Stat 201: Statistics I
Week 1

May 15, 2017

What is statisitcs?

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

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.

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
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: 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: 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.

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?
- Do the results have practical significance?

Example

A clinical triel shows a new drug lowers systolic blood pressure by an average of 5 mmHg. Results are statistically significant, but may not be practically significant.

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: Reported results

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

Potential pitfalls: Small samples

Sample size is important. Be wary of results drawn from very small samples.

Potential pitfalls: Loaded questions

Loaded questions are those designed to elicit a particular response or to influence the subject.

- Also known as: push polls

Potential pitfalls: Order of questions

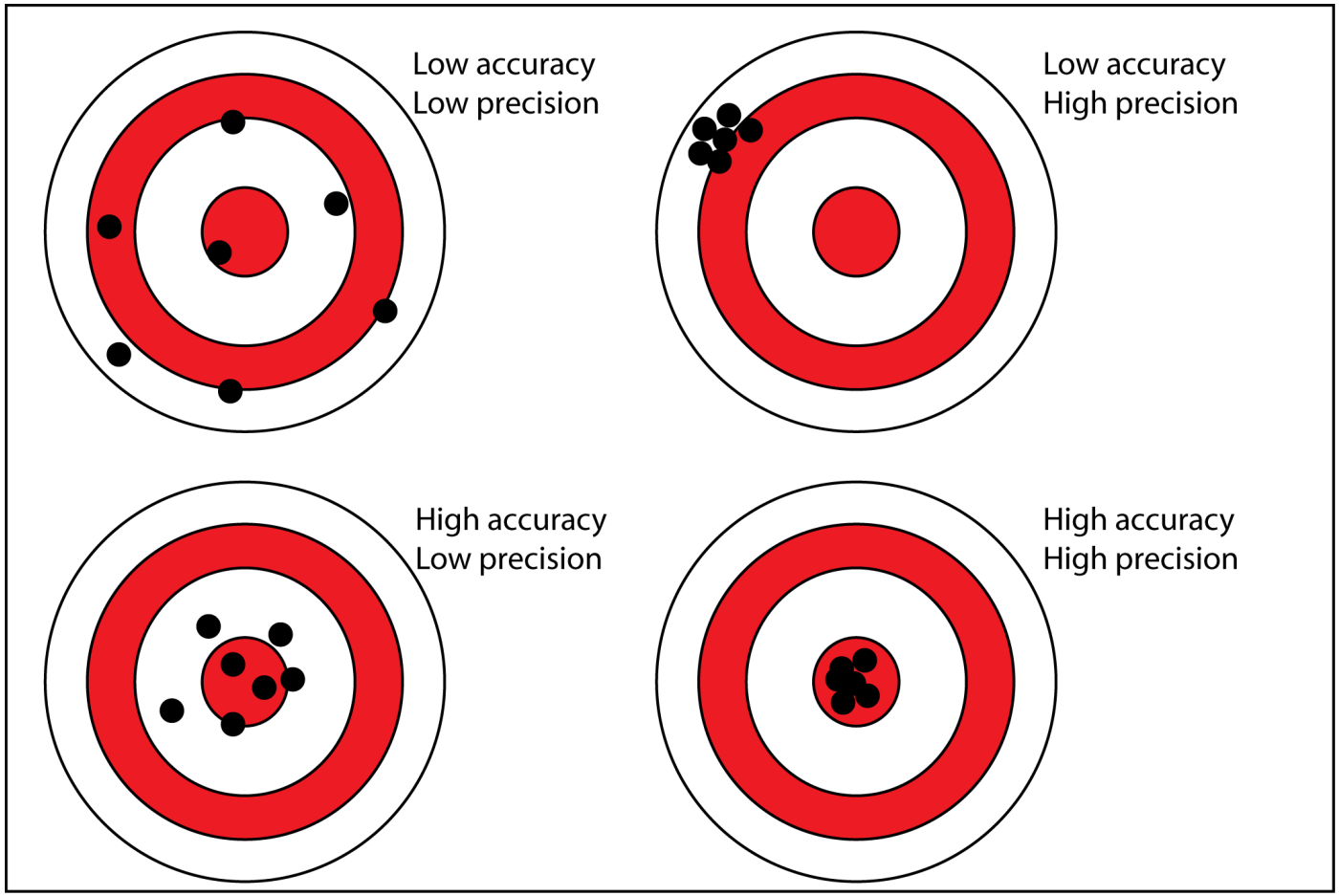
The order of questions can influence responses.

