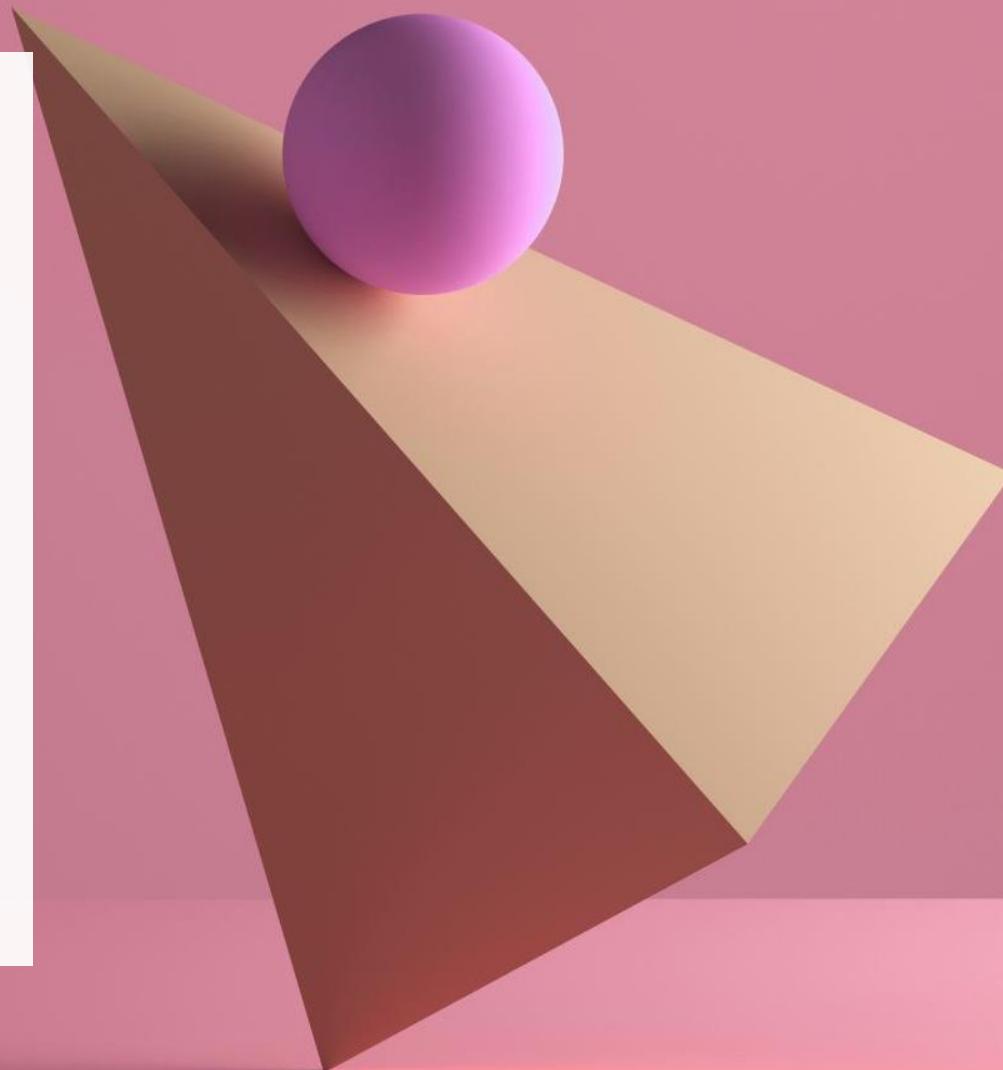


Behavioral Research: Statistical Methods

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.

IF YOU DON'T CONTROL FOR
CONFOUNDING VARIABLES,
THEY'LL MASK THE REAL
EFFECT AND MISLEAD YOU.



BUT IF YOU CONTROL FOR
TOO MANY VARIABLES,
YOUR CHOICES WILL SHAPE
THE DATA, AND YOU'LL
MISLEAD YOURSELF.



SOMEWHERE IN THE MIDDLE IS
THE SWEET SPOT WHERE YOU DO
BOTH, MAKING YOU DOUBLY WRONG.
STATS ARE A FARCE AND TRUTH IS
UNKNOWNABLE. SEE YOU NEXT WEEK!



Why do statistics?

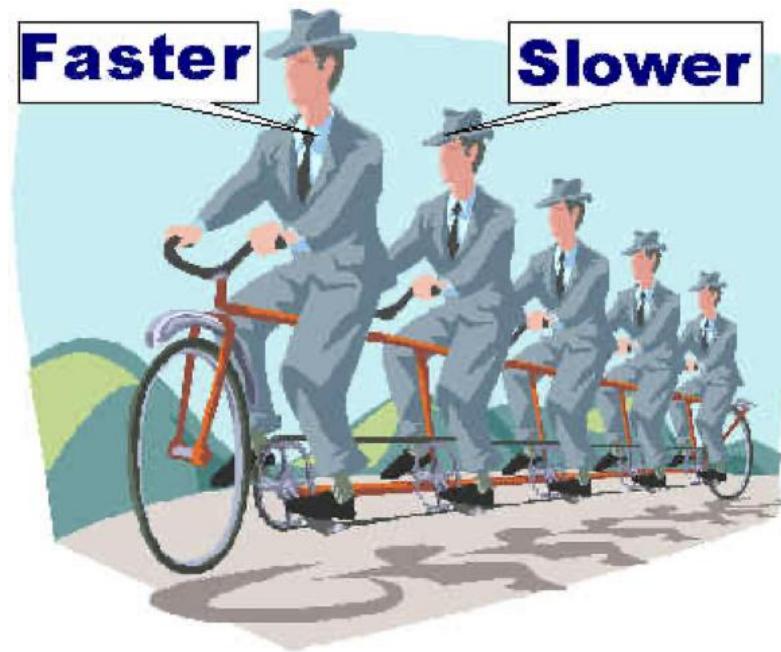
Why do statistics, why not use common sense?

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.

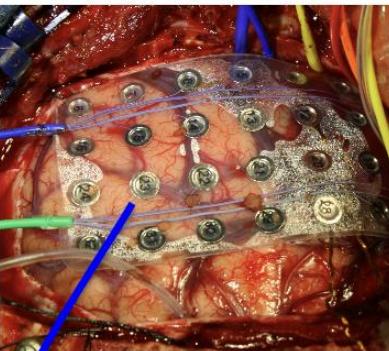
You might have encountered many such claims, especially during the early days of COVID. There is also currently a proliferation of pseudoscientific thinking in India. An education in basic statistics and research design will hopefully help you see through some of the issues with such claims.

Human-beings

- Complexity
- Variability
- Reactivity

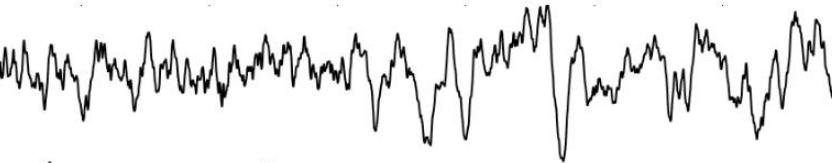


Brains



Memorize:

"Red"



"Face"



"Sign"



1 second

Related statistical pitfalls

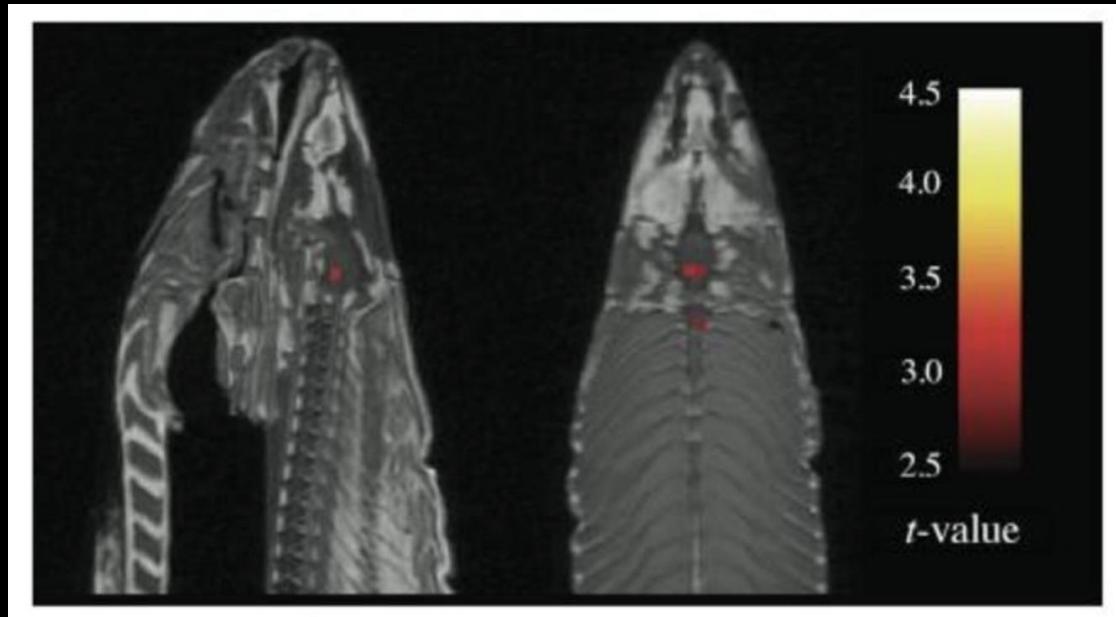


Replication Crisis

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)

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

Evans, Barston, &
Pollard (1983)

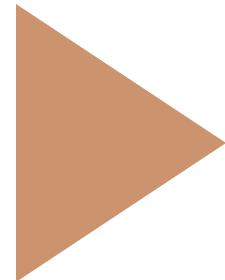
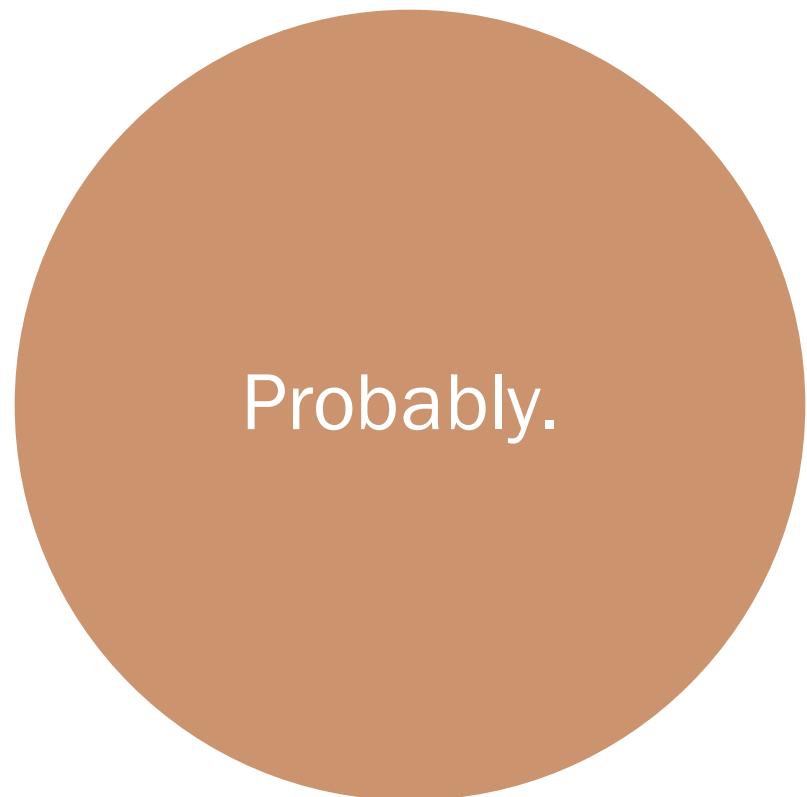
when the structure of the argument
was in line with pre-existing beliefs
and biases

	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%

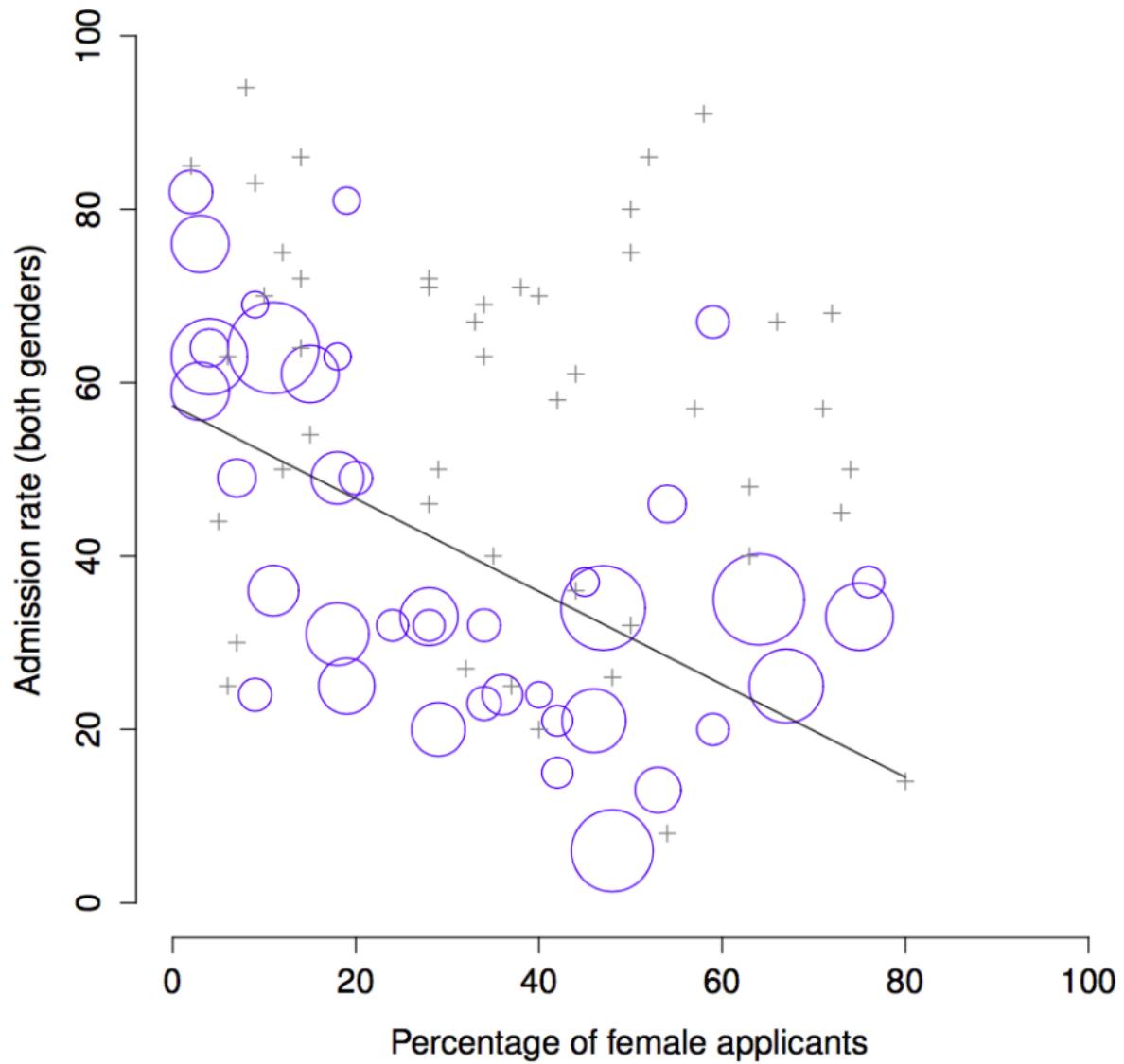
Department	Applicants	Percent admitted	Applicants	Percent admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

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



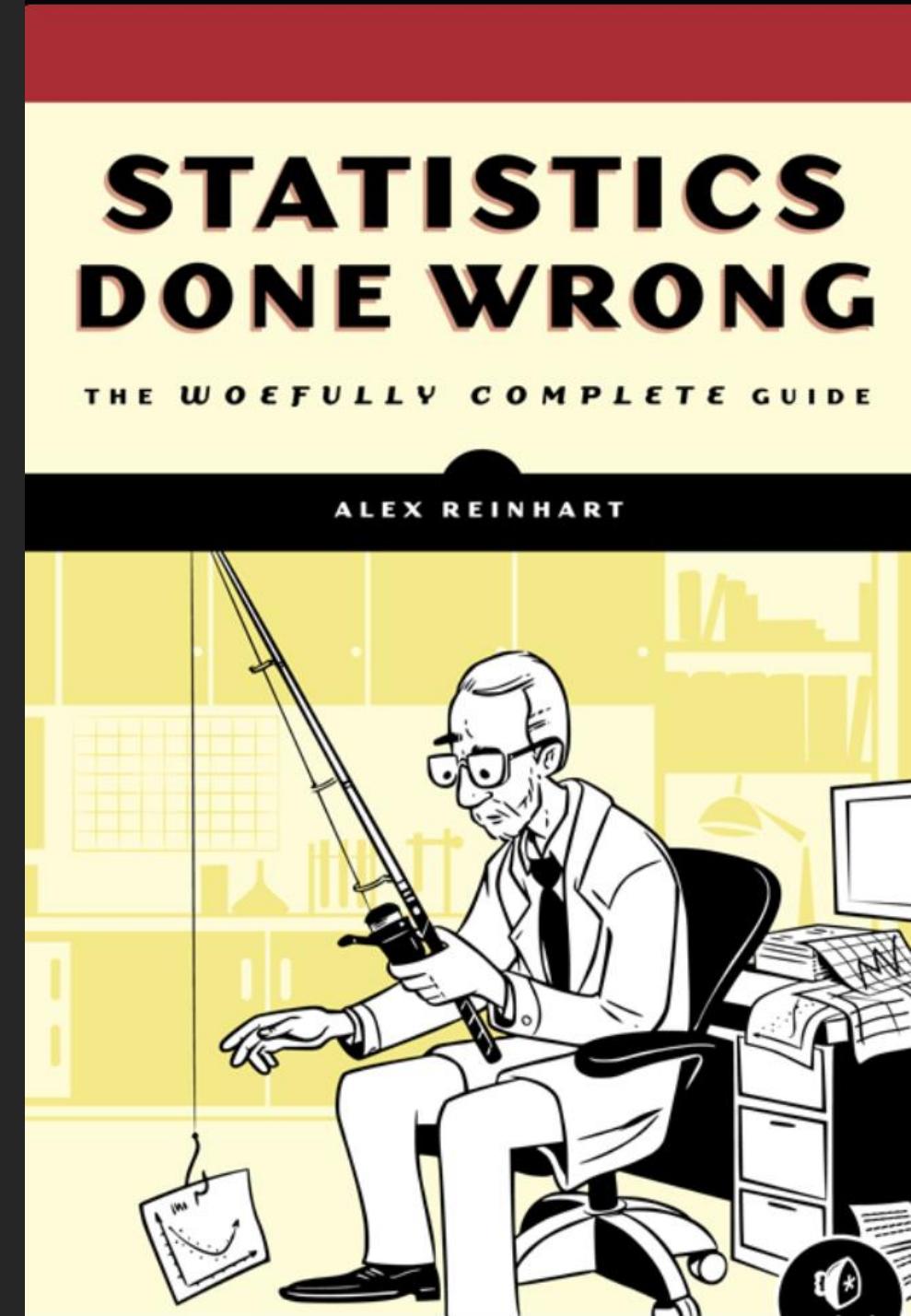
OFTEN, REPORTERS MAKE FUNDAMENTAL
ERRORS WHEN THEY REPORT NUMBERS (E.G.
NOT TAKING INTO ACCOUNT BASE RATES)

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.

$(7/77) \times 100 = 9\%$.

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:

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

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

Bayes' rule

$$P(\text{cancer}|\text{positive test}) = P(\text{positive test}|\text{cancer}) * P(\text{cancer}) / P(\text{positive test})$$

What we know:

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

$$P(\text{positive test}) = 77/1000 = 0.077 \text{ (from the last slide)}$$

The other way to calculate $P(\text{positive test}) = P(\text{positive test}|\text{cancer}) * P(\text{cancer}) + P(\text{positive test}|\text{not cancer}) * P(\text{not cancer}) = 0.9 * 0.008 + 0.07 * 0.992 = 0.07664 \approx 0.077$

$$P(\text{cancer}|\text{positive test}) = 0.9 * 0.008 / 0.077 = 0.09$$

$P(\text{cancer}) = \text{base rate}$

$P(\text{positive test}|\text{not cancer}) = \text{false positive probability}$

Many fail
to give the
right
answer



2/3rds of doctors fail this test



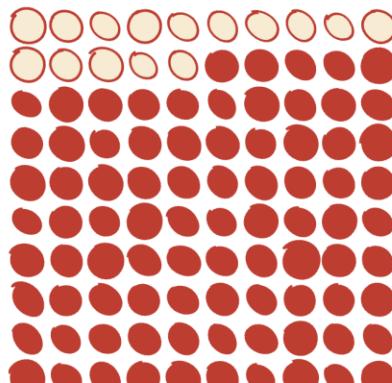
1/3rds of statistics and
methodology instructors like myself
and statistics students.

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.

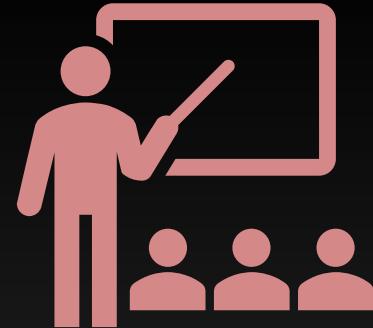
Knowledge of basic statistical and probabilistic pitfalls can help you evaluate these claims better.

