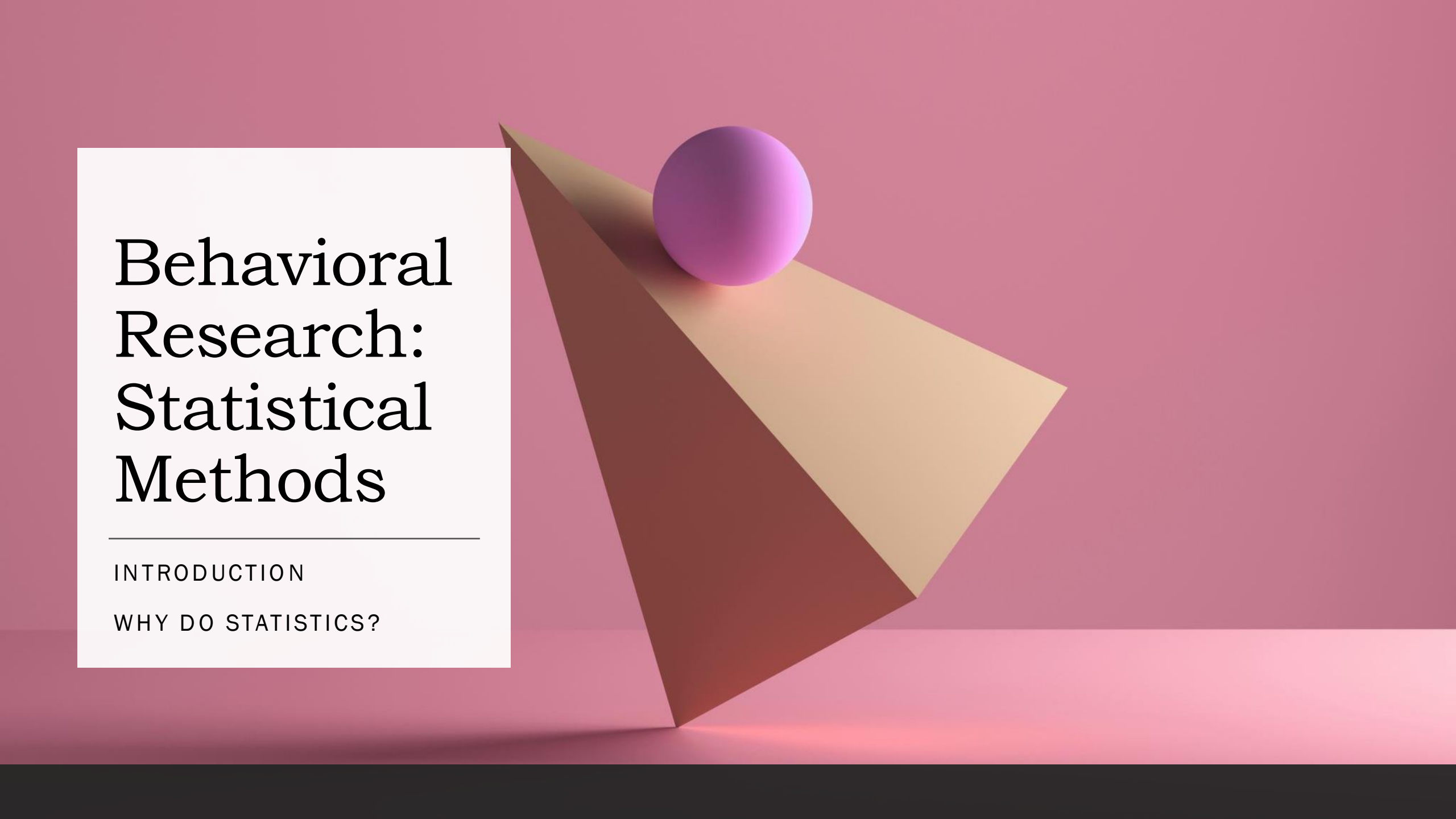
Behavioral Research: Statistical Methods

INTRODUCTION

WHY DO STATISTICS?



Agenda

Syllabus and related questions

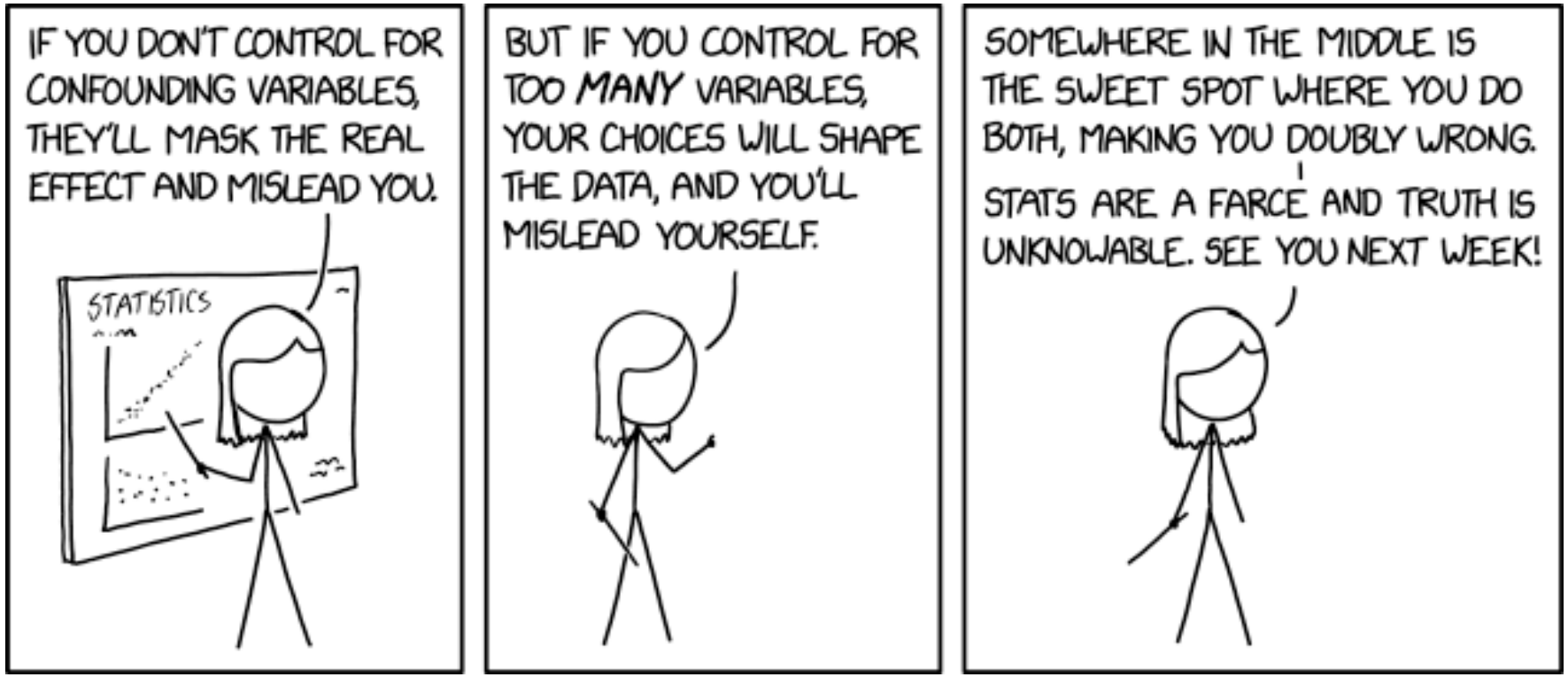
Syllabus

- Please read carefully!
- Uploaded on Moodle.
- Some big changes this semester due to the high enrollment.

Question about coding language

The reference textbook for the course uses R. Many of our problem and practice sets will include R code snippets. So you will need a laptop with R and RStudio installed.

You can however use any language of your choice (MATLAB, Python, etc) to complete your assignments and projects.



Why do statistics?

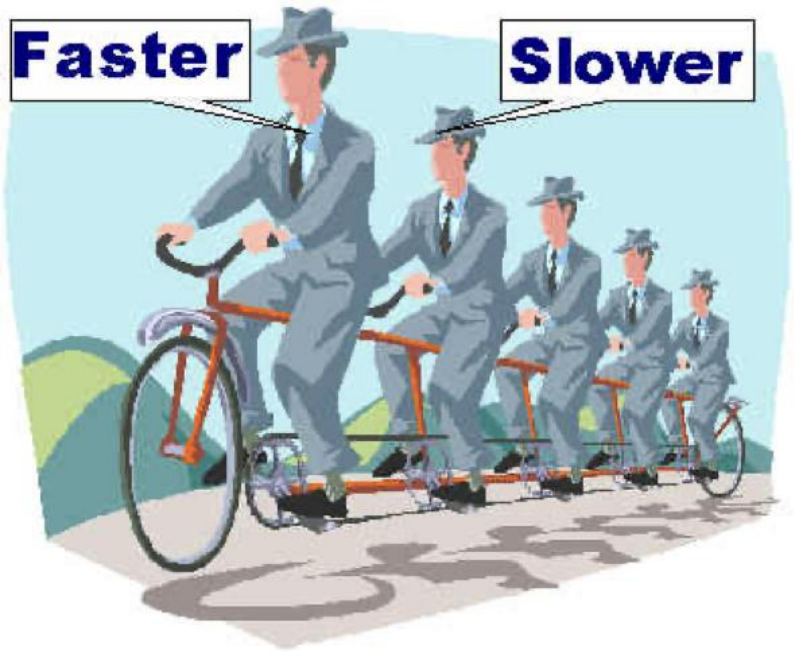
Why do statistics, why not use common sense?

My mom: drink milk with turmeric, it will cure you of sore throat. I have experienced this, 3 days of drinking it and my sore throat is gone. My friends have also experienced it.

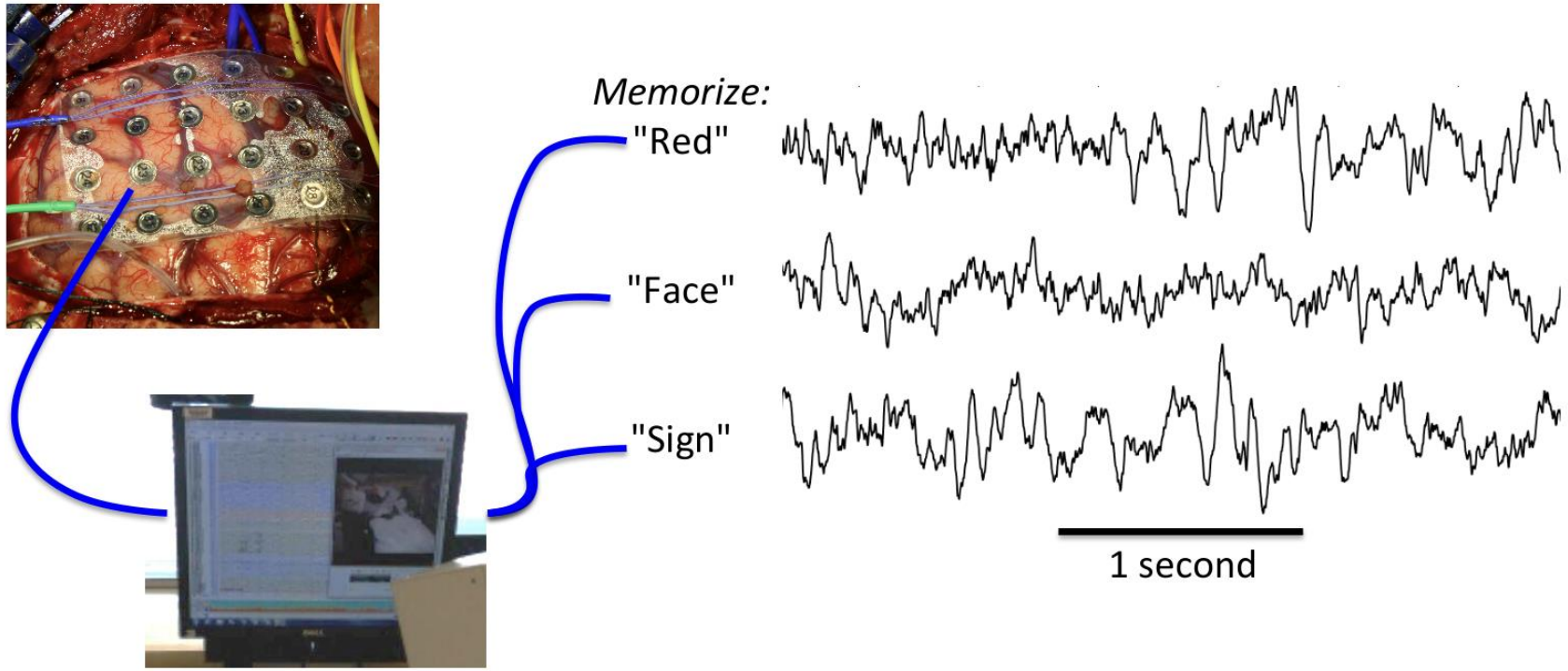
You might have encountered many such claims, especially during the early days of COVID. There is also currently a proliferation of pseudoscientific thinking in India. An education in basic statistics and research design will hopefully help you see through some of the issues with such claims.

Human-beings

- Complexity
- Variability
- Reactivity



Brains



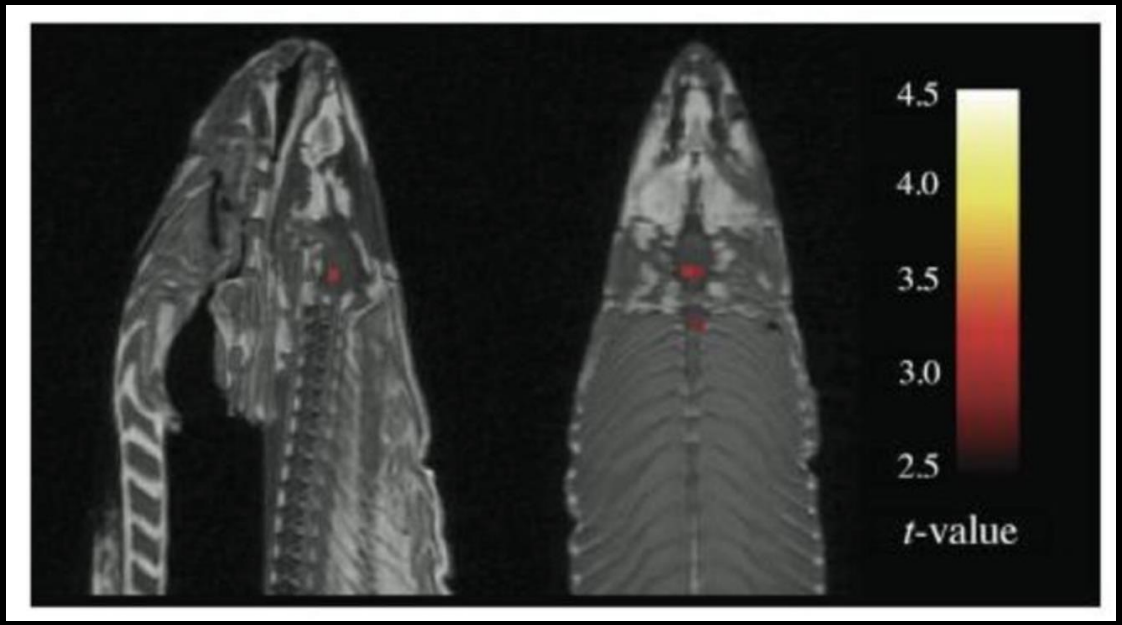
Related statistical pitfalls

Replication Crisis

Reproducibility Crisis

Reviewed by Psychology Today Staff

The replication crisis in psychology refers to concerns about the credibility of findings in psychological science. The term, which originated in the early 2010s, denotes that findings in behavioral science often cannot be replicated: Researchers do not obtain results comparable to the original, peer-reviewed study when repeating that study using similar procedures. For this reason, many scientists question the accuracy of published findings and now call for increased scrutiny of research practices in psychology.



IgNobel Prize in Neuroscience: The dead salmon study

Human biases

Belief bias

We tend to be swayed by the "believability" of the conclusion even when we are trying to deduce the conclusion from certain premises in a logical fashion (I.e., assuming the premises are true, is the conclusion valid?).

Believable conclusion and valid argument

- No cigarettes are inexpensive (Premise 1)
- Some addictive things are inexpensive (Premise 2)
- Therefore, some addictive things are not cigarettes (Conclusion)

Unbelievable conclusion but valid argument

- No addictive things are inexpensive (Premise 1)
- Some cigarettes are inexpensive (Premise 2)
- Therefore, some cigarettes are not addictive (Conclusion)

Believable conclusion but invalid argument

- No addictive things are inexpensive (Premise 1)
- Some cigarettes are inexpensive (Premise 2)
- Therefore, some addictive things are not cigarettes (Conclusion)

Unbelievable conclusion and invalid argument

- No cigarettes are inexpensive (Premise 1)
- Some addictive things are inexpensive (Premise 2)
- Therefore, some cigarettes are not addictive (Conclusion)

	conlusion feels true	conclusion feels false
argument is valid	92% say “valid”	–
argument is invalid	–	8% say “valid”

Evans, Barston, & Pollard (1983)

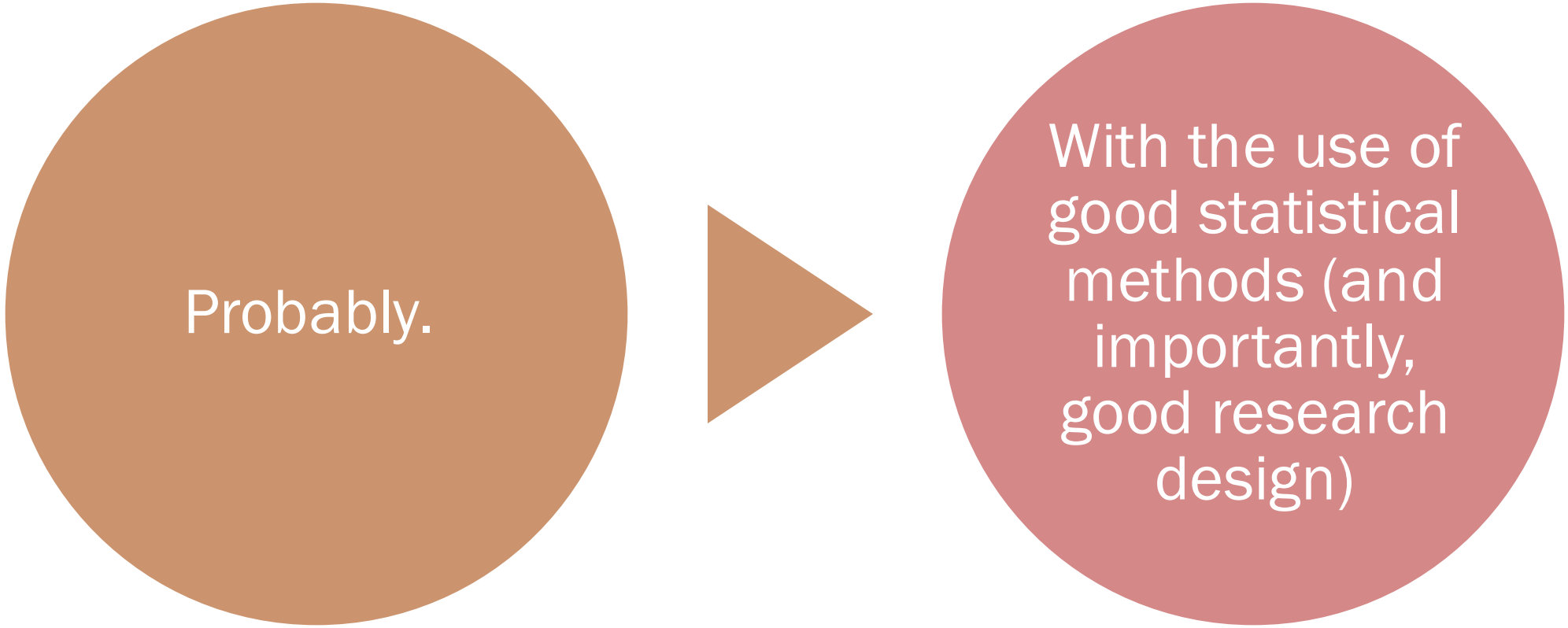
when the structure of the argument was in line with pre-existing beliefs and biases

	conlusion feels true	conclusion feels false
argument is valid	92% say “valid”	**46% say “valid”**
argument is invalid	**92% say “valid”**	8% say “valid”

Evans, Barston, & Pollard (1983)

when the structure of the argument contradicted pre-existing beliefs and biases

Can we improve our chances of being correct from 60% to 90+%?



