

CMSE 890-310 | BMB 960-003

Gaps, Missteps, & Errors in Data Analysis

Arjun Krishnan

arjun@msu.edu | @compbiologist | thekrishnanlab.org

Day 01

Welcome, Topics 1 and 2

- Welcome, overview
- Scientific method, Critically reading literature, Cognitive biases
- Estimation of error & uncertainty
- What's due next week?

Land acknowledgement

Michigan State University occupies the ancestral, traditional and contemporary lands of the Anishinaabeg – Three Fires Confederacy of Ojibwe, Odawa and Potawatomi peoples. The university resides on land ceded in the 1819 Treaty of Saginaw.

Map: <https://www.canr.msu.edu/nai/about/land-acknowledgements>

Congratulations on surviving 2020 & 2021!

I'm amazed by and grateful for the seemingly limitless courage, hard work, and sacrifice of healthcare workers, other essential workers, and the people who stood up for social justice.

These two years have been incredibly tough for students, care givers (esp. parents of young children), and daily-wage workers -- difficulties that were hugely compounded due to also belonging to systematically minoritized/underrepresented/disadvantaged groups.

In addition to these difficulties, my heart goes out to those who also have suffered the loss of loved ones.

Introductions

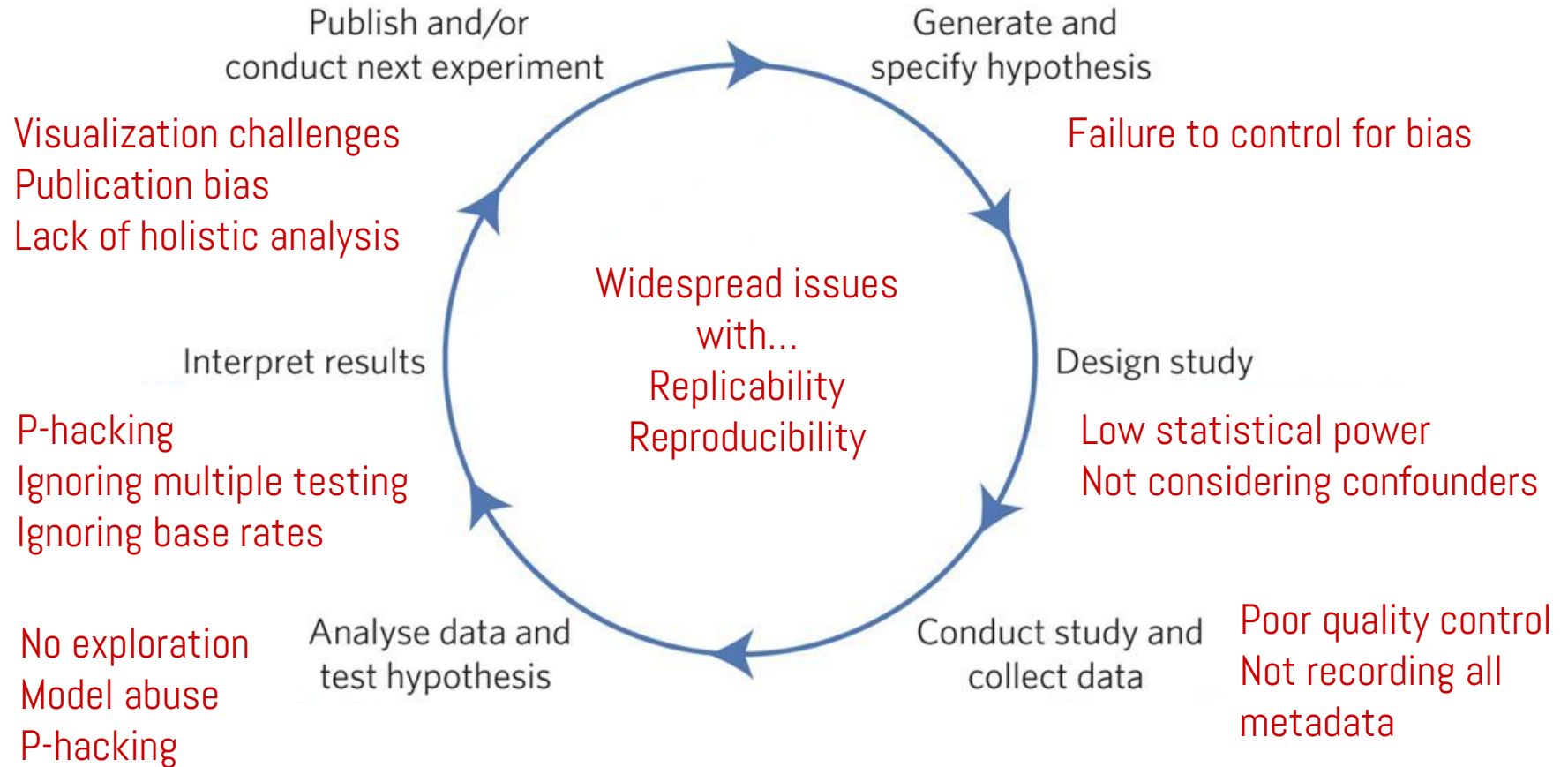
- Please call me 'Arjun'.
- **arjun**@msu.edu | the**krishnan**lab.org | @**comp**biologist
- Assistant Professor
 - Dept. Computational Mathematics, Science, and Engineering
 - Dept. Biochemistry and Molecular Biology
- Research Interests: Computational genomics, Biomedical data science, Biological networks, Natural language analysis, Data integration, Machine learning

Introductions

Introduce yourself to everyone in this class on the #welcome channel on Slack with:

- Name:
- Pronouns:
- Three words/phrases to describe you/your-interests:
- Research/interests in emojis!

What's this course about?



What's this course about?

Questionable requests that biostatisticians commonly receive:

- Altering some data to support hypothesis
- Interpreting findings on basis of expectation
- Not reporting missing data
- Ignoring violations of assumptions

[These requests are reported more frequently by younger statisticians.]

Trainees...

- Pressured by a PI or collaborator to produce “positive” data
- Pressure to publish influences the way they report data.

What's this course about?

Applying statistics to novel research questions and new complex datasets:

1. What the practical definitions of various concepts and their relationships are?
2. When and why they are applied in certain situations and not others?
3. What is a robust sequence of actions to take when applying them to data?
4. How to judiciously interpret the results?
5. How to clearly and transparently communicate the findings?

Most students piece together a mental model of the acceptable/standard/best practices in their field based on bits information from their mentors, peers, and often published papers.

What's this course about?

This is an advanced short (1-credit) course designed to:

- Discuss common misunderstandings & typical errors in the practice of statistical data analysis.
- Provide a mental toolkit for critical thinking and enquiry of analytical methods and results.

