Organizing & Describing Data:

Terminology and visualization

Reminders

- Sections start this week.
- Participation starts today, so have Canvas open.
- Slides are posted on Canvas before lecture. They should facilitate note-taking.
- Please complete the course survey (if you haven't already).
- The book should be available on Canvas.
- Don't fall behind in the class. The early material is easier/review, but gets harder as we go.

Please note: all lectures, including comments, are recorded.

Today

Organizing and describing data (with some review)

- Data terminology
  - Variable sources, types, and levels of measurement
- Data Visualization
  - Bar charts
  - Line graphs
  - Scatter plots
  - Infographics
  - Misleading visualizations
- Frequency Distributions
  - Tables & Histograms
  - Ungrouped vs. Grouped
  - Outliers
  - Relative & Cumulative Frequency Distributions

I remember like it was three days ago…

Memory check

What is the area of statistics most involved with displaying sample data?

A. Inferential statistics
B. Descriptive statistics
C. Observational statistics
D. Experimental statistics
E. The what now?

I remember like it was three days ago…

Memory check

What is the area of statistics most involved with displaying sample data?

A. Inferential statistics
B. Descriptive statistics
C. Observational statistics
D. Experimental statistics
E. The what now?

B. Summarizing and displaying collected data.

I remember like it was last year…

Memory check

You design an experiment to measure the reaction times of human participants placed into one of 3 different stress conditions: high, medium, and low. What is the level of measurement for the dependent variable?

A. Nominal
B. Ordinal
C. Interval
D. Ratio
E. I don’t remember

I remember like it was last year…

Memory check

You design an experiment to measure the reaction times of human participants placed into one of 3 different stress conditions: high, medium, and low. What is the level of measurement for the dependent variable?

A. Nominal

B. Ordinal

C. Interval

D. Ratio

E. I don't remember

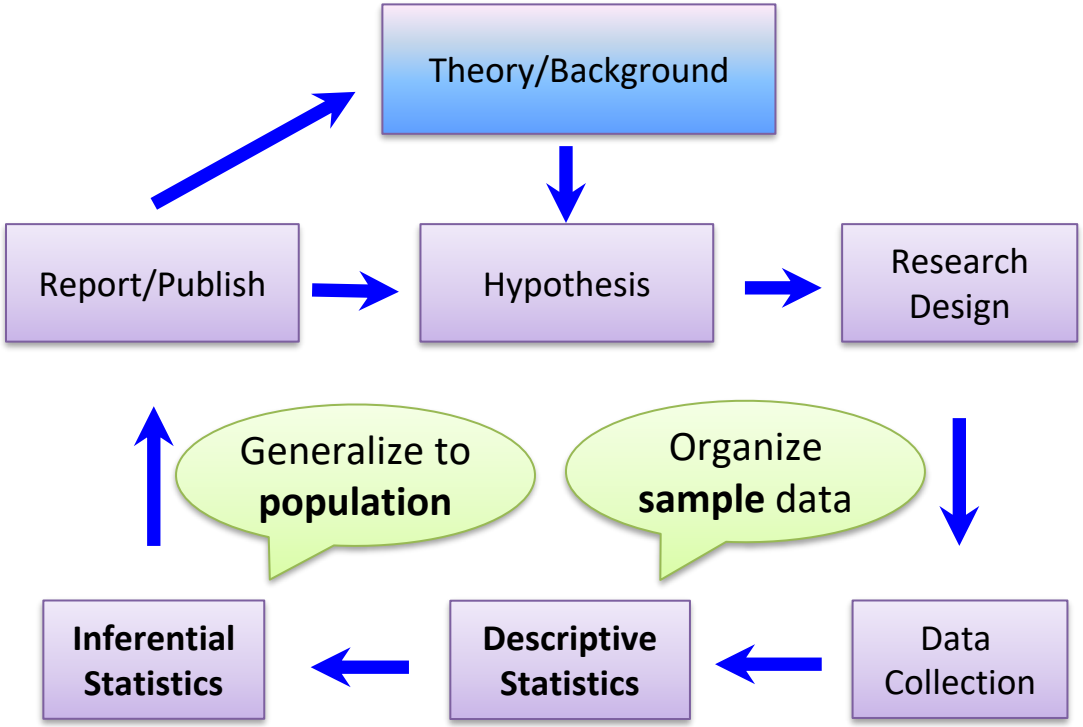
D. Reaction time is the thing being measured and (by hypothesis) is *possibly* affected by the level of stress.