Simpson's paradox

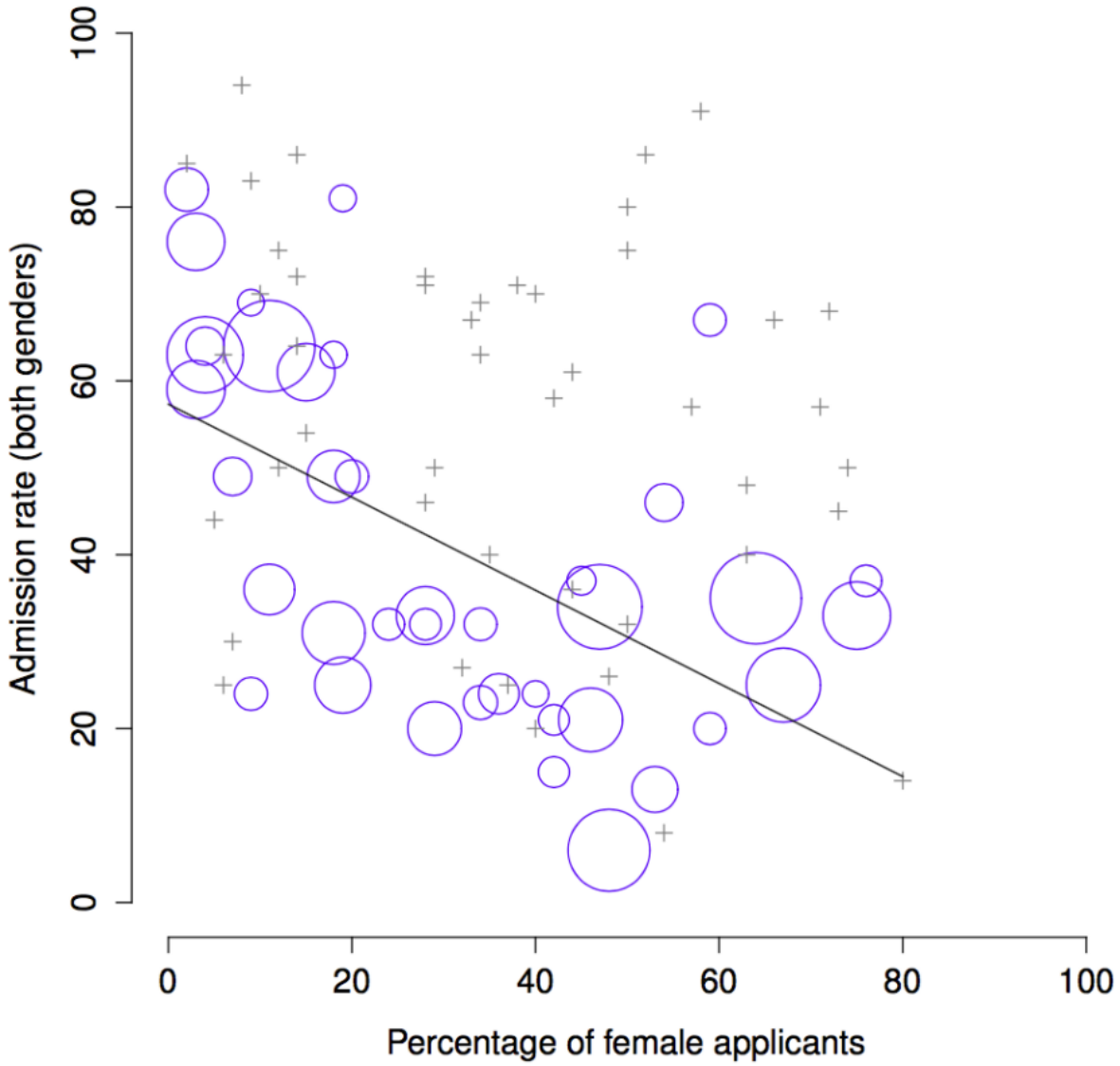
	Number of applicants	Percent admitted
Males	8442	44%
Females	4321	35%

Department	Applicants	Percent admitted	Applicants	Percent admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Counter-intuitive but of practical relevance

The overall rate of admission was lower for females than males but in individual departments, it was the opposite!!

The textbook says this is a rare example, but this is actually quite applicable in many scenarios where instead of departments in this example, you have data from different human subjects. These subjects do slightly different things but you try to make a conclusion about the whole population with some average measure. What the average tells you in some cases may be misleading. We need to have strong foundations in statistics to be aware of such cases.



Data visualization

Data interpretation

Once you do the statistics, it is time to interpret the results of your analysis

Is there gender bias in admissions?

Based on the departmental data?

Based on what criteria? This now is where you bring your theories to bear upon the data. For example, does the theory care about systemic issues that make females apply less frequently to say the engineering departments (explaining why the total number of applicants are distributed differently across the departments for males and females)?

Statistics in everyday life

WE SEE CLAIMS EVERY DAY IN THE MEDIA USING STATISTICS

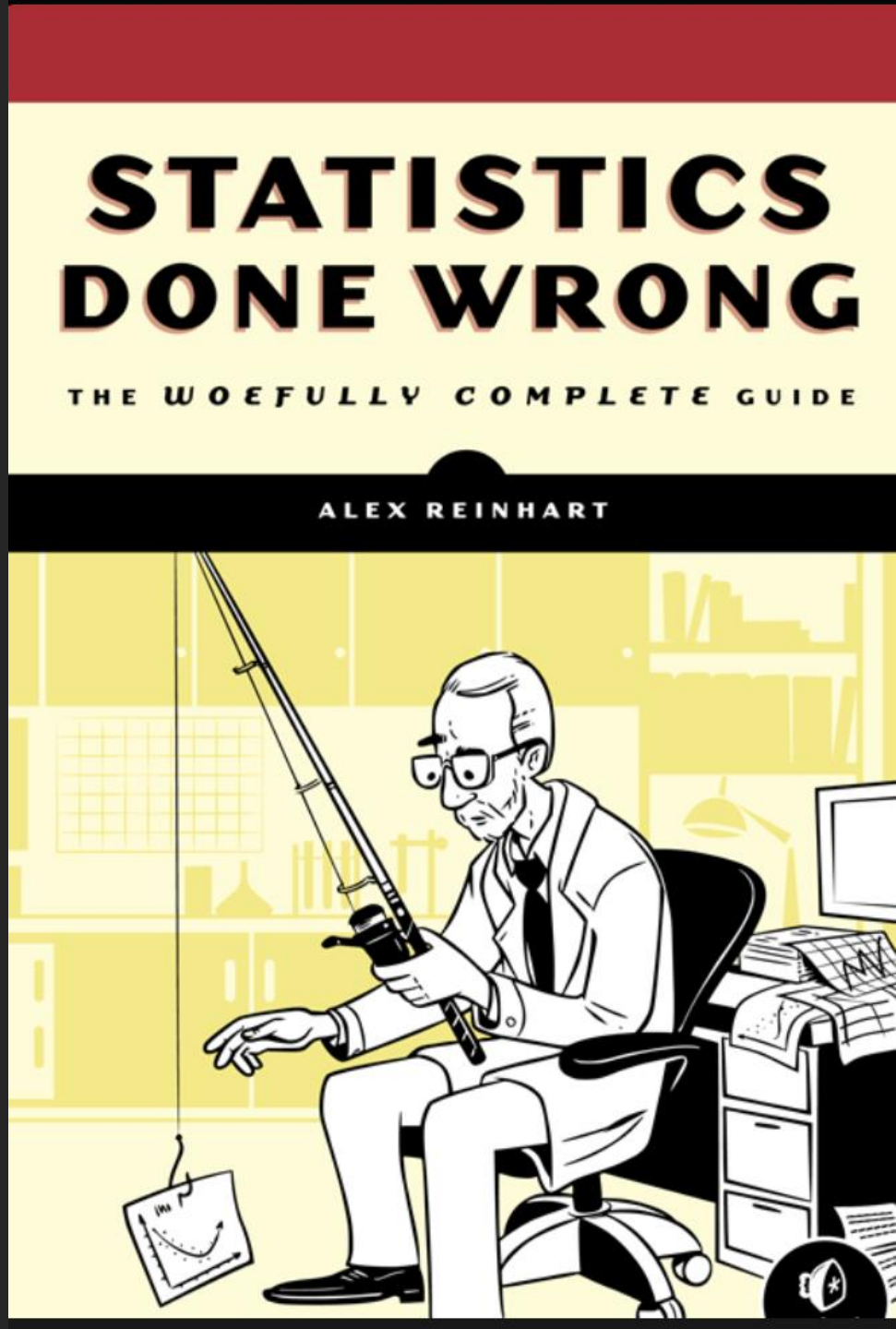
OFTEN, REPORTERS MAKE FUNDAMENTAL ERRORS WHEN THEY REPORT NUMBERS (E.G. NOT TAKING INTO ACCOUNT BASE RATES)

Some pitfalls

Misinterpreting p values

Misinterpreting confidence intervals (I estimate the mean height of boys in this class to be 5'7" with a 95% CI of [5'3", 5'11"])

Base rate fallacy (a lack of understanding of Bayesian probability)



The base rate fallacy

Let's say you or your relative/friend gets a positive mammogram result.

How likely is it that they have cancer?

Some relevant info (premises)

- 0.8% of all women who get mammograms have breast cancer
- In 90% of these women with breast cancer, a mammogram will correctly detect it (defined as the **statistical power**)
- However 7% of women without cancer will get a false positive mammogram
- How likely is it that a positive test indicates cancer?

Imagine 1000 tests

8 of them have cancer

7/8 of them will get a positive mammogram (due to the 90% power of this test)

992 with no breast cancer

7% false positive = ~70 women incorrectly told they have breast cancer

Now, how many total positive mammograms do we have?? 70 + 7 = 77

Only 7 of them actually have breast cancer.

(7/77)x100 = 9%.

So the probability that given a positive mammogram, someone actually has breast cancer = 0.09 or 9%

Bayes' rule

Bayes' Theorem

We can turn the process above into an equation, which is Bayes' Theorem. It lets you take the test results and correct for the "skew" introduced by false positives. You get the real chance of having the event. Here's the equation:

$$\Pr(\mathrm{H}|\mathrm{E}) = \frac{\Pr(\mathrm{E}|\mathrm{H})\,\Pr(\mathrm{H})}{\Pr(\mathrm{E}|\mathrm{H})\,\Pr(\mathrm{H}) + \Pr(\mathrm{E}|\text{not H})\,\Pr(\text{not H})}$$

The chance evidence is real (supports a hypothesis)
is the chance of a true positive among
all positives (true or false)

Bayes' rule

P(cancer|positive test) = P(positive test|cancer) * P(cancer)/P(positive test)

What we know:

1. P(positive test|not cancer) = 0.07 (false positive probability)
2. P(cancer) = 0.8% = 0.008
3. P(positive test|cancer) = 0.9

P(positive test) = 77/1000 = 0.077 (from the last slide)

The other way to calculate P(positive test) = P(positive test|cancer)*P(cancer) + P(positive test|not cancer) * P(not cancer) = 0.9*0.008 + 0.07*0.992 = 0.07664 ~= 0.077

P(cancer|positive test) = 0.9 *0.008/0.077 = 0.09

P(cancer) = base rate P(positive test|not cancer = false positive probability

Many fail to give the right answer

2/3rds of doctors fail this test

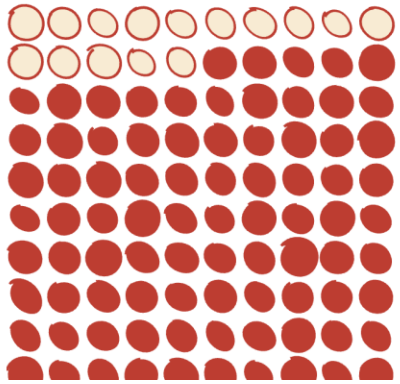
1/3rds of statistics and methodology instructors like myself and statistics students.

Just from recent news:

The New York Times

When They Warn of Rare Disorders, These Prenatal Tests Are Usually Wrong

Some of the tests look for missing snippets of chromosomes. For every 15 times they correctly find a problem ○ ...



Genetic counselors who have dealt with false positives say some doctors may not understand how poorly the tests work. And even when caregivers do correctly interpret the information, patients may still be inclined to believe the confident-sounding results sheets.

When Cloey Canida, 25, got a positive result from Roche's Harmony test in September, the result sheet seemed clear: It said her daughter had a "greater than 99/100" probability of being born with Patau syndrome, a condition that babies often do not survive beyond a week.

“I wish that we would have been informed of the false positive rate before I agreed to the test,” she said. “I was given zero information about that.”

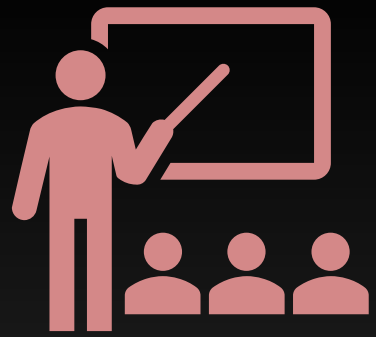
Basic knowledge of statistics and probability can help you in everyday life as well

You read a story in the newspaper about a certain group of people (with certain attributes: religion, caste, etc) and how prone they are to violence based on some numbers. Knowledge of basic statistical and probabilistic pitfalls can help you evaluate these claims better.

You read a story about COVID-19 and Omicron, and the probability of getting seriously sick based on hospital admission numbers. You know about confounding variables, you know about base rates, etc – evaluate the claims calmly and logically.

False positives, false negatives, etc in statistics for psychology

We as researchers too conduct tests. Every test has some chance of a false positive, false negative, etc. To make inferences from the data, we need to compute numbers. There are many statistical tools available to do this.



This will be a major topic of this course.

15 min homework

PLEASE READ CHAPTER 1: HTTPS://LEARNINGSTATISTICSWITHR.COM/BOOK/WHY-DO-WE-LEARN-STATISTICS.HTML

> *To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*
>
> – Sir Ronald Fisher[6]

Research Design

BRSM

Measurement in the behavioral sciences

Measurement

Define the property you want to study

Find a way to *detect* that property

Examples

Aggression (*operational definition*? *Measure*?)

Intelligence (*operational definition*? *Measure*?)

Productivity in the office (*operational definition*? *Measure*?)

Age (how do you measure this? Depends.. Developmental psych? Consumer research?)

Operational definition

- A working definition of what a researcher is measuring

In this task, "on target" is +/- 10% of the goal distance.

Golf Slides from Lassiter Speller (OSU).

Terminology

- ***A theoretical construct***. This is the thing that you're trying to take a measurement of, like "age", "gender" or an "opinion". A theoretical construct can't be directly observed, and often they're actually a bit vague.
- ***A measure***. The measure refers to the method or the tool that you use to make your observations. A question in a survey, a behavioural observation or a brain scan could all count as a measure.
- ***An operationalisation***. The term "operationalisation" refers to the logical connection between the measure and the theoretical construct, or to the process by which we try to derive a measure from a theoretical construct.
- ***A variable***. Finally, a new term. A variable is what we end up with when we apply our measure to something in the world. That is, variables are the actual "data" that we end up with in our data sets.

Variable types: scales of measurement

- Nominal
- Ordinal
- Interval
- Ratio

Nominal scale

Categorical

e.g. Eye color, sex

Does not make sense to say one is greater than the other

Also does not make sense to average them (e.g. average eye color?!)

Nominal scale

Transportation	Number of people
(1) Train	12
(2) Bus	30
(3) Car	48
(4) Bicycle	10

Ordinal Scale

- Slightly more structured than nominal: now you can order the variables in some sensible way

Here's an more psychologically interesting example. Suppose I'm interested in people's attitudes to climate change, and I ask them to pick one of these four statements that most closely matches their beliefs:

> 1. Temperatures are rising, because of human activity
> 2. Temperatures are rising, but we don't know why
> 3. Temperatures are rising, but not because of humans
> 4. Temperatures are not rising

Natural ordering of the options

- Relative to some ground truth (e.g. scientific evidence), statement 1>2>3>4

So, let's suppose I asked 100 people these questions, and got the following answers:

	Number
(1) Temperatures are rising, because of human activity	51
(2) Temperatures are rising, but we don't know why	20
(3) Temperatures are rising, but not because of humans	10
(4) Temperatures are not rising	19

- How do we group these responses for analysis?
- If it is an ordinal scale measurement, there are some sensible ways to do this and others that don't make sense
- Again, the average does not make sense: the average endorsed statement here is 1.97

Interval scale

Both interval and ratio scales: numerical value now can be interpreted directly

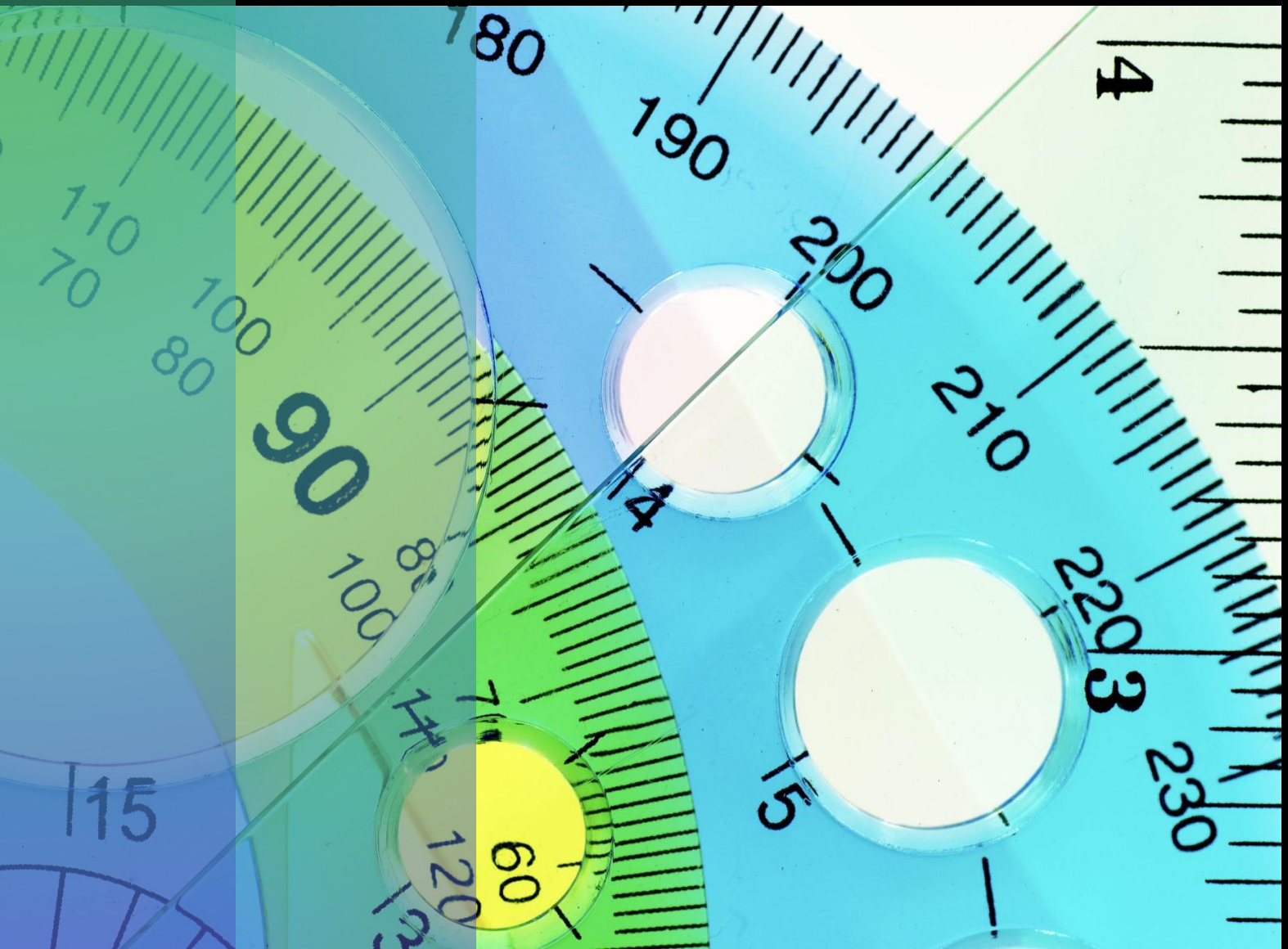
Interval: differences between numbers make sense, but there is no natural "zero" on this scale

Addition and subtraction make sense, but not multiplication or division

e.g. temperature. Difference between 20 and 17 deg celsius = 3 degrees. The same as the difference between 30 and 27 degrees. There is no natural zero, just an arbitrary point (freezing point) chosen as a reference.