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://LEARNIN
GSTATISTICSWITHR.
COM/BOOK/WHY-
DO-WE-LEARN-
STATISTICS.HTML](https://learninstatisticswithr.com/book/why-do-we-learn-statistics.html)

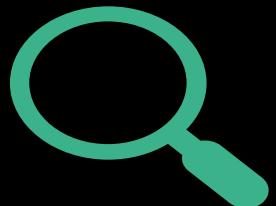
To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

– Sir Ronald Fisher⁶

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.



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

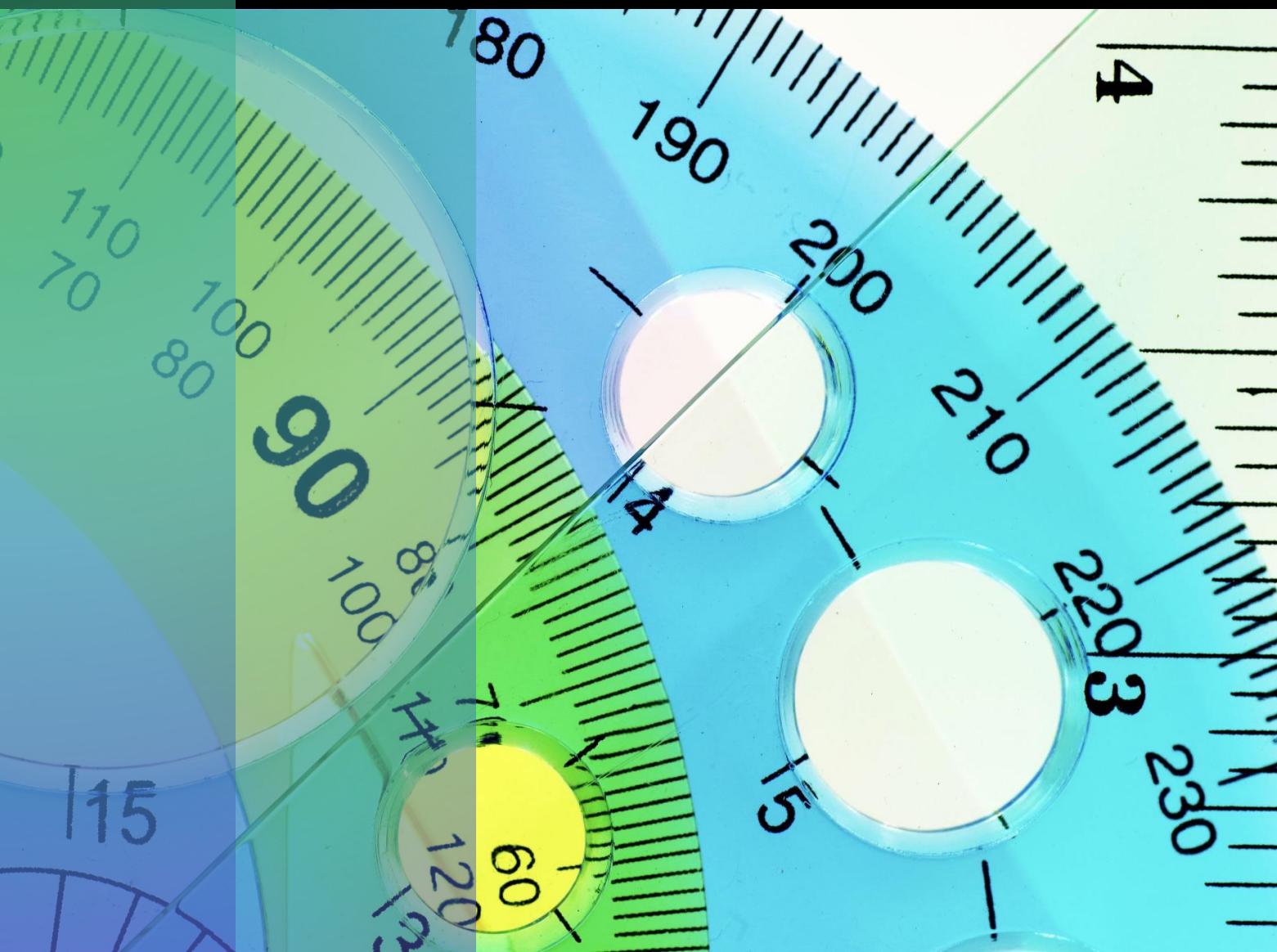
e.g. temperature. Difference
between 20 and 17 deg celsius = 3
degrees. The same as the
difference between 30 and 27
degrees. There is no natural zero,
just an arbitrary point (freezing
point) chosen as a reference.

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

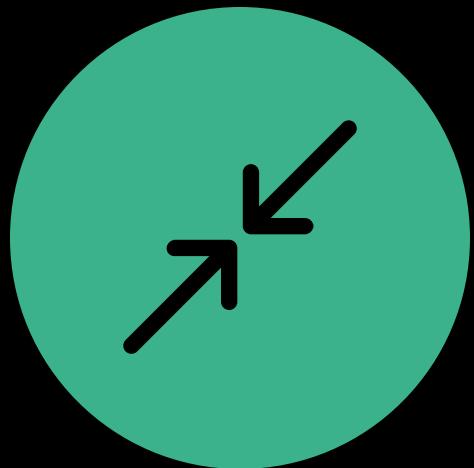
Table 2.1: The relationship between the scales of measurement and the discrete/continuity distinction.
Cells with a tick mark correspond to things that are possible.

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

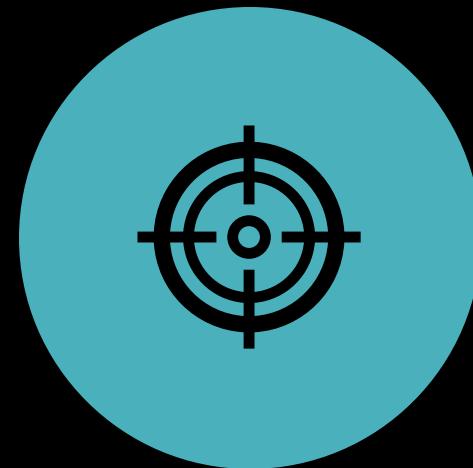
Real world variables may not always adhere to these classifications

- Likert scale
 - Choose from the following options. You feel happy today:
 - What scale is this?
 - 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).
- 1. Strongly disagree
 - 2. Disagree
 - 3. Neutral
 - 4. Agree
 - 5. Strongly agree

Is the measurement any good?



RELIABILITY: HOW REPEATABLE?



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

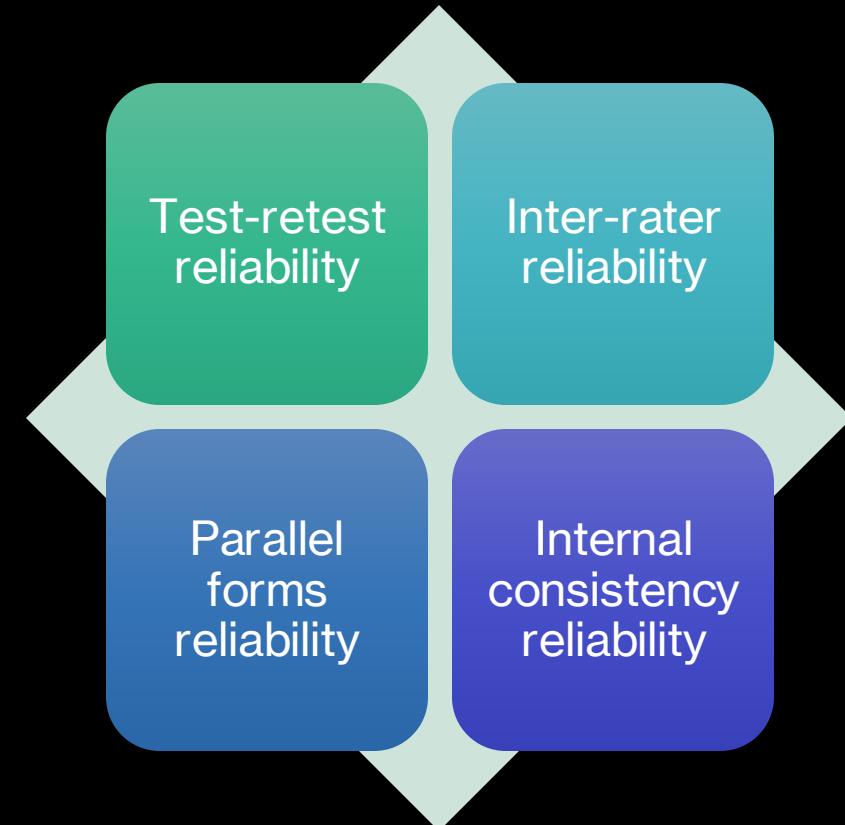
Reliability

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?

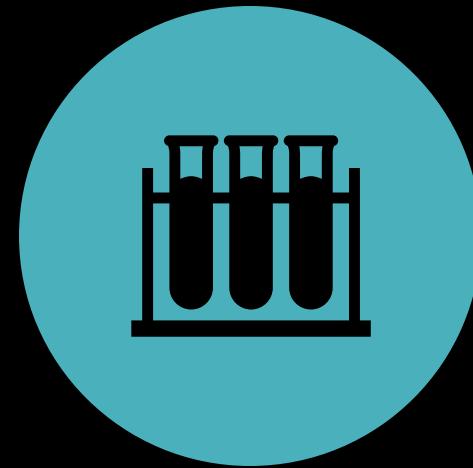
Different ways to measure reliability



Test-retest reliability



CONSISTENCY OVER TIME



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

Inter-rater reliability

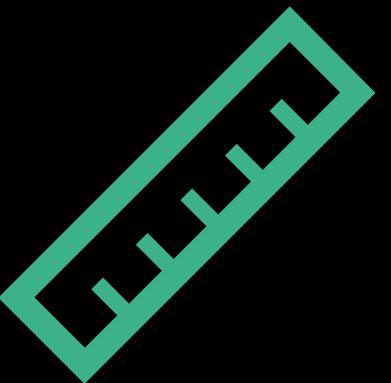


CONSISTENCY ACROSS PEOPLE

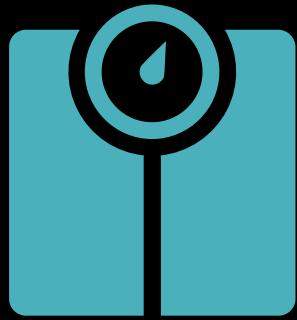


IF SOMEONE ELSE DOES THE
MEASUREMENT, WILL WE GET
THE SAME RESULT?

Parallel forms reliability

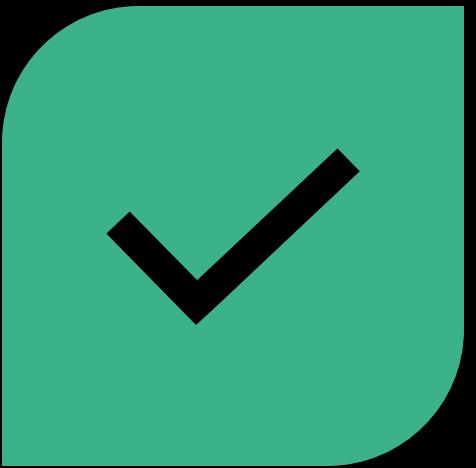


Consistency across theoretically-equivalent measurements

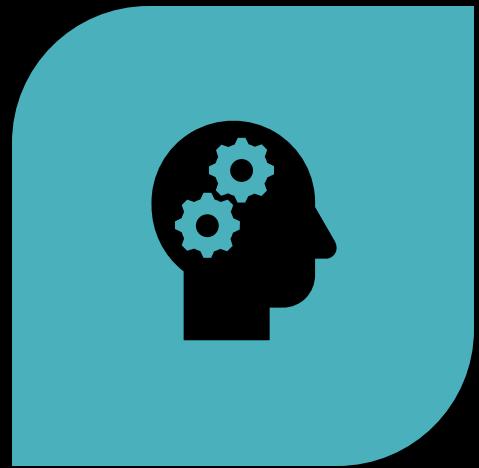


If I use a different weighing scale, do I get the same weight measurement?

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

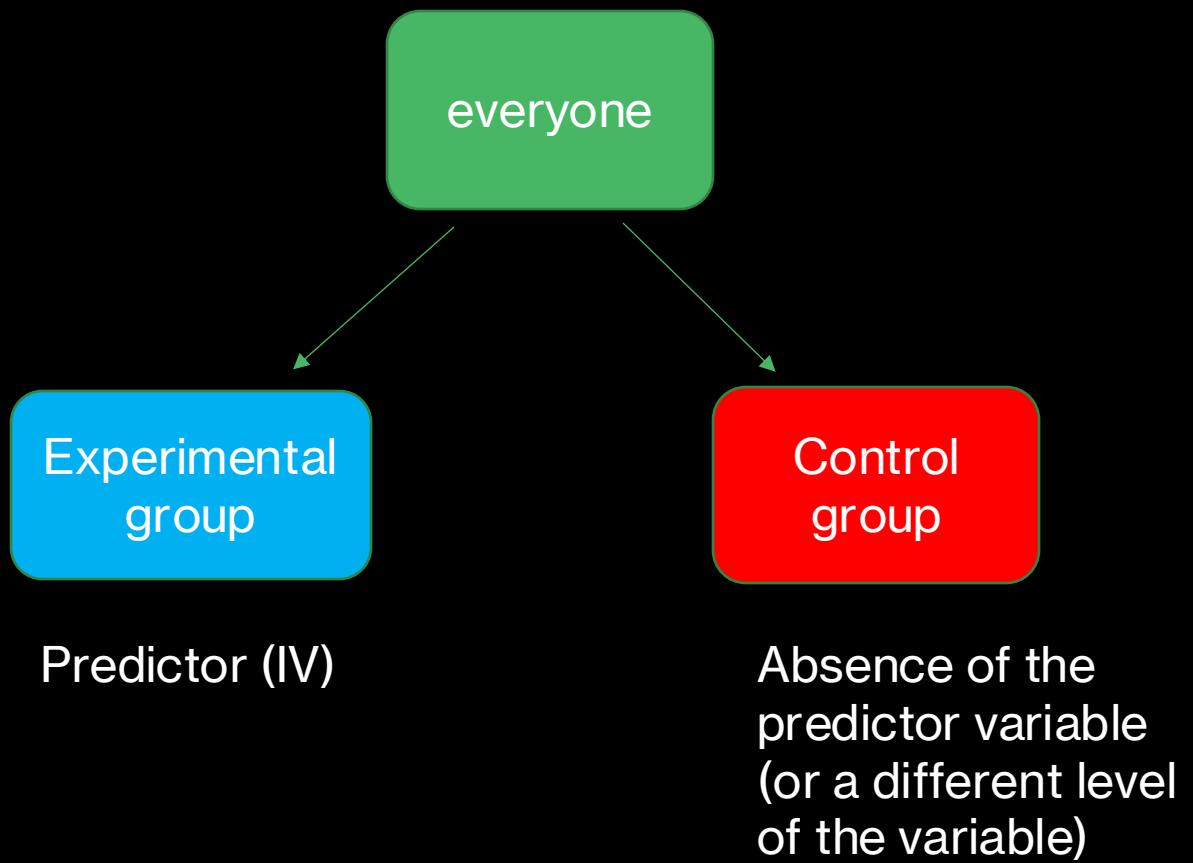


We cannot possibly think of ALL the predictors that can influence the outcome



How do we solve this issue?

Randomization



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?

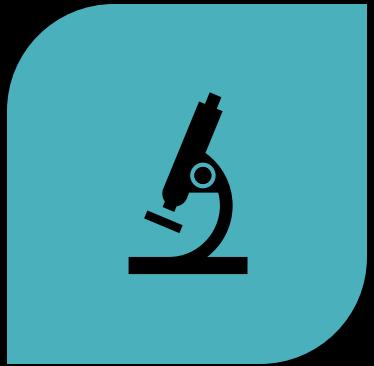
Construct validity

- 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?

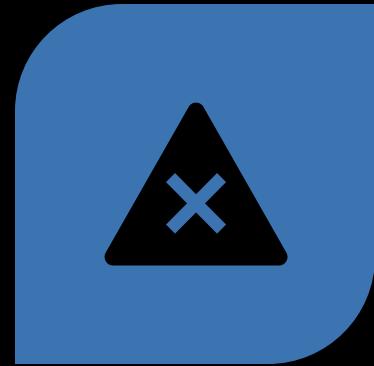
Face validity



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

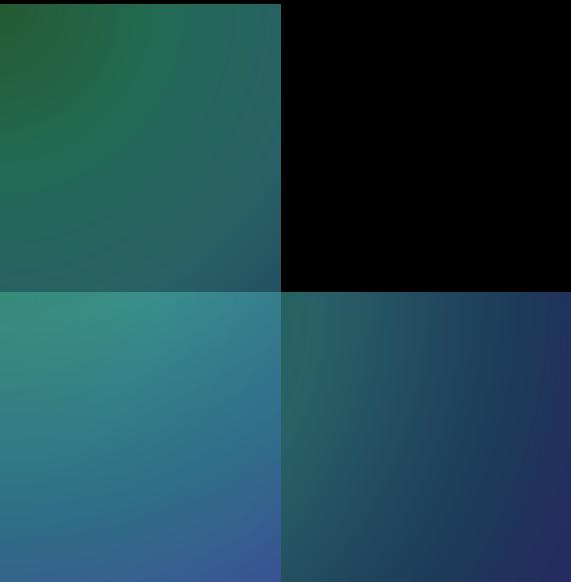


Something that happens during the study (or preceding) that can influence the results



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

A stack of five colorful envelopes (yellow, white, green, pink, blue) is arranged on a background with horizontal color gradients. The envelopes are partially visible, showing their flaps. The colors transition from green on the left to yellow, white, green, pink, and blue on the right.

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 & Tversky, 1973)

Experimenter Bias



Oskar Pfungst: 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



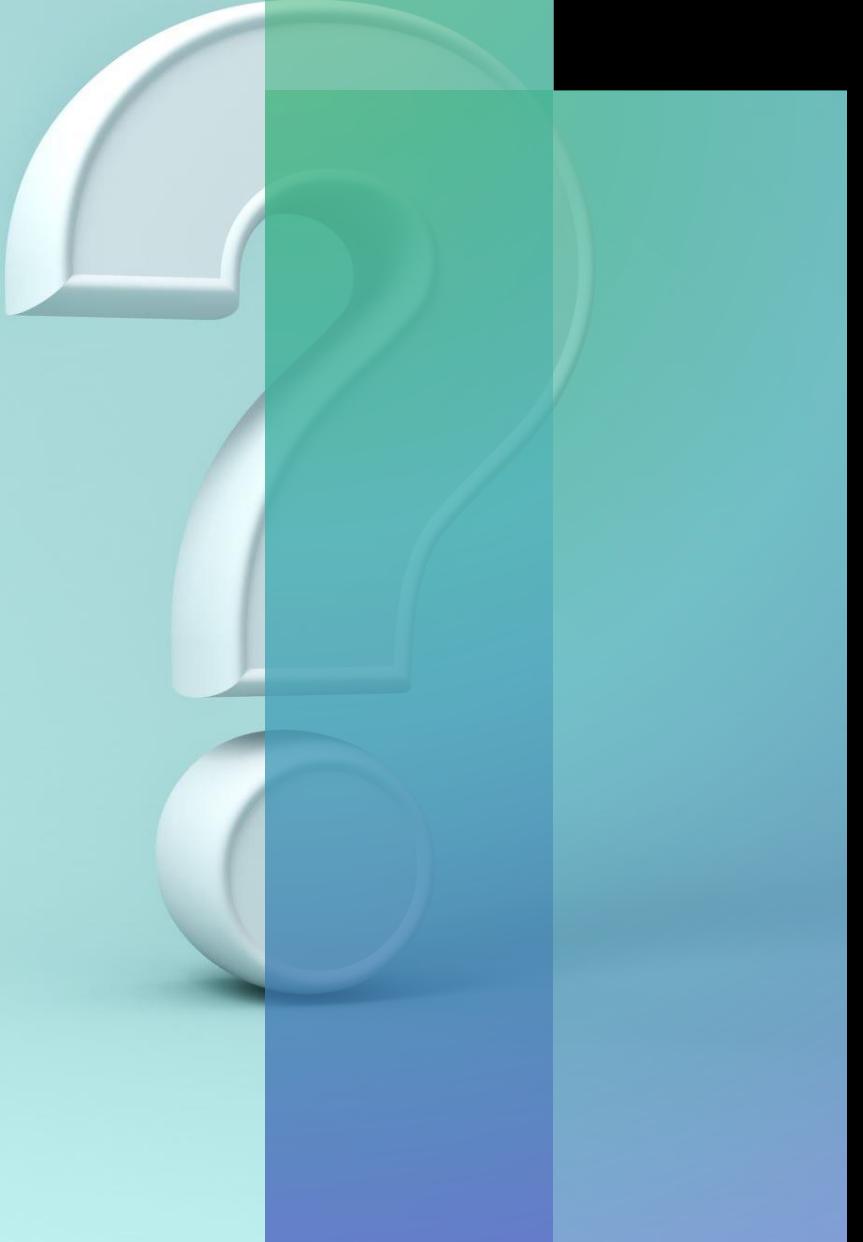
- "Hawthorne" effect



- 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

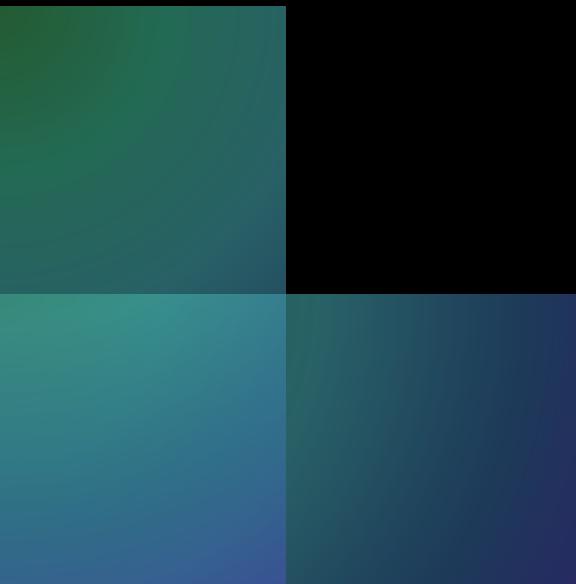
Placebo effects

- 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/>

The background of the slide features a complex, abstract pattern of numerous thin, wavy lines in various colors, including green, orange, yellow, and blue, creating a sense of depth and motion.