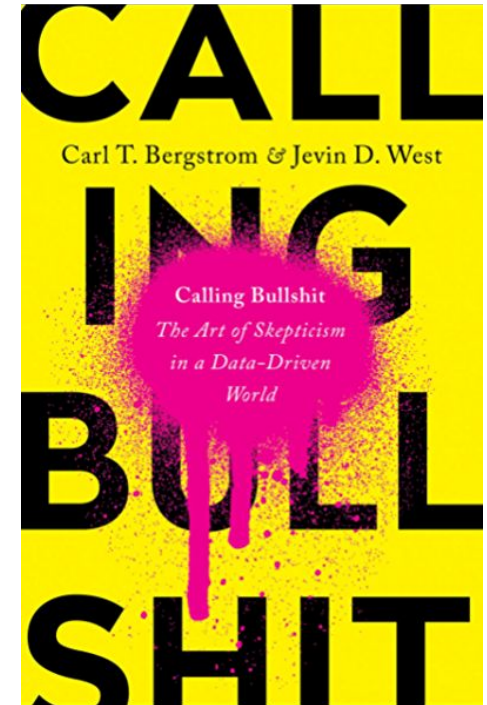
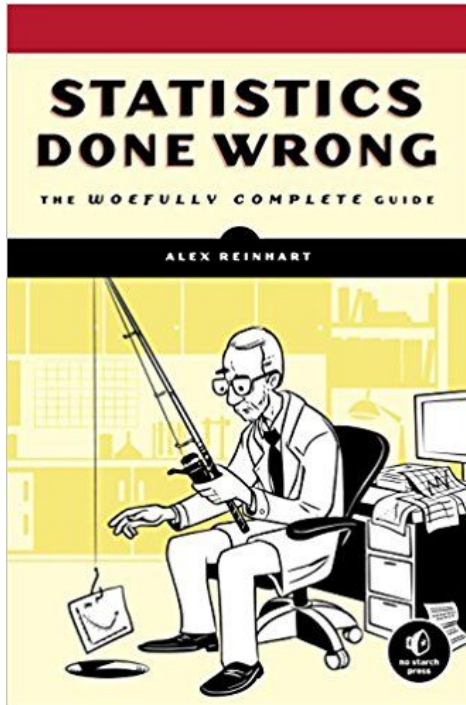
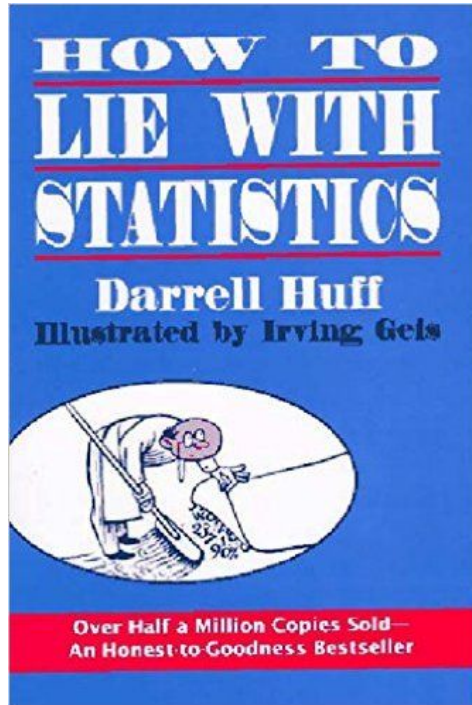
Prerequisites

- 1) Familiarity with basic statistics & probability
- 2) Ability to do basic data wrangling, analysis, & visualization using R or Python.

What's this course about?

- Scientific method, Critically reading literature, Cognitive biases
- Estimation of error & uncertainty
- P-value, Multiple testing, P-hacking, Publication bias
- Types of inference errors, Statistical power
- Sampling biases, Pseudoreplication, Confounding variables
- Circular analysis, Regression to the mean
- Base rates, Conditional probabilities, Bayesian reasoning
- Measuring association with continuous variables
- High dimensional data issues, Machine learning pitfalls
- Challenges in data presentation & visualization, Communicating statistics
- Data management, Code management, Reproducible research

Resources



Original research articles | Reviews | Blog posts | Podcasts

Some general thoughts

This class is not a cynical take on data analysis.


On the contrary, I *firmly* believe in the power of statistical enquiry, data analysis, and visualization.

The point is, because **many of the ideas involved are complex and unintuitive**, we need to develop a new set of skills to carefully use this power.

Some general thoughts

- Conscious ignorance: from unknown unknown → known unknown
 - Dunning-Kruger effect: knowing that something is unknown is as hard as knowing that thing!
 - The importance of feeling stupid: threshold of learning something new!
- Intelligent persistence
 - I don't understand this → What about this don't I understand?
 - Gaps in my knowledge → Gaps in collective knowledge

bit.ly/statgaps2021

- Contact information
 - Course outline and materials 
 - Schedule, location, calendar, & offline hours
 - Website and communication
 - Course activities
 - Grading information
 - Attendance, conduct, honesty, and accommodations
- Lecture slides
 - Learning materials
 - Assignments
 - Notes

statgaps2021.slack.com

- The primary mode of communication in this course (including major announcements) will be the course Slack account.
- All of you should have invitations to join this account in your MSU email.