Psych e.g. student attitudes as a function of time elapsed since joining date – the year of entry is an interval scale measurement

Averages, medians, etc make sense: the average temperature for the month



Ratio scale

Zero means zero
Can divide
e.g. Reaction times (e.g. I'm twice as fast as you)

Continuous vs discrete variables

- A ***continuous variable*** is one in which, for any two values that you can think of, it's always logically possible to have another value in between.
- A ***discrete variable*** is, in effect, a variable that isn't continuous. For a discrete variable, it's sometimes the case that there's nothing in the middle.

Examples? -- what type of scale? Discrete or continuous?

- RTs?
- Year in which participants were born?
- Temperature?
- Your mode of transport to work?
- Place attained in a race?

- RTs – ratio scale and continuous
- Year in which participants were born – interval scale and discrete
- Temperature – interval scale and continuous
- Your mode of transport to work? - nominal and discrete
- Place attained in a race? - ordinal and discrete

Continuous vs discrete variables

Table 2.1: The relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible.

	continuous	discrete
nominal		✓
ordinal		✓
interval	✓	✓
ratio	✓	✓

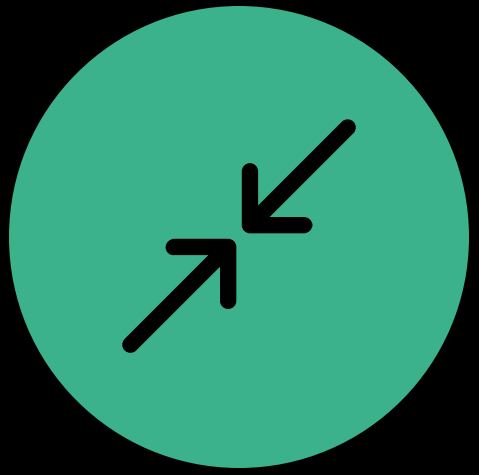
Real world variables may not always adhere to these classifications

- Likert scale
- Choose from the following options. You feel happy today:
- What scale is this?

1. Strongly disagree
2. Disagree
3. Neutral
4. Agree
5. Strongly agree

- Nominal? (hint: is there a natural ordering? If so, it can't be nominal)
- Ratio? (hint: is there a natural "zero"?)
- Ordinal or interval. Which one is it?
- Can we prove that everybody treats the difference between 1. and 2. the same as the difference between 4. and 5.?
- In practice, most people treat the likert scale as an interval scale since many participants treat the entire scale seriously (but this is very much dependent on the task and context).

Is the measurement any good?



RELIABILITY: HOW REPEATABLE?

VALIDITY: HOW ACCURATE IS IT IN RELATION TO WHAT YOU WANT TO MEASURE?

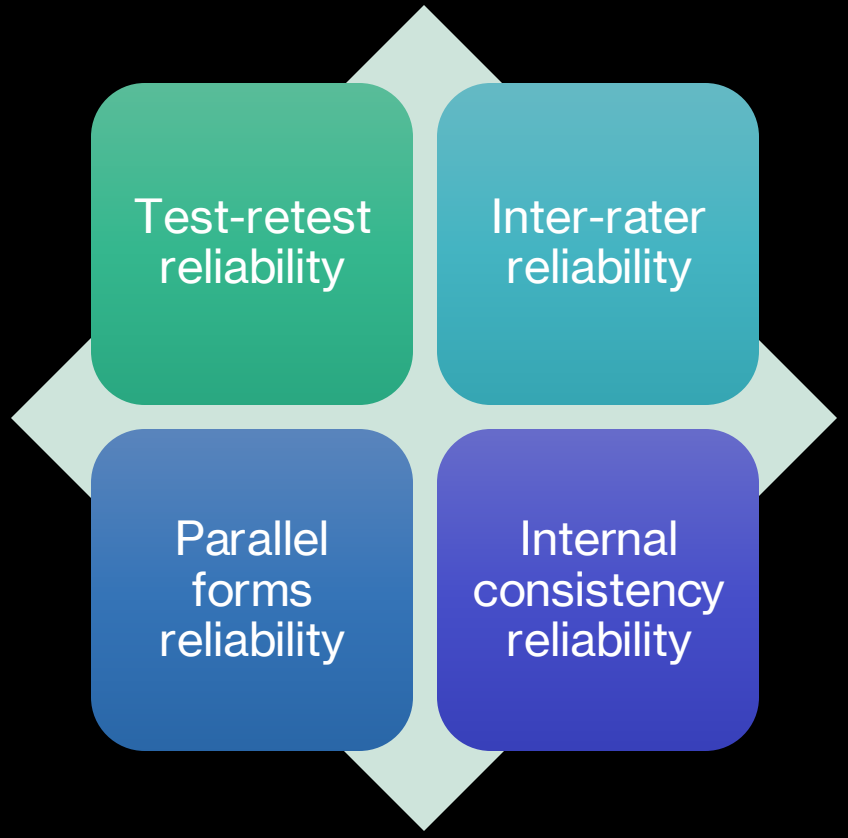
Reliability

Will we produce the same thing repeatedly?

E.g.

- Weighing machine: day 1 = 90 kgs, day 2 = 110 kgs unreliable!
- Psychology example:
  - Want to measure depression
  - Operational definition: Number of times you hang out with family and friends (lower = depression)
  - Measurement in July vs Nov
  - Reliable?

Different ways to measure reliability



Test-retest reliability

CONSISTENCY OVER TIME

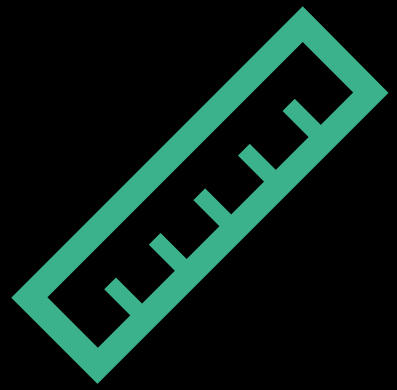
DO WE GET THE SAME RESULTS WHEN WE TEST AT ANOTHER TIME?

Inter-rater reliability

CONSISTENCY ACROSS PEOPLE

IF SOMEONE ELSE DOES THE MEASUREMENT, WILL WE GET THE SAME RESULT?

Parallel forms reliability



Consistency across theoretically-equivalent measurements

If I use a different weighing scale, do I get the same weight measurement?

Internal consistency reliability

CONSISTENCY ACROSS DIFFERENT PARTS WITH THE SAME FUNCTION

IF QUESTIONS ON FLUID INTELLIGENCE SPREAD ACROSS THE IQ TEST ALL GIVE SIMILAR ESTIMATES OF MY INTELLIGENCE, THE TEST HAS INTERNAL CONSISTENCY

Think about the evaluation components of this course

How good is the internal consistency of the evaluations? (problem sets + quizzes + projects)

How about within quizzes or any given component?

Experimental variables

Independent variable: (IV) the variable that
is manipulated Examples: amount of light, exposure to a loud noise, drug

Dependent variable: (DV) the variable that is measured to see if the independent variable had an effect. Examples: Plant growth, change in heart rate, anxiety scores

Table 2.2: The terminology used to distinguish between different roles that a variable can play when analysing a data set. Note that this book will tend to avoid the classical terminology in favour of the newer names.

role of the variable	classical name	modern name
to be explained	dependent variable (DV)	outcome
to do the explaining	independent variable (IV)	predictor

Modern terminology

- We're using the predictors to make guesses about the outcome

Experimental Research

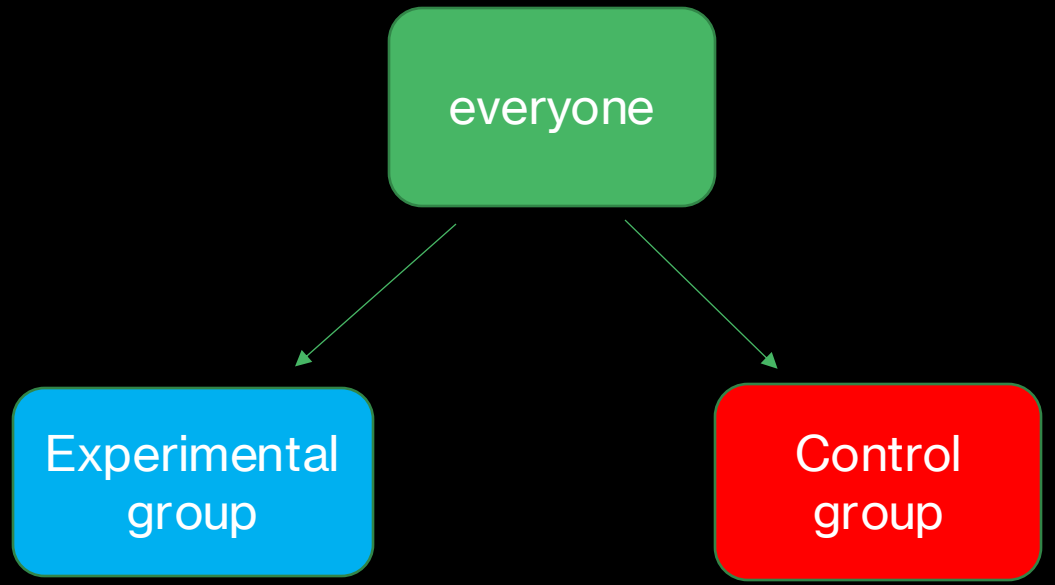
- The experimenter controls everything
- Manipulates the predictors and sees how the outcome changes

Practical issues

We cannot possibly think of ALL the predictors that can influence the outcome

How do we solve this issue?

Randomization



Predictor (IV)

Absence of the predictor variable (or a different level of the variable)

Then compare the outcomes in the two groups

Discussion: Effect of playing violent video games on aggression

Examine the database of players provided by a gaming company

Get criminal records

Test for a difference in the records between game players and non-players

Any problems with this?

The role of confounds

- Perhaps the people playing violent video games as young children are also ones without proper parental support
- In the previous study, there was no consideration of this potential confound

The ideal experiment?

- Take a random sample from the population
- Randomly assign them into violent game-play vs peaceful game-play groups
- Monitor their lives for a few decades
- Get criminal records
- This is not exactly feasible though

So what do we do then?

Use statistics!

Incorporate confounds as covariates in your statistical models!

I.e., we still want to understand how the outcome (aggression) varies as the predictor value is changed (violent game play) but now we will first take into account what amount of the outcome is affected by the confounding variables

Validity

Confounds affect the validity of your study

Many more factors that affect the validity of a study

Important to examine those before we delve into statistical methods

Validity

- Internal validity
- External validity
- Construct validity
- Face validity
- Ecological validity

Internal validity

- The ability to draw cause and effect inferences from the data
- The effect of covid (Delta) on IQ.
- Recruit govt hospital patients. Compare with healthy controls who responded to online ads for your study.
- Internal validity?

External validity

- Generalizability of your findings
- Govt hospital COVID patients and their cognitive issues: generalizable to the rest of the population?
- A basic perception study with college undergrads?
- A study on attitudes towards psychotherapy based on CogSci students at IIITH?

Construct validity

- Are you really measuring what you want to be measuring?
- I want to understand the prevalence of depression in the student population
- I post a tweet and ask people with depression to like the tweet and others to retweet. The proportion fo students who liked the tweet = my answer. How good is my construct validity?

Face validity

DOES YOUR TEST "APPEAR" TO BE DOING THE JOB IT SAYS IT WILL DO?

DOESN'T REALLY MATTER FOR SCIENTISTS.

CAN MATTER IF YOU'RE TRYING TO CONVINCE POLICY MAKERS FOR EXAMPLE. THEN THEIR PERCEPTION ABOUT THE TEST WOULD MATTER.

Ecological validity

- Does the experiment closely mimic real-world scenarios?
- Related to external validity in that ecological validity is supposed to help us generalize the findings to real-world scenarios
- Though that is not guaranteed
- e.g. eye-witness studies in the lab lack ecological validity
- e.g. Word memory experiments
- However, insights from word memory experiments may (and do) generalize to more ecologically valid settings

Threats to validity

- Confounds – related to both predictors and outcomes in some systematic way. A threat to internal validity. Why?
- Artifacts – something about the way you did the experiment that gave you the result. A threat to external validity (but probably also internal). Why?

History effects

Something that happens during the study (or preceding) that can influence the results

Hospital stay, patient testing, 3rd day compared to 7th day. Electrode rearrangement surgery on day 5.

Maturational effects

Something that changes naturally over time that can influence your results

One big effect in psych lab experiments: waning attention, fatigue, which increases over the course of the experiment. How do you know that primacy effects are not driven by such maturational effects?

(Repeated) testing effects

Practice effects

Familiarity with the test

Better scores in session 2 compared to session 1

Selection bias

- Refers to anything that makes the groups being compared different in some potentially critical aspect
- Different proportions of males/females in the two groups in a study on aggression
- No more internal validity

Differential attrition

- If you do a long study, or a longitudinal study or any study that requires quite a bit of effort from the participants, this may be relevant.
- People drop out.
- The people dropping out are not random people.

Homogeneous vs heterogeneous attrition

- The rates of attrition can be the same across groups you're comparing – homogeneous attrition
- But they can also be different! - heterogeneous attrition
- Older people for instance may not carry on with a demanding task, and if you have a critical comparison between age groups, this can be a major issue

Non-response bias

- You work for a company
- You send out a survey to 1000 randomly selected email ids from your database
- Only 200 respond
- You say you chose the initial emails at random, so what's the problem?
- Again, the people who choose to respond are NOT random!

Regression to the mean

- When you select data based on an extreme value of some measure, a subsequent measurement will tend to "regress to the mean"
- Good examples in the textbook
- The children of tall people will tend to be taller than average but shorter than the parents but the children of short parents tend to be taller than the parents.
- Early studies suggested that people learn better from negative feedback than positive feedback
- But not really, it was also an artifact of regression to the mean (Kahneman & Tverseky, 1973)

Experimenter Bias

Oskar Pgungst: student at the Psychological Institute at the University of Berlin, through careful experiments, showed that Clever Hans was responding to subtle, involuntary cues from von Osten. Classic early example of experimental design in behavioral Psychology

Demand and reactivity effects

"Hawthorne" effect

The influence of lighting on factory worker productivity

But results were driven by the fact that workers did better when they thought they were being observed

Solution to both experimenter bias and reactivity effects

Double blind studies

Placebo effects

- The expectation of a positive effect even from an inert drug will sometimes make people feel better

Fraud and deception

- This part is important as they are very much related to statistical methods, inappropriate use of methods (sometimes intentionally, in order to deceive)

Data fabrication

See https://retractionwatch.com/

People make up data! Including some very high profile researchers

There are data science sleuths who detect fraud using statistical methods

Study misdesigns

Issues with study design that don't get reported

Results may be artifacts of such misdesign

e.g. surveys that are self-evident, sit back and let reactivity decide your results for you. If reviewers don't see the full surveys, this may not get detected

Data mining and post-hoc hypothesizing

- Data mining: I run 50 different variations of a model. Report only the one that worked.
- If you are honest, your statistical methods would "correct" for the 50 times you touched the data because we want to know that the result obtained is a true one that is not likely to have come about due to mere chance.
- Post-hoc hypothesizing: my initial hypothesis didn't work but as part of the data mining effort above, I found something else and reported that I had actually hypothesized it.
- Huge statistical issue when you do this because many frequentist statistical methods depend on assumptions made about the null hypothesis

Publication Bias

- Journals as well as authors do not publish negative findings
- Distorts the literature which comes to be dominated by small N but "significant" studies
- Partly led to the "replication crisis" in Psychology
- Also limits what you can learn from meta-analyses/reviews.

Summary

1. Be aware of all the different ways in which the data from a study may have issues with reliability/validity
2. Be aware of potential confounds
3. Address the confounds using statistical methods
4. Be aware of dubious practices such as data mining and post-hoc hypothesizing

Advanced topics

Article | Open Access | Published: 12 November 2020

Collider bias undermines our understanding of COVID-19 disease risk and severity

Gareth J. Griffith, Tim T. Morris, Matthew J. Tudball, Annie Herbert, Giulia Mancano, Lindsey Pike, Gemma C. Sharp, Jonathan Sterne, Tom M. Palmer, George Davey Smith, Kate Tilling, Luisa Zuccolo, Neil M. Davies & Gibran Hemani

Nature Communications **11**, Article number: 5749 (2020) | Cite this article

39k Accesses | **159** Citations | **334** Altmetric | Metrics

Abstract

Numerous observational studies have attempted to identify risk factors for infection with SARS-CoV-2 and COVID-19 disease outcomes. Studies have used datasets sampled from patients admitted to hospital, people tested for active infection, or people who volunteered to participate. Here, we highlight the challenge of interpreting observational evidence from such non-representative samples. Collider bias can induce associations between two or more variables which affect the likelihood of an individual being sampled, distorting associations between these variables in the sample. Analysing UK Biobank data, compared

Install R and RStudio

- http://cran.r-project.org/
- RStudio: http://www.RStudio.org/

Probability Distributions

BRSM

The role of assumptions in statistics

Before the match, Fischer had won 3 games, Taimanov had won 2 games, and 1 game was drawn.

We bet on the winner of the next game, after each round.

The limits of logic in everyday life.

What is statistical inference?

- Polling company
- Randomly call 1000 people
- 35% said they'd vote for XYZ party
- The result comes out. The number actually is 26%
- The question is: how surprised (or not) should we be by this result?
- To do this, we need tools for statistical inference
- Each tool makes some assumptions about the data
- We need to understand probabilities and probability distributions first

What is the difference between probability and statistics?

- What is the probability that in two successive coin tosses, you get both tails?
- You have the model of the world here (e.g. it is a fair coin, P(H) = 0.5), but no data and are asked to come up with the probability of a hypothetical event
- Going back to Fischer-Taimanov, after 3 rounds and 3 wins to Fischer, we are to make an inference about what model is correct, given the 3 win data. Is P(Fischer) really 0.5 or is it something else? This is the realm of inferential statistics.

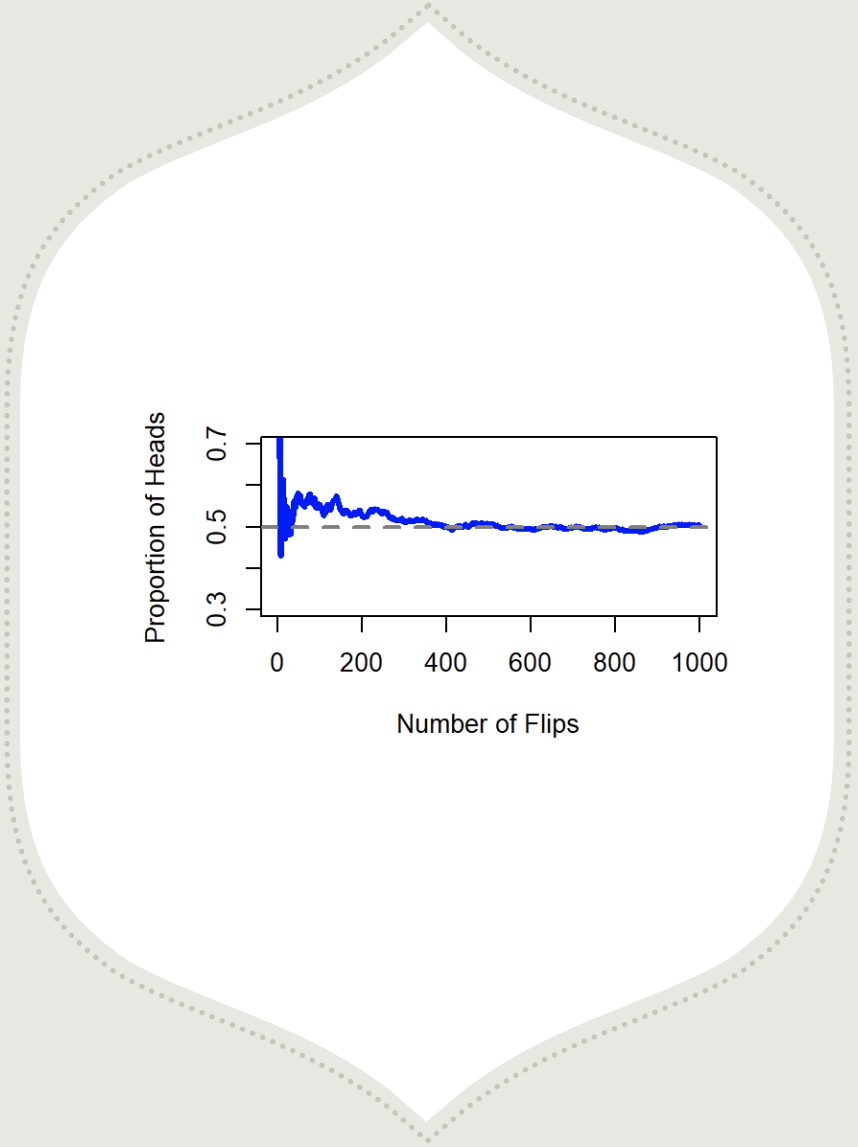
What is a probability?