Stress is manipulated and known beforehand by the experimenter. [Independent variable is ordinal.]

From last time

Applying the scientific method

- A series of steps describing how the scientific method is used in experiments.



Where do the data come from?

Population: A *complete* collection of **observations** (data) or *potential* observations for all individuals or *units of interest*.

- This is defined by the investigator. The population can be very targeted (e.g. language ability in recovering stroke patients or a single stroke patient) or very broad (e.g. language ability in all young children).

Sample: A *partial* set of observations taken from the population.

- Typically much smaller than the population (e.g. 20 stroke patients or 20 toddlers).
- The partial set of data is sometimes called a **convenience sample**.
- Should be *representative* of the population.
- Many (!) different samples from a single population are possible.

Variables from the population vs. sample

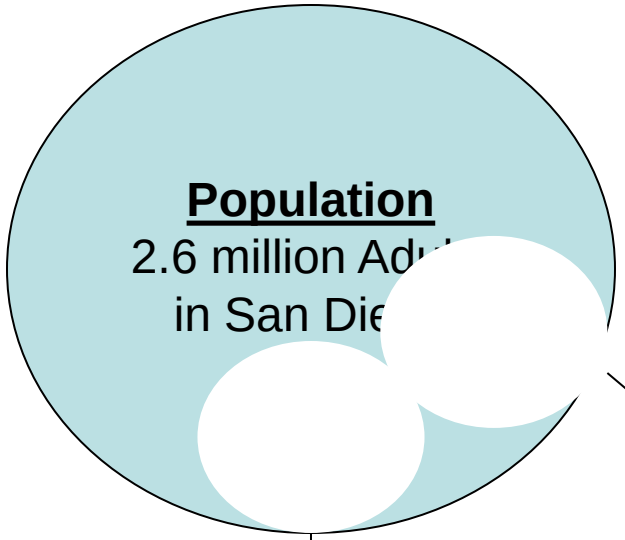
Parameter: A value that reflects something in the entire *population* of interest.

- e.g. Average height of all adults in San Diego.

Statistic: A value that reflects something from a *sample*. It can be said to *estimate* the population parameter.

- e.g. Average height of 100 UCSD students.

Parameters vs. Statistics



Population Parameters
Average height: 66 ½ ins.
48% Male, 52% Female

Random sampling: All potential observations in the population have an *equal chance* of being selected in a sample. (Done here?)

Sample 1
100 students

Sample Statistics
Average height: 64 ¾ ins.
43% Male, 57% Female

Sample 2
100 students

Sample Statistics
Average height: 67 ½ ins.
48% Male, 52% Female

Sample statistics will vary *by chance* due to random **sampling error**.

Statistics: Descriptive vs. Inferential

Descriptive versus Inferential statistics

- Descriptive: Statistical procedures and visualization tools used to organize, summarize, and simplify data from a collection of *actual observations* (i.e. your sample or population).
  - Average UCSD GPA is 3.05
  - 39% of UCSD undergrads lived on campus last year