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 $P(\text{Fischer})$ really 0.5 or is it something else? This is the realm of inferential statistics.

What is a probability?

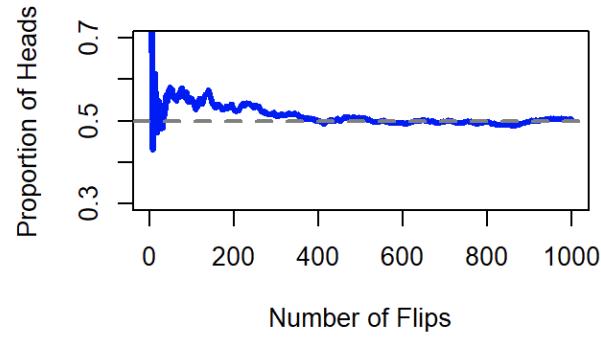
- Means slightly different things if you are a frequentist statistician vs if you are a Bayesian
- Carlsen has a 70% chance of winning a game against Nepomniachtchi: what does this mean to you?
- If they play a 10 game match, Carlsen is expected to win 7?
- If I bet Rs 100 on Nepomniachtchi, I should get a reward of Rs 233 ($700/3$) if Neko 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 Neko.

Frequentist probability

FLIP A COIN MANY TIMES AND COUNT THE
PROPORTION OF HEADS



- As $N \rightarrow \infty$, 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 $N \rightarrow \infty$ 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

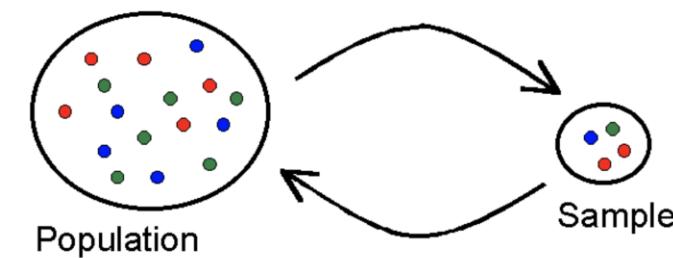
- You will often hear things like "variable x is i.i.d"
- Independently and identically distributed
- Say Y_i are dice throws for $i=1:n$
- The outcome of each different set of (n throws) is a random variable itself
- The outcome of each throw has the same distribution (uniform over 6 possibilities):
 Y_1, Y_2, \dots, Y_n are identically distributed
- Y_1 is independent of Y_2 and so on.
- Therefore, iid.

A function applied on the sample

- Y_i 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

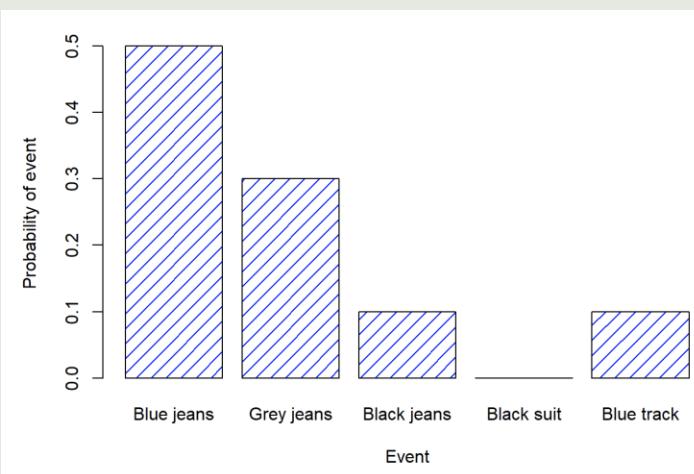
SAMPLING DISTRIBUTION OF A STATISTIC: THE DISTRIBUTION OF A STATISTIC (OR A FUNCTION) APPLIED ON THE SAMPLES

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)

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

Defined for continuous random variables

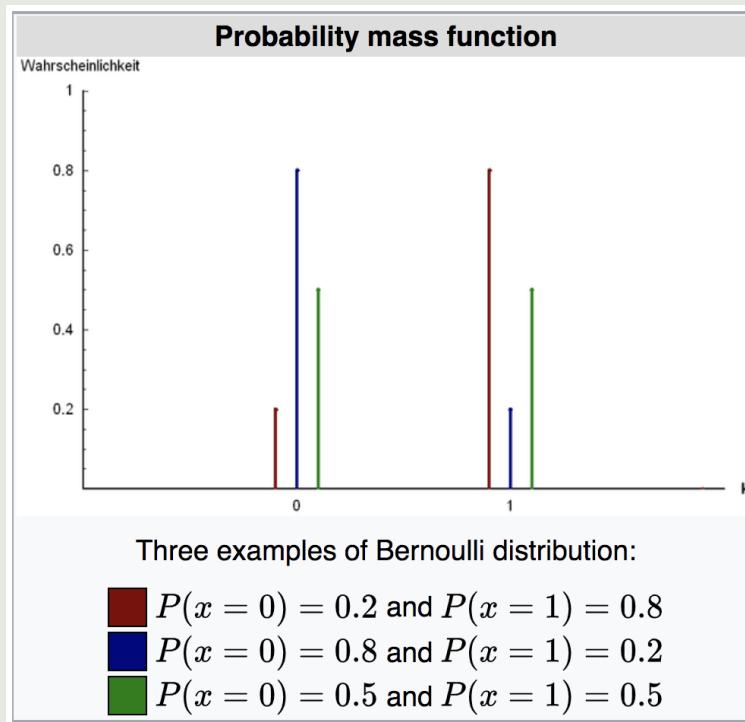
The probability that $x = \text{an exact value} = 0$ for continuous variables because $a = b$ in this integral

Cumulative Distribution Function (CDF)

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

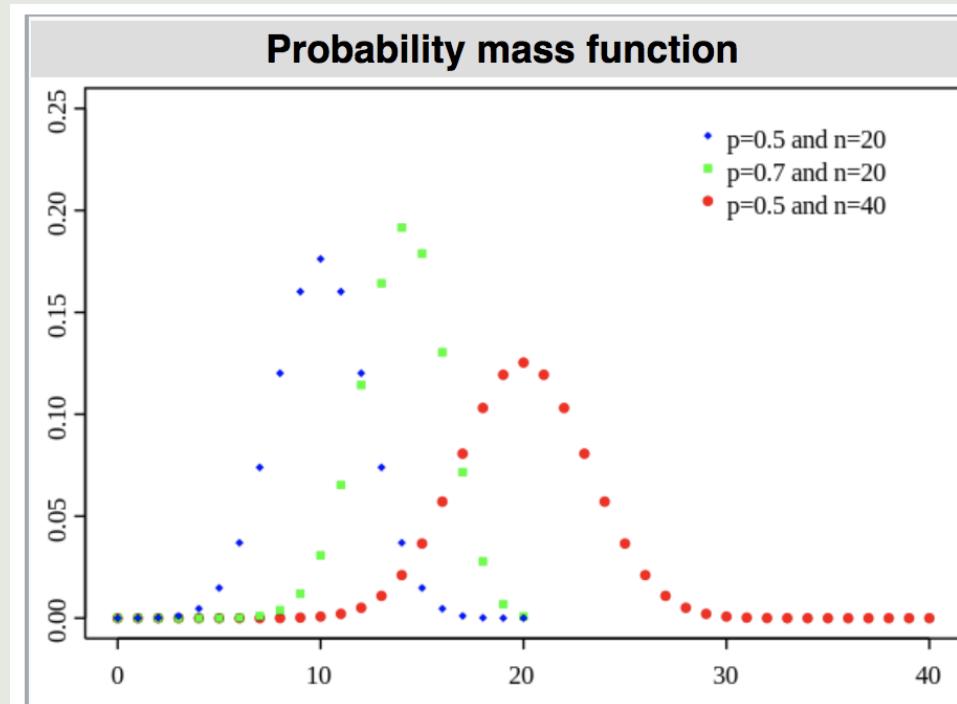
Discrete variables: Bernoulli Distribution

- 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 $(1-p)$.



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 $\text{Binom}(n,p)$



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()
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.

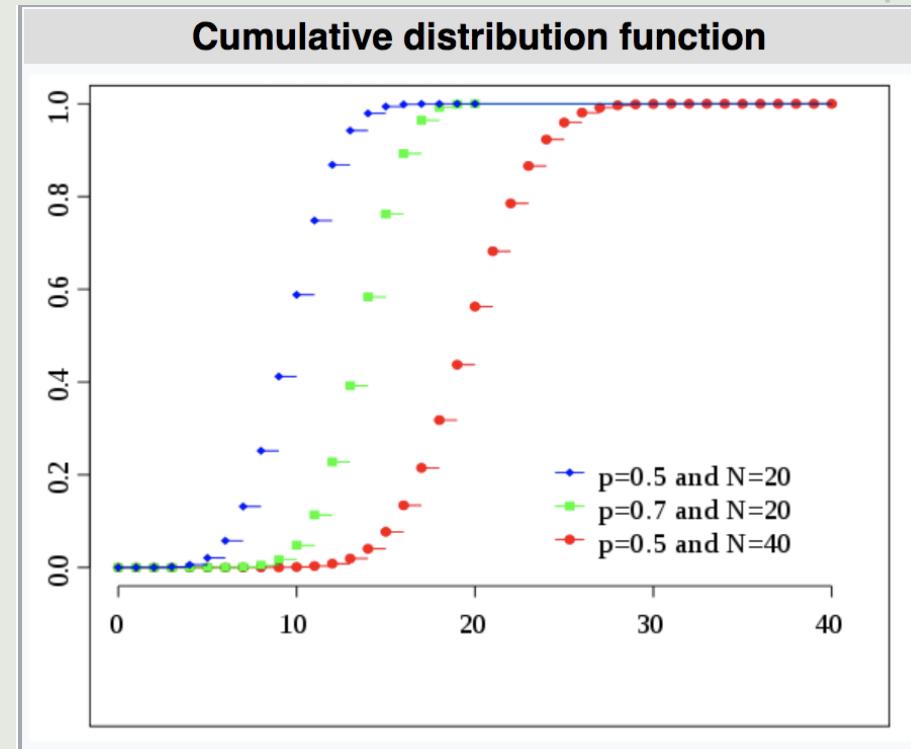
The r form is a *random number generator*: specifically, it generates n random outcomes from the distribution

10 coin tosses

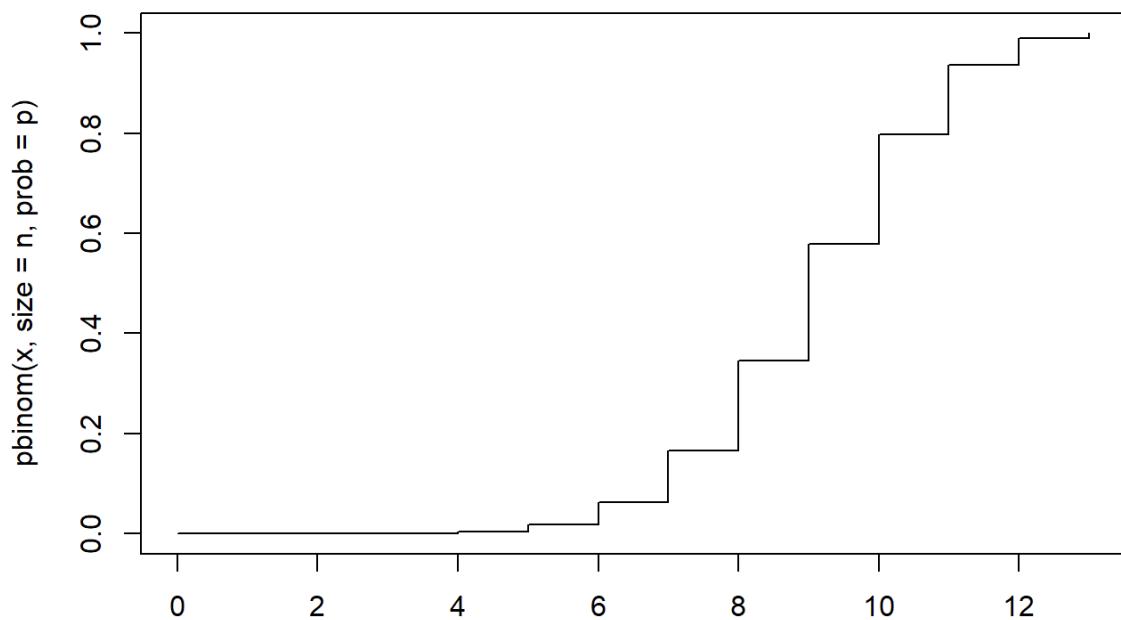
- Probability that I get ≤ 4 heads?
- $P(1) + P(2) + P(3) + P(4) = \text{dbinom}(x = 1, \text{size} = 10, \text{prob} = 0.7) + \text{dbinom}(x = 2, \text{size} = 10, \text{prob} = 0.7) + \text{dbinom}(x = 3, \text{size} = 10, \text{prob} = 0.7) + \text{dbinom}(x = 4, \text{size} = 10, \text{prob} = 0.7)$
- 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

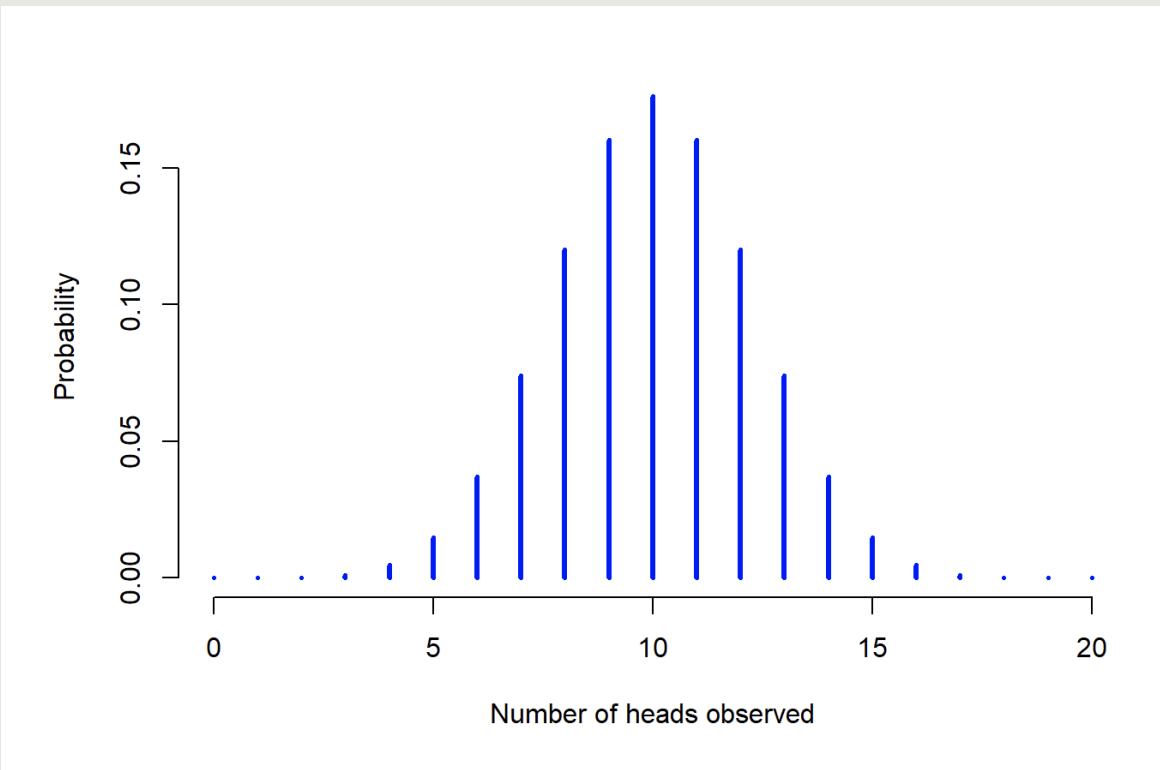
- Supported only on countable numbers
- So only some percentiles on the Y axis ->
- If you provide it any other percentile, the R function will round upwards.
- Not a problem for continuous distributions



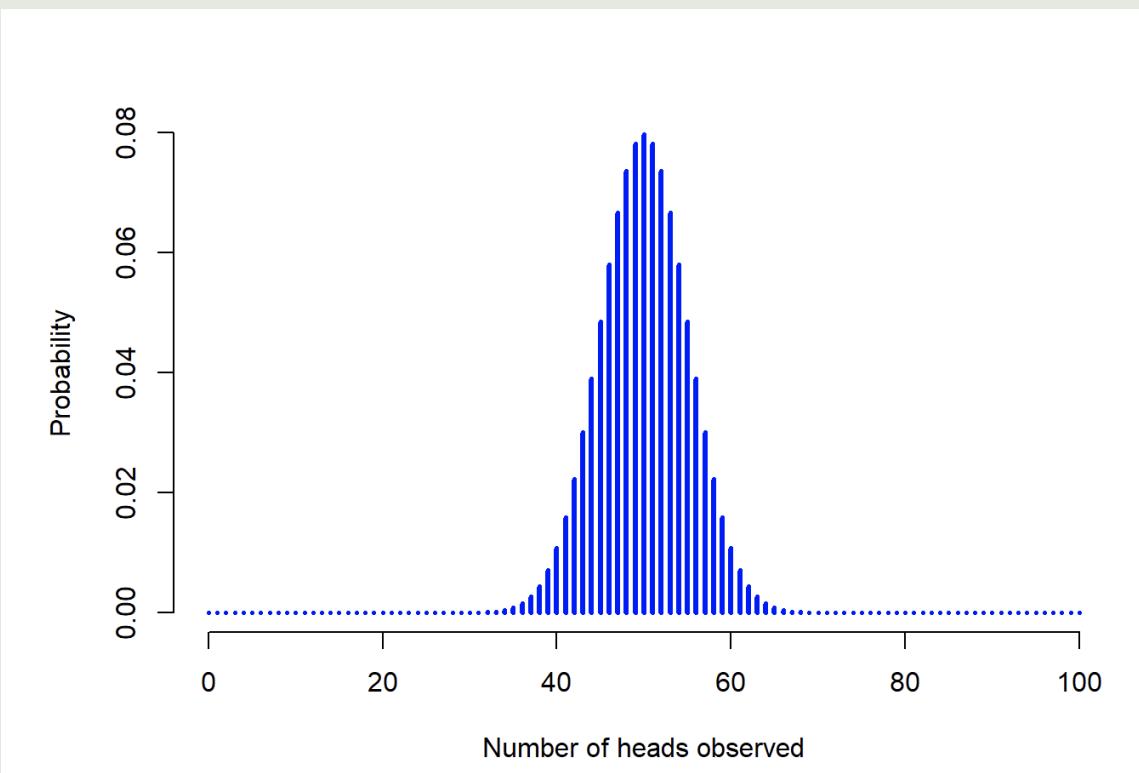
Cumulative distribution function for Bin(13,0.7)