- Means slightly different things if you are a frequentist statistician vs if you are a Bayesian
- Carlsen has a 70% chance of winning a game against Nepomniachtchi: what does this mean to you?
- If they play a 10 game match, Carlsen is expected to win 7?
- If I bet Rs 100 on Nepomniachtchi, I should get a reward of Rs 233 (700/3) if Nepo wins against your bet of Rs 233 on Carlsen (and if Carlsen wins, you get Rs 100).
- 70% reflects my subjective belief of how much stronger Carlsen is compared to Nepo.

Frequentist probability

FLIP A COIN MANY TIMES AND COUNT THE PROPORTION OF HEADS

- As N -> infinity, the probability converges to the true probability
- Frequentist statistics rely on assumptions about how you sample the data (just like a coin toss), and cares about long-run proportions of a certain result (e.g. heads) in such hypothetical future samples.



Frequentist statistics

- Pros: objective because anyone following the same "sampling plan" will observe a similar proportion over the long run.
- Cons: The equivalent of flipping a coin infinite times to understand a probability can be counterintuitive in practice: "There is 80% chance of rain today." We can intuitively somehow understand what this means.
- The interpretation in frequentist terms: "There is a class of day for which if we observe across N--> infinite days, it rained on 80% of those days".
- This type of conundrum is exactly what you will see drives debates in statistical methods between frequentists and Bayesians.

Bayesian probability

Subjective

Minority view amongst statistical practitioners

Degree of subjective belief assigned to an event

Bayesian probability

- Pros:
  - You can assign probabilities to non-repeatable events
  - You can legitimately interpret the probability as degree of belief (similar probabilities in the frequentist world will have more convoluted interpretations leading to the sorts of pitfalls we discussed/will discuss about p-values, confidence intervals, etc).
- Cons:
  - Not objective
  - Depends on priors (background knowledge), which can be subjective

Independent Events

- Two events A and B are independent if
- $P(AB) = P(A).P(B)$
- $P(A|B) = P(AB)/(P(B)) = P(A)$

Variables and their distributions

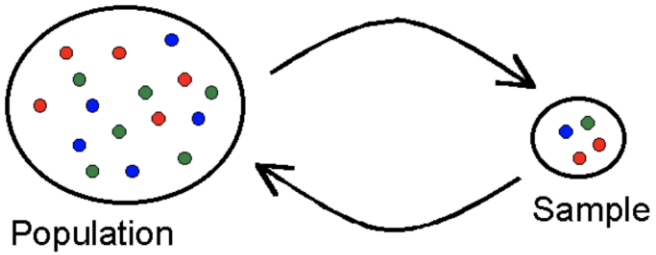
- You will often hear things like "variable x is i.i.d"
- Independently and identically distributed
- Say Yi are dice throws for i=1:n
- The outcome of each different set of (n throws) is a random variable itself
- The outcome of each throw has the same distribution (uniform over 6 possibilities): Y1, Y2, ...,Yn are identically distributed
- Y1 is independent of Y2 and so on.
- Therefore, iid.

A function applied on the sample

- Yi is iid
- Now, if we apply a function on the sample, such as a sum or an average, this is also a random variable
- We can also talk about distributions of such variables!
- This is an important concept in statistics: **sampling distribution of some statistic**

Sample vs population

- Sample (data sample) : e.g. one particular "sample" of N throws or one particular sample of 1000 people in an exit poll in Punjab
- Population: e.g. The universal set of all possible N throw outcomes or all voters in Punjab



Distribution of what? Be clear

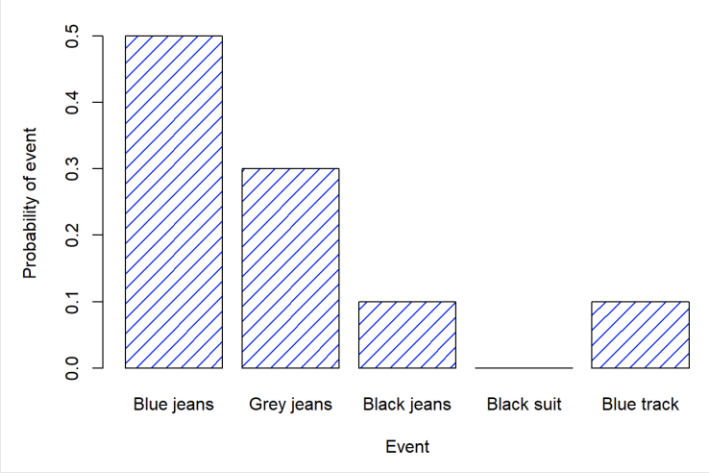
SAMPLING DISTRIBUTION OF A STATISTIC: THE DISTRIBUTION OF A STATISTIC (OR A FUNCTION) APPLIED ON THE SAMPLES

POPULATION: WHAT IS THE DISTRIBUTION OF VOTING PREFERENCES TAKEN FROM THE ENTIRE POPULATION OF PUNJAB?

NEED TO BE CLEAR ABOUT THE DISTINCTIONS

Probability distribution

Which.pants	Blue.jeans	Grey.jeans	Black.jeans	Black.suit	Blue.tracksuit
Label	X_1	X_2	X_3	X_4	X_5
Probability	$P(X_1) = .5$	$P(X_2) = .3$	$P(X_3) = .1$	$P(X_4) = 0$	$P(X_5) = .1$



Probability density function (PDF)

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

Defined for continuous random variables

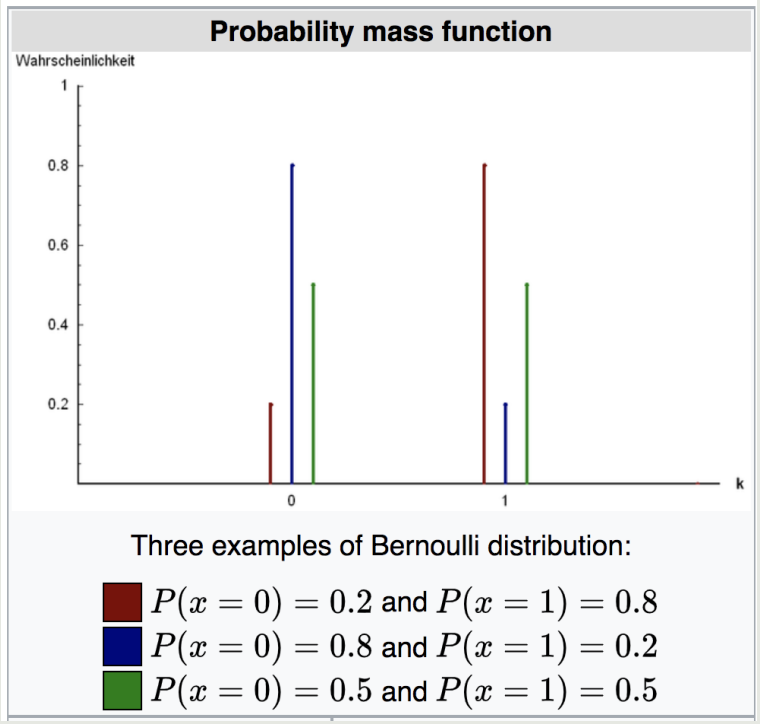
The probability that x = an exact value = 0 for continuous variables because a = b in this integral

Cumulative Distribution Function (CDF)

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

Discrete variables: Bernoulli Distribution

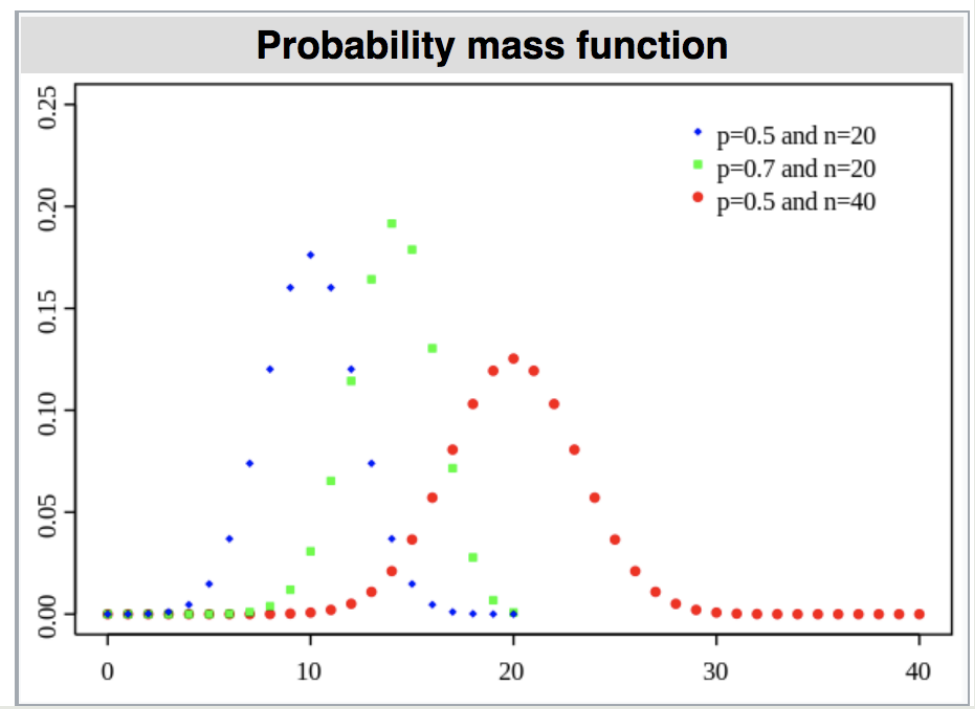
- The Bernoulli distribution is the discrete probability distribution of a random variable which takes a binary, boolean output: 1 with probability p, and 0 with probability (1-p).



Wikipedia

Binomial distribution

- If there is a series of n i.i.d Bernoulli trials (all trials have a success probability of p), then the sum of outcomes is distributed as Binom(n,p)



Wikipedia

Notation

$$X \sim \text{Binomial}(\theta, N)$$

Working with distributions in R

Table 9.3: The naming system for R probability distribution functions. Every probability distribution implemented in R is actually associated with four separate functions, and there is a pretty standardised way for naming these functions.

What.it.does	Prefix	Normal.distribution	Binomial.distribution
probability (density) of	d	dnorm()	dbinom()
cumulative probability of	p	dnorm() pnorm()	pbinom()
generate random number from	r	rnorm()	rbinom()
q qnorm() qbinom()	q	qnorm()	qbinom(

What is the probability of observing 6 heads in 10 coin tosses given an unfair coin?

- P = 0.7
- **dbinom(** x = 6, size = 10, prob = 0.7)
- 0.2001209

R distributions

The d form we've already seen: you specify a particular outcome x, and the output is the probability of obtaining exactly that outcome. (the "d" is short for ***density***, but ignore that for now).

The p form calculates the ***cumulative probability***. You specify a particular value q, and it tells you the probability of obtaining an outcome *smaller than or equal to* q.

The q form calculates the ***quantiles*** of the distribution. You specify a probability value p, and gives you the corresponding percentile. That is, the value of the variable for which there's a probability p of obtaining an outcome lower than that value.

The r form is a ***random number generator***: specifically, it generates n random outcomes from the distribution

10 coin tosses

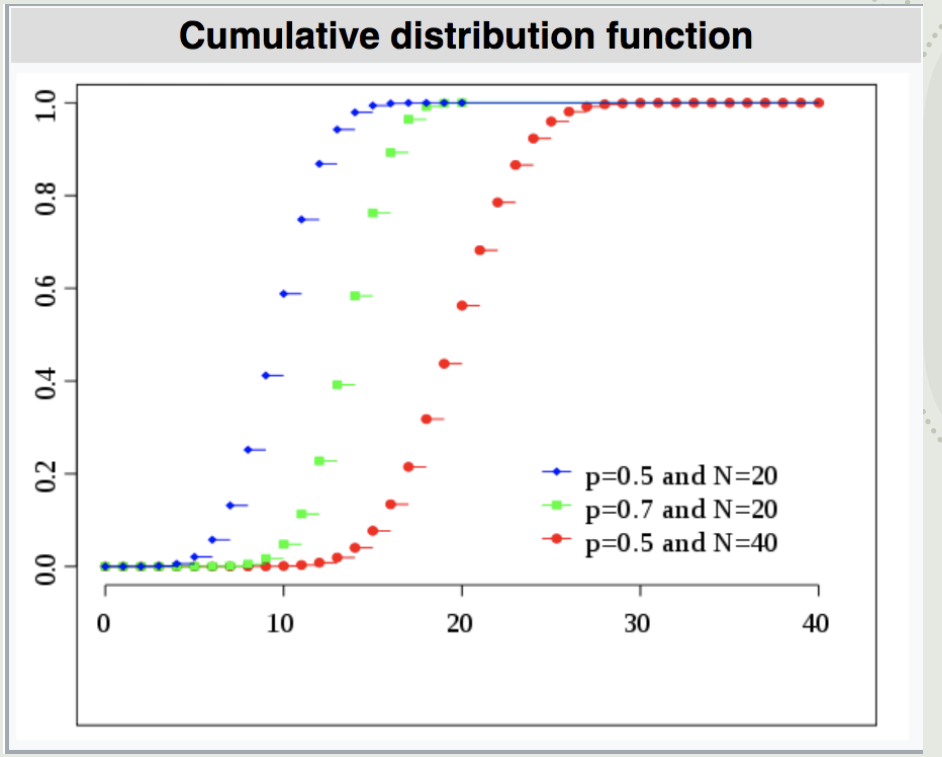
- Probability that I get <= 4 heads?
- P(1) + P(2) + P(3) + P(4) = **dbinom**(x = 1, size = 10, prob = 0.7) + **dbinom**(x = 2, size = 10, prob = 0.7) + **dbinom**(x = 3, size = 10, prob = 0.7) + **dbinom**(x = 4, size = 10, prob = 0.7)
- 0.04734308

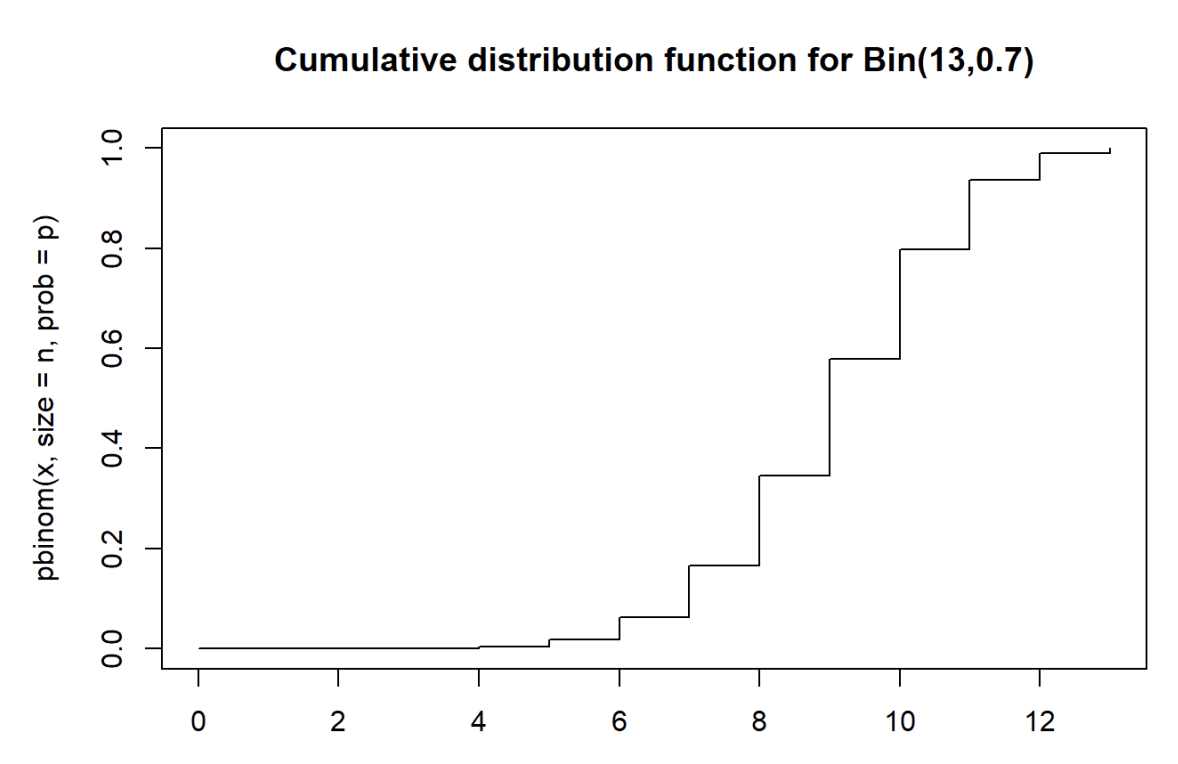
- Easier way: **pbinom**(q= 4, size = 10, prob = 0.7)
- 0.04734899 (4 is the 4.7 th percentile of the Binomial data or 4.7% of the values fall under 4)

- qbinom(p = 0.04, size = 10, prob = 0.7)
- 4 (the 4 th percentile of the data is 4)
- Wait, how can the 4th percentile also be 4??
- The Binomial distribution here doesn't really have a 4th percentile.