#announcements

#topical-discussions

#lectures-assignments

#fun-wellness

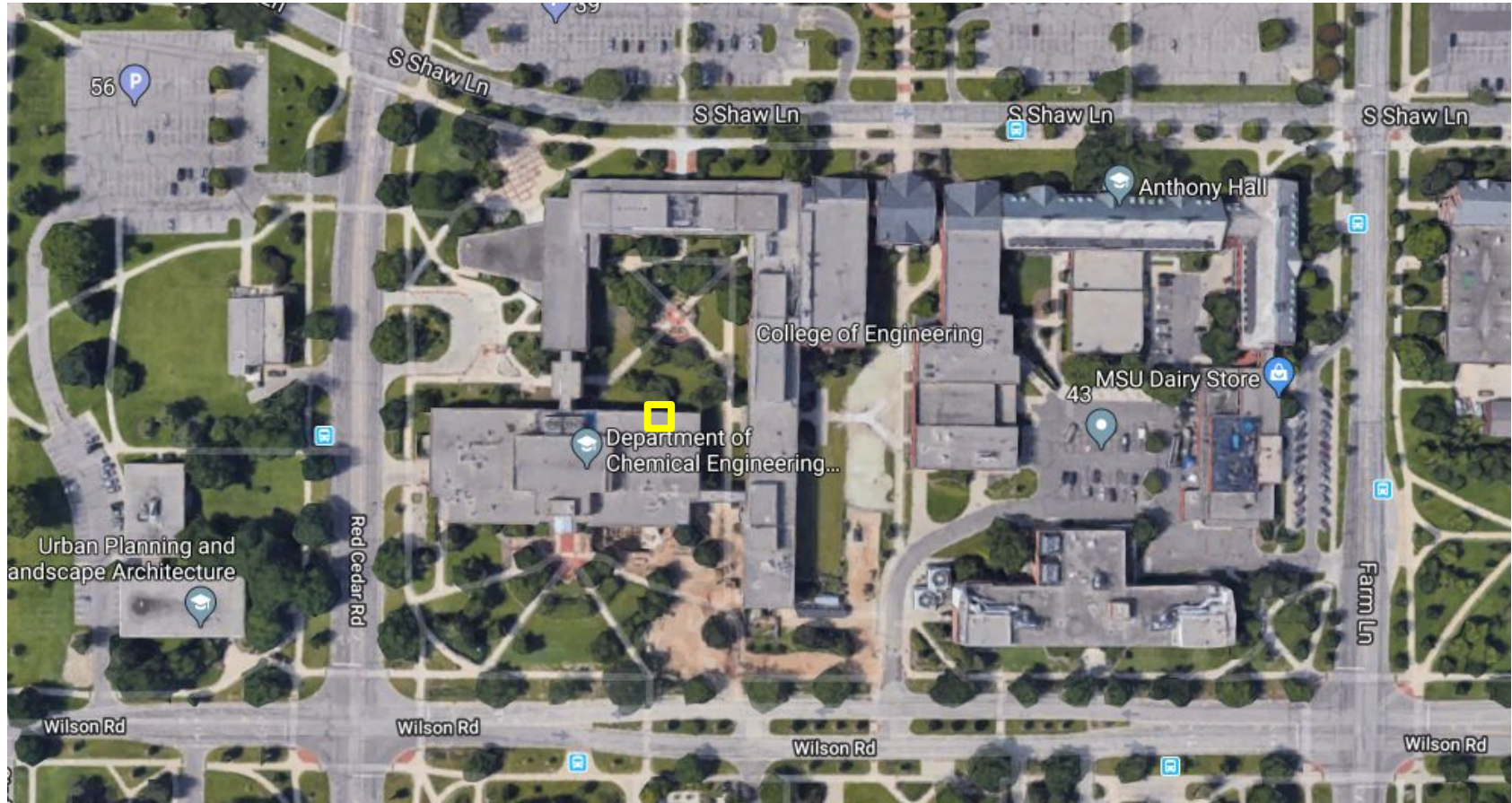
#papers-articles

#random

bit.ly/statgaps2021_incoming

- Select convenient hours for offline discussion
 - Will give preference to enrolled students when picking the time.
 - Even if you're not able to make it to the designated hours, just messaging on Slack with your questions/concerns will work as well.

My office: 2507H Engineering Building (2nd floor)



Course activities

- Assignments: 40%
- Class participation: 60%

Just like in the real world:

- There are no tests of memory. I strongly encourage you to talk to your fellow learners, peers, mentors, and me. Also, everything is open-internet. You can refer to anything you like.

Assignments

- Will be posted on Slack a week before it is due.
- The goal is to recap previous lecture and prepare for the discussions the following week.
- Reflecting on concepts in statistics / data-analysis.
- R or Python code to play with.

Coding

You will be working with code to:

- read-in existing datasets or generate mock datasets,
- wrangle them into a convenient format,
- call common statistical functions from standard packages/libraries to calculate mean, std. deviation, quantiles, correlation, etc.
- implement some simulations/tests
 - random number generation
 - writing for/while loops
- make plots (scatterplot, histograms, boxplots, etc.)

Language, IDE, Notebook
Pre-built external packages
Scientific computing
Data wrangling & visualization

- R | RStudio | R Notebook
- CRAN, Bioconductor
- In-built + Hundreds of packages
- Tidyverse

- Python | Rodeo | Jupyter
- PyPI, Biopython
- NumPy, SciPy + Hundreds of packages
- Pandas, Seaborn

Class participation

- Do the assignments and additional readings.
- Show up to class.
- Work in groups during in-class discussion sessions.
- Contribute to material in-class and on slack.
- No one will have the perfect background + the topics are all non-straightforward at all.
 - [Ask questions](#) about statistical or biological concepts.
- Postdocs, researchers, & faculty-members: I'm asking for your active engagement with the class & its materials, along with any feedback.

Class participation

Stop me to ask questions. I love getting Qs from you!

I will extremely sad if I don't get Qs :(

In this class, **there's nothing wrong about being wrong.**

- Being wrong is an opportunity to learn something.
- So, when I pose questions, think about it and always take a shot.
- It not only helps you learn a new piece of information, but it also helps you calibrate how to think about that information, which is way more valuable.

- Things to note:
 - I do not have a PhD in Statistics. I consider myself as an almost-power-user!
 - I will tell what parts of my understanding of these topics/ideas are works in progress and, hence, known-incomplete. I will try to be explicit about where the limits of my knowledge & understanding are.
 - I have no problem saying "Hmm, I'm not sure. Let me think about this & get back to you" or "I have no clue now but, if you're interested, we can read a couple of sources together & revisit this."
 - Correct me if/when I'm wrong.

Teaching philosophy

Each class → an inclusive collaborative learning community.

- My goal is to make sure that our class is a space where all of you can:
 - Join in,
 - Breathe,
 - Be seen & heard,
 - Be curious, and
 - Openly engage with the ideas.

Teaching philosophy

You absolutely belong in this community and you will be valued and respected.

My point of view is as follows:

- **Past:** Your unique background, training, and life experiences are your strengths that I and others can always learn from.
- **Present:** You have a life much bigger and multifaceted than your academic life within classrooms.
- **Future:** You are going to be my future colleagues within or outside academia.

Teaching philosophy

Finally, I design and teach classes that:

- Maximize my learning,
- Help me identify gaps in my knowledge, and
- Find better ways of discussing each of the many complex/interesting ideas.

Critically reading literature

StatGaps2021

Procedure

Looking out for biases

Arjun Krishnan | arjun@msu.edu | thekrishnanlab.org | [@compbio101](https://twitter.com/compbio101)

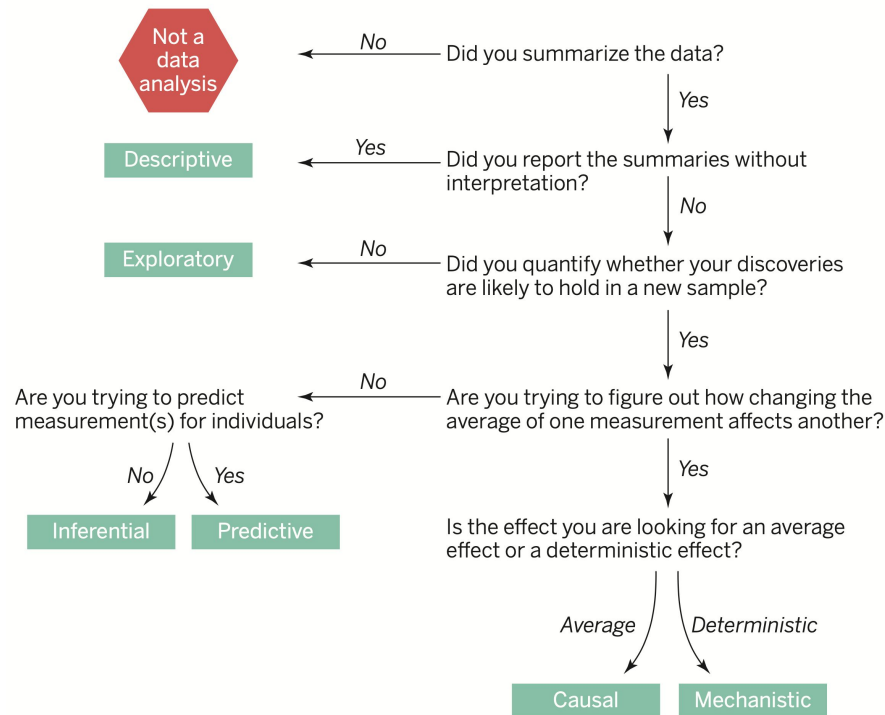
Reading primary research papers

1. Use **Title & Abstract** for only selecting paper.

- Don't be swayed by high-profile papers, media hype, or current dogma.