Flip a fair coin 20 times



Flip a fair coin 100 times



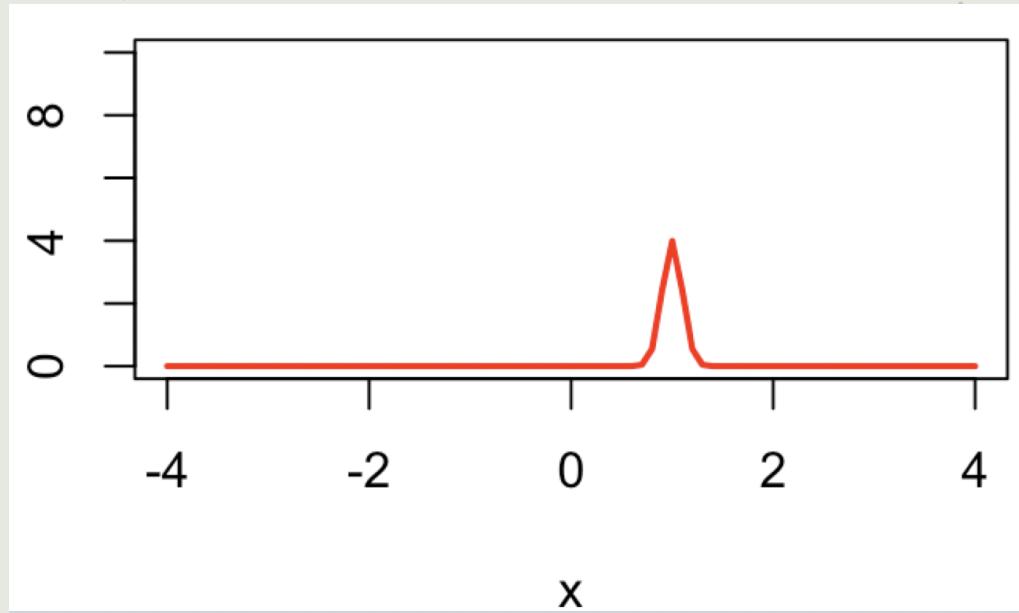
Normal Distribution

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

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")
```

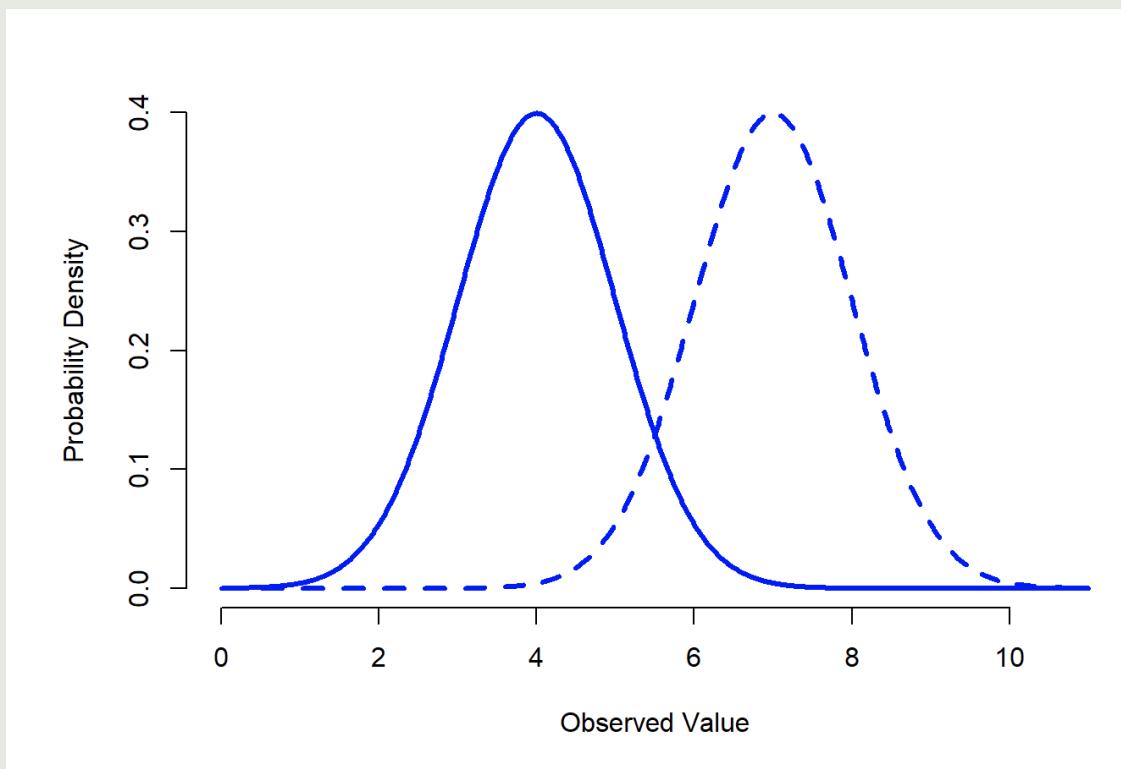


Normal PDF

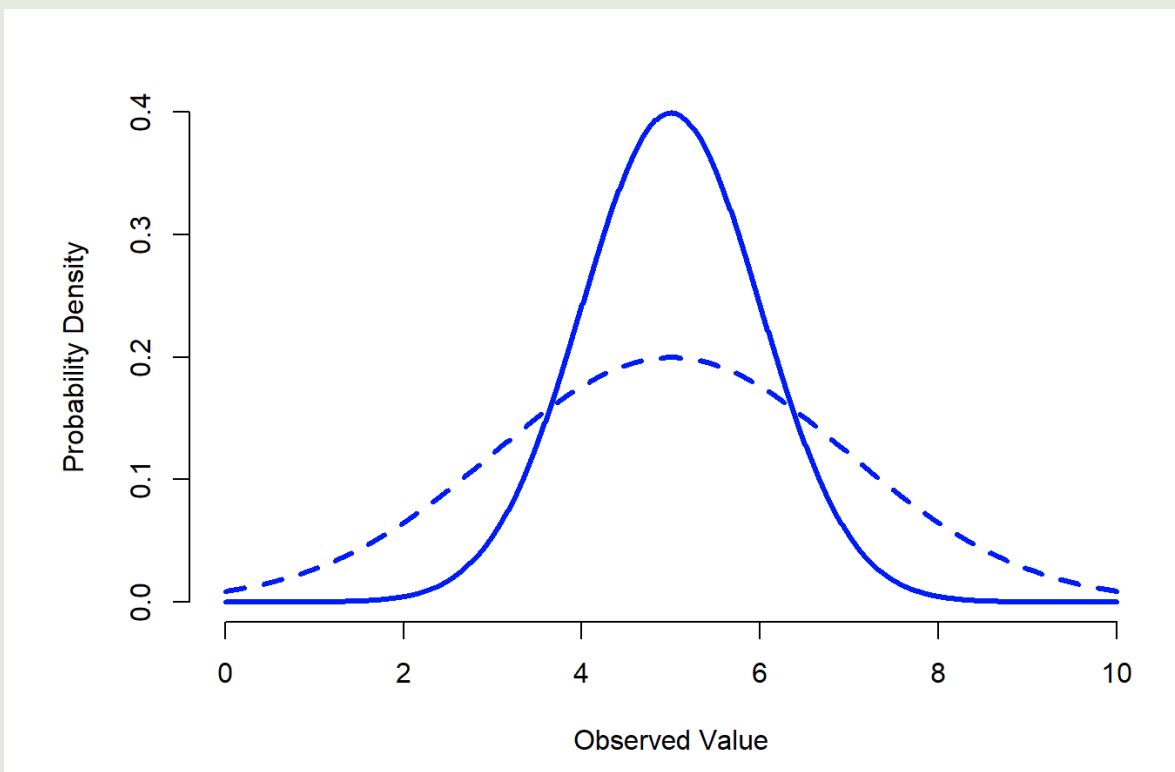
Q: What is the probability that $x = 1$?

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

Different means, same standard deviation
("width")



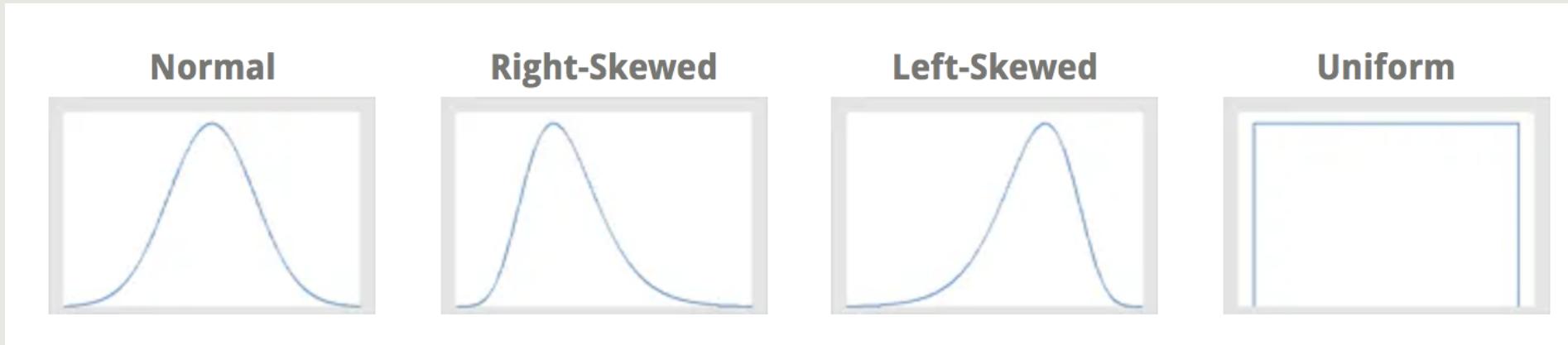
Same mean, different widths



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

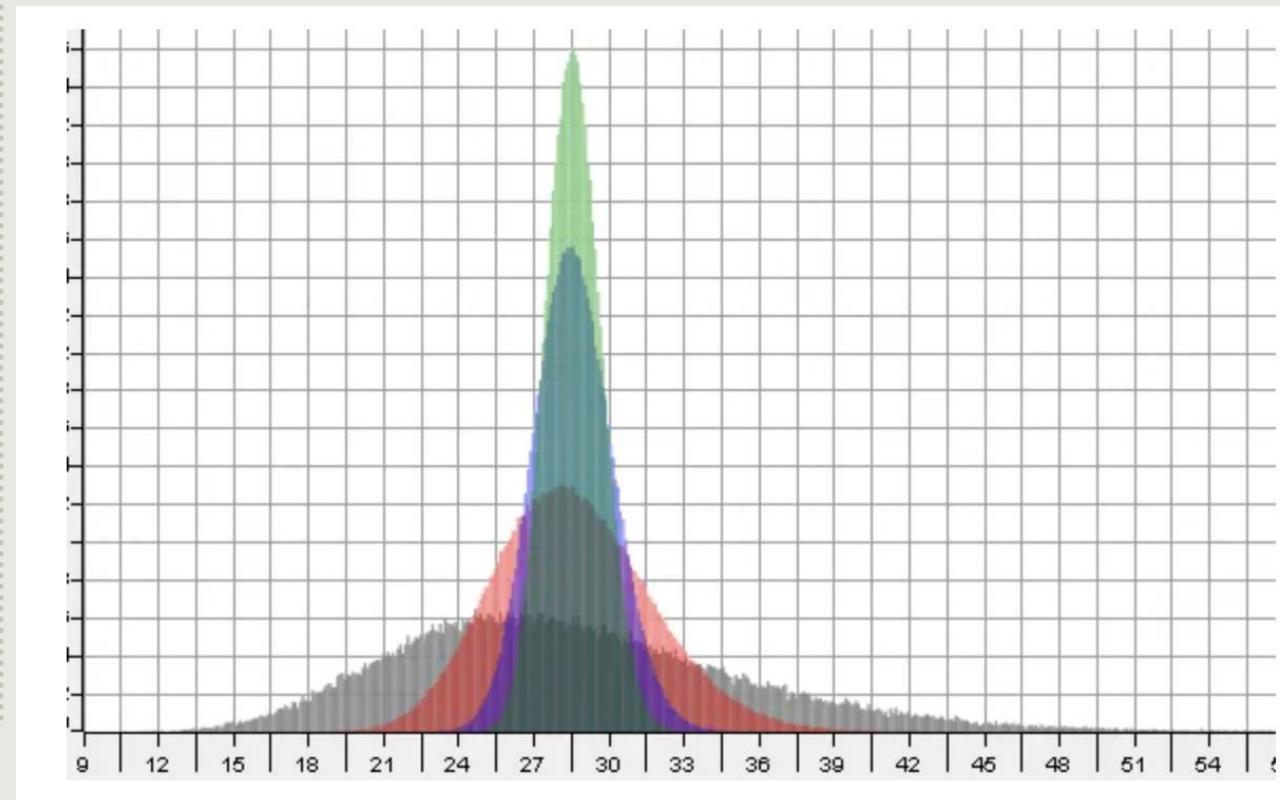
Sampling distribution of...

- 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

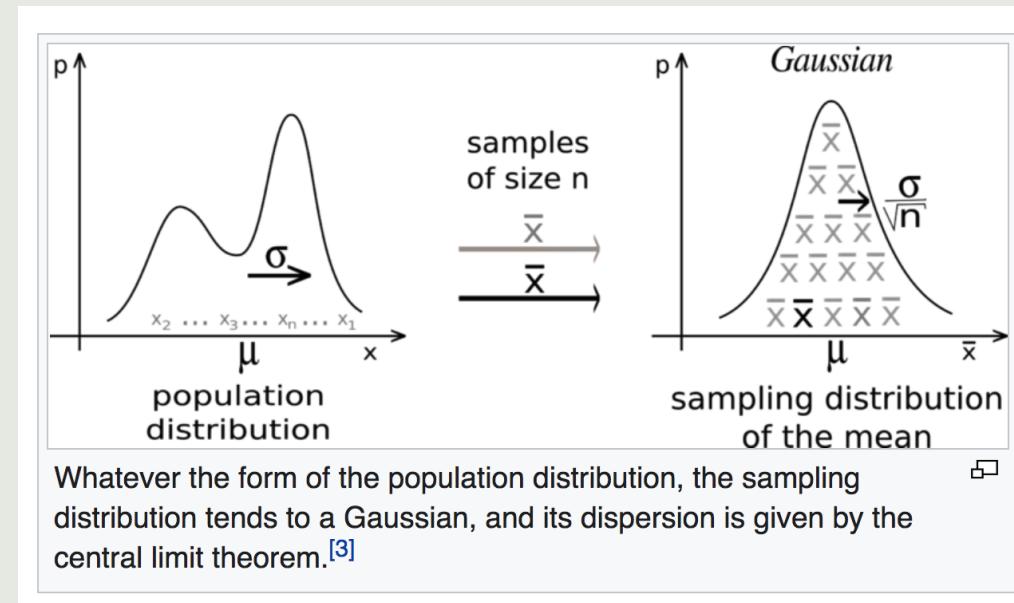
- For CLT to work, we need a sufficient sample size when we randomly draw samples **with replacement** from the population. The exact number will depend on the population distribution. Skewed distributions tend to need higher n.
- The sample mean will be equal to the population mean

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



Lindeberg–Lévy CLT. Suppose $\{X_1, \dots, X_n\}$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2 < \infty$. 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]

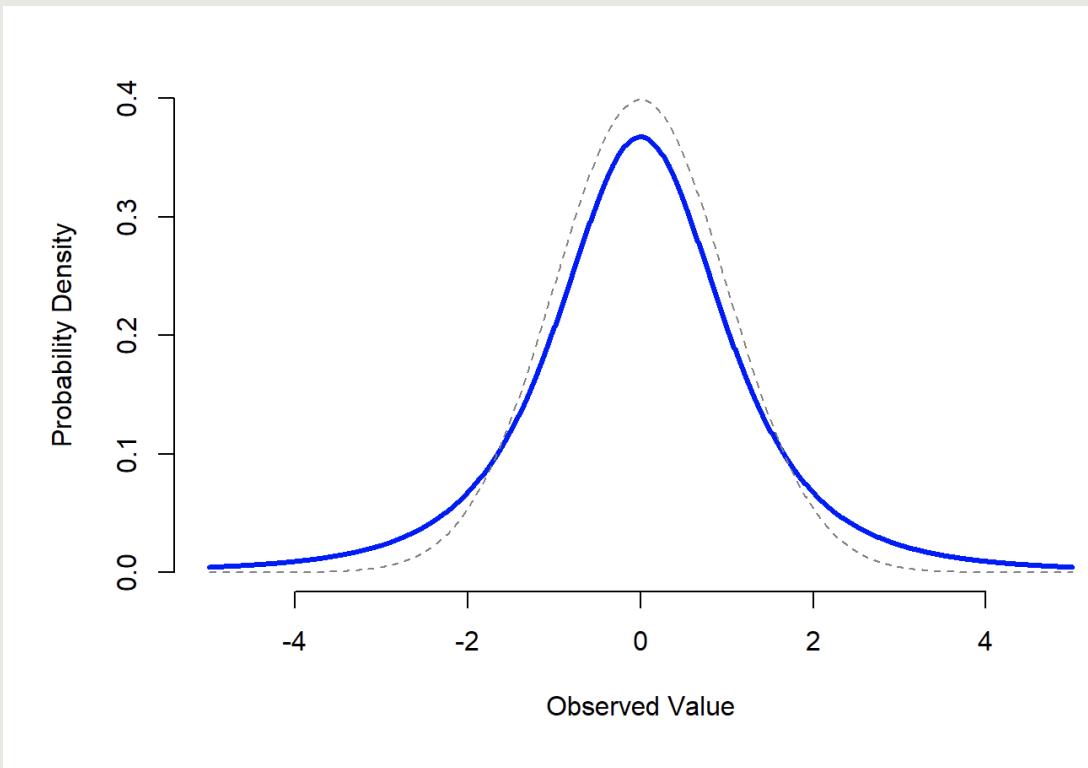
$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$



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

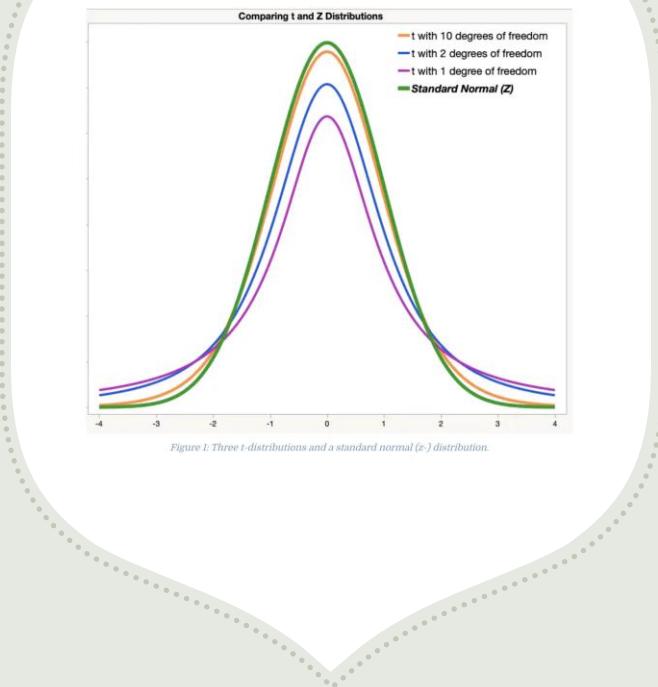


Heavy-tailed

Arises in smaller n situations and when you don't know the population s.d.
As $n \rightarrow \infty$, t-distribution begins to look more like a Normal.

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).



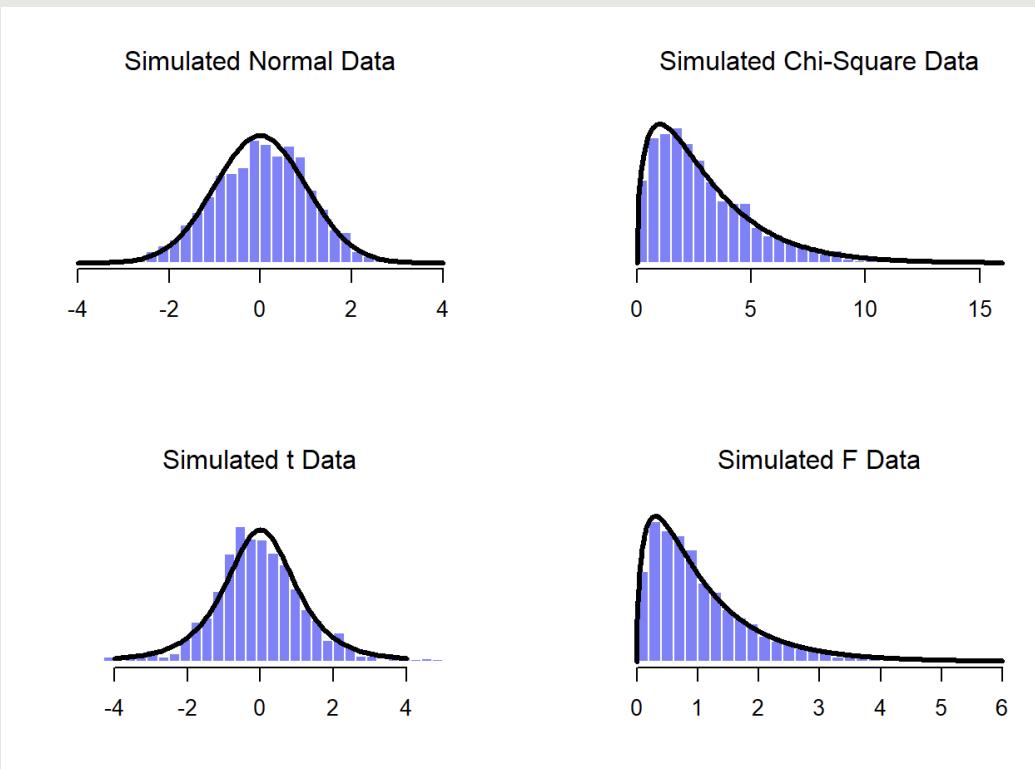
T-distributions and k

The use of t-distributions later

Suppose $x_i \sim N(\mu, \sigma^2)$ 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
```