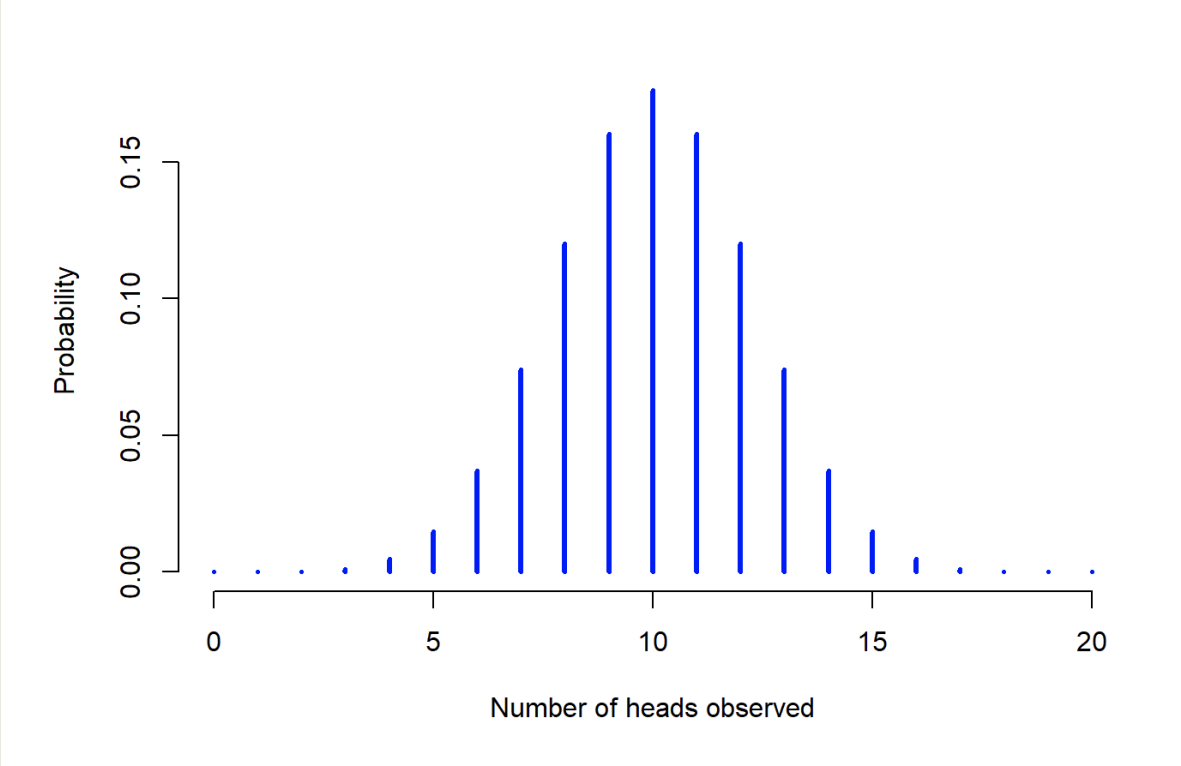
Warning: discrete variables and cumulative distribution functions

- Supported only on countable numbers
- So only some percentiles on the Y axis ->
- If you provide it any other percentile, the R function will round upwards.
- Not a problem for continuous distributions

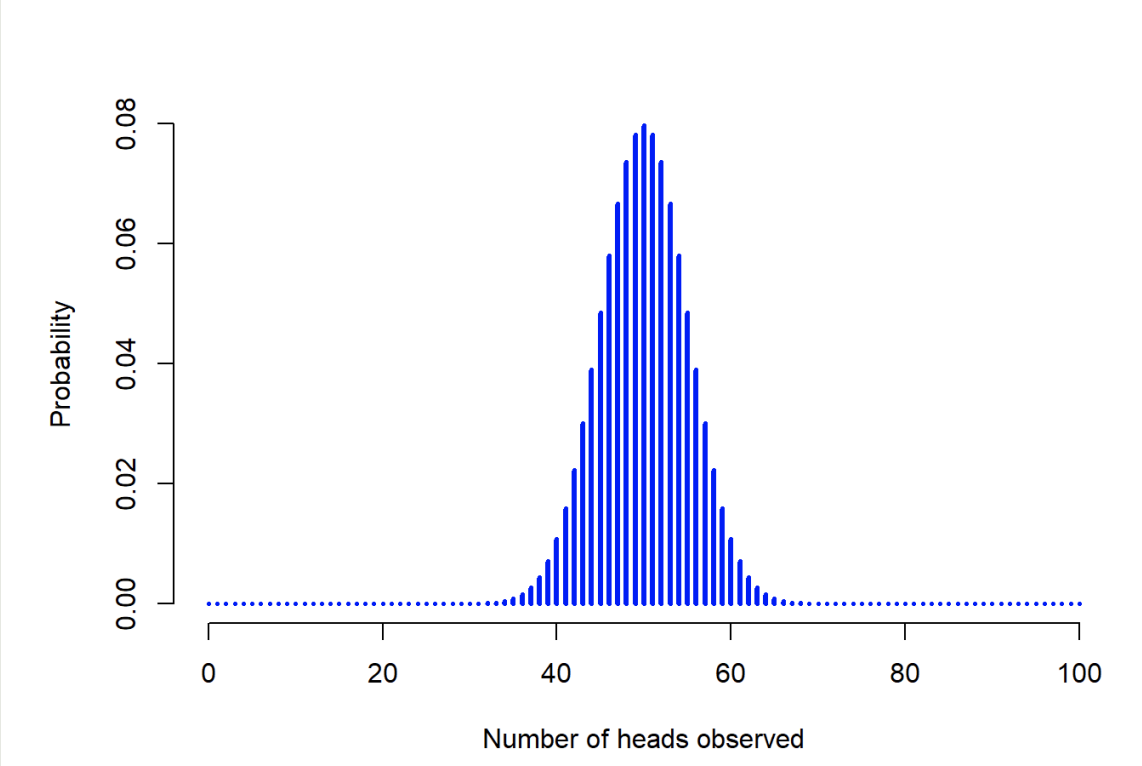



Cumulative distribution function for Bin(13,0.7)
pbinom(x, size = n, prob = p)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12

Flip a fair coin 20 times



Flip a fair coin 100 times



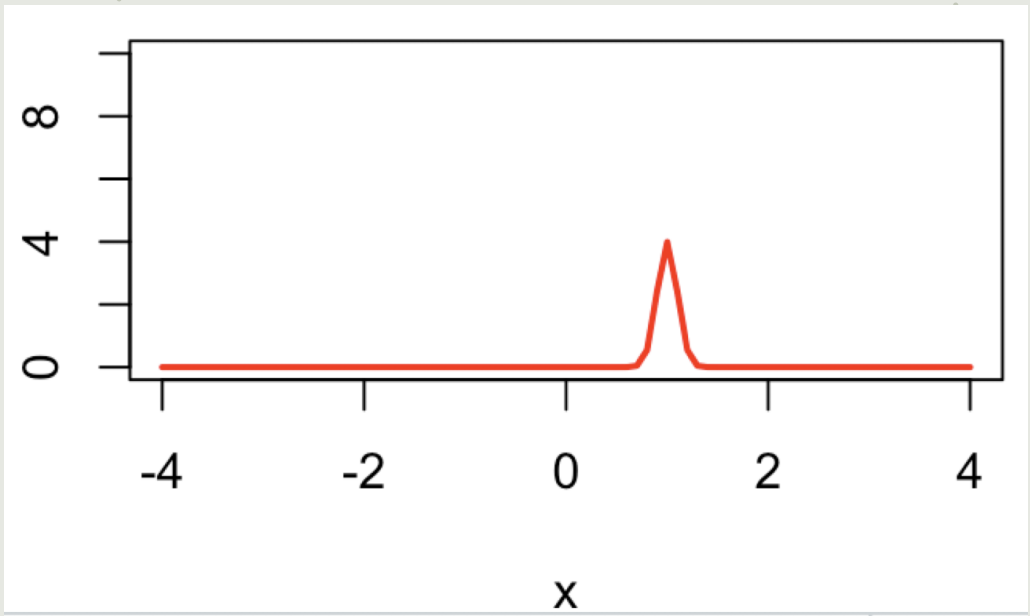
Normal Distribution

$$X \sim \text{Normal}(\mu, \sigma)$$

Normal

$$p(X|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$$

```
plot(x, dnorm(x, mean = 1, sd = 0.1), type = "l",
     ylim = c(0, 10), ylab = "", lwd = 2, col = "red")
```

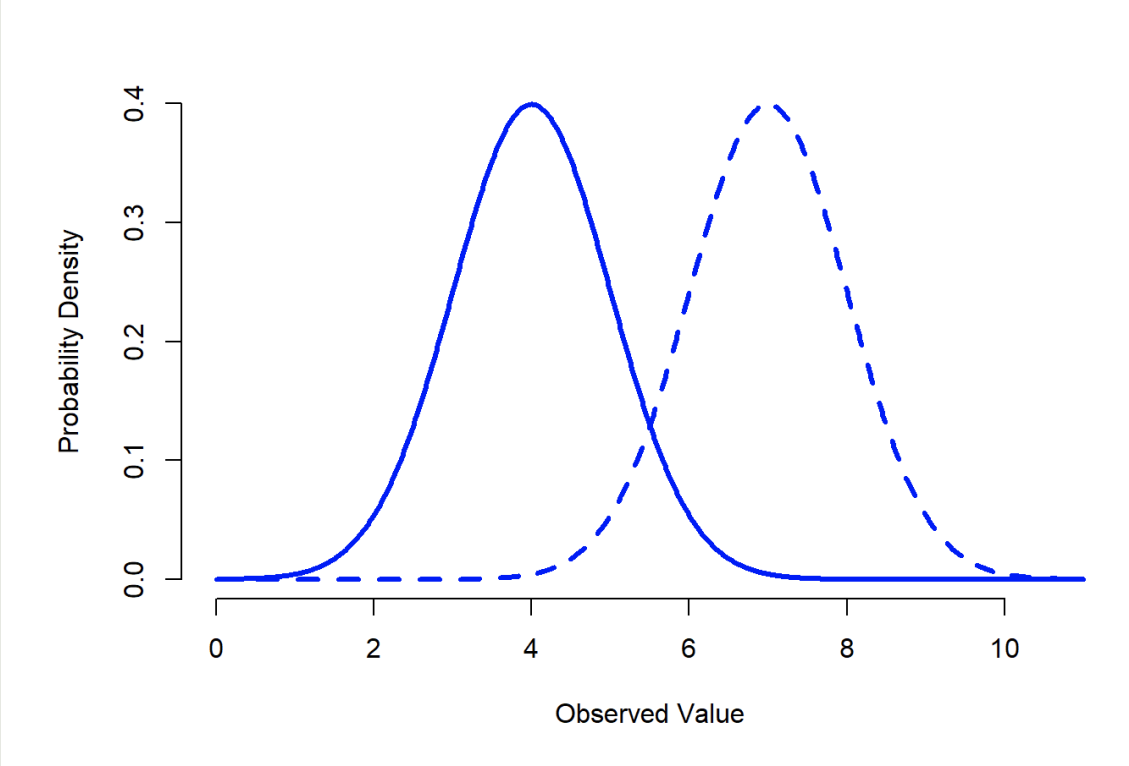


Normal PDF

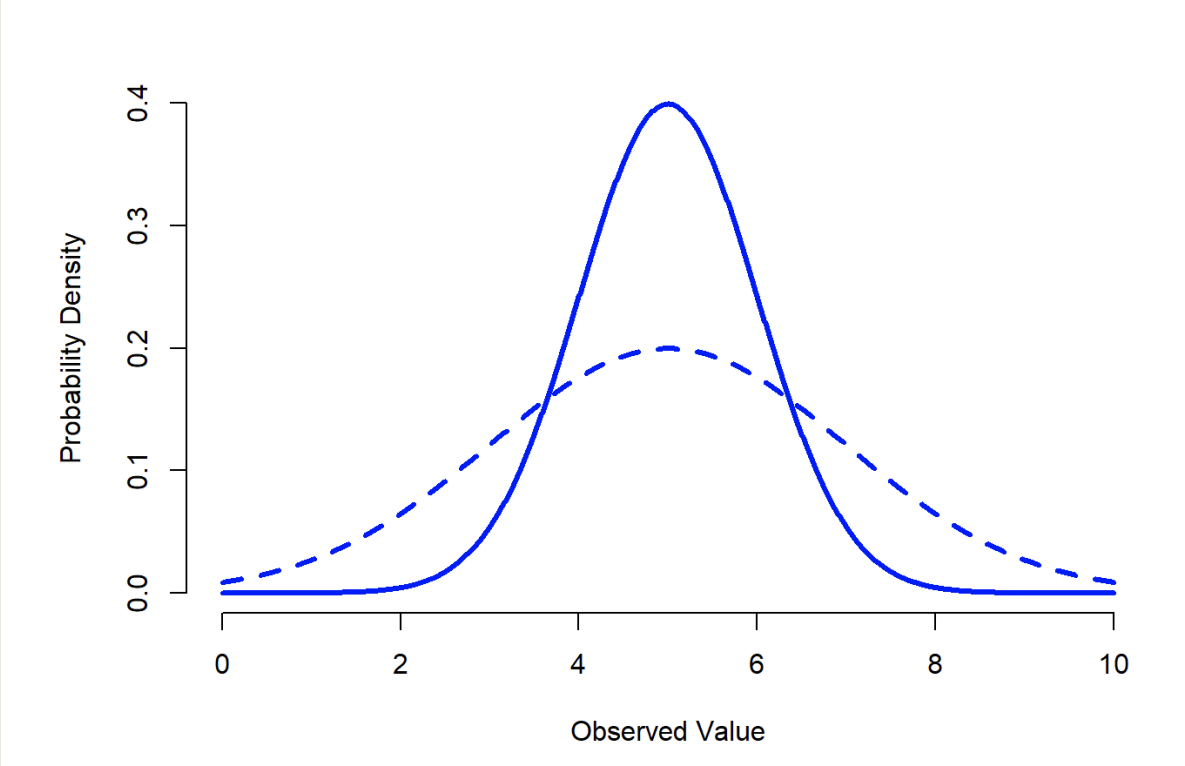
Q: What is the probability that x = 1?

```
> dnorm( x = 1, mean = 1, sd = 0.1 )
[1] 3.989423
```

Different means, same standard deviation ("width")



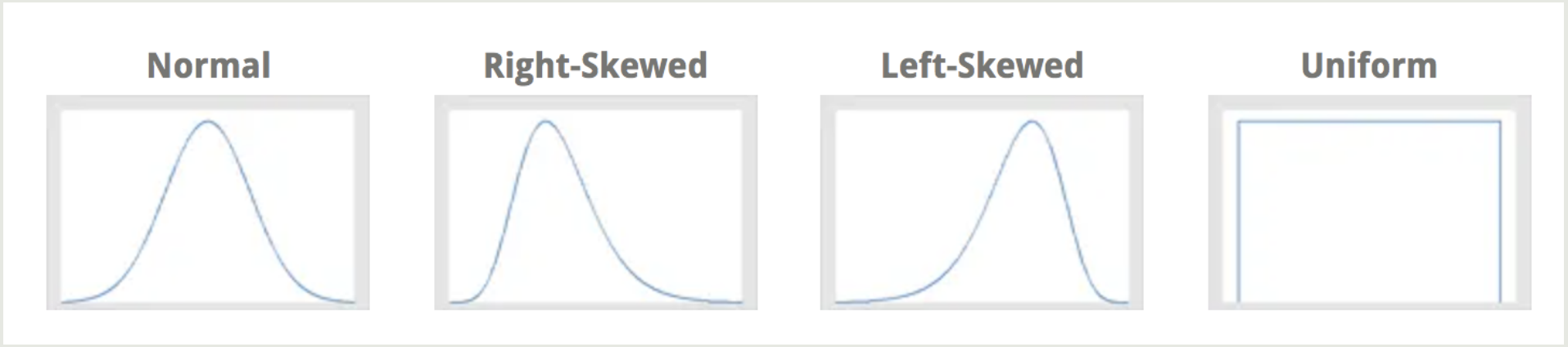
Same mean, different widths



Central Limit Theorem

- The central limit theorem states that, given a sufficiently large sample size, the sampling distribution of the mean for a variable will approximate a normal distribution regardless of that variable's distribution in the population.

Applies to almost all probability distributions of the population



The above is the distribution of the variable in the population!
Now you draw a random sample of size n from this.

The only requirement: the population distribution must have finite variance

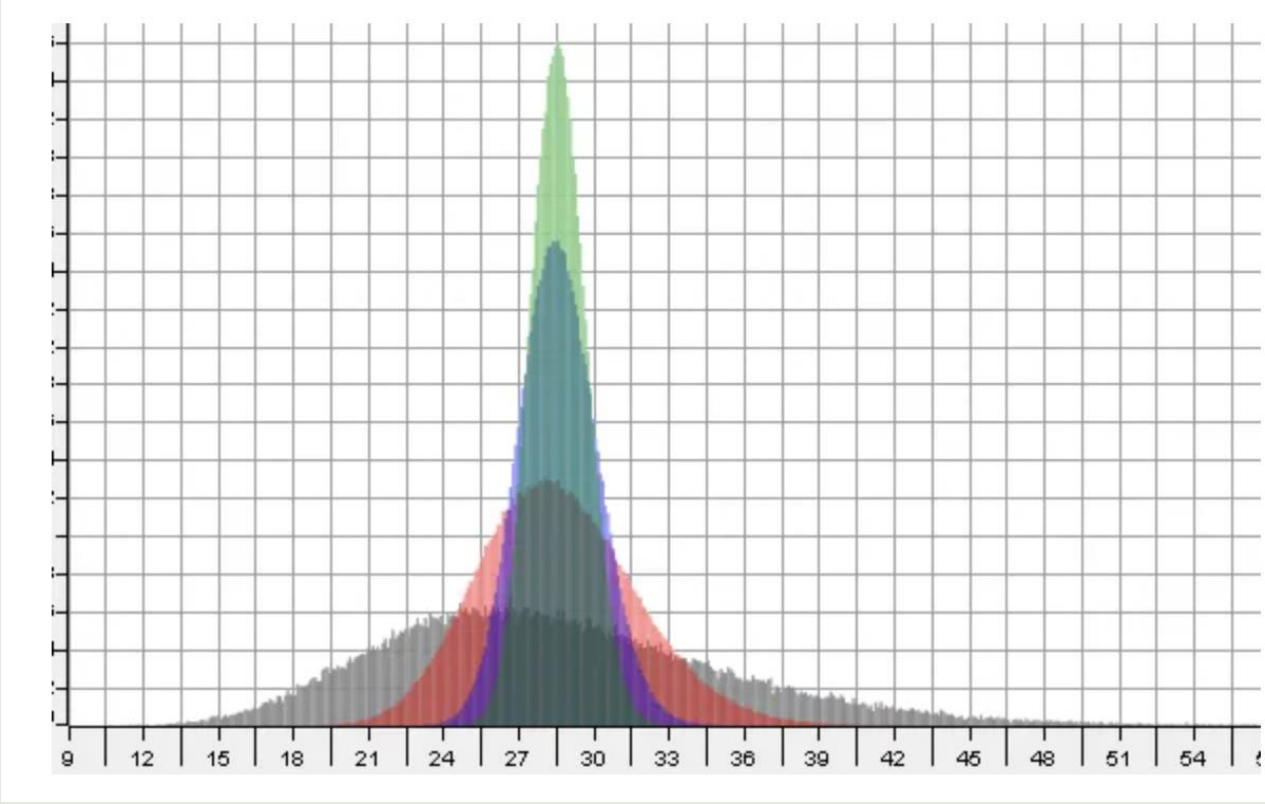
Sampling distribution of...

- the mean, is what CLT deals with
- For each sample, take the mean. Accumulate across say 1000 random draws
- Plot the distribution of these sample means = sampling distribution of the mean

Sample size

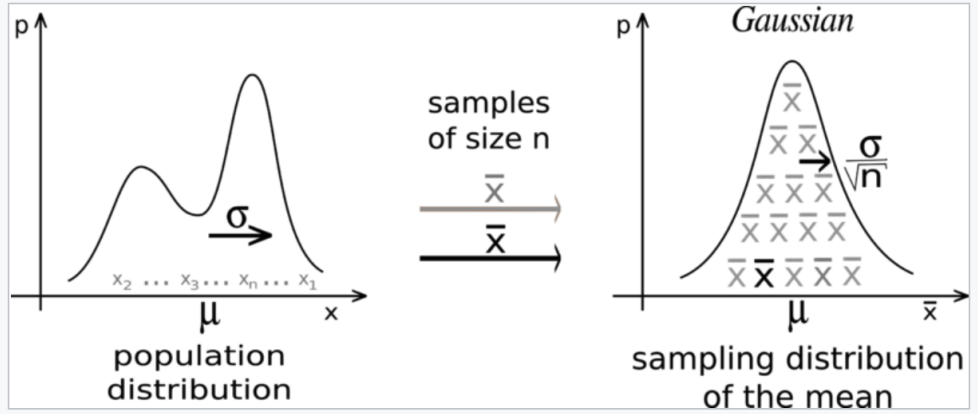
- For CLT to work, we need a sufficient sample size when we randomly draw samples **with replacement** from the population. The exact number will depend on the population distribution. Skewed distributions tend to need higher n.
- The sample mean will be equal to the population mean

Grey = population
Red = sample n = 5
Blue = sample n = 10
Green = sample n = 20



Lindeberg–Lévy CLT. Suppose $\{X_1, \ldots, X_n\}$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{n}(\bar{X}_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$:[4]

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$



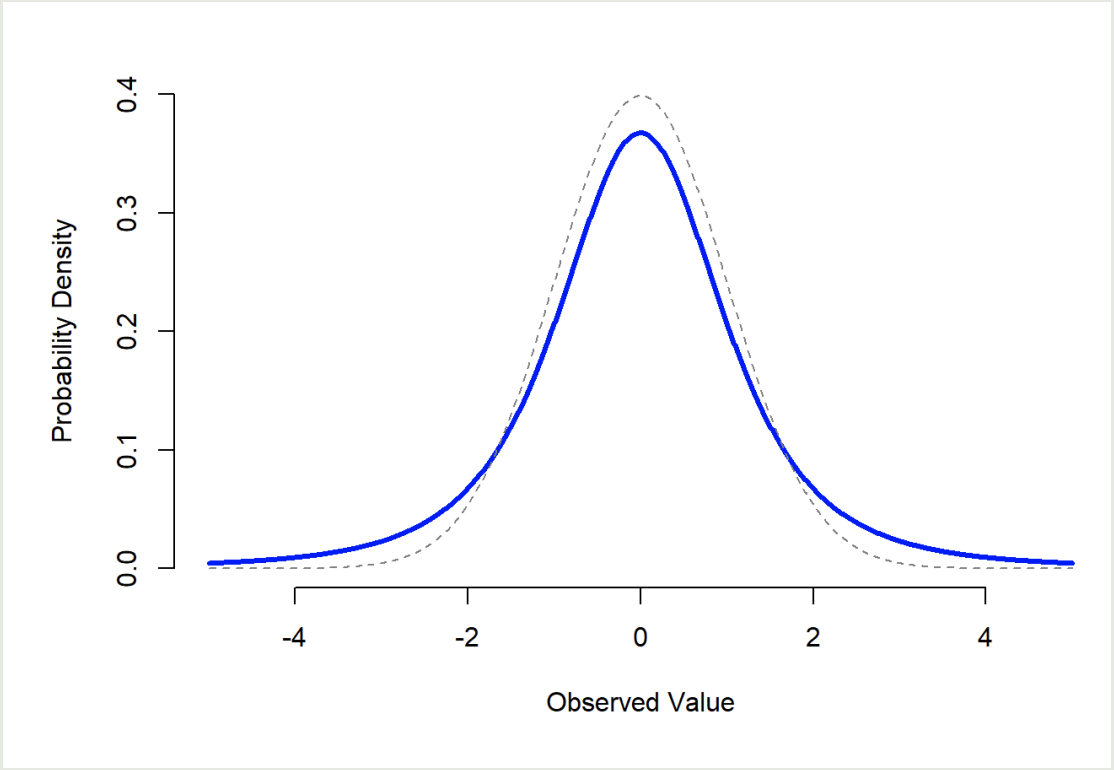
Whatever the form of the population distribution, the sampling distribution tends to a Gaussian, and its dispersion is given by the central limit theorem.[3]

See Wikipedia for an extended introduction to the various forms of the central limit theorem

Why is the central limit theorem important?