2. Read the **Introduction**:

- Identify *the* question. What is the big challenge the authors are trying to solve?
- What are the *specific* questions this paper is going to answered?



Reading primary research papers

3. Read **Data & Methods**: [Be critical!]

- a. For each specific Q, note data (type & source) & method (algorithms/techniques, software, & approach).
- b. Are the data & methods describes sufficient to answer the Qs raised in the Intro?
- c. Make detailed notes on: 1) what's unclear, 2) what you might do differently.

4. ALWAYS read the **Supplementary Materials**

These days much of the good stuff is in here!

Reading primary research papers

5. Read the **Results**: [Be critical!]

- a. Go figure-by-figure, panel-by-panel. Based on your reading of Data & Methods, is there enough information to know/reproduce that analysis?
- b. Try to interpret each figure/panel, then read the figure legend and the part of the results that explains it. [**Supplemental figures/tables** abound!]
 - i. Do your interpretations match that of the authors'?
 - ii. Are the results answering the specific Qs?
- c. Make detailed notes on: 1) what's unclear, 2) what you might do differently.

6. Read the **Discussion/Conclusions, Title, & Abstract**:

- a. Step back to think about contributions, limitations, open Qs, & next steps.

7. Read what other researchers (**papers that cite this paper**) say about this paper.

Puzzle

A Quick Puzzle to Test Your Problem Solving

By **DAVID LEONHARDT** and **YOU** JULY 2, 2015

A short game sheds light on government policy, corporate America and why no one likes to be wrong. [RELATED ARTICLE](#)

Here's how it works:

We've chosen a rule that some sequences of three numbers obey — and some do not. Your job is to guess what the rule is.

We'll start by telling you that the sequence 2, 4, 8 obeys the rule:

2

4

8

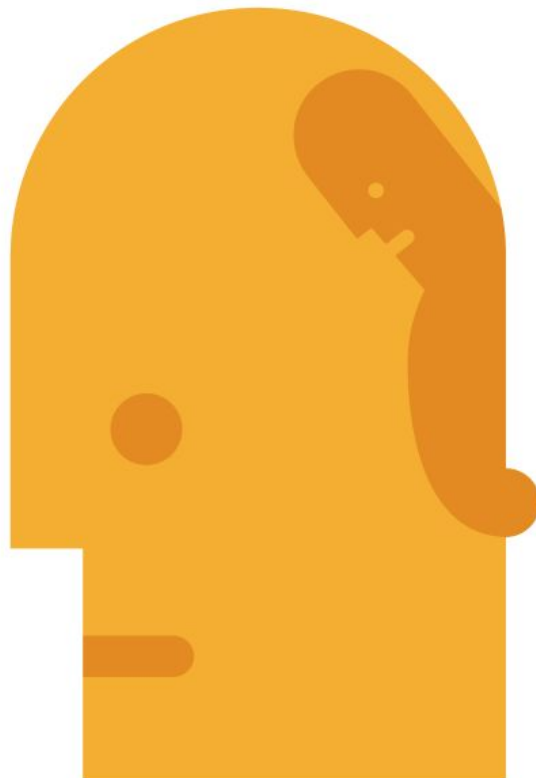
Obeys the rule

Now it's your turn. Enter a number sequence in the boxes below, and we'll tell you whether it satisfies the rule or not. You can test as many sequences as you want.

Enter your first sequence here:

Check

[I don't want to play; just tell me the answer.](#)



Confirmation bias

The tendency to search for, interpret, favor, and recall information in a way that confirms one's beliefs or hypotheses.

- Scientists rate studies that report findings consistent with their prior beliefs more favorably than studies reporting findings inconsistent with their previous beliefs.
- Data that conflict with the experimenter's expectations may be more readily discarded as unreliable.

"It is the peculiar and perpetual error of the human intellect to be more moved and excited by affirmatives than by negatives; whereas it ought properly to hold itself indifferently disposed towards both alike."

– *Francis Bacon*

Cognitive biases in research



HYPOTHESIS MYOPIA

Collecting evidence to support a hypothesis, not looking for evidence against it, and ignoring other explanations.



TEXAS SHARPSHOOTER

Seizing on random patterns in the data and mistaking them for interesting findings.



ASYMMETRIC ATTENTION

Rigorously checking unexpected results, but giving expected ones a free pass.



JUST-SO STORYTELLING

Finding stories after the fact to rationalize whatever the results turn out to be.

Cognitive biases in research – Debiasing techniques



HYPOTHESIS MYOPIA

Collecting evidence to support a hypothesis, not looking for evidence against it, and ignoring other explanations.



TEXAS SHARPSHOOTER

Seizing on random patterns in the data and mistaking them for interesting findings.



ASYMMETRIC ATTENTION

Rigorously checking unexpected results, but giving expected ones a free pass.



JUST-SO STORYTELLING

Finding stories after the fact to rationalize whatever the results turn out to be.



DEVIL'S ADVOCACY

Explicitly consider alternative hypotheses — then test them out head-to-head.



PRE-COMMITMENT

Publicly declare a data collection and analysis plan before starting the study.



TEAM OF RIVALS

Invite your academic adversaries to collaborate with you on a study.



BLIND DATA ANALYSIS

Analyse data that look real but are not exactly what you collected — and then lift the blind.

Uncertainty, Error

Standard deviation

Standard error

Confidence interval

Whether we are right vs. the chances of being wrong

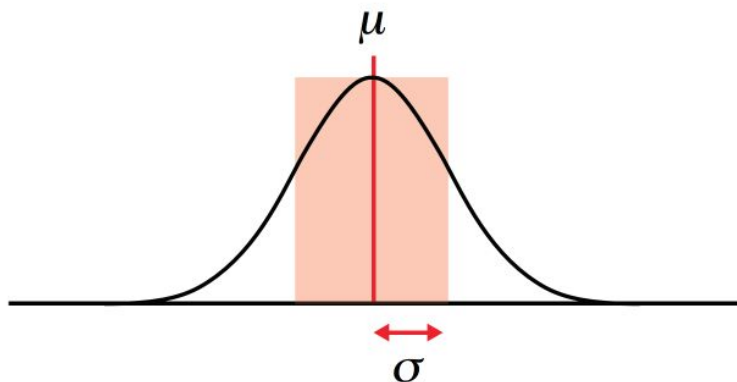
Repeated measurements \rightarrow Range of values.