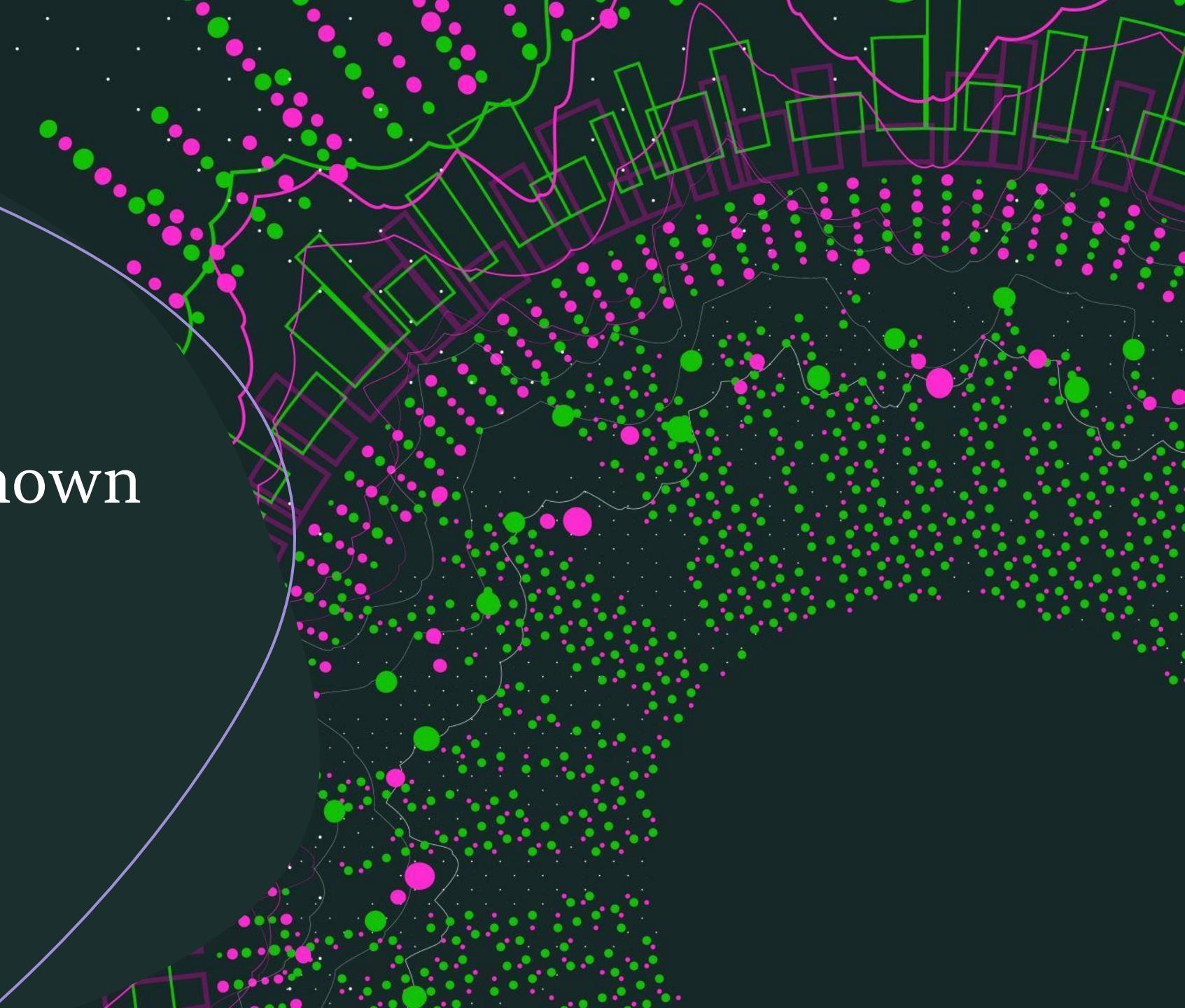
```
normal.c <- rnorm( n=1000 ) # and another!
```

```
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



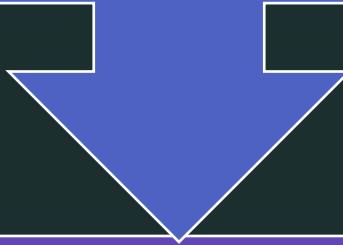
Population: cogsci questions

- 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

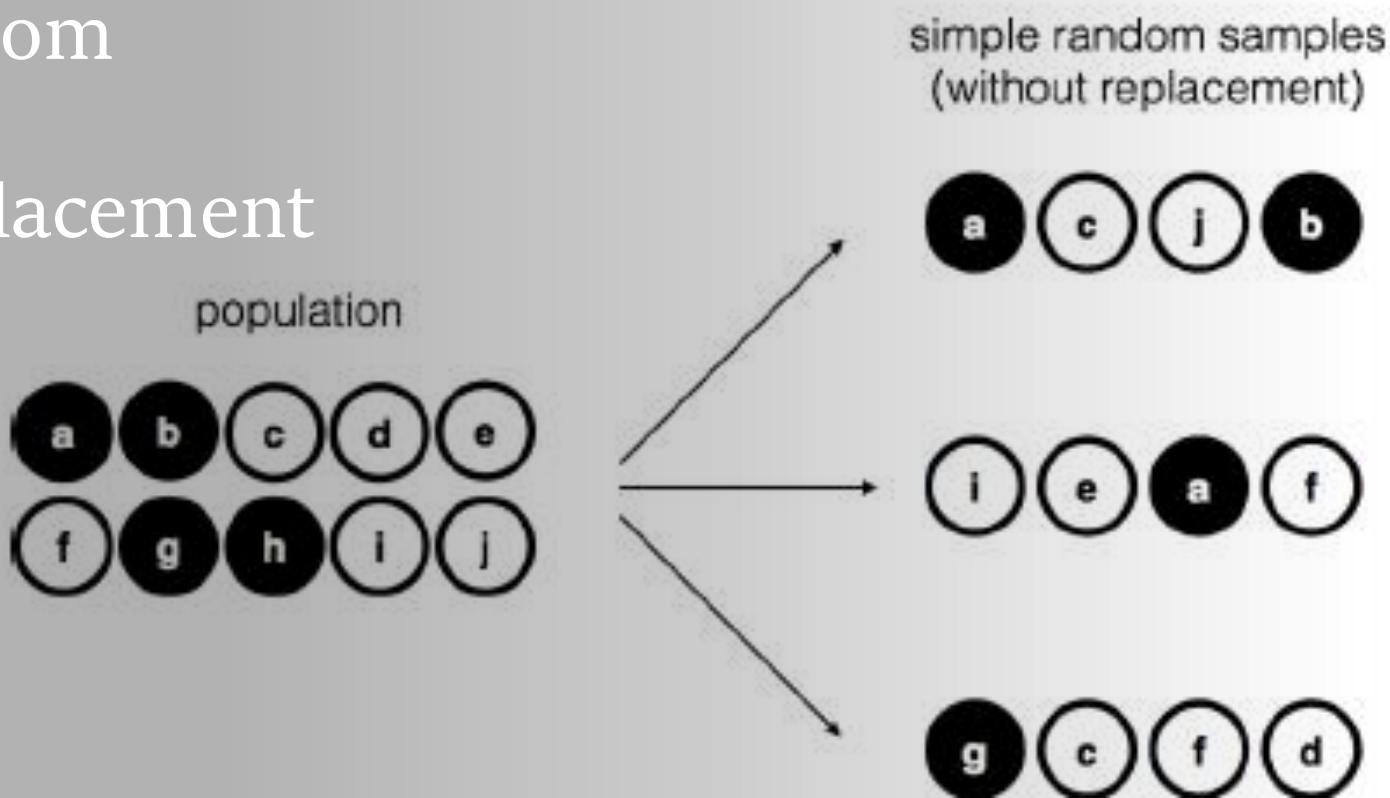


Figure 10.1: Simple random sampling without replacement from a finite population

Biased sampling without replacement

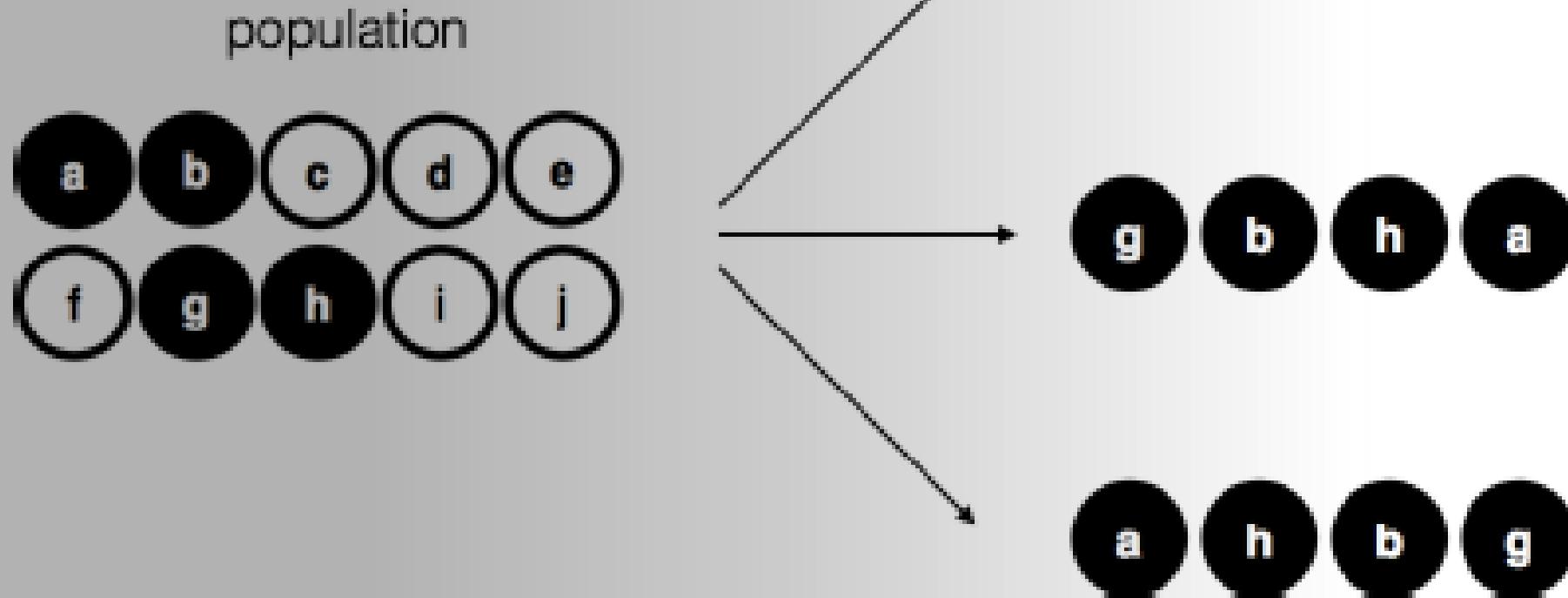
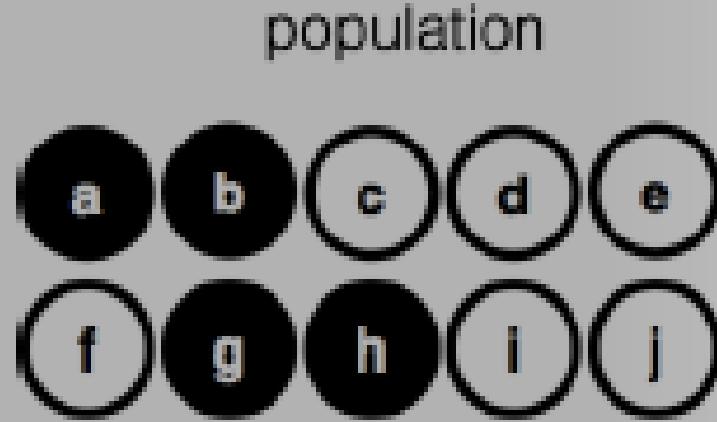


Figure 10.2: Biased sampling without replacement from a finite population

Simple random sampling with replacement



simple random samples
(with replacement)

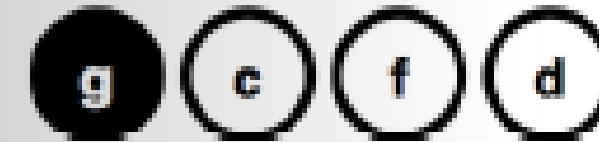
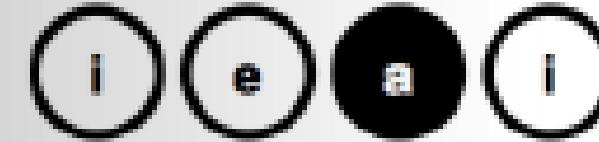


Figure 10.3: Simple random sampling *with replacement* from a finite population

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?

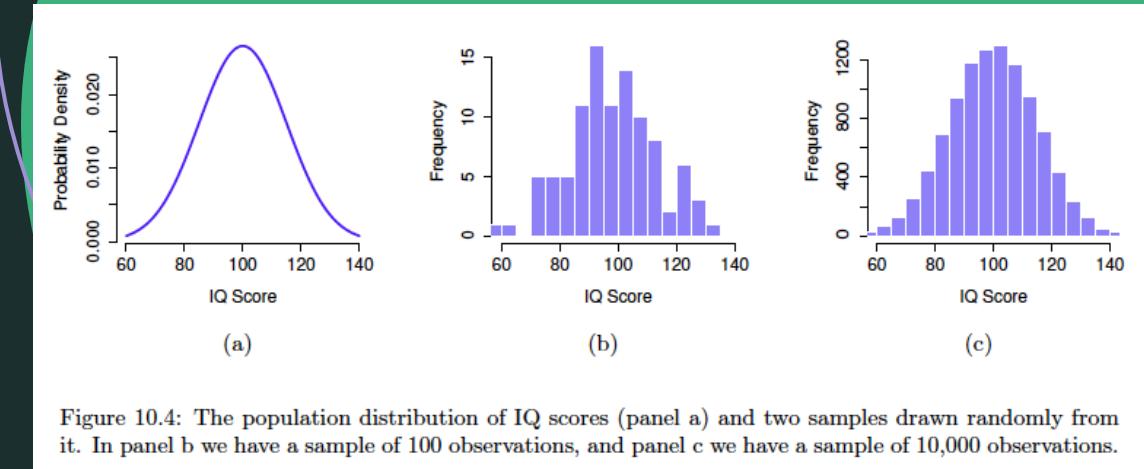


Figure 10.4: The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and panel c we have a sample of 10,000 observations.

The law of large numbers

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

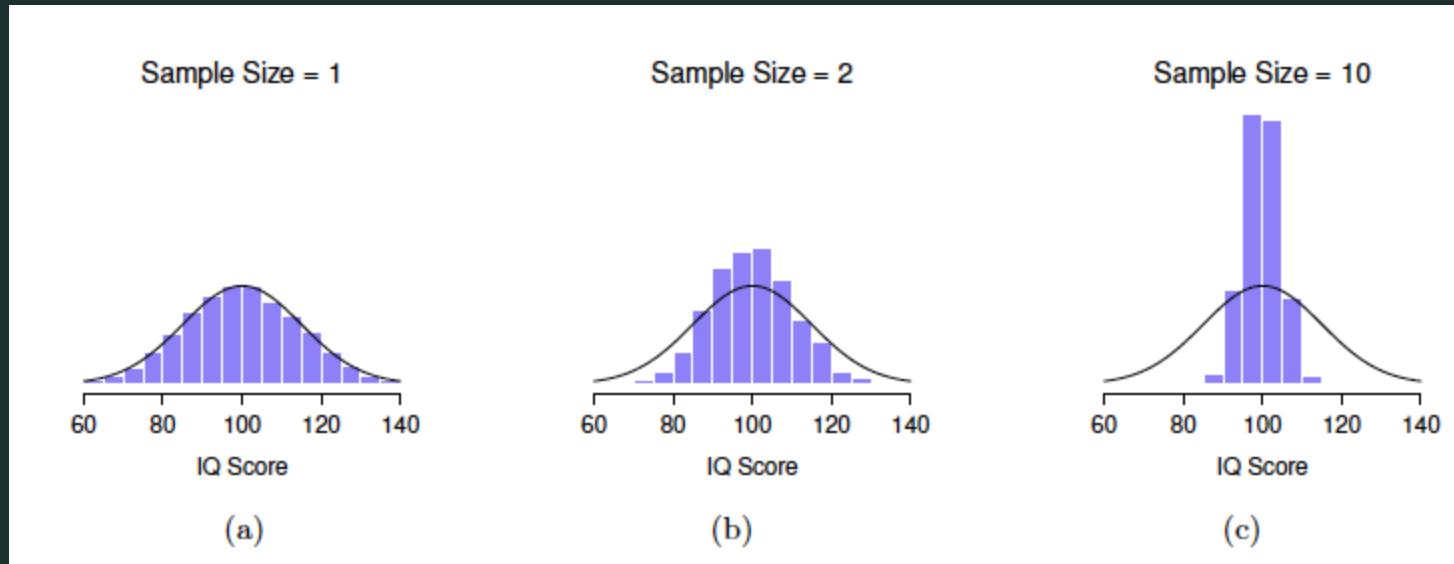
This law of large numbers applies to many statistics, but easiest to demonstrate is as a law of averages (sampling distribution of the mean, which we saw in the probability distribution lecture)

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 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

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

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

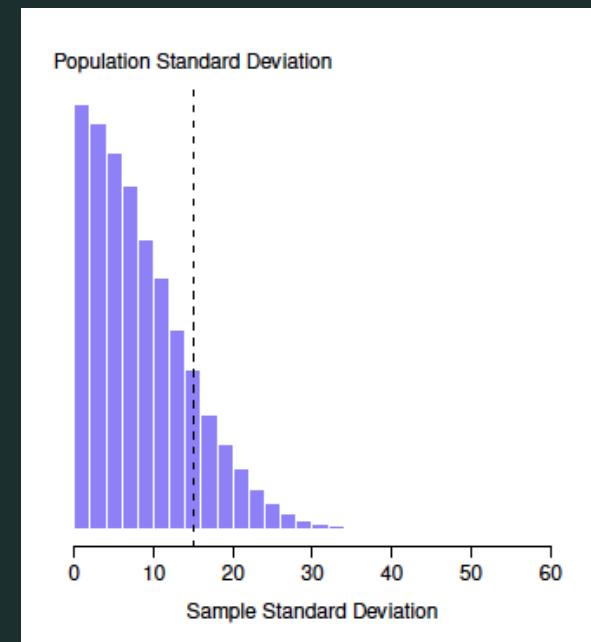
Remember: this only works when CLT applies, so need a sufficiently large sample size, otherwise your estimate will not be very accurate

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

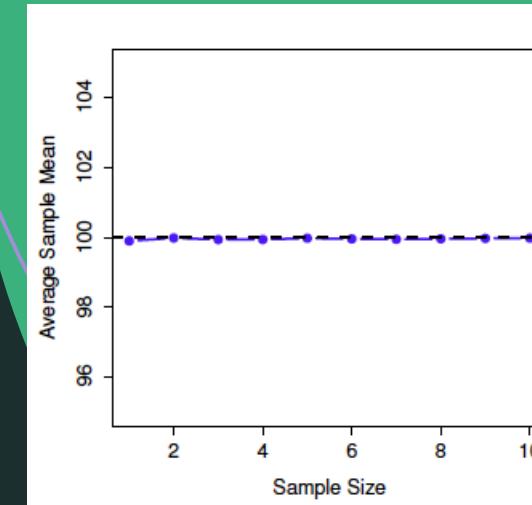
How about population standard deviation?

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

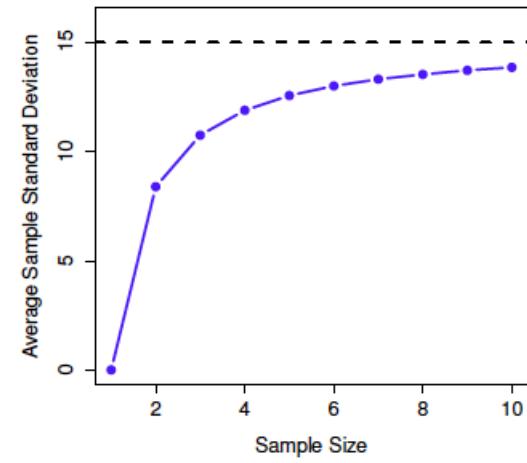


Intuition: Sample s.d. is a biased estimator of population s.d.

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



(a)



(b)



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