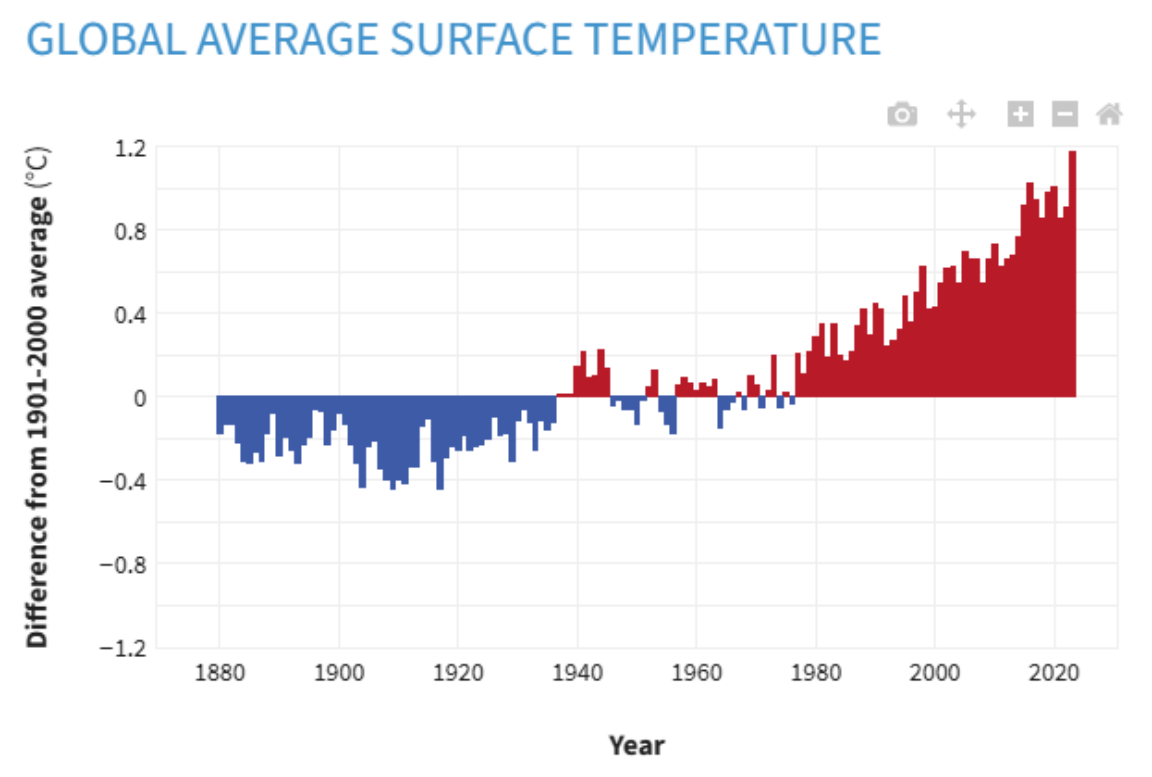
  (Charts, tables, and graphs function to summarize data. These are typically considered descriptive.)

- Inferential: Statistical tools that facilitate inferences *beyond samples* to estimate the *unobserved* population parameters.
  - Based on current data, average GPA in the Fall will be 3.06
  - Approximately 38.5% (+/- 3%) of public university students in San Diego live on campus.

Statistics: Descriptive vs. Inferential

Intuition check

How would we characterize this **graph** of actual observations? [Explain]



Climate Change: Global Temperature | NOAA Climate.gov

A. Descriptive statistics

B. Inferential statistics

A. Summarizing, organizing, and displaying data that we have.

Statistics: Descriptive

- Summary statistics:
  - Measures of central tendency: Means, medians, and modes.
  - Measure of variance: Range, interquartile range (IQR), and standard deviation.

- Data visualization (basic):
  - Bar charts
  - Line graphs
  - Tables and charts
  - Infographics

In all cases, you are trying to provide *accurate, concise, and clear representations* of your data *to a specific audience*.

Statistics: Descriptive vs. Inferential

Examples

- On average, students currently taking COGS 14B are 20.2 years old.
- The population of the world exceeded 7.8 billion as of January 2020, according to the United Nations.
- Four years has been the most frequent term of office served by American presidents.
- A recent poll indicates that 58.4% of recent UCSD graduates would “definitely” recommend prospective students attend UCSD.

Statistics: Descriptive vs. Inferential

Examples

- On average, students currently taking COGS 14B are 20.2 years old. [**Descriptive**]
- The population of the world exceeded 7.8 billion as of January 2020, according to the United Nations. [**Inferential**]
- Four years has been the most frequent term of office served by American presidents. [**Descriptive**]
- A recent poll indicates that 58.4% of recent UCSD graduates would “definitely” recommend prospective students attend UCSD. [**Inferential**]

Reporting/Publishing statistics

Where to generally find (or report) statistical analysis in a scientific paper.

Paper section	Steps in scientific method
Introduction	Hypotheses and theoretical background
Methods	Research design and data collection [Some **descriptive statistics** describing your participants]
Results	**Descriptive and Inferential statistics**
Discussion	Statement of findings and their theoretical importance

Variable types (review)

Discrete versus Continuous variables

- Discrete variables consist of isolated numbers or categories, with no values between neighboring categories.
  - Number of siblings, children, vocabulary size, political party, etc.

