided into sections or clusters. Then, a number of clusters are randomly selected and all members of the clusters are included in the sample.

- More convenient than some methods, but better randomization the pure convenience sampling.

Types of samples: Cluster sampling

Cluster sampling is a method where the population is divided into sections or clusters. Then, a number of clusters are randomly selected and all members of the clusters are included in the sample.

- More convenient than some methods, but better randomization the pure convenience sampling.

Example

- Choose 5 random classes, and survey all the students in those classes.

Types of samples: Multistage sampling

Multistage sampling is a when a combination of methods are used to produce a sample.

Types of samples: Multistage sampling

Multistage sampling is a when a combination of methods are used to produce a sample.

Example

- Choose random classes by cluster sampling, and then take a simple random sample of students from each chosen class.

Types of studies

In an **observational study** data is collected from a sample without trying to modify behavior or results.

Types of studies

In an **observational study** data is collected from a sample without trying to modify behavior or results.

In an **experiment** a change (treatment) is made to some or all of sample and then data is collected in order to detect changes.

Types of observational studies

A **cross-sectional** study measures and collects data from one point in time (the present).

Types of observational studies

A **cross-sectional** study measures and collects data from one point in time (the present).

A **retrospective** study collects data from the past, whether from recollections or by examining records.

- Also know as: case-control

Types of observational studies

A **cross-sectional** study measures and collects data from one point in time (the present).

A **retrospective** study collects data from the past, whether from recollections or by examining records.

- Also know as: case-control

A **prospective** study follows subjects into the future to measure and collect data.

- Also know as: longitudinal study, cohort study

Experiment design: Replication

Replication is the repetition of the experiment on more than one individual or in more than one study.

Experiment design: Replication

Replication is the repetition of the experiment on more than one individual or in more than one study.

- Experimental studies should have adequate sample sizes to ensure that observed effects are "true" effects and not due to individual characteristics or chance.

Experiment design: Replication

Replication is the repetition of the experiment on more than one individual or in more than one study.

- Experimental studies should have adequate sample sizes to ensure that observed effects are "true" effects and not due to individual characteristics or chance.
- Experimental studies should be, but rarely are, repeated by different researchers to verify results.

Experimental design: Blinding

Blinding is the process of hiding which treatment or lack of treatment a subject is receiving from one or more groups of study participants. This is done in order to reduce bias in the results.

Experimental design: Blinding

Blinding is the process of hiding which treatment or lack of treatment a subject is receiving from one or more groups of study participants. This is done in order to reduce bias in the results.

- A **single blinded** study is one where the subjects don't know what treatments they are receiving.

Experimental design: Blinding

Blinding is the process of hiding which treatment or lack of treatment a subject is receiving from one or more groups of study participants. This is done in order to reduce bias in the results.

- A **single blinded** study is one where the subjects don't know what treatments they are receiving.
- A **double blinded** study is one where both the subjects and the researchers administrating treatment and gathering results don't know which treatment the subjects are receiving.

Experimental design: Blinding

Blinding is the process of hiding which treatment or lack of treatment a subject is receiving from one or more groups of study participants. This is done in order to reduce bias in the results.

- A **single blinded** study is one where the subjects don't know what treatments they are receiving.
- A **double blinded** study is one where both the subjects and the researchers administrating treatment and gathering results don't know which treatment the subjects are receiving.

The **placebo effect** is a phenomenon where people who believe they are being treated demonstrate improvement.

Experimental design: Randomization

Randomization is the process selecting samples and assigning treatment groups randomly. This is done to ensure that samples are representative of populations and that characteristics are evenly distributed among treatment groups.

Experimental design: Randomization

Randomization is the process selecting samples and assigning treatment groups randomly. This is done to ensure that samples are representative of populations and that characteristics are evenly distributed among treatment groups.

Confounding variables (or just confounders) are unmeasured and possible unknown factors that affect the experimental outcome.

Group work

- Get back into your groups.
- Discuss and note some answers to part (c) for each question.
- After about 10 minutes we will discuss as a class.