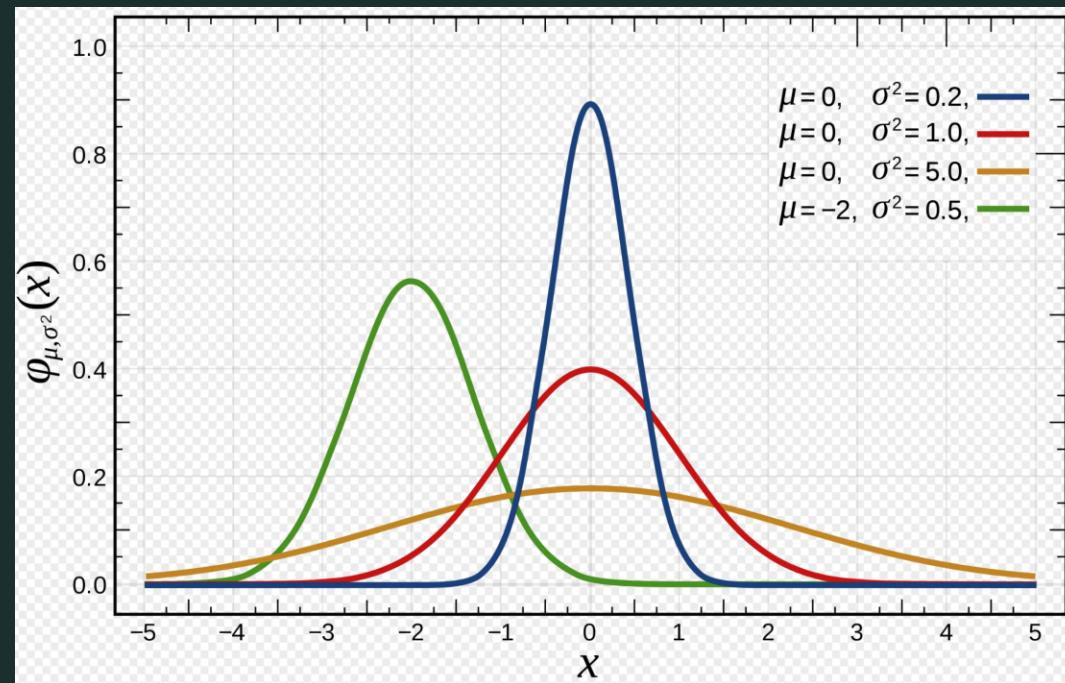
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Mu = 0, var = 1

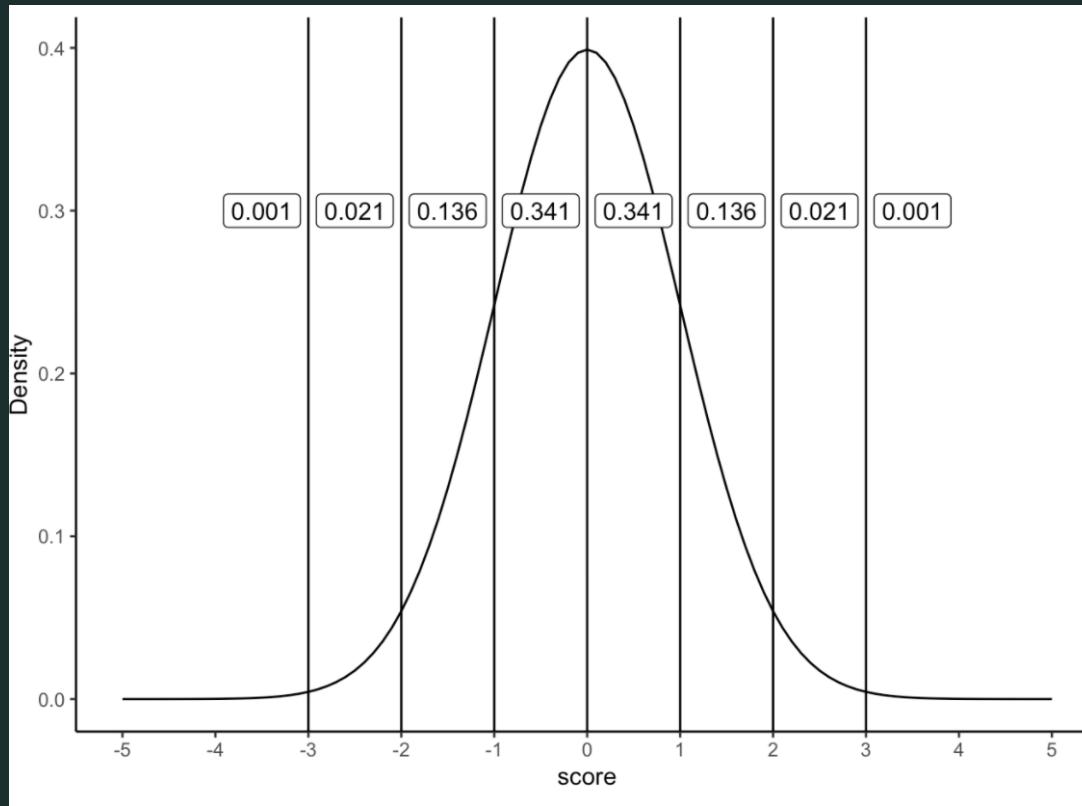


$$\varphi(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$$

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

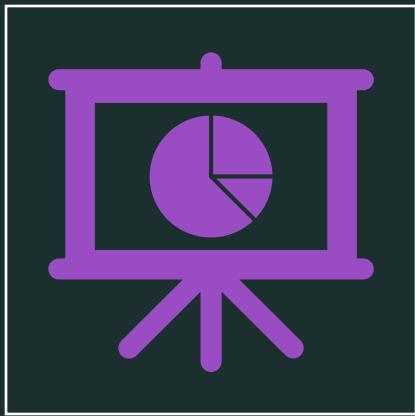
95% of the values lie within 2 s.d. of the mean

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 = μ and s.d. = σ
- 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?

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

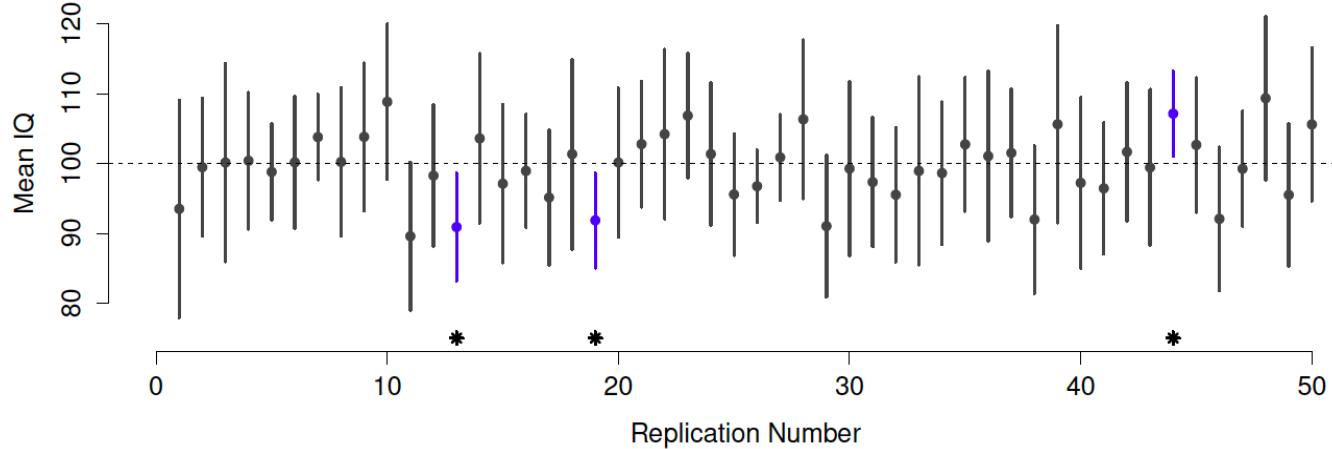
$$\bar{X} - (1.96 \times \text{SEM}) \leq \mu \leq \bar{X} + (1.96 \times \text{SEM})$$

- 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

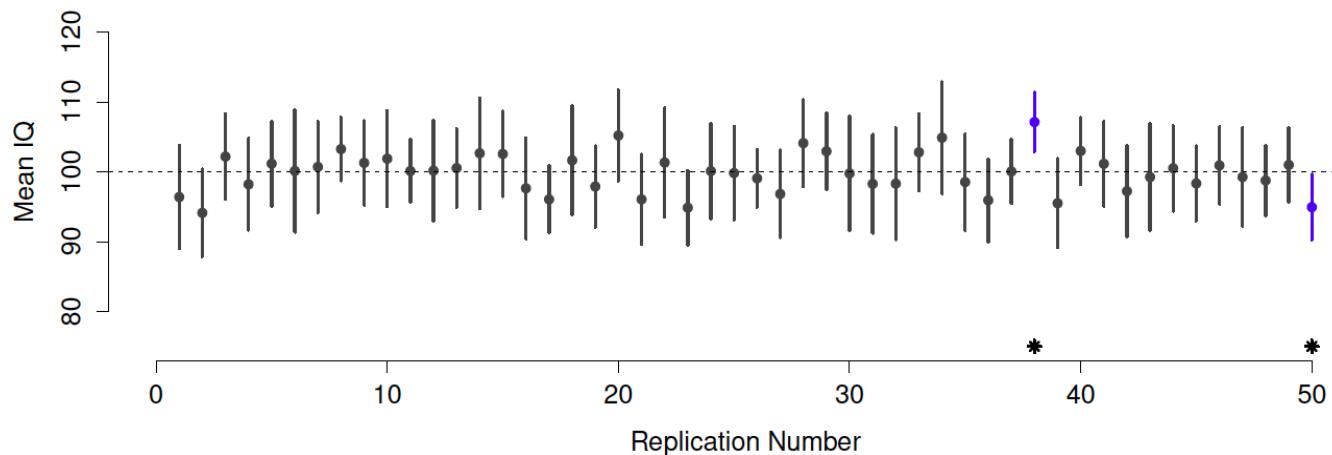
The confidence interval changes from sample to sample

- 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



Sample Size = 25



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

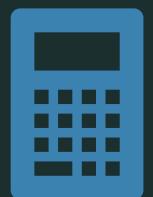
```
N <- 10 # suppose our sample size is 10  
qt( p = .975, df = N-1) # calculate the 97.5th quantile of  
the t-dist
```

[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



Estimating means and standard deviations



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



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

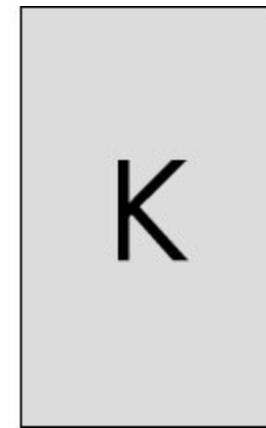
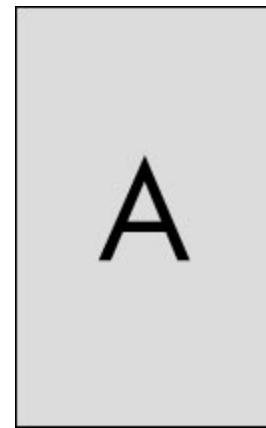
The importance of falsifying in theory testing

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

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

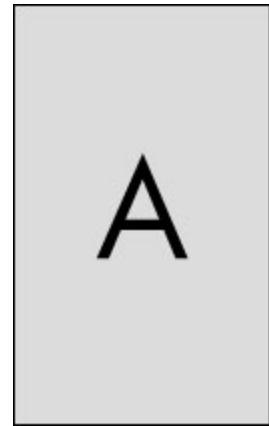
- A. Theory proven?
- B. Theory not disproven?

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]



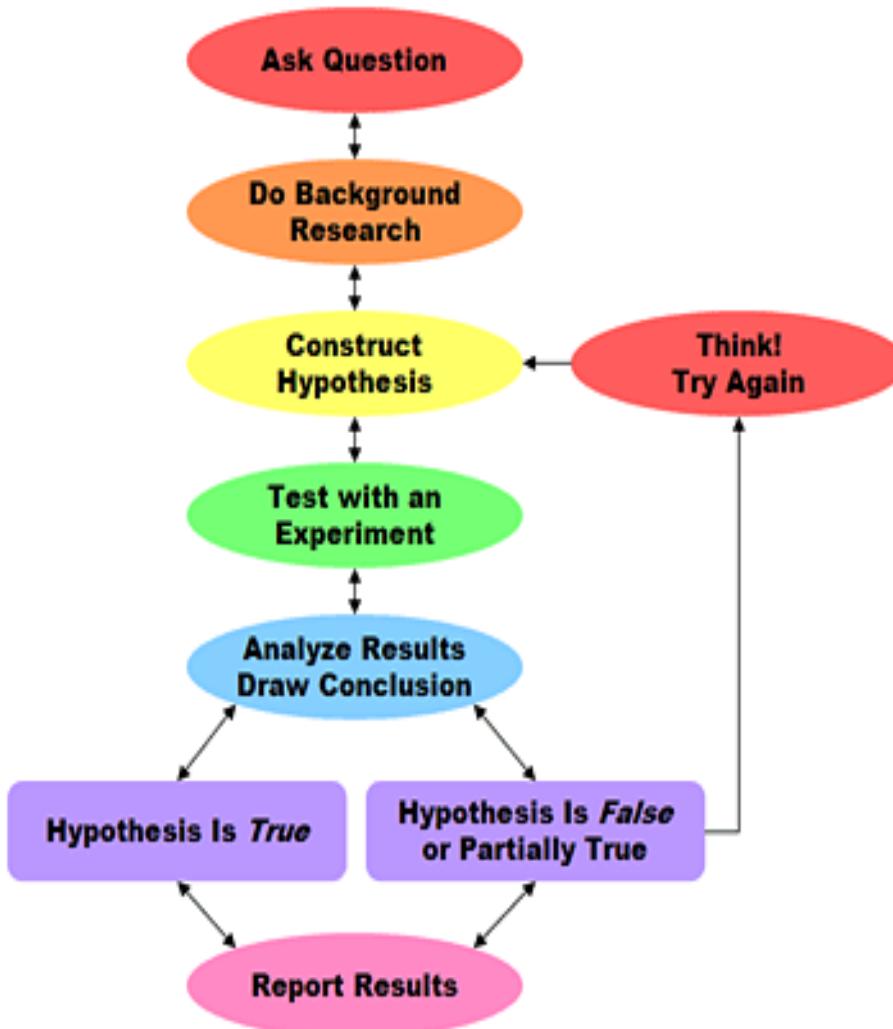
[L]

- 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.

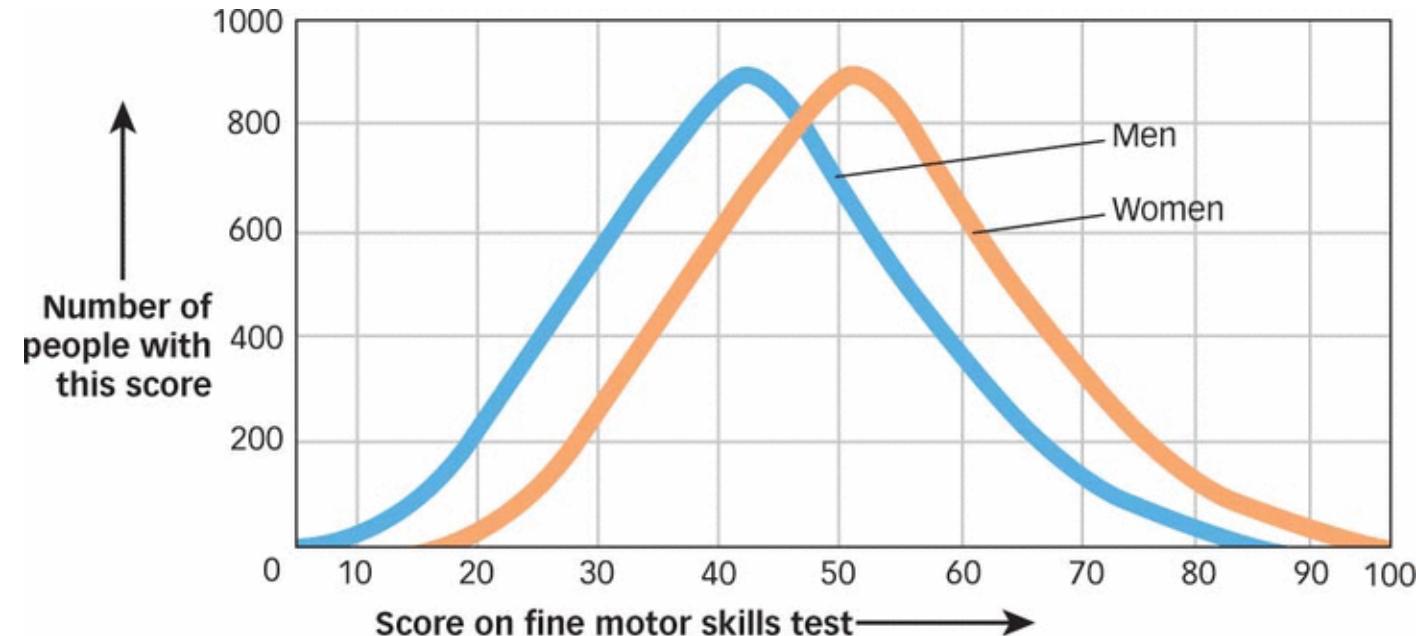
The Scientific Method – A Review

- 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 $\alpha = 0.05$?

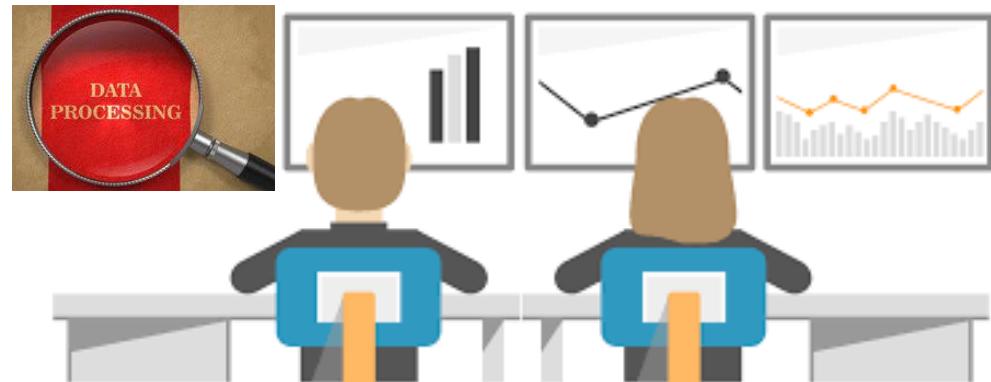
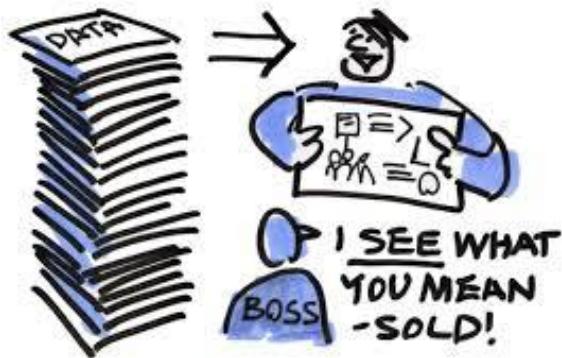
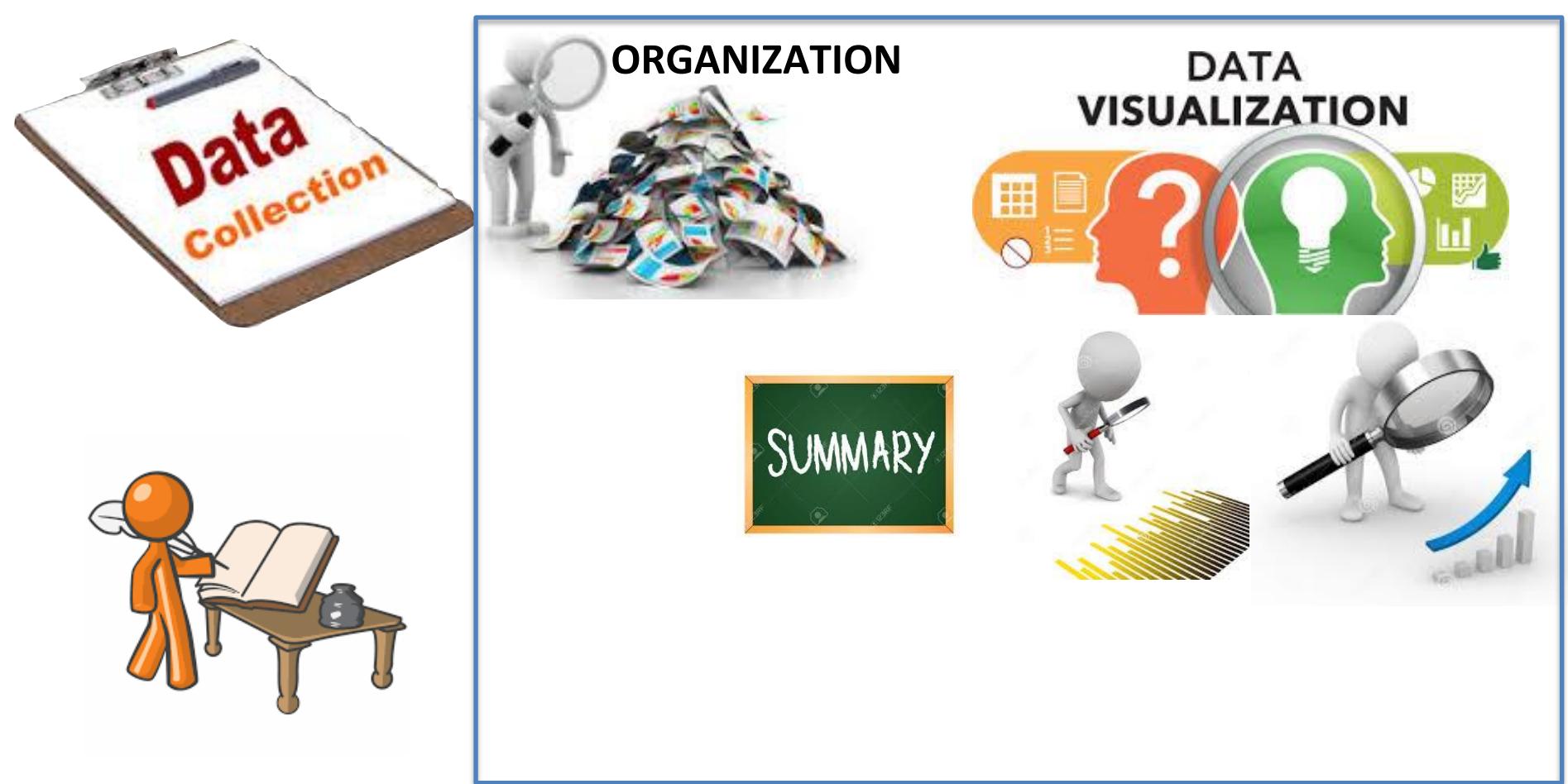
Choose 1 answer:

-
- A Fail to reject H_0
 - B Reject H_0 and accept H_a
 - C Accept H_0
-

BRSM

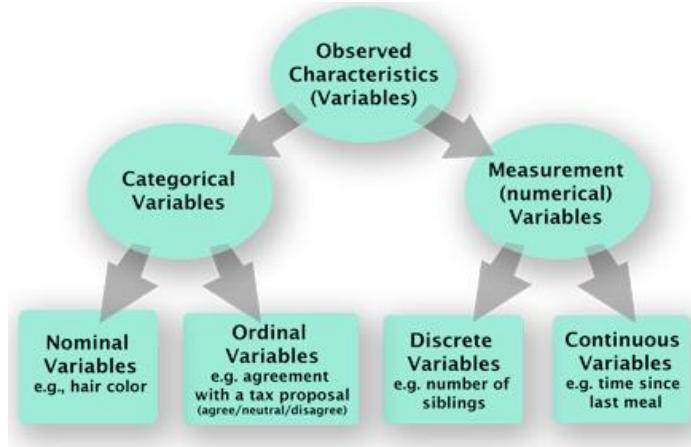