- When we test hypotheses about the means of samples (e.g. did healthy adults have a better average performance on my memory task than older adults with MCI?), the tests are often based on the assumption of normality of sampling distributions of the mean.
- CLT says that even if you violate normality assumptions of the variable in the population, as long as you have a sufficiently large sample size, your statistical methods will often be robust to violations of the normality assumptions.

Other distributions: t-distribution



Heavy-tailed
Arises in smaller n situations and when you don't know the population s.d.
As n--> inf, t-distribution begins to look more like a Normal.

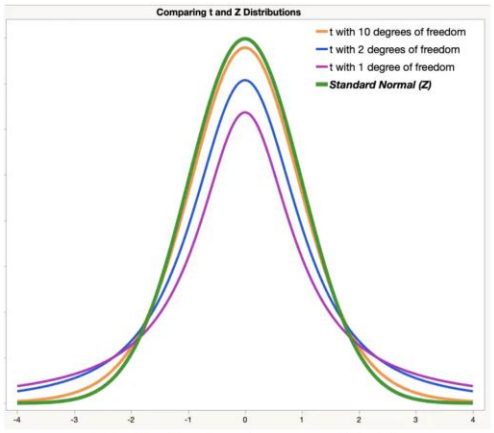


Figure 1: Three t-distributions and a standard normal (z-) distribution.

T-distributions and k

Degrees of freedom, k, is related to sample size

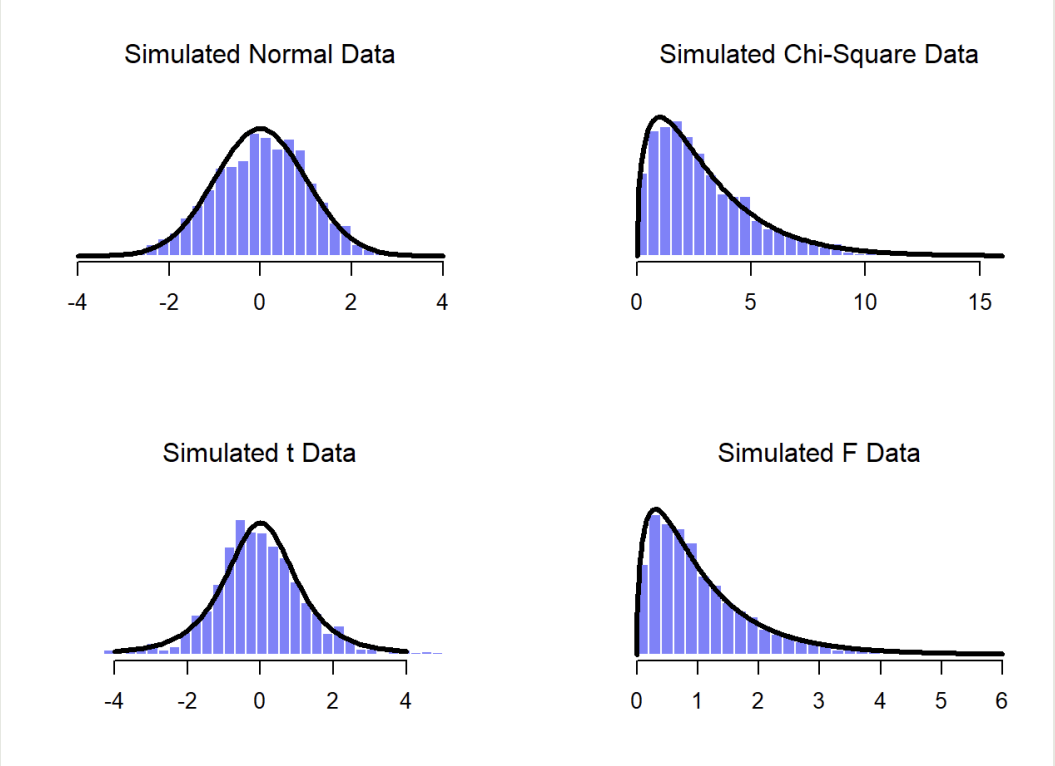
You can appreciate that as k increases, the shape looks more like a Normal (or the tail gets less heavy).

The use of t-distributions later

Suppose $x_i \sim \mathrm{N}(\mu, \sigma^2)$ and we want to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

Assuming we do not know sigma, we will construct a statistic which is where we will encounter the t-distribution to use to construct confidence intervals and p-values to test the above hypothesis

Other distributions



Sum of squares of normally distributed variables: Chi-square

Comparing chi-square distributions: F distributions

Chi-square

- All these other distributions we talk about now are related to the Normal
- chi-square distribution with k degrees of freedom is what you get when you take k normally-distributed variables (with mean 0 and standard deviation 1), square them, and add them up.

```
normal.a <- rnorm( n=1000, mean=0, sd=1 )

normal.b <- rnorm( n=1000 )  # another set of normally distributed data

normal.c <- rnorm( n=1000 )  # and another!

chi.sq.3 <- (normal.a)^2 + (normal.b)^2 + (normal.c)^2
```

R exercises

Sampling and estimating unknown quantities from samples

Making assumptions

- About the data
- Sampling theory: will help us specify the assumptions upon which our statistical methods rely

Inferences about what and based on what?

- Inferences about the population
- Based on the sample

Population: cogsci questions

- All of the undergraduate students at IIITH?
- Undergraduate students in general, anywhere in the world?
- Indians currently living?
- Indians of similar ages to my sample?
- Anyone currently alive?
- Any human being, past, present or future?
- Any biological organism with a sufficient degree of intelligence operating in a terrestrial environment?
- Any intelligent being?

What is the population?

Not always clear

This is probably the first assumption you will make: "My study will reveal phenomenon X as it pertains to the population of ____"

Sample

- Different sampling schemes: how you gather a data sample from the population

Simple random sampling without replacement

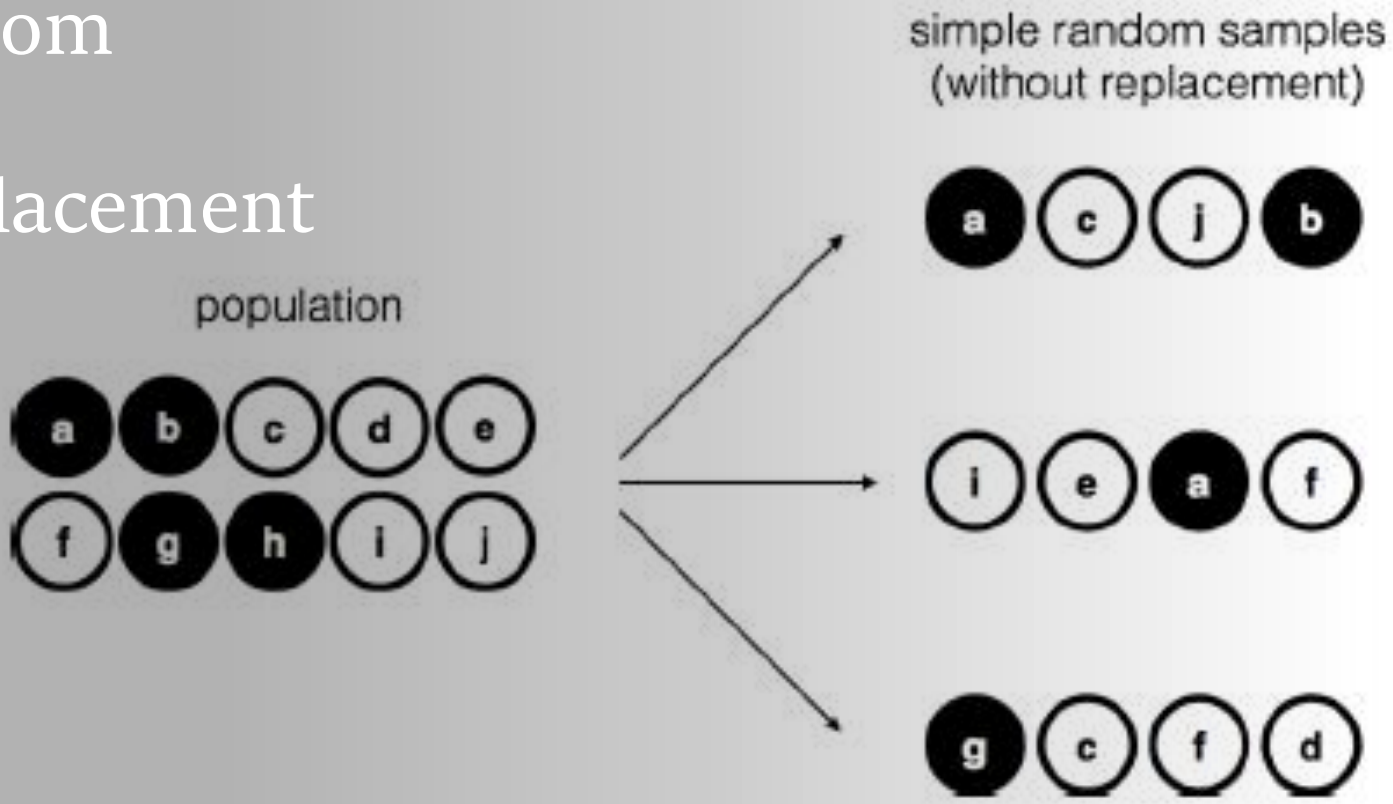


Figure 10.1: Simple random sampling without replacement from a finite population

Biased sampling without replacement

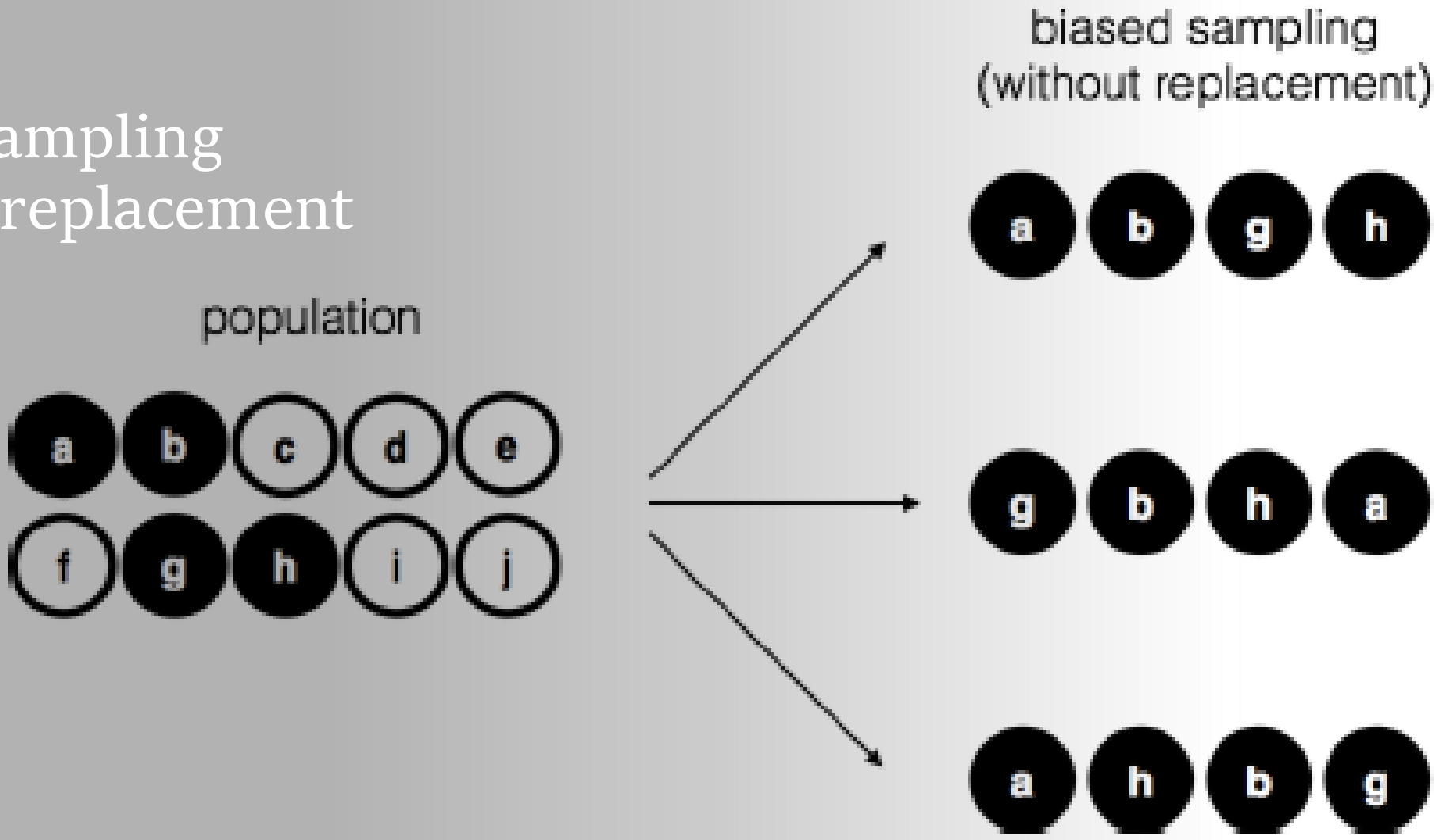
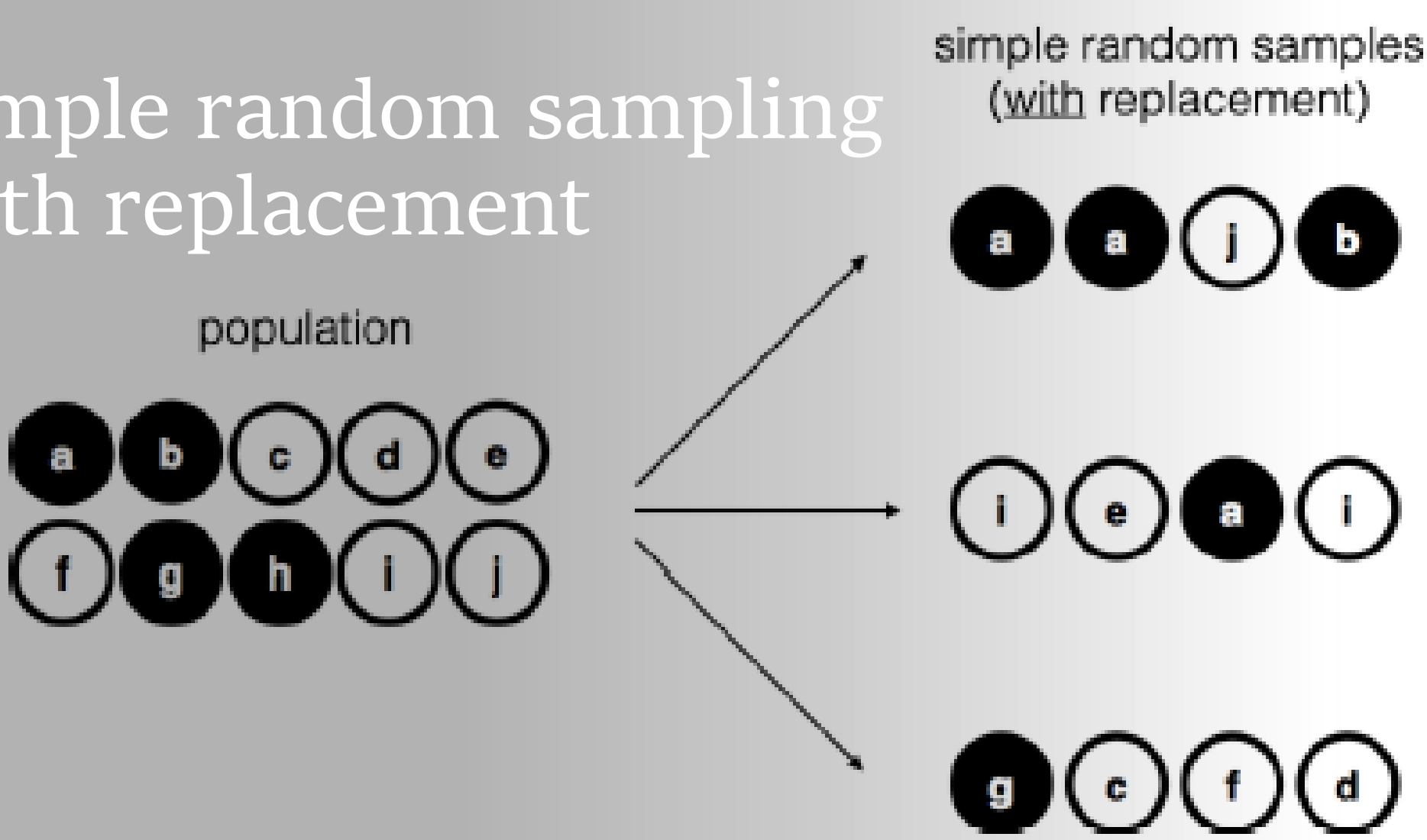


Figure 10.2: Biased sampling without replacement from a finite population

Simple random sampling with replacement



re 10.3: Simple random sampling *with* replacement from a finite population

Real world behavioral experiments

- Samples with or without replacement?
- Most statistical methods assume sampling with replacement
- For a large enough number of participants, the difference does not matter too much
- Biased vs unbiased (simple random) sampling needs attention though

Stratified Sampling

- A natural strata structure (e.g. schizophrenia or Alzheimer's study)
- Simple random sampling within each strata: why?
- If you attempt to randomly sample from the whole population, you will get skewed numbers across the groups of interest.
- For example: Nielson et al., 2015, *PNAS* – all 9 participants in our fMRI study were female as we recruited by placing ads around our psychology building.
- Solution: oversample from the rare strata to equate the numbers

Convenience Sampling

- A convenient sample, not random usually
- e.g. recruit from the undergrad population of iiith
- Not always a problem but depends on the study question and goals
- Most psychology studies involve convenience sampling, which is why people have realized the importance of at least occasionally doing large N replication studies, using more ecologically valid frameworks to test psychology theories that were developed in the lab.