Statistics helps us by helping with:

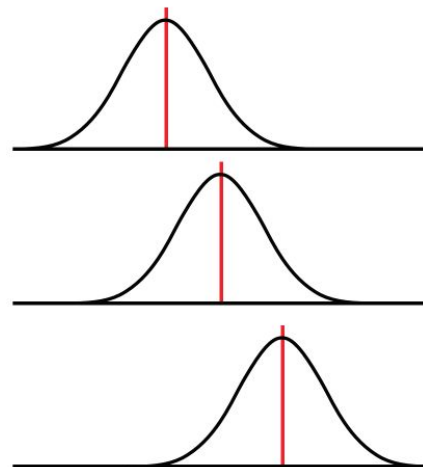
- Modeling the role of chance
- Represent data as estimates with errors

Population distribution

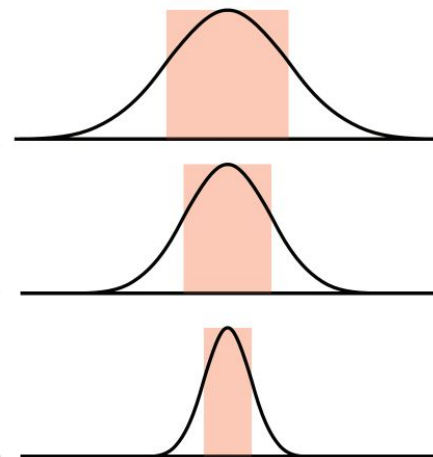
Population distribution



Location



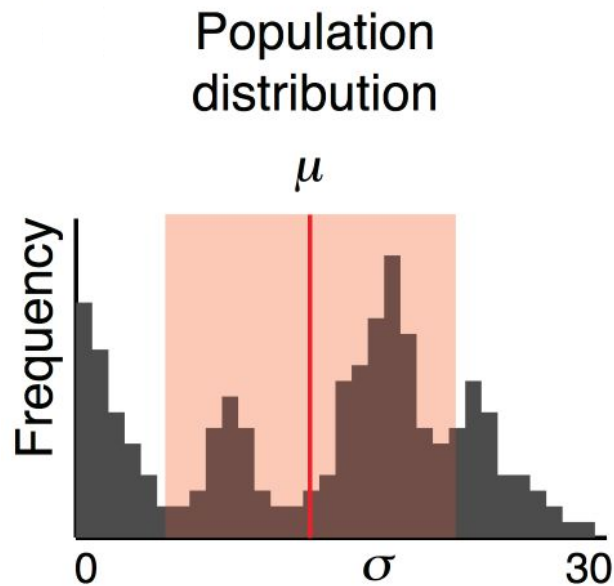
Spread



μ : Population mean | σ : Population standard deviation

These are, of course, hard to calculate because it is hard to collect data about the entire population.

Estimating population parameters by sampling



Samples

$X_1 = [1, 9, 17, 20, 26]$

$X_2 = [8, 11, 16, 24, 25]$

$X_3 = [16, 17, 18, 20, 24]$

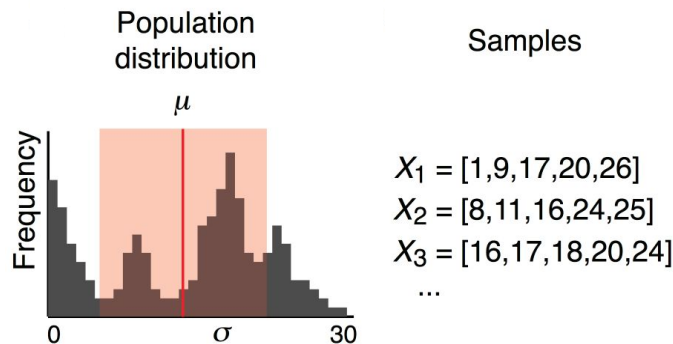
...

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Standard deviation

- Error bars based on **s.d.** → spread of your data.
- Useful as predictors of the range of new samples.
- Only indirectly supports visual assessment of differences in values:
 - **s.d.** bars reflect the variation of the data
 - They do not reflect the error in your measurement.

