Stat 201: Statistics I
Week 1

Metropolitan
State University

Week 1
Introduction to Statistics and Data

Section 1.1
Statistical and Critical Thinking

Why study 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
- ...

Why study statistics? (cont)

Some of us will **do** statistics.

Statistics are an important component of many of the fastest growing careers. As of November 2018,

- Statistician is the second fastest growing career in Minnesota.
- Computer and information research scientist, operations research analyst, actuary and market research analyst are all in the top 20.

All of us are **consumers** of statistics.

- Many fields make extensive use of statistics and research to make decisions.
- Much of the news and information we encounter on a daily basis involves statistics.

What is statistics?

Statistics is the science of using data from samples to learn about populations.

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

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

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

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

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 process

- **Prepare**
  - Identify research question
  - Design the study
  - Collect data
- **Analyze**
  - Graph and explore the data
  - Produce informative summaries of data
  - Conduct statistical tests or other analyses
- **Conclude**
  - Report results in the context of the research question

Possible pitfalls: Prepare

- What is the goal of the study? Is the research question clear?
- What do the data mean? Can the data answer the research question?
- Are the data or is the study from a source with a special interest so that there is pressure to obtain results that are favorable to the source?
- What sampling method was used? Were the data collected in a way that is biased?
- Are the data self-reported?
- Is the sample size adequate?
- What questions were used to elicit responses?

Can the data answer the research question?

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.

Are the data from a source with a special interest?

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.

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
- Internet comment or review sections
- Trending on twitter

Other issues with prepare step

- **Self-reported results** are data provided by the subjects of a study, rather than measured directly. Self-reported data can be inaccurate and/or imprecise.
- **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.

Possible pitfalls: Analyze

- Are there any outliers?
- Are there missing data?
- How are the data distributed?
- Was an appropriate test used?

The rest of the course is primarily focused in the analyze step. We will discuss most of these issues as we go.

Are there missing data?

Missing data can result from subjects dropping out of the study or skipping scheduled data collection appointments. Almost every study has to deal with missing data. The existence of missing data is not necessarily concerning unless,

- A large proportion of the the original sample is missing or has missing data
- The pattern of missing data is biased (known as **informative missingness**)

Example

An experimental drug trial with a large proportion of drop-outs from group taking the new drug would be cause for concern. It could be caused by subjects who are experiencing severe side effects and/or ineffective treatment leaving the study. Any conclusions drawn only from the remaining subjects would then be misleading.

Possible pitfalls: Conclude

- Do the results have statistical significance? Do the results have practical significance?
- Are the conclusions justified by the data?
- Is there a confusion of precision for accuracy?
- Are percentages used clearly?

Are the results significant?

- 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.

Are the conclusions justified by the data?

It is common to mistake an association or relationship between two factors for one factor causing the other.

!!!

Correlation does not imply causation.

Example

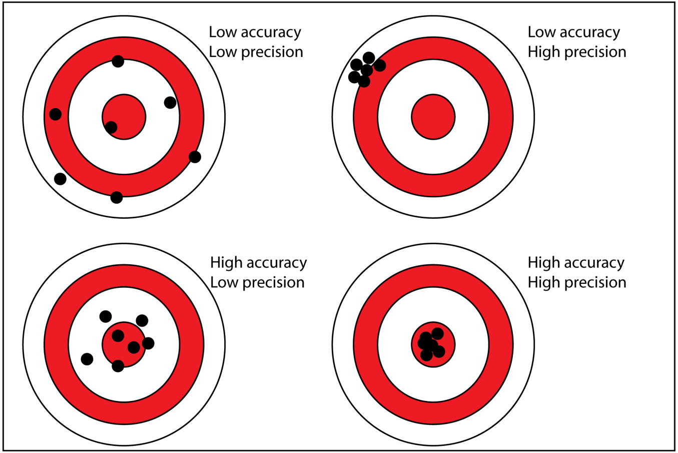
Recall the sleeping naked study.

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”.

Is there a confusion of precision for accuracy?

Precision is not the same thing as accuracy.



Are percentages used clearly?

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.

Example

Suppose a company sold 10 widgets last year. This year they report an increase in widget sales of 150%. How many widgets did they sell this year?

- Answers of 15 or 25 widgets could both be reasonably justified.

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:

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

To find the percentage a quantity represents:

Section 1.2
Introduction to Data

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**

Organization of data

ID	AGE	SEX	LENGTH	WEIGHT	CHEST
1030	21	Male	61.0	150	34.0
1004	115	Male	72.0	348	49.0
1050	17	Female	52.5	76	28.0
1040	58	Male	70.5	365	50.0
1020	177	Male	72.0	436	48.0
1010	53	Female	58.0	144	31.0

Data is often presented as a table or **data matrix**.

- Each row represents one **case** or **subject**. Also known as a **unit of observation**.
- Each column represents one characteristic or **variable**. Also known as a **factor**.
- Ideally, a data matrix will be accompanied by a codebook, a document that explains each variable.

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

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.

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.

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".

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”.

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.