Snowball sampling

- Typically to be used when you want to recruit hard-to-locate participant groups
- e.g. a study on trans health, you do not have many personal contacts, and recruiting from the whole population might be too expensive if you have to discard the majority of the data
- So you get the few contacts you have, ask them to provide other contacts, and so on = snowball sampling
- Fraught with issues: privacy, ethical, highly non-random samples in ways that are hard to control
- However, this is often the only way you can get a sufficient number of participants for such studies
- Snowball sampling is a type of convenience sampling

What to do when you don't have a random sample

If you know exactly how you sampled and what bias you introduced, there are advanced statistical methods to correct for bias (e.g. in stratified sampling).

Otherwise, if you have a random sample, you only need to worry about randomness in sampling certain features that are relevant for the concept being studied.

Memory study

- Options:
- Sample from the Indian population
- Sample from many different countries but restricted to people born on a Sunday
- Goal: to make conclusions about how memory works in all humans
- Both are random samples, but one is better than the other: where the randomness is in a feature that doesn't matter for the concept being studied given the generalization goal

Population parameters, sample statistics

- Plot b: mean IQ of 98.5, and the standard deviation of 15.9, N = 100 – sample statistics
- An approximation of the population parameters.
- Our goal: how can we estimate population parameters based on sample statistics?
- Also, can we come up with a measure of "confidence" in our estimates?

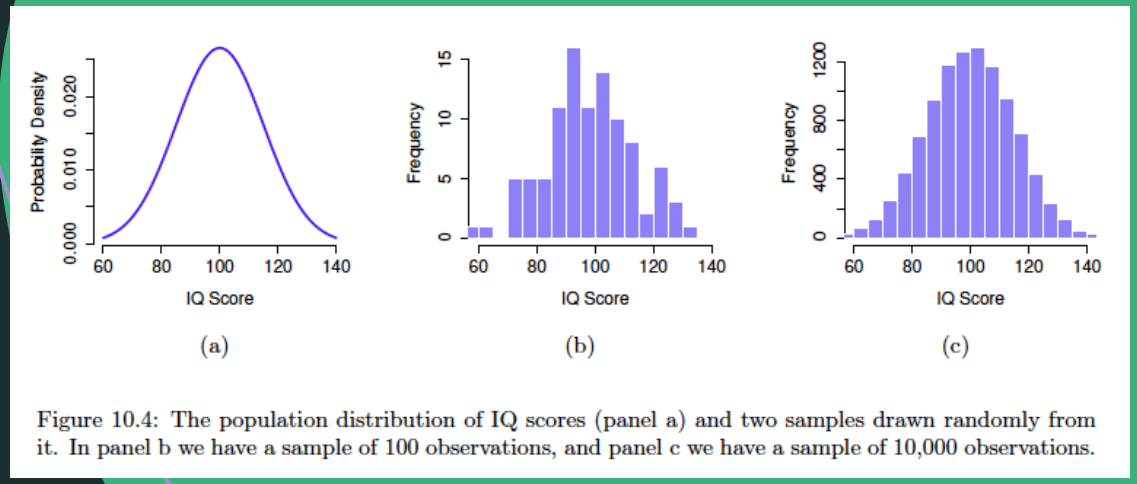


Figure 10.4: The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and panel c we have a sample of 10,000 observations.

The law of large numbers

Previous slide: plot c with a greater N (10k) provided a closer approximation to the population parameters

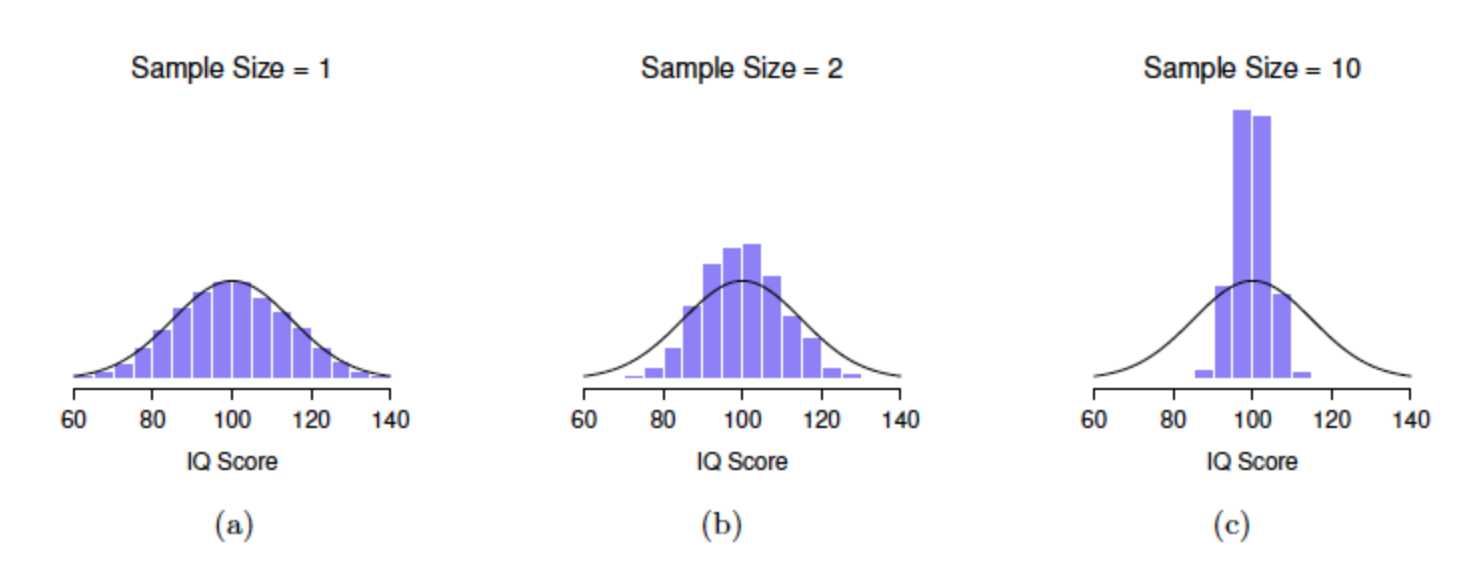
This law of large numbers applies to many statistics, but easiest to demonstrate is as a law of averages (sampling distribution of the mean, which we saw in the probability distribution lecture)

Revisiting the central limit theorem

Table 10.1: Ten replications of the IQ experiment, each with a sample size of $N = 5$.

	Person 1	Person 2	Person 3	Person 4	Person 5	Sample Mean
Replication 1	90	82	94	99	110	95.0
Replication 2	78	88	111	111	117	101.0
Replication 3	111	122	91	98	86	101.6
Replication 4	98	96	119	99	107	103.8
Replication 5	105	113	103	103	98	104.4
Replication 6	81	89	93	85	114	92.4
Replication 7	100	93	108	98	133	106.4
Replication 8	107	100	105	117	85	102.8
Replication 9	86	119	108	73	116	100.4
Replication 10	95	126	112	120	76	105.8

10k sample means



The black line is the true population distribution
What observation do can you make about the mean of any single sample and how that relates to the population mean across different values of sample size?

Other observations

- The mean of the sampling distribution is the same as the mean of the population
- The standard deviation of the sampling distribution (i.e., the standard error) gets smaller as the sample size increases
- The shape of the sampling distribution becomes normal as the sample size increases

Standard error of the mean

- SEM
- Sampling distribution of the mean
- The standard deviation of the sampling distribution (the standard error), or the standard error of the mean (SEM, in this case) relates to the population standard deviation sigma as

$$\mathrm{SEM} = \frac{\sigma}{\sqrt{N}}$$

IQ test in a small village of Andhra

- We're not sure the true population mean is 100 (as "defined" by the test makers)
- We need to provide a best guess about the population mean based on say 50 villagers who agreed to take the test
- I conduct the test and the mean in this sample of 50 comes out to be 97
- What is my best guess about the population mean?
- CLT, sampling distribution of the mean exercises earlier --> my best guess is 97!

Estimating population mean from the sample mean

Symbol	What.is.it	Do.we.know.what.it.is
$\bar{X}$	Sample mean	Yes calculated from the raw data
μ	True population mean	Almost never known for sure
$\hat{\mu}$	Estimate of the population mean	Yes identical to the sample mean

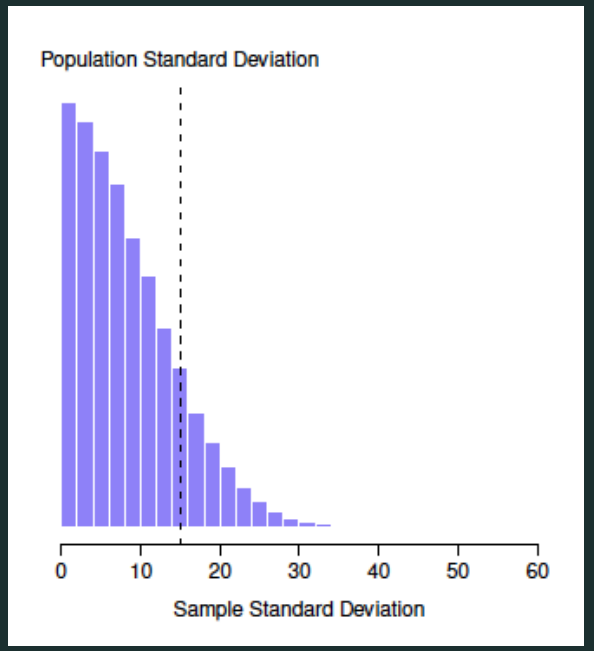
Remember: this only works when CLT applies, so need a sufficiently large sample size, otherwise your estimate will not be very accurate

How about population standard deviation?

- Say N = 1, IQ = 120 (IIITH student)
- What is your best guess about the IIITH mean IQ?
- 120 is the best guess you can make based on your data, you wouldn't be very confident but you can make a guess
- However, what is the population standard deviation?
- No idea! With our sample of 1, the standard deviation is 0 but it would not make any sense to say that about the population as we know it is going to be a wrong guess, so can't say this is the best guess possible

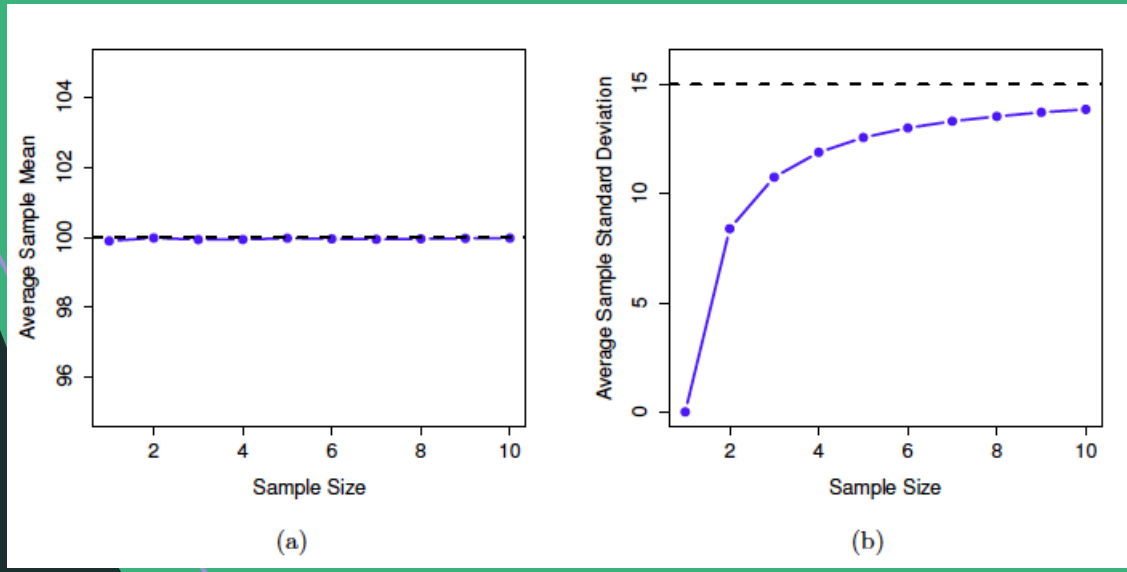
How about population standard deviation?

- N = 2, say s.d. (of the sample) = 8.5
- Intuition: the sample s.d. is a biased estimator of the population s.d.



Intuition: Sample s.d. is a biased estimator of population s.d.

- Demonstrate this?
- Simulate N = 10, N = 100, etc?
- s.d. is systematically smaller than the population s.d.



Biased and unbiased estimators

The sample mean is an unbiased estimator of the population mean

The sample s.d. is a biased estimator of the population s.d.

How do we fix this bias?

Sample variance

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

- Also a biased estimator (of the population variance)
- A minor tweak in the formula can make it an unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

- This is what the R var function calculates, not the sample variance but the unbiased estimator (dividing by N-1 instead of N)
- Similarly for the s.d.

Reporting

- When you calculate the sample s.d. (dividing by N), it should be referred to as the sample s.d.
- When dividing by N-1, this is an unbiased estimator of the population s.d. -- I.e., your best guess about the population s.d. parameter!
- Many people use the unbiased estimator (the output of R std and var functions) and refer to them as the sample s.d. and sample variance
- This is technically incorrect

Estimating the population s.d. and variance: Summary

Symbol	What is it?	Do we know what it is?
s	Sample standard deviation	Yes, calculated from the raw data
σ	Population standard deviation	Almost never known for sure
$\hat{\sigma}$	Estimate of the population standard deviation	Yes, but not the same as the sample standard deviation

Symbol	What is it?	Do we know what it is?
s^2	Sample variance	Yes, calculated from the raw data
σ^2	Population variance	Almost never known for sure
$\hat{\sigma}^2$	Estimate of the population variance	Yes, but not the same as the sample variance

Standard normal distribution

- If you know the population mean and s.d., you can normalize your variable:

$$\frac{X - \mu}{\sigma}$$

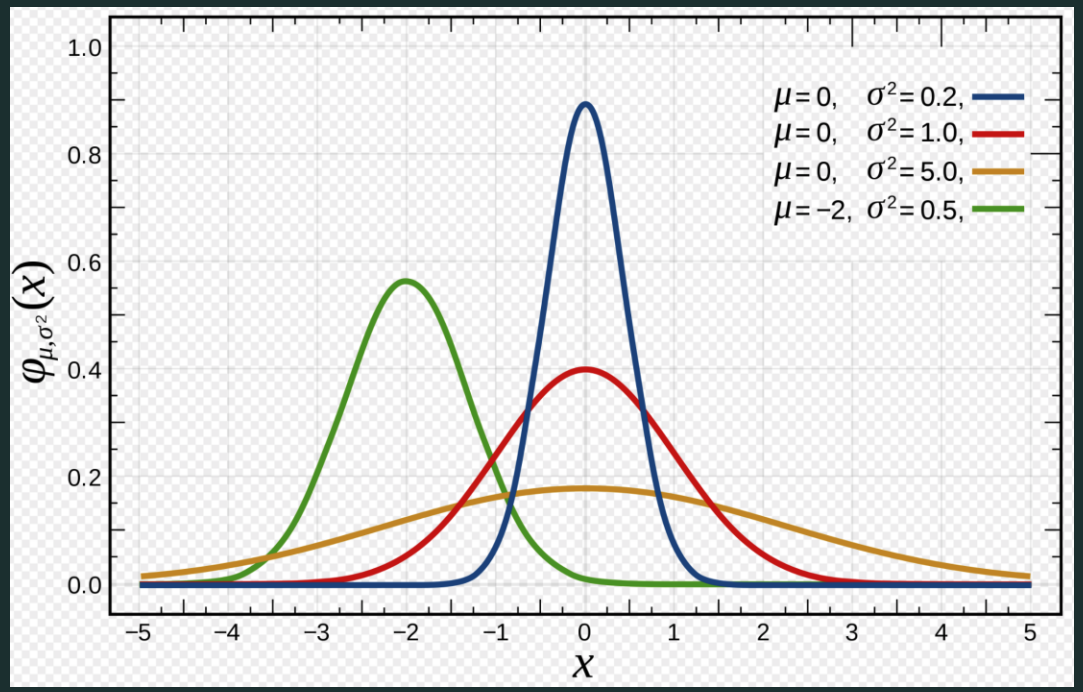
Standard normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

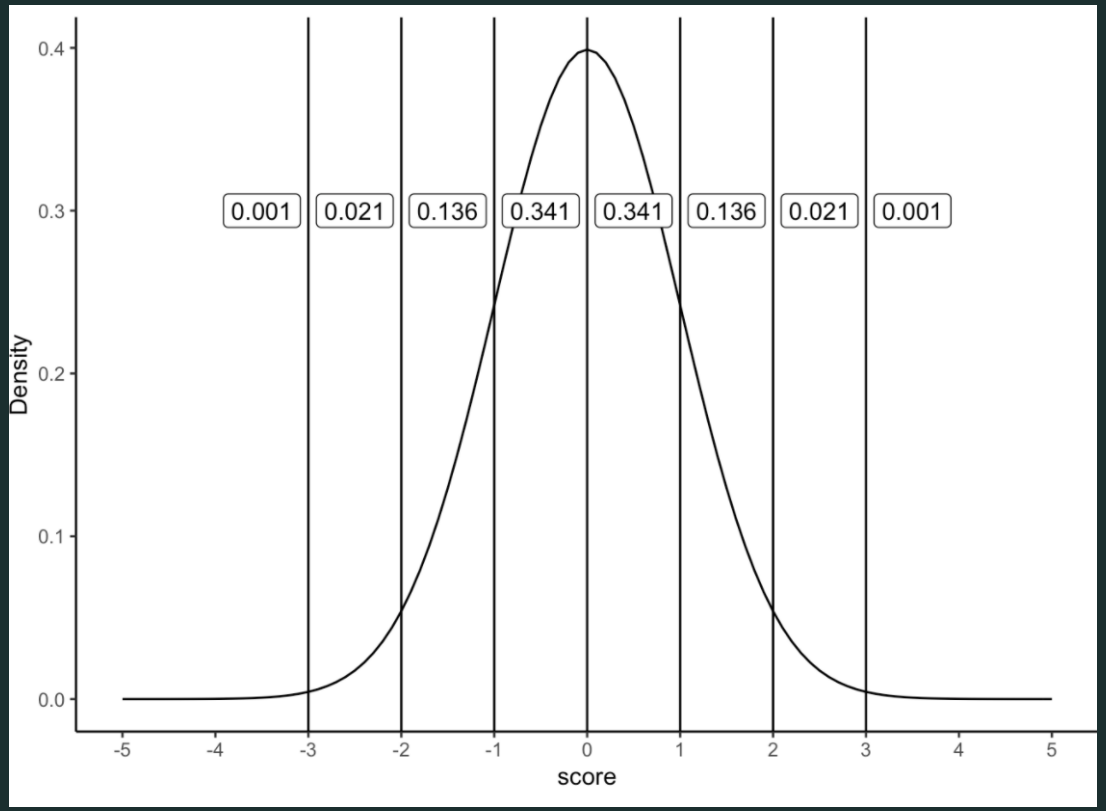
Mu = 0, var = 1

$$\varphi(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$$

Standard normal distribution



The one in red is the standard normal distribution



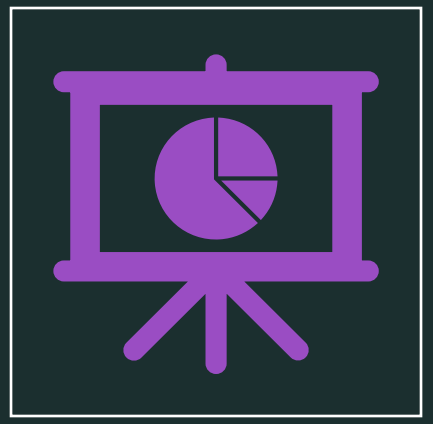
Normal distributions:
Approx: 68% of the values lie within 1 s.d. of the mean
95% of the values lie within 2 s.d. of the mean

The standard normal quantiles

p	z_p
0.80	1.281 551 565 545
0.90	1.644 853 626 951
0.95	1.959 963 984 540
0.98	2.326 347 874 041
0.99	2.575 829 303 549
0.995	2.807 033 768 344
0.998	3.090 232 306 168

p	z_p
0.999	3.290 526 731 492
0.9999	3.890 591 886 413
0.99999	4.417 173 413 469
0.999999	4.891 638 475 699
0.9999999	5.326 723 886 384
0.99999999	5.730 728 868 236
0.999999999	6.109 410 204 869

Confidence intervals



Ok, so now you've made a guess about the population parameters from your data sample

How confident are you about your guess? (recall, that this probability need not be an intuitive probability, like we discussed, it is a frequentist probability)

Confidence Intervals (CIs)

- My best guess of the mean IQ of IIITH students is 120 based on a sample of 100
- The 95% confidence interval is 110–130
- Compared to the 95% CI of 100-140
- The margin of error is lower in the former case
- Intuition: you can reduce margins of error by using more people in your sample

How do we construct CIs?