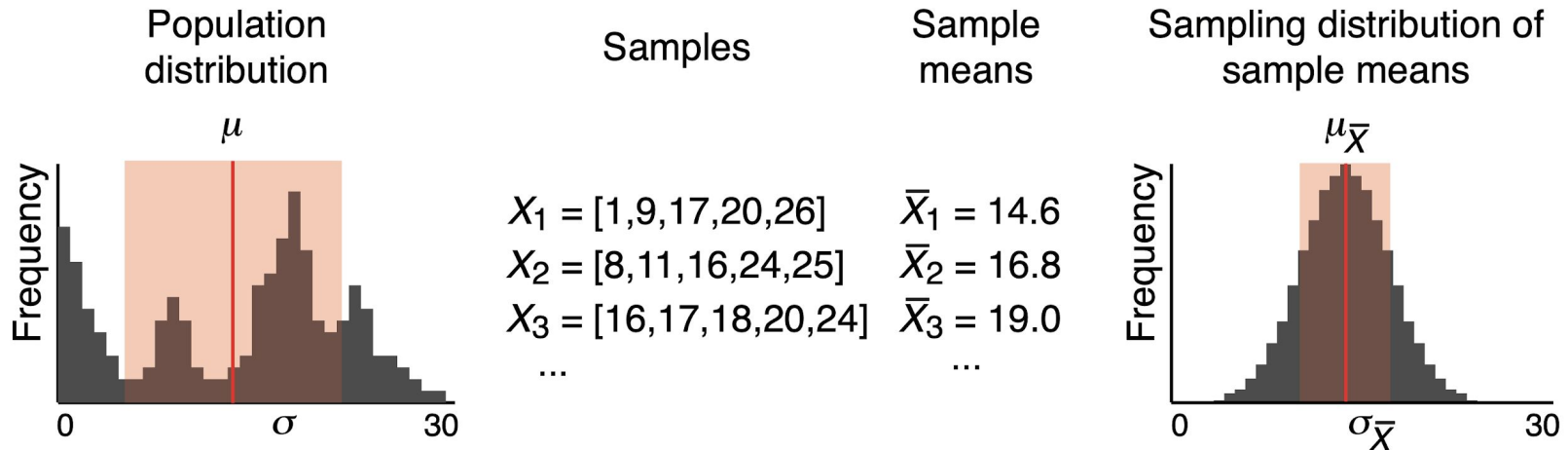


$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Standard error of the mean

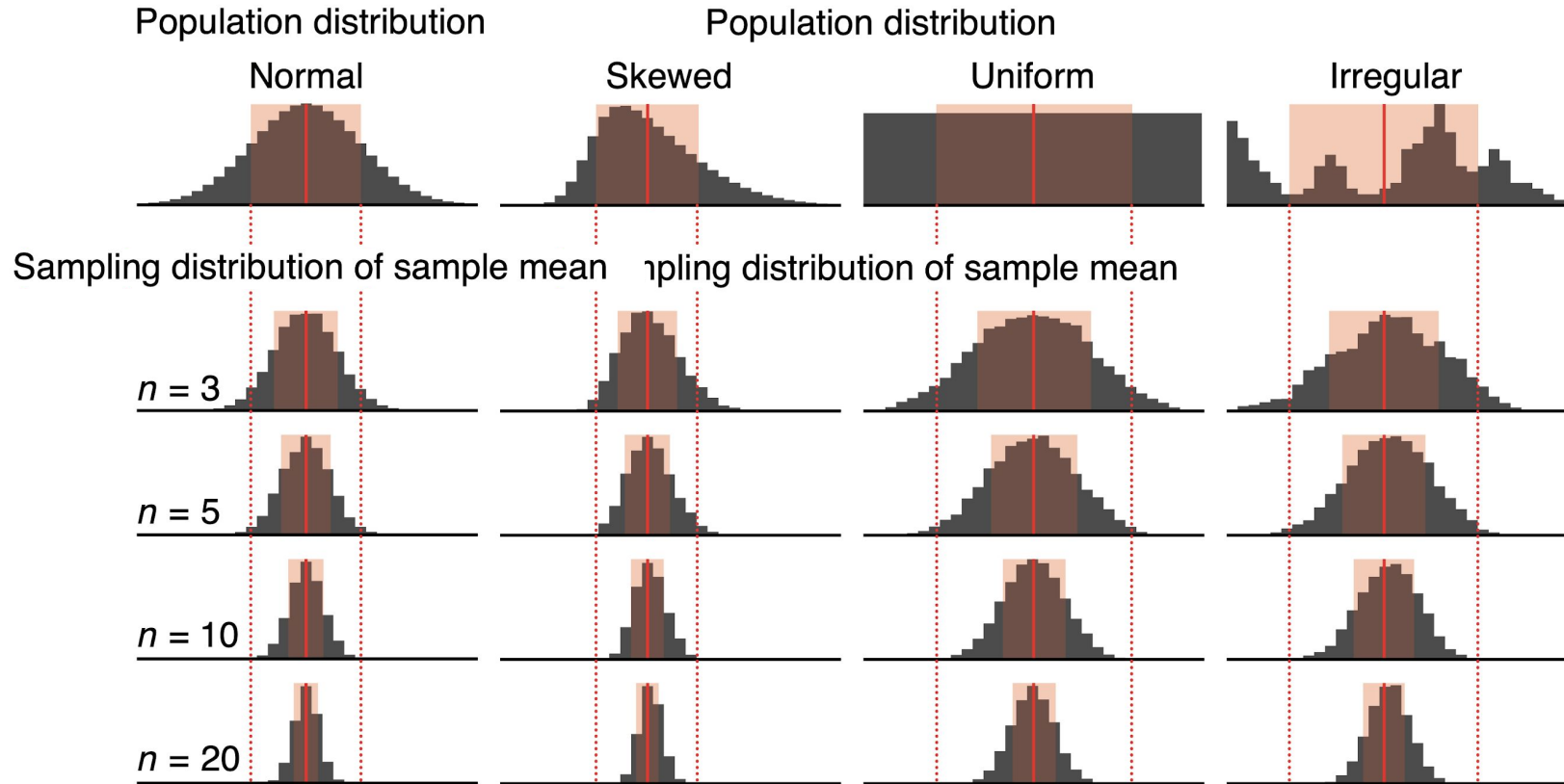
- Error bars based on **s.e.m.** → spread of the *means* of independent measurement samples, not the sample you collected (your data).
- s.e.m. = standard deviation of the means



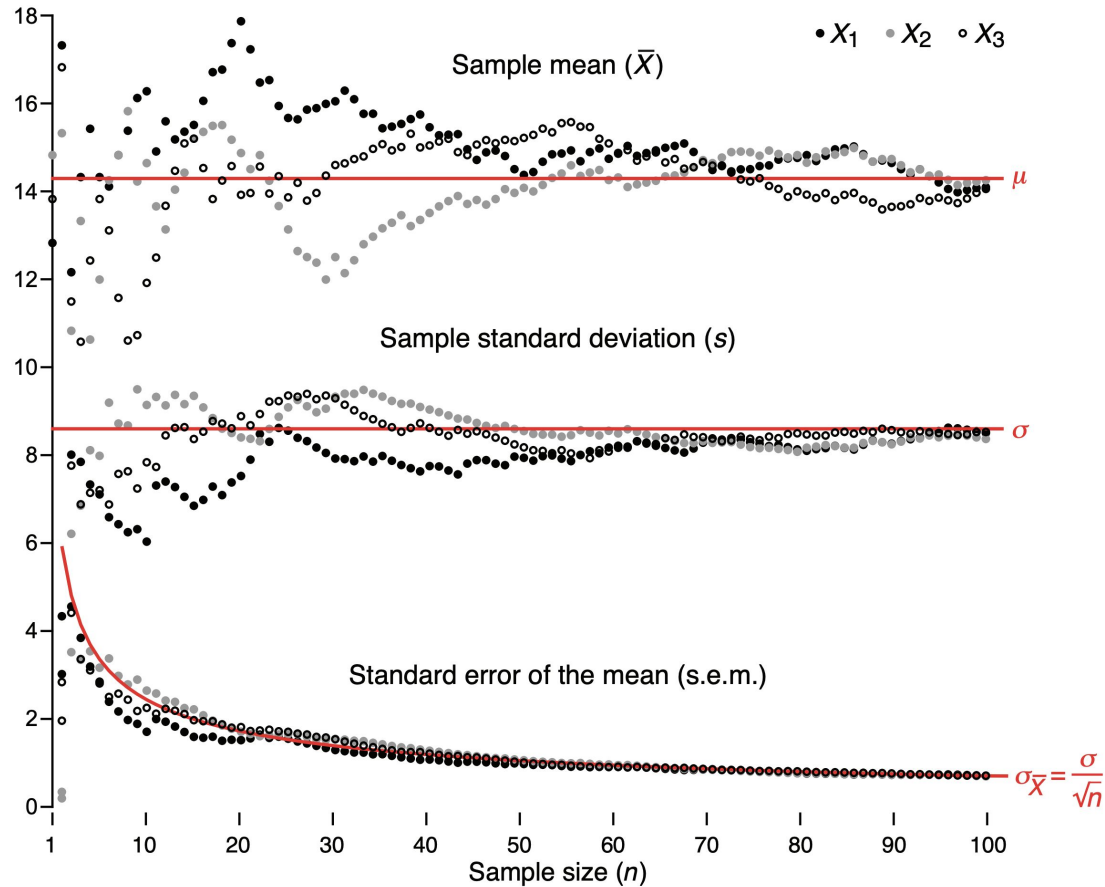
Standard error of the mean

- Error bars based on **s.e.m.** → spread of the ***means*** of independent measurement samples, not the sample you collected (your data).
- s.e.m. = standard deviation of the means
- s.e.m. \ll s.d. of individual samples
- In rare cases, can be estimated using a formula: $\text{s.e.m.} = \text{s.d.} / \sqrt{n}$
 - Rest of the times, use bootstrapping.
- Dependent on sample size:
 - Shrinks as we perform more measurements.

Standard error of the mean



Standard error of the mean



Let's write code to calculate mean, s.d., and s.e.m. of a sample

Instructions

1. Generate 1000 random numbers from a normal distribution with mean = 0 & s.d. = 1.
Let these 1000 numbers represent the population.
2. Randomly choose 10 numbers from these 1000.
These 10 numbers represent a sample from the population.
3. Calculate the sample **mean**.
4. Calculate the sample **s.d.**
5. Calculate **s.e.m.** using the formula $(\text{s.d.} / \sqrt{n})$.

Let's write code to empirically calculate s.e.m

Instructions

1. Generate 1000 random numbers from a normal distribution with mean = 0 & s.d. = 1
2. Repeat the following a 100 times:
 - a. Randomly sample 10 numbers from the population of 1000 numbers
 - b. Record their means
3. Calculate the s.d. of these 100 means.

What does this give you?

Recall: **s.e.m.** → spread of the ***means*** of independent measurement samples.

Let's write code to empirically calculate s.e.m. of a sample

Given 10 numbers that represent a sample.
We have no access to the entire population.

30, 37, 36, 43, 42, 43, 43, 46, 41, 42

Instructions

1. Create 1000 *bootstrap* samples:
 - a. Each time, sample 10 numbers *with replacement*
 - b. Calculate the mean of each bootstrap sample
2. Calculate the **s.d.** of these means.

43 36 46 30 43 43 43 37 42 42 43 37 36 42 43 43 42 43 42 43
43 41 37 37 43 43 46 36 41 43 43 42 41 43 46 36 43 43 43 42
42 43 37 43 46 37 36 41 36 43 41 36 37 30 46 46 42 36 36 43
37 42 43 41 41 42 36 42 42 43 42 43 41 43 36 43 43 41 42 46
42 36 43 43 42 37 42 42 42 46 30 43 36 43 43 42 37 36 42 30
36 36 42 42 36 36 43 41 30 42 37 43 41 41 43 43 42 46 43 37
43 37 41 43 41 42 43 46 46 36 43 42 43 30 41 46 43 46 30 43
41 42 30 42 37 43 43 42 43 43 46 43 30 42 30 42 30 43 43 42
46 42 42 43 41 42 30 37 30 42 43 42 43 37 37 37 42 43 43 46
42 43 43 41 42 36 43 30 37 43 42 43 41 36 37 41 43 42 43 43

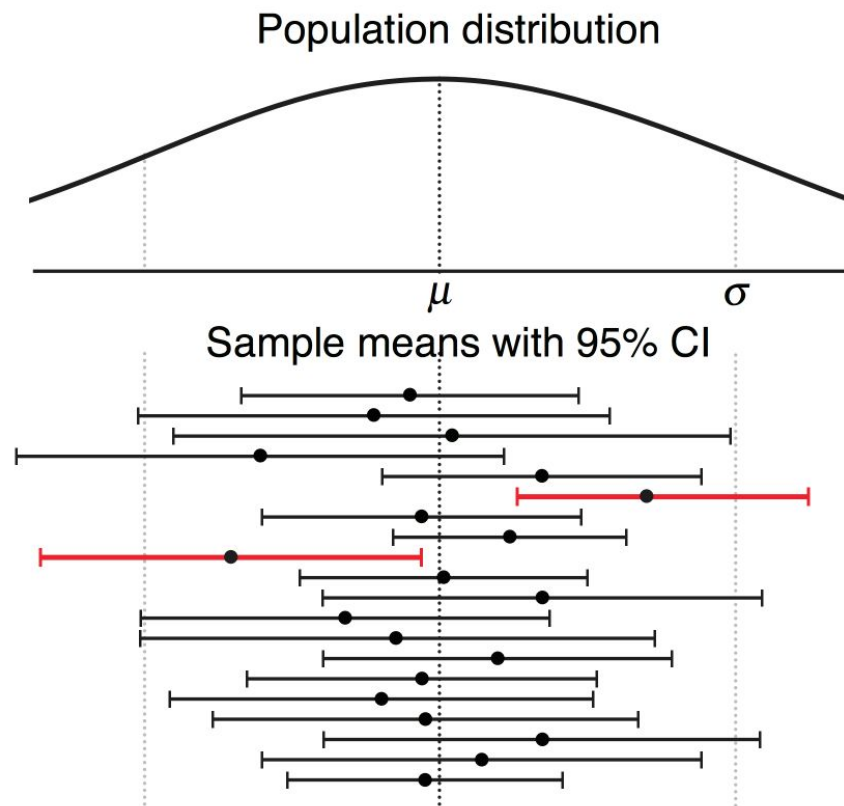
This is the **s.e.m.** of your sample estimated using bootstrapping!

Confidence interval

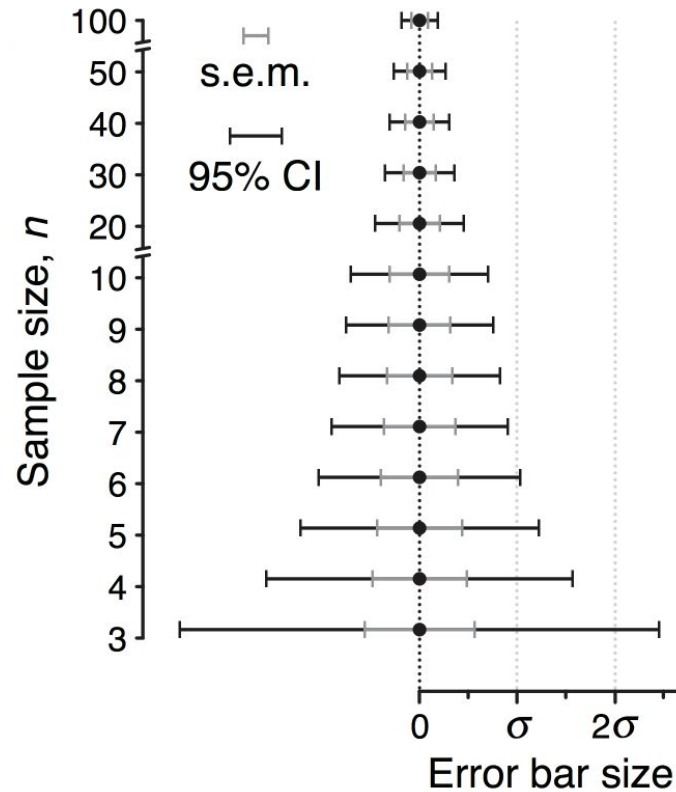
- **CI** is an interval estimate that indicates the reliability of a measurement.
 - The 95% CI bar captures the population mean 95% of the time.
- Just like s.e.m., CI can be calculated using a bootstrapping technique.