Data Visualization & Summarization

Vinoo Alluri & Bapi Raju





Data Organization



- identify variables (IV, DV) and respective types
- identify different levels of measurement

- missing data?
 - replace with mean
 - remove

20-25 years = 1
26-30 years = 2
31-35 years = 3
36-40 years = 4
41-45 years = 5
46 years and older= 6

Continuous



Categorical

Table format: XY

	X	A		
	minutes	Test group A		
	X	A:Y1	A:Y2	A:Y3
1	Title	0	0.0	0.0
2	Title	2	3	5.611248
3	Title	4	2	5.5560017
4	Title	5	3	4.5405
5	Title	6	4	5.236287
6	Title	7	5	5.9417286
7	Title	8	6	5.4199543
8	Title	9	7	4.4019384
9	Title	10	8	5.1843286
10	Title	11	9	5.3209386
11	Title	12	10	3.9951186
				5.300459
				5.080454
				5.2821956
				5.1487226
				5.308297
				5.6112976
				5.461459
				4.35598
				5.340737
				4.5518494
				6.075916
				5.3088202
				5.288509
				5.5335727



Summarize

to tell, in your own words,
what has happened in the



Summarize

How?



What information does it give ???

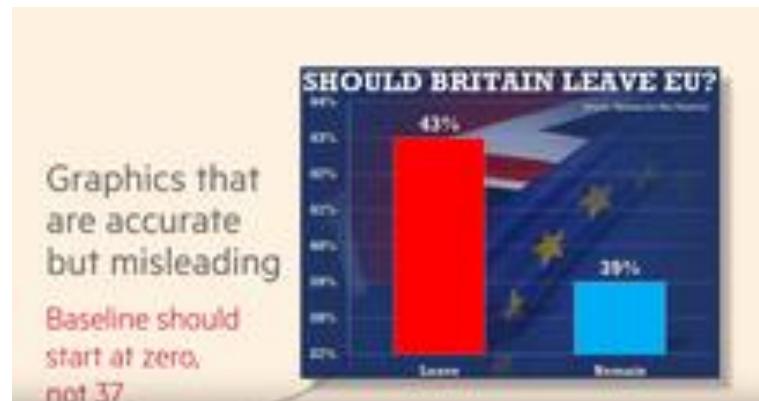
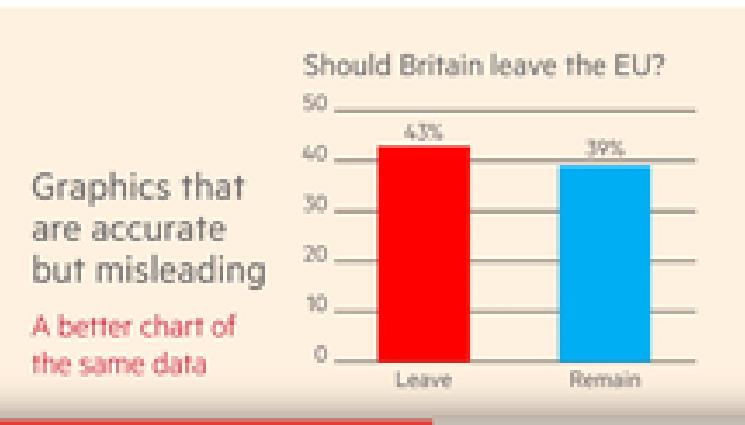
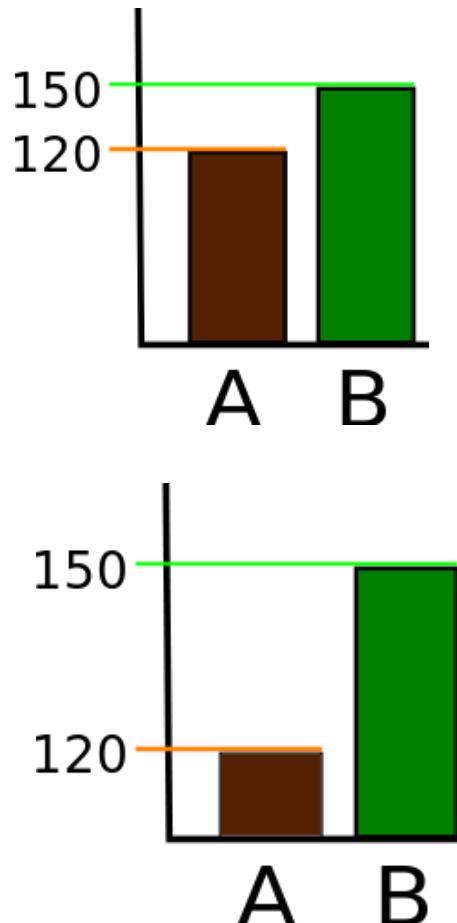


Outline

- **Visualization**
 - why we visualise
 - how to pick a plot
 - initial data vs final results visualization (some examples)
 - bad designs and misleading graphs
- **Summarization**
 - measures of central tendency & dispersion
 - which measure to pick

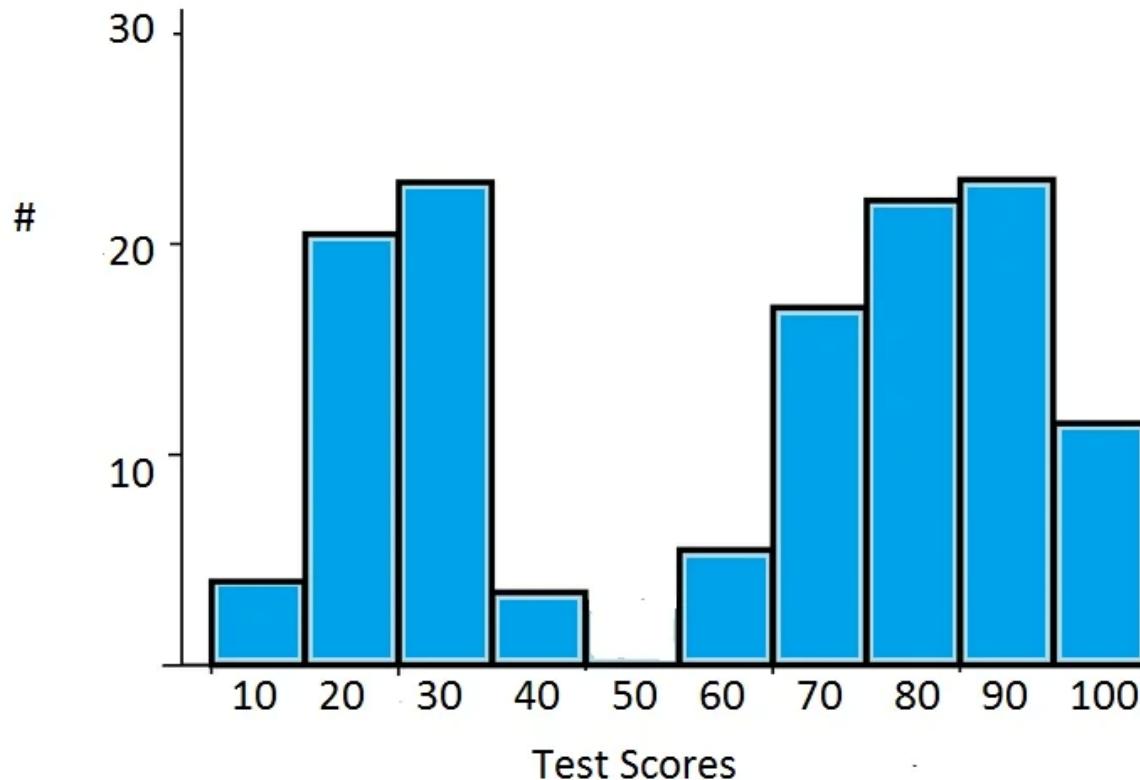


Data Visualisation



EXAMPLE

Mean Mid-Sem Test Score = 65.5



How can i summarise this data?



Anscombe's Quartet

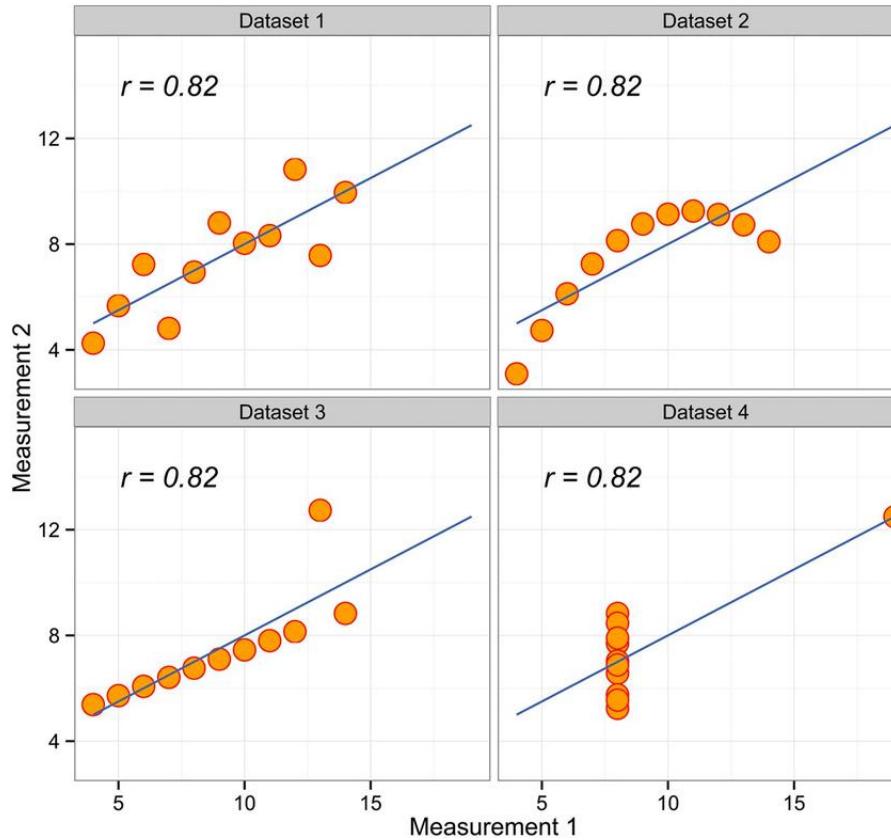
- same mean, std, correlation, regression line

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	5.76
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	8.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	7.26	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Anscombe's Quartet

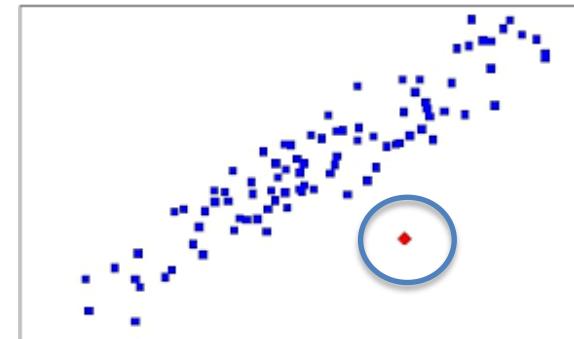
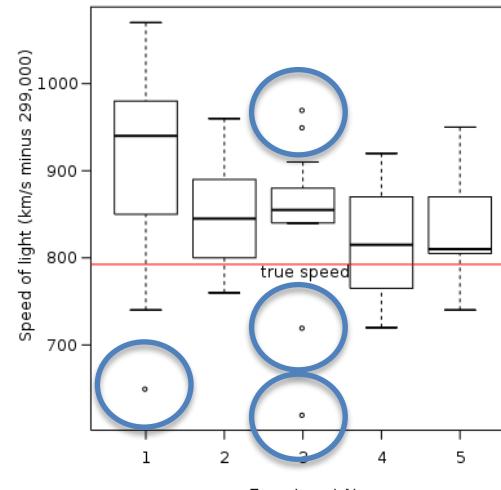
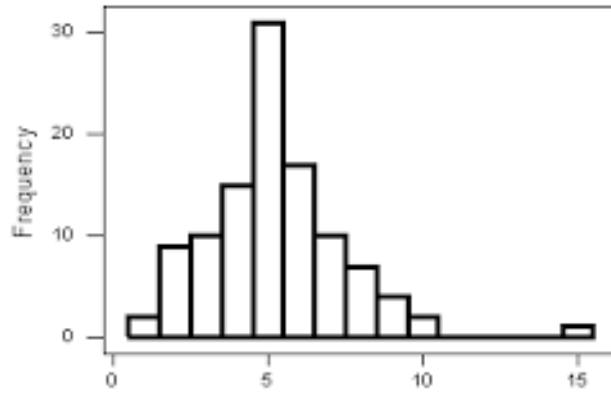
- same mean, std, correlation, regression line





Why do we visualise?

- allows for initial guesses of data distribution
- direction of effect
- error detection (eg: missing, NaNs)
- outlier detection
- present results