Potential pitfalls: Nonresponse / missing data

When subjects refuse to participate or data is not collected for any reason, bias can be introduced 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).
- 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:

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

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

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 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

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.

Samples

- Remember: 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 we are collecting data for the Metro State Annual Taco Survey (MSATS). It would be difficult, if not impossible, to ask every student about their taco eating habits. We need a sample.

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.

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.

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.

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.

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.

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.

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).

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

Frequency distributions

A **frequency** is the number of times a particular value occurs in a set of data, i.e. the count.

A **frequency distribution** (or **frequency table**) summarizes a set of data by listing the frequencies of data in categories or classes (groups).

- For categorical data, the categories are simply the possible values of the data.
- For quantitative data, the classes are usually ranges of possible values.

Frequency distribution for categorical data

Example

Favorite kind of taco $= \{$Chicken, Fish, Fish, Veggie, Chicken, Beef $\}$

Kind of taco	Frequency
Beef	1
Chicken	2
Pork	0
Fish	2
Veggie	1

Frequency distribution for quantitative data

Example

Tacos eaten $= \{3, 0, 17, 6, 4, 3, 5\}$

Number of tacos eaten	Frequency
0 - 4	4
5 - 9	2
10 - 14	0
15 -20	1

Relative frequency

Relative frequency is the proportion (fraction) of the whole data set that resides in each category or class. When expressed as a percent it is called **percentage frequency**.

To calculate: For each class,

$$\text{Relative frequency} = \frac{\text{class frequency}}{\text{total count}}$$

$$\text{Percentage frequency} = \frac{\text{class frequency}}{\text{total count}} \times 100$$

Relative frequency example

Example

Tacos eaten	Frequency	Relative	Percentage
0 - 4	4	0.5714	57.14 %
5 - 9	2	0.2857	28.57 %
10 - 14	0	0	0 %
15 -20	1	0.1428	14.28 %
Total	7	1	100 %

Cumulative frequency

Cumulative frequency is the frequency for a class and *all previous classes*.

Example

Tacos eaten	Frequency	Cumulative
0 - 4	4	4
5 - 9	2	6
10 - 14	0	6
15 -20	1	7

Normal distributions

A **normal distribution** can be identified from a frequency table that has the following characteristics:

- The frequencies start low, increase to a high point and then decrease to low frequencies at the end
- The frequencies are approximately symmetric around the high point.

Normal distributions, example

Example

Normal

IQ	Frequency
80 - 89	1
90 - 99	5
100 - 109	11
110 - 119	10
120 - 129	4
130 - 139	2

Not normal

IQ	Frequency
80 - 89	2
90 - 99	9
100 - 109	13
110 - 119	4
120 - 129	3
130 - 139	1

Gaps in frequency tables

A **gap** of frequencies in a table indicates that the data probably come from two different populations.

The converse is not necessarily true. Data from two different populations might not display a gap.

Example

- Pennies made before 1983 are 95% copper and 5% zinc.
- Pennies made after 1983 are 2.5% copper and 97.5% zinc.

Gaps in frequency tables, example

Example, cont.

Weight (g) of penny	Frequency
2.40 - 2.49	18
2.50 - 2.59	19
2.60 - 2.69	0
2.70 - 2.79	0
2.80 - 2.89	0
2.90 - 2.99	2
3.00 - 3.09	25
3.10 - 3.19	8