30, 37, 36, 43, 42, 43, 43, 46, 41, 42

43 36 46 30 43 43 43 37 42 42 43 37 36 42 43 43 42 43 42 43
43 41 37 37 43 43 46 36 41 43 43 42 41 43 46 36 43 43 42
42 43 37 43 46 37 36 41 36 43 41 36 37 30 46 46 42 36 36 43
37 42 43 41 41 42 36 42 42 43 42 43 41 43 36 43 43 41 42 46
42 36 43 43 42 37 42 42 42 46 30 43 36 43 43 42 37 36 42 30
36 36 42 42 36 36 43 41 30 42 37 43 41 41 43 43 42 46 43 37
43 37 41 43 41 42 43 46 46 36 43 42 43 30 41 46 43 46 30 43
41 42 30 42 37 43 43 42 43 43 46 43 30 42 30 42 30 43 43 42
46 42 42 43 41 42 30 37 30 42 43 42 43 37 37 37 42 43 43 46
42 43 43 41 42 36 43 30 37 43 42 43 41 36 37 41 43 42 43 43

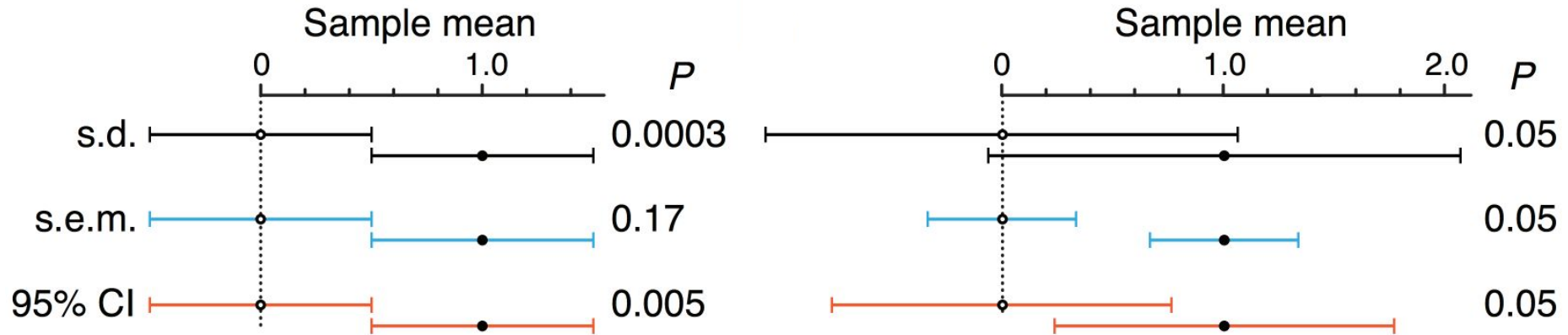


Confidence interval

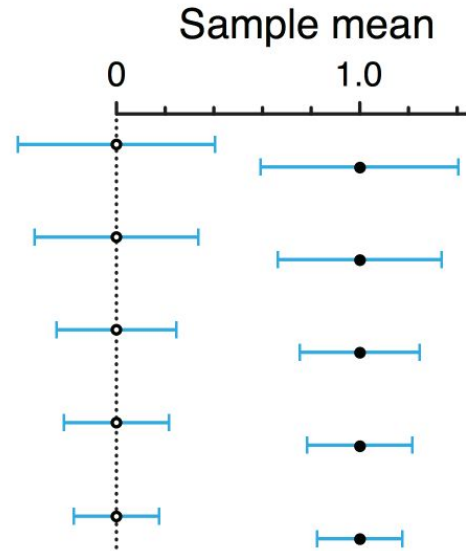


The type of “spread” matters

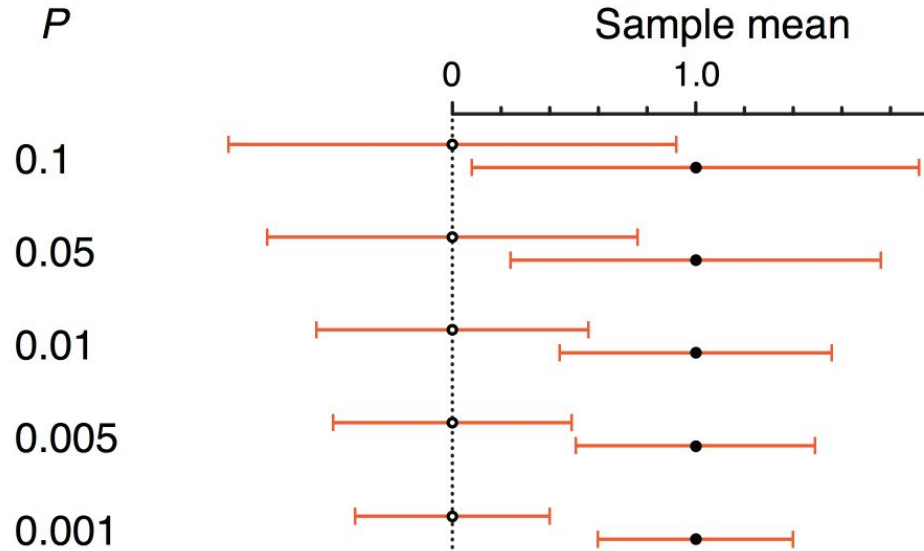
- Non-overlapping \neq “significant” difference
- Overlapping \neq not “significant” difference
- It depends on the type of the error bar.



Standard error of the mean, Confidence interval



s.e.m. error bars



95% CI error bars

What you need to do before the next class

PART 1

Complete the incoming survey: bit.ly/statgaps2021_incoming

Among other things, this will help in finding a time for offline discussions.

What you need to do before the next class

PART 2

Install R or Python

- Install R, RStudio, and Tidyverse (package); Get familiar with R Notebooks, or
- Install Anaconda, Python 3.7, Jupyter Notebooks

Resources with detailed instructions are on the class website.

What you need to do before the next class

PART 3

- A major goal of this course is to prepare your ability to perform and critique statistical data analysis and to present your ideas and results effectively.
- bit.ly/statgaps2021_assignment01
- This assignment will give you an opportunity to revisit many statistical concepts and set the tone for this course.

1. Data properties and study design

- Sample size, sample collection, and replication
- Choice of sample size, inclusion/exclusion of samples
- Study design, randomization, blinding

2. Data analysis

- Justification for each test (question + assumptions)
- Detailed description to reproduce each analysis

3. Data reporting

- Definition of each statistical method and measure

4. Data and code availability

How to pick an analysis/result to focus on?