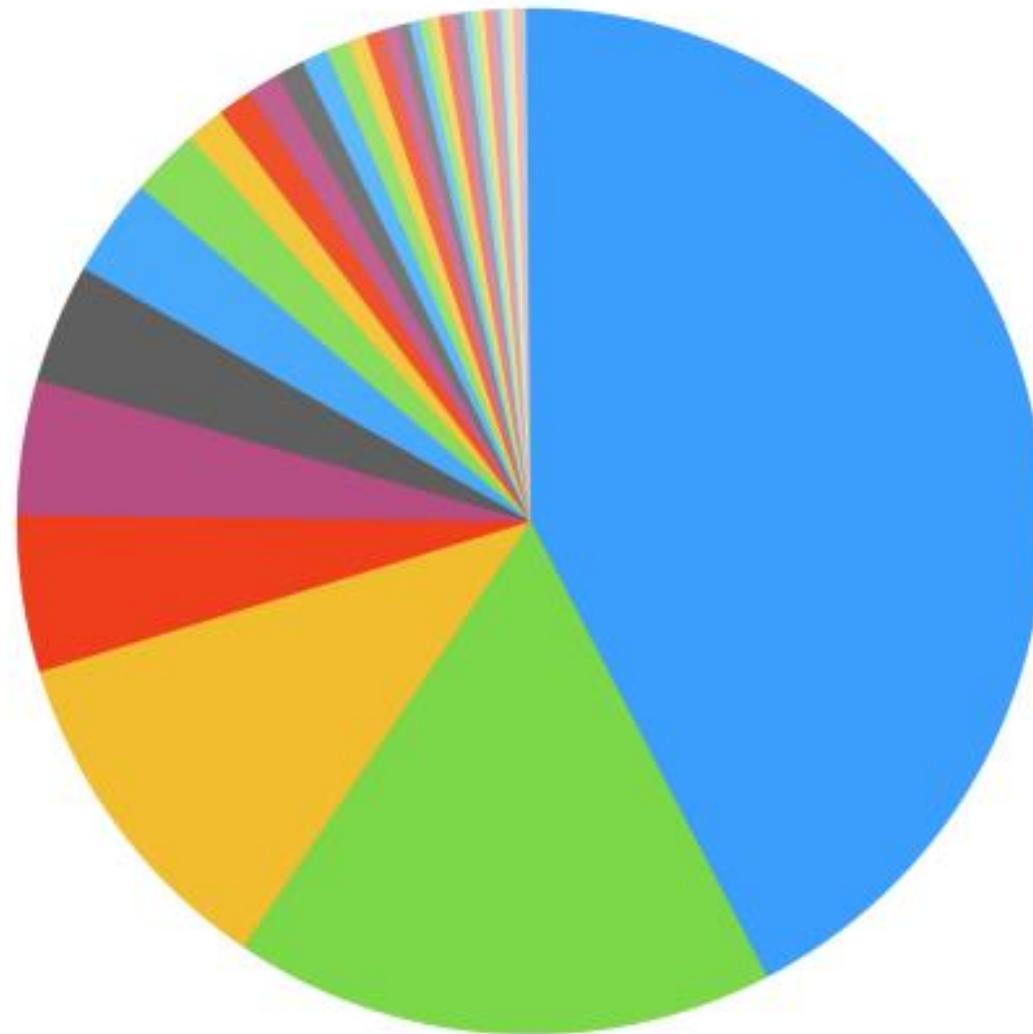


what makes them “good” or “bad”?

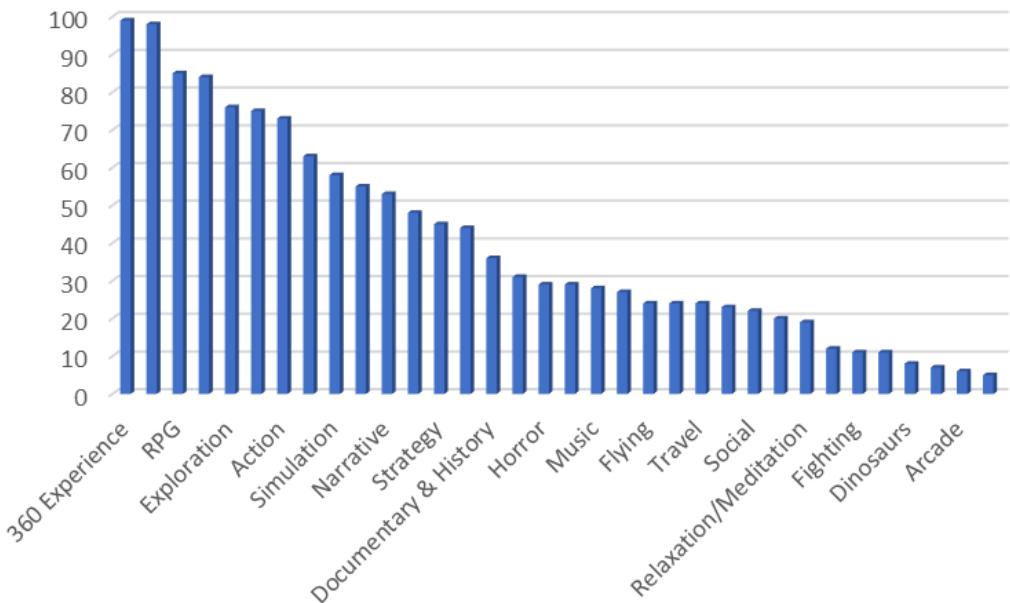
comment on these visualizations

Which game(s) have you played the most?

3,994 responses



- Zelda
- The Legend of Zelda: Breath of the Wild
- Breath of the Wild
- BOTW
- Botw
- Breath of the wild
- BotW
- zelda
- Legend of Zelda: Breath of the Wild
- Legend of Zelda
- Zelda BOTW
- BoTW
- botw
- Zelda: Breath of the Wild
- Zelda BotW
- Zelda Breath of the Wild
- The Legend of Zelda
- Breath of The Wild
- The Legend of Zelda Breath of the Wild
- Zelda: BOTW
- Zelda: BotW
- Breath of the Wild
- Zelda breath of the wild
- Breath Of The Wild
- Legend of Zelda Breath of the Wild
- LoZ
- LoZ: BotW
- Zelda botw
- zelda botw
- breath of the wild
- Legend of zelda
- legend of zelda
- LoZ BOTW
- The Legend of Zelda: Breath of The Wild
- The legend of Zelda: breath of the wild
- ZELDA
- Zelda: BoTW



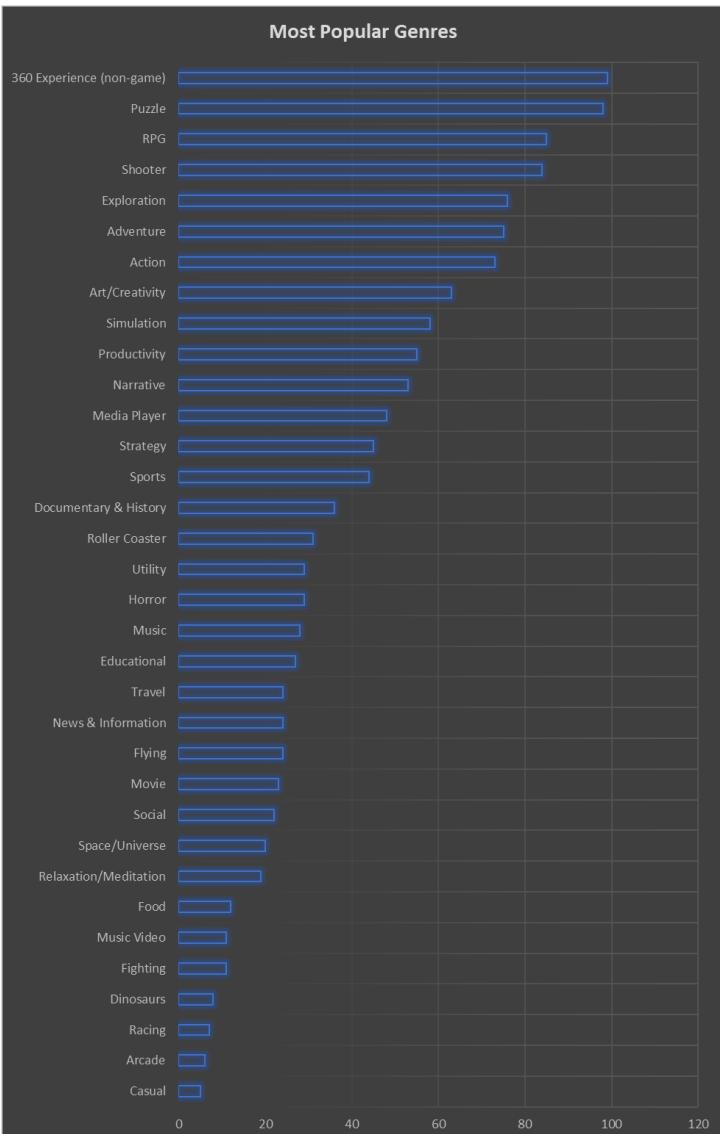
MOST WICKETS IN DEATH OVERS IN ODIS

SINCE THE START OF JANUARY 2017

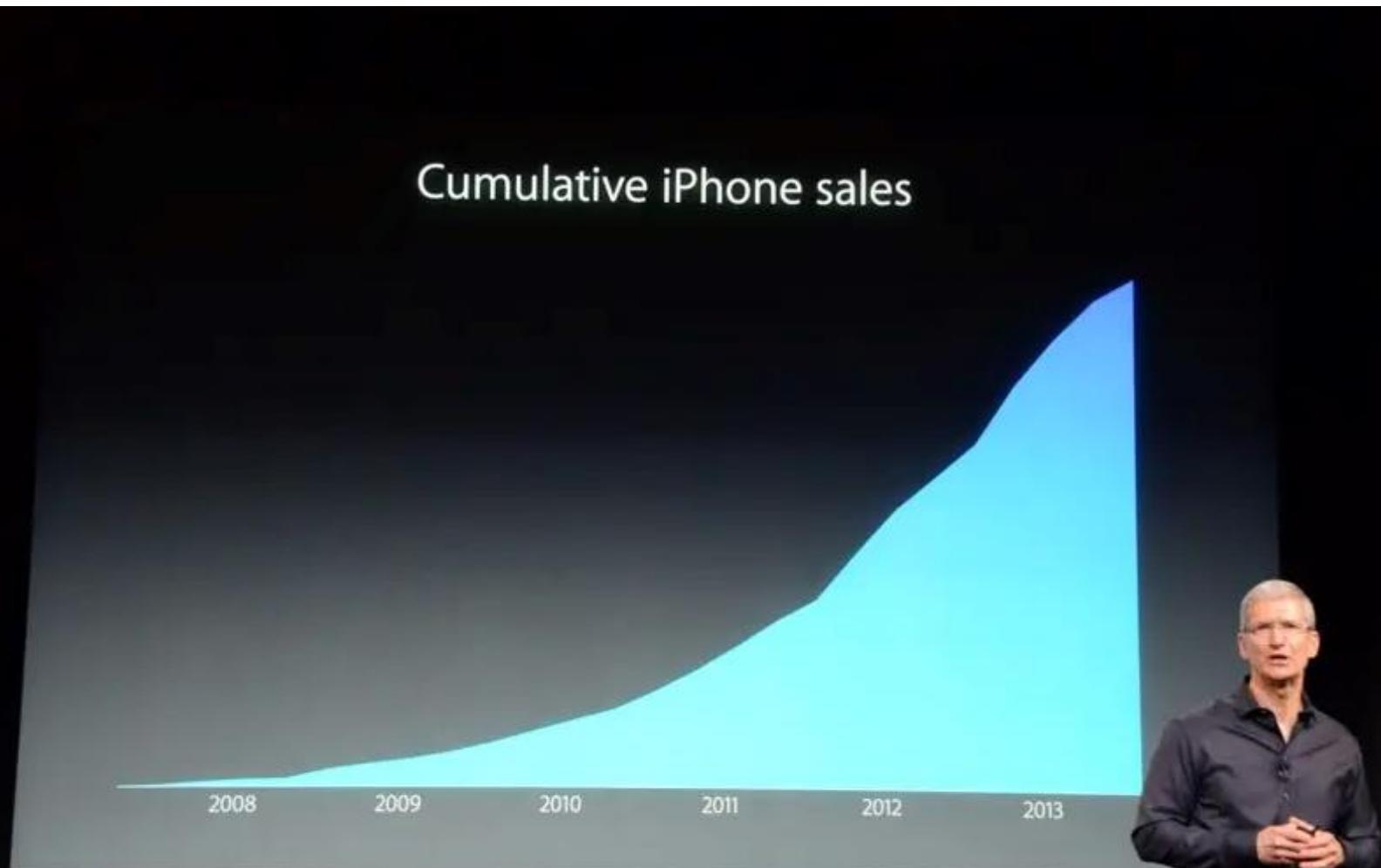
■ WKTS ■ AVE

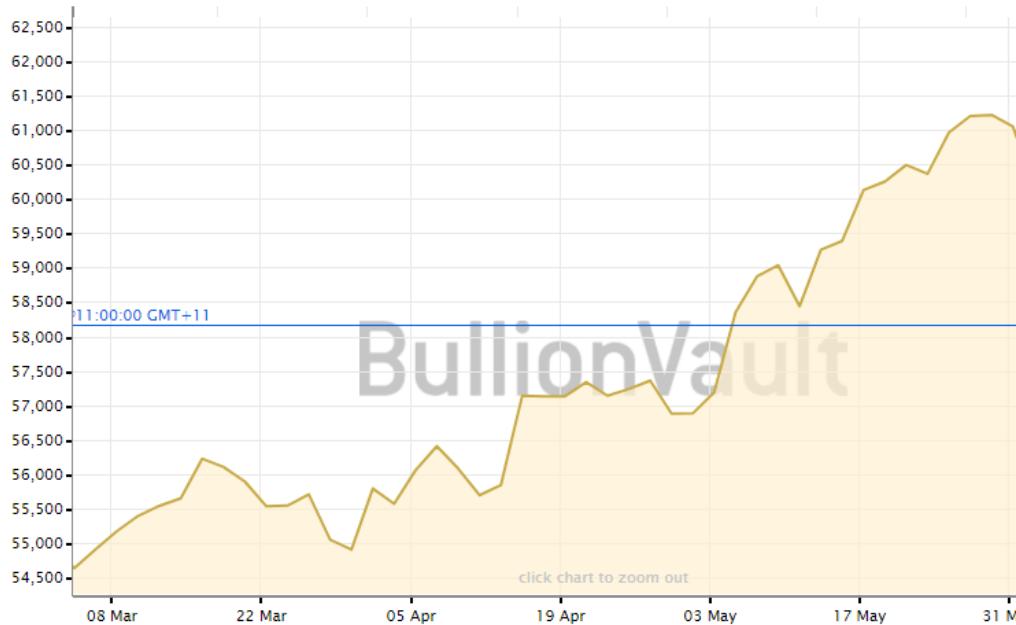
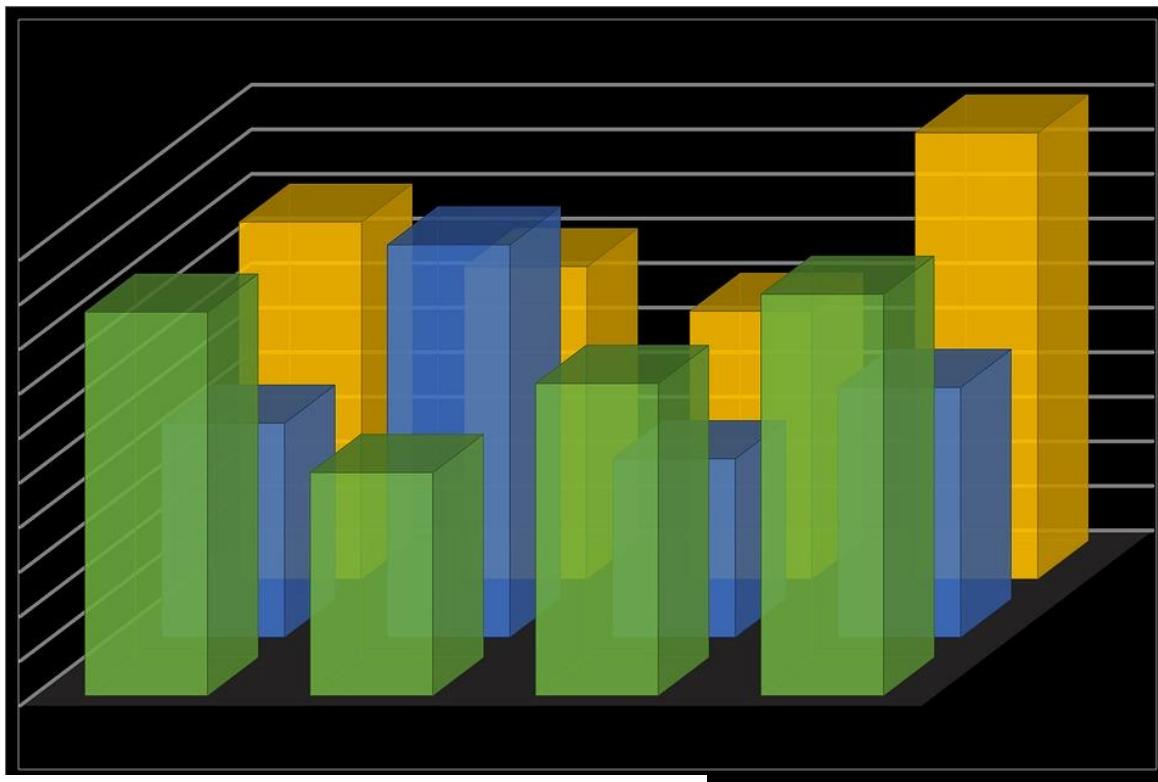
JASPRIT BUMRAH	37	14.48
RASHID KHAN	30	10.63
LIAM PLUNKETT	29	12.20
HASAN ALI	24	19.87
MUSTAFIZUR RAHMAN	23	17.43
BHUVNESHWAR KUMAR	21	29.09
PAT CUMMINS	20	15.65
ADIL RASHID	20	20.55
YUVVENDRA CHAHAL	19	13.89
TENDAI CHATARA	19	20.31

NUMBERS UPDATED TILL MAY 14, 2019

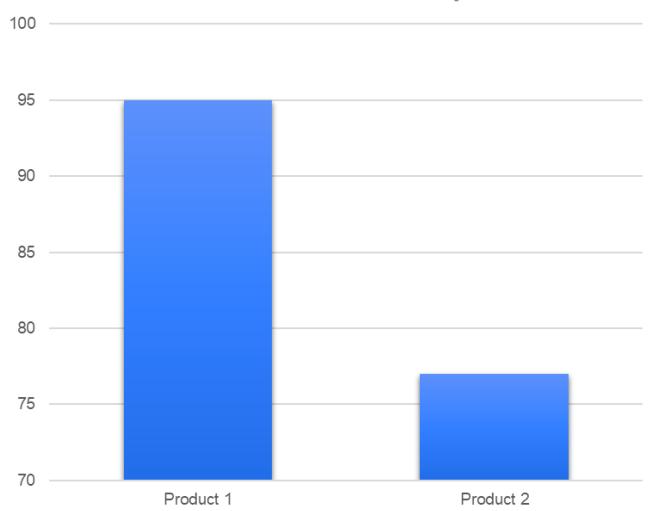


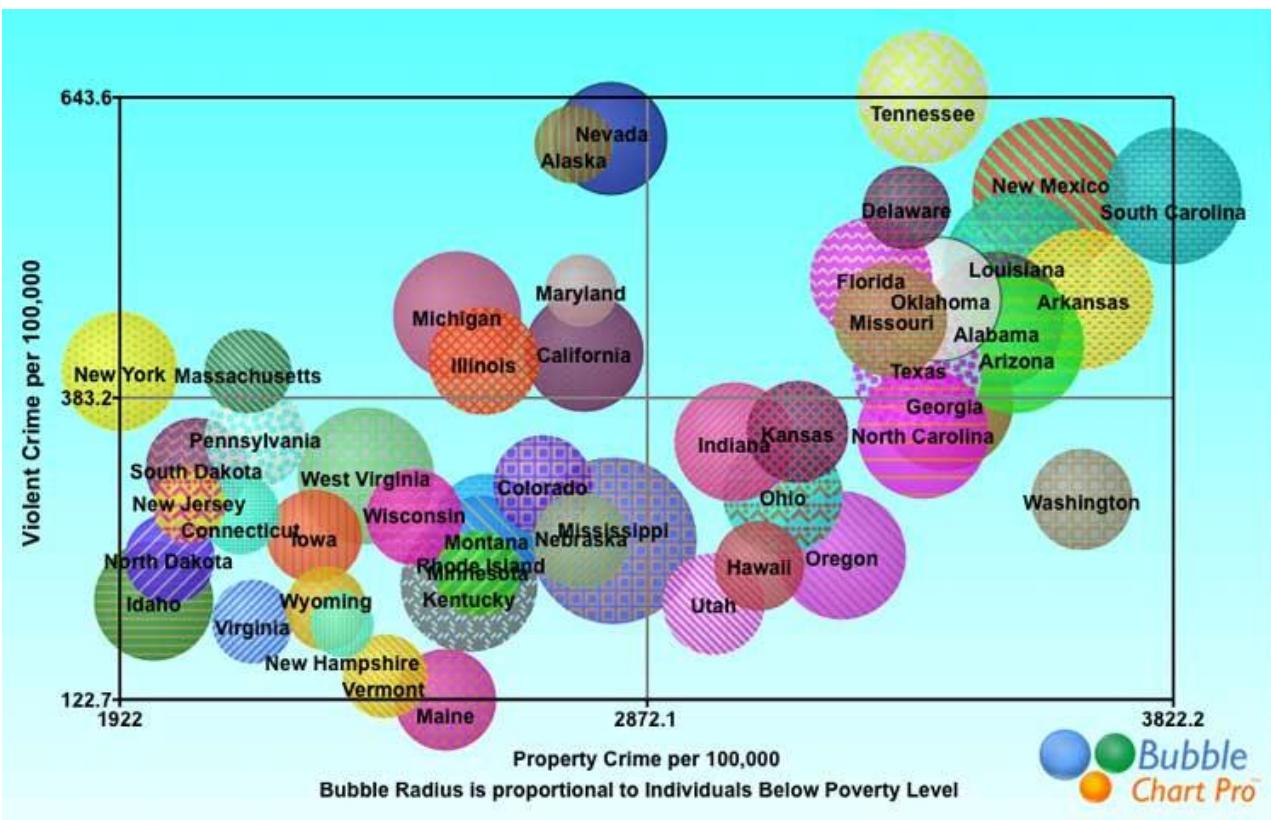
Tim Cook used the particular chart to showcase the rising sale of iPads between the years 2008-2013.

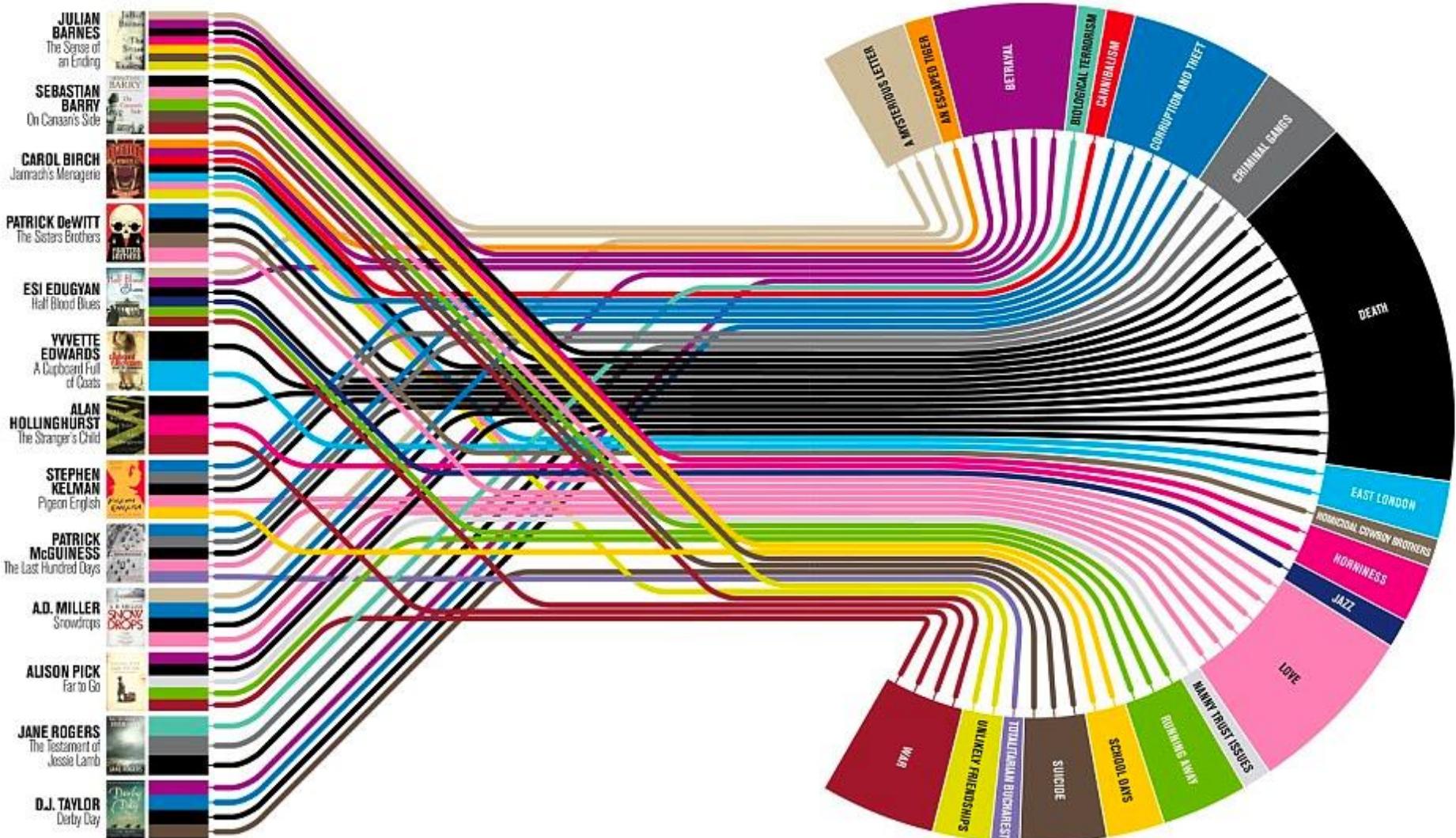




Number of sales for each product







Plot lines

What makes a prize-winning novel? As Julian Barnes wins the Booker Prize, Delayed Gratification's Johanna Kamradt charts the themes of this year's longlisters.

What makes a good visualisation?

- reduce cognitive Load
 - simplicity
 - relevancy
 - less is more
- storytelling
 - ability to support the reader during their journey
 - convince the reader

Remove
to improve
(the **data-ink** ratio)

Created by Darkhorse Analytics

www.darkhorseanalytics.com

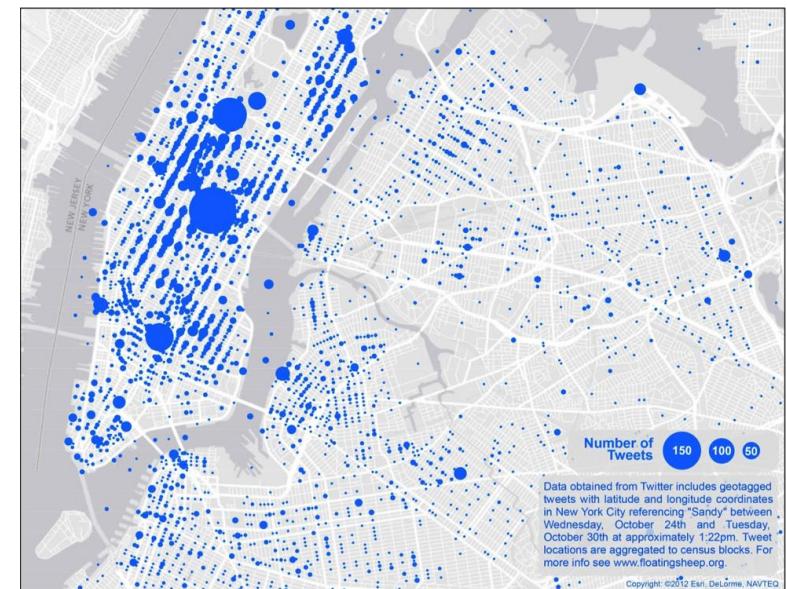