- Continuous variables have potentially infinite values between any two observed values. Finer measurement could yield more precise values.
  - Height, weight, reaction time, temperature, interest rates, etc.

Variable levels (review)

We will look at three levels of measurement:

- Nominal: Consists of labels or categories that classify *qualitative* data.
  - Sex, favorite food, political groups, blood type.
- Ordinal (ranked): A set of categories that are organized in an ordered or *ranked* sequence.
  - Clothing size, race results, letter grades.
- Interval/Ratio: A set of ordered categories where the categories form intervals of equal size. Interval has no true zero point, while ratio does. These are *quantitative*.
  - Interval: IQ score, GPA, temperature in ° Fahrenheit.
  - Ratio: Distance, weight, income, temp. in ° Kelvin

Variable levels (review)

Level	Properties	Observations reflect	Example	Type of data
Nominal	Classification	Differences in kind	Favorite food	Qualitative
Ordinal	Ordered classification	Differences in degree	Letter grade	Ranked
Interval/Ratio*	• Equal intervals • Ordered classification	Differences in total amount	Height	Quantitative

*Some sources distinguish these from each other – ratio has true zero point and expressing a value as a ratio makes sense.

How variables are used (review)

- Variable: A characteristic or property that changes or can take on different values for different individuals. This implies multiple observations. Can be discrete or continuous.
  - e.g. Height, age, gender, political affiliation, etc.

- Independent variable: A characteristic or property *manipulated* by the investigator in an experiment.
  - e.g. Different amounts of caffeine given to participants in memory exp.

- Dependent variable: A characteristic or property which may change in *response* to manipulation in the independent variable. Also called the dependent measure.
  - e.g. number of recalled items for caff. vs. non-caff. groups.

- Constant: A characteristic or property that can take on only one value. This is usually determined by the research design.
  - During an experiment, this may be a variable that is 'held' constant across conditions (e.g. sex, age, or IQ).

See section 1.6 in text.

Descriptive Statistics

What are some good ways to organize, summarize, and simplify data from a collection of *actual observations* (i.e. your sample or population)?

- Bar charts
- Line charts
- Scatter plots
- Histograms

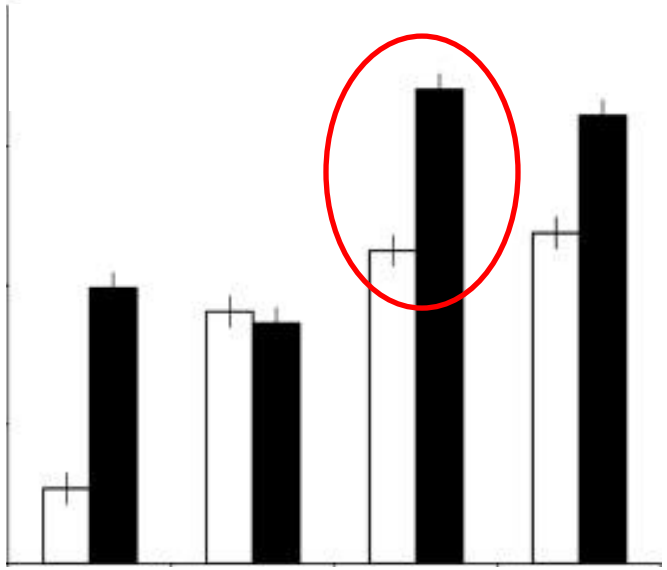
Note: The goal in all cases is to be *accurate, concise, and clear*. How this is done may differ by discipline and audience.

Descriptive Statistics: Graphs

Bar charts

- Bar length is proportional to measured quantity on y-axis.
- Best used when x-axis variable is **discrete** and **nominal** (*qualitative* data).
- Differences (if they exist) are easy to see.
- Gaps between bars emphasize the *discontinuous* nature of the x-axis variables.
- Can be used to present raw data or summary statistics.

Error bars *suggest* a *significant* difference in values.



Boroditsky *et al*. (2011) Do English and Mandarin speakers think about time differently?

Next time

- More visualizations
- Misleading Data Visualizations
- Frequency Distributions (Chapter 2)
  - Tables & histograms
  - Ungrouped vs. Grouped
  - Outliers
  - Relative & Cumulative Frequency Distributions