- Assume true population mean = mu and s.d. = sigma
- We know from the central limit theorem that the sampling distribution of the mean is normal, and that for Normal distributions, 95% of the values lie within 1.96 (had approximated it to 2 earlier) standard deviations from the mean.
- Check for yourself using qnorm(p = c(.025, .975))

How do we construct CIs?

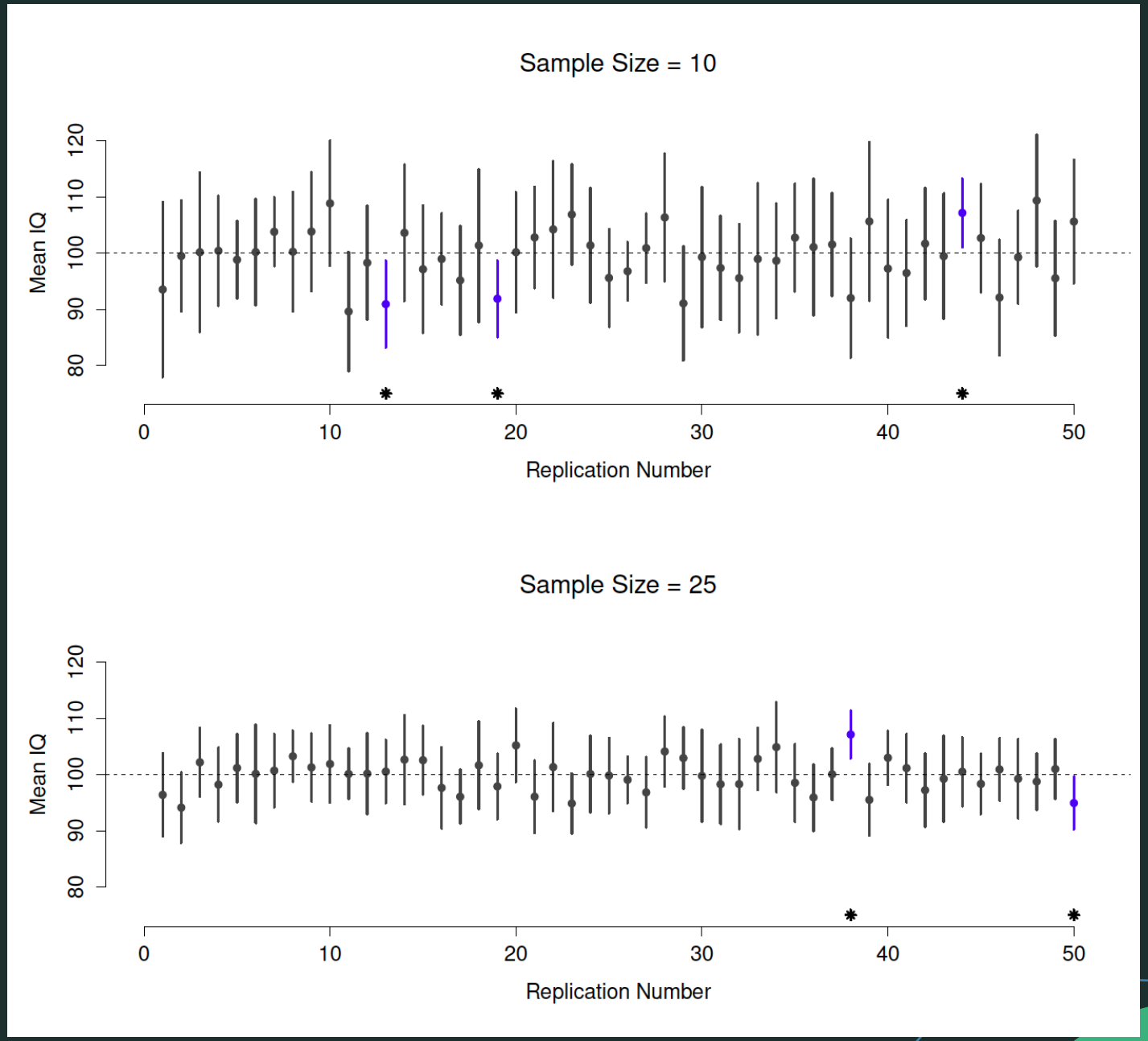
$$\mu - (1.96 \times \text{SEM}) \leqslant \bar{X} \leqslant \mu + (1.96 \times \text{SEM})$$

$$\bar{X} - (1.96 \times \text{SEM}) \leqslant \mu \leqslant \bar{X} + (1.96 \times \text{SEM})$$

- SEM is used because note that we are referring to the sampling distribution of the mean, so the s.d. that matters is the standard error of the mean. $\text{SEM} = \frac{\sigma}{\sqrt{N}}$
- "The population mean has a 95% chance of falling into this range." -- the textbook says this when it introduces this concept of a 95% coverage area in the Normal distribution but see the final section on interpretation

The confidence interval changes from sample to sample

- So if I say I'm 95% confident true population mean IQ lies between 110-130, and then the very next sample, I say I'm 95% confident the true population mean IQ lies between 100-140, there is something quirky about this.
- Also, note that the calculation on the previous slide used the standard error, we do not know the population s.d.
- What it works out to be is that if you repeat this procedure many times, 95% of the confidence intervals you construct would be expected to contain the population mean – this is the correct interpretation of a 95% CI.


Sample Size = 10
Mean IQ
80
90
100
110
120
0
10
20
30
40
50
Replication Number
Sample Size = 25
Mean IQ
80
90
100
110
120
0
10
20
30
40
50
Replication Number

Frequentist CI

- A population mean is not a repeatable random variable
- Repeatable: very important for frequentist probability interpretation
- What is repeatable is the CI (in different samples)
- So a frequentist is not allowed to make probabilistic statements about the probability of the population mean (I.e., there is a 95% chance the population mean lies in a certain range) but is allowed to make probabilistic statements about the CI across many samples (I.e., that 95% of such CIs will contain the true population mean).

Does the interpretation matter practically?

- The Bayesian version = credible intervals (will be covered in the last lecture)
- Under some conditions, credible intervals and frequentist CIs can look very different. So the interpretation differences matter in these cases.

An additional issue

In the SEM formula, we used the population s.d. but we do not know the population s.d.!

So we have to use an estimate

We also know that the SE (i.e., the s.d of the sampling dist) changes as the sample size from all the simulations we did

What is a probability distribution that looks very much like the Normal distribution but has a dependence on sample size?

The T-distribution!

So instead of using the standard normal quantiles, we will use quantiles from the T-distribution

```
N <- 10000   # suppose our sample size is 10,000
qt( p = .975, df = N-1)   # calculate the 97.5th quantile of the t-dist
```

```
1.960201
```

```
N <- 10   # suppose our sample size is 10
qt( p = .975, df = N-1)   # calculate the 97.5th quantile of the t-dist
```

```
## [1] 2.262157
```

Captures the intuition that with smaller sample sizes, our margin of error should be larger

Summary

Basic ideas about samples, sampling and populations

Statistical theory of sampling: the law of large numbers, sampling distributions and the central limit theorem

Estimating means and standard deviations

Estimating a confidence interval

The Scientific Method

- A set of principles about the appropriate relationship between ideas and evidence.
  1. Develop theories (i.e. ideas)
  2. Derive hypotheses from the theories and test them (i.e., evidence)
  3. Modify your theory based on evidence
  4. Repeat 2-3 or if required restart at 1

Theory and Hypothesis

- Theory: A GENERAL hypothetical explanation of a natural phenomenon
- Hypothesis: A SPECIFIC falsifiable prediction made by a theory

Determine if the following are theories/hypotheses:

1. If we give plant A acid while we give plant B water then plant B will grow to be taller than plant A.
2. Any two particles of matter attract one another with a force directly proportional to the product of their masses and inversely proportional to the square of the distance between them.
3. If I throw this ball at 2 m/s at an angle of 45 deg, it will take 0.3s to hit the ground.

Theory vs. Hypothesis

- What's the difference?
  - Hypothesis: specific prediction for a single event
    - "If I throw this ball at 2 m/s at an angle of 45 deg, it will take 0.3s to hit the ground"
  - Theory: framework for understanding a larger phenomenon
    - ie: theory of gravity
    - Can derive many hypotheses from a theory

Falsifiability

“No amount of experimentation can ever prove me right, but a single experiment can prove me wrong”

- Albert Einstein

The importance of falsifying in theory testing

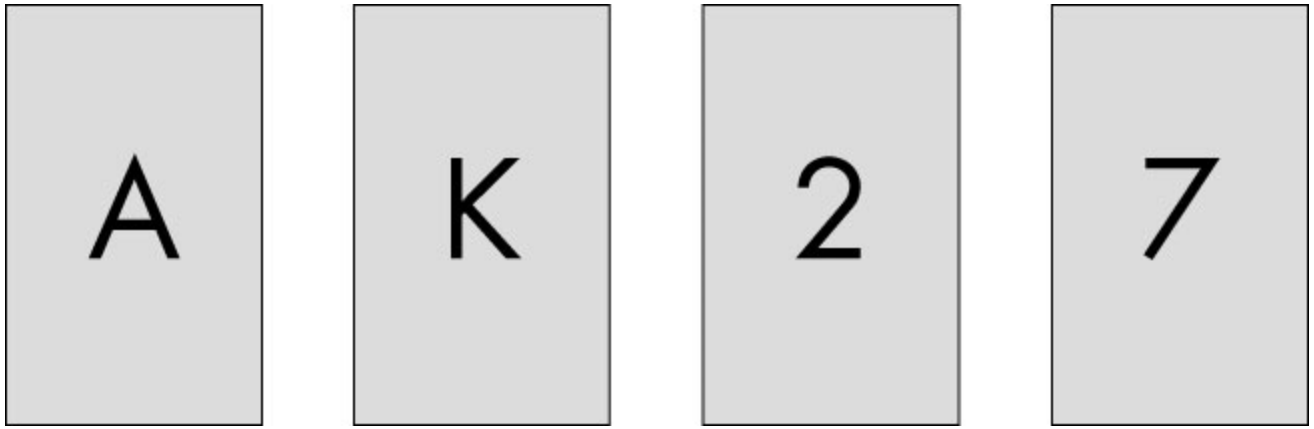
e.g. Hypothesis (derived from some theory): All swans are white.

Observation/Evidence: Observed 100 swans and all were white.

A. Theory proven?

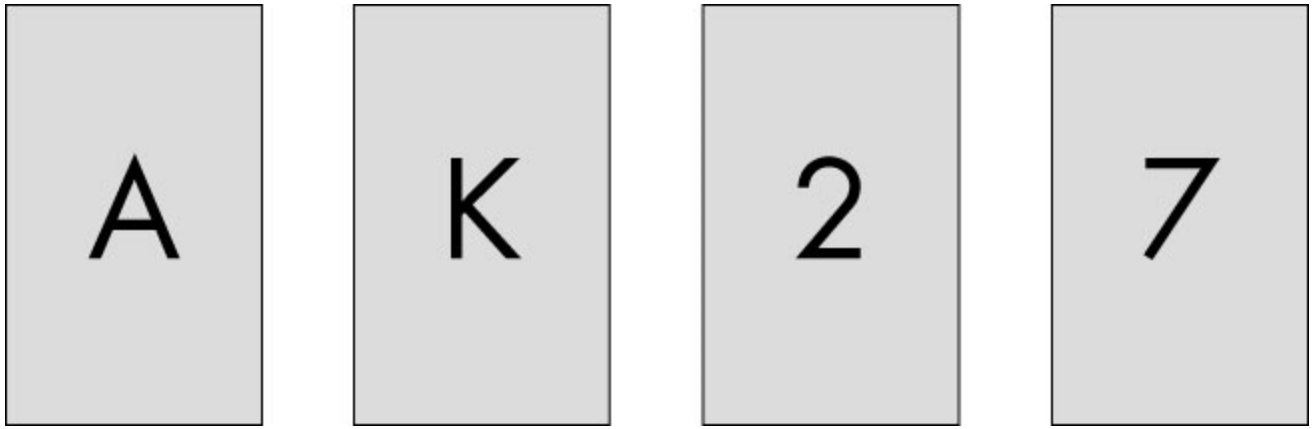
B. Theory not disproven?

If a card has an odd number on one side, then it has a vowel on the other side.



Flip the two best cards to test your hypothesis

If a card has an odd number on one side, then it has a vowel on the other side



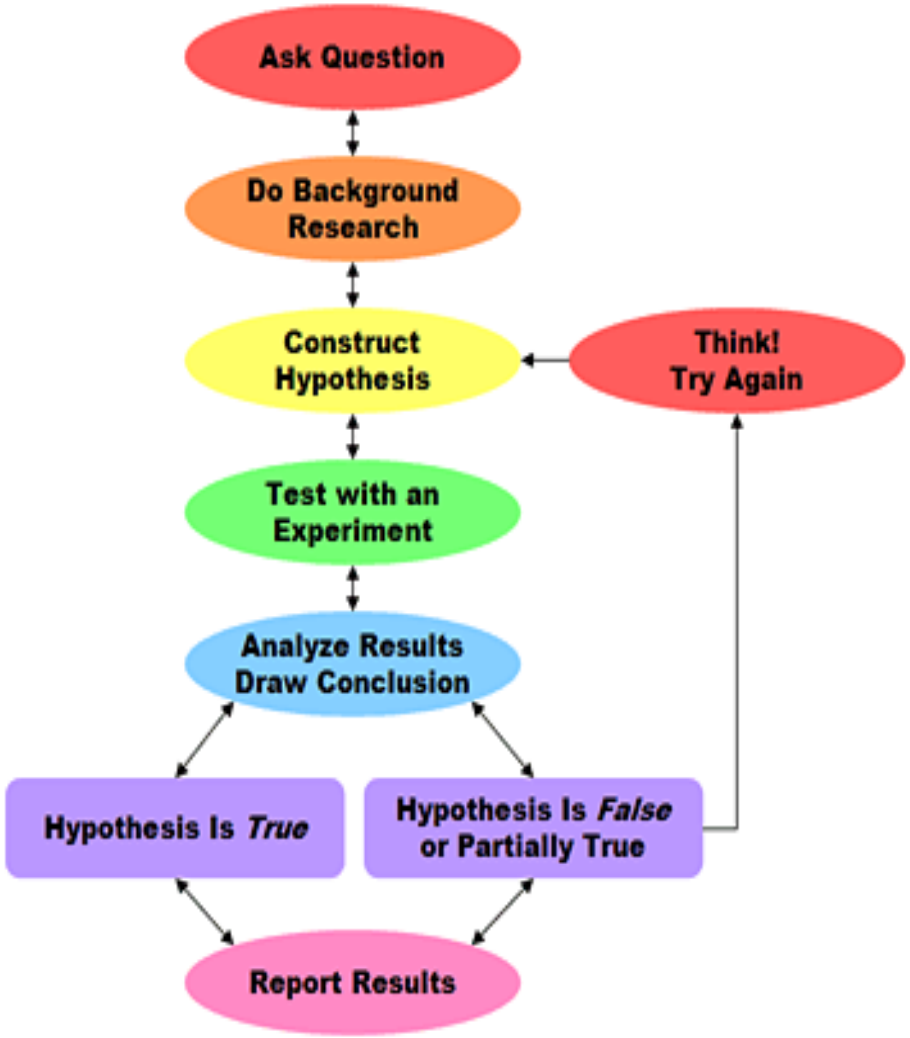
[2] [3] [D] [L]

- How many of you chose:
  - A?
  - 2?
  - 7?
  - K?

- Confirmation bias if you pick A
- To falsify the hypothesis, you need to pick cards 7 and K.
- Falsified if you find a consonant on the other side of 7.
- Falsified if you find an odd number on the other side of K.

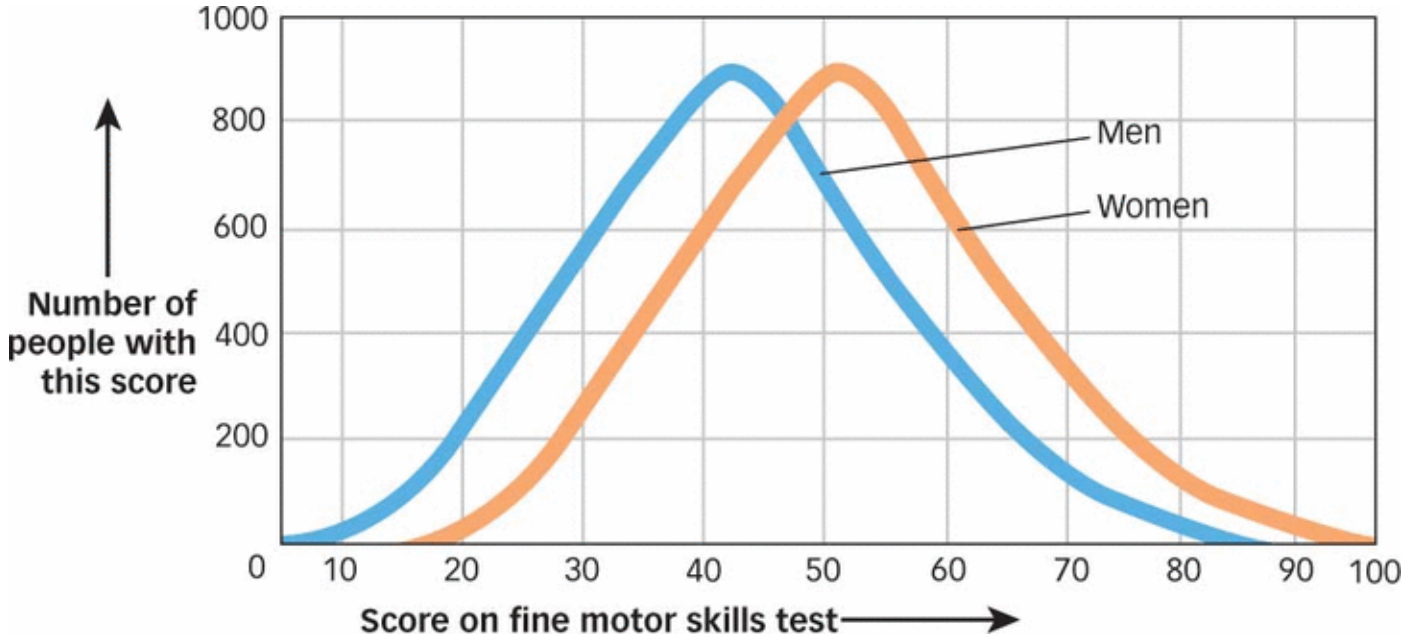
The Scientific Method – A Review

- Provides a logical framework for examining scientific questions.
- Allows other researchers to replicate studies.



Normal Distributions

- Graphic Representations
  - Frequency Distributions
  - **Normal (Gaussian) Distributions**



Hypothesis testing

- Are the means different or do they come from the same population the same mean?
- Consider a test that we *think* is 99% accurate (i.e., the null hypothesis).
- Get 100 people and administer the test. Knowing the truth, let's say we observe that it was actually accurate in 98/100 people (i.e., our sample). The Q is should we believe our "null hypothesis" that the test is 99% accurate?
- How about if our sample shows 97/100 is accurate? Etc etc.

Example

Alessandra designed an experiment where subjects tasted water from four different cups and attempted to identify which cup contained bottled water. Each subject was given three cups that contained regular tap water and one cup that contained bottled water (the order was randomized). She wanted to test if the subjects could do better than simply guessing when identifying the bottled water.

Her hypotheses were $H_0 : p = 0.25$ vs. $H_a : p > 0.25$ (where p is the true likelihood of these subjects identifying the bottled water).

The experiment showed that 20 of the 60 subjects correctly identified the bottle water. Alessandra calculated that the statistic $\hat{p} = \frac{20}{60} = 0.\bar{3}$ had an associated P-value of approximately 0.068.

QUESTION A (EXAMPLE 1)

What conclusion should be made using a significance level of $\alpha = 0.05$?

Choose 1 answer:

(A) Fail to reject H_0

(B) Reject H_0 and accept H_a

(C) Accept H_0

https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-categorical-proportions/idea-significance-tests/a/p-value-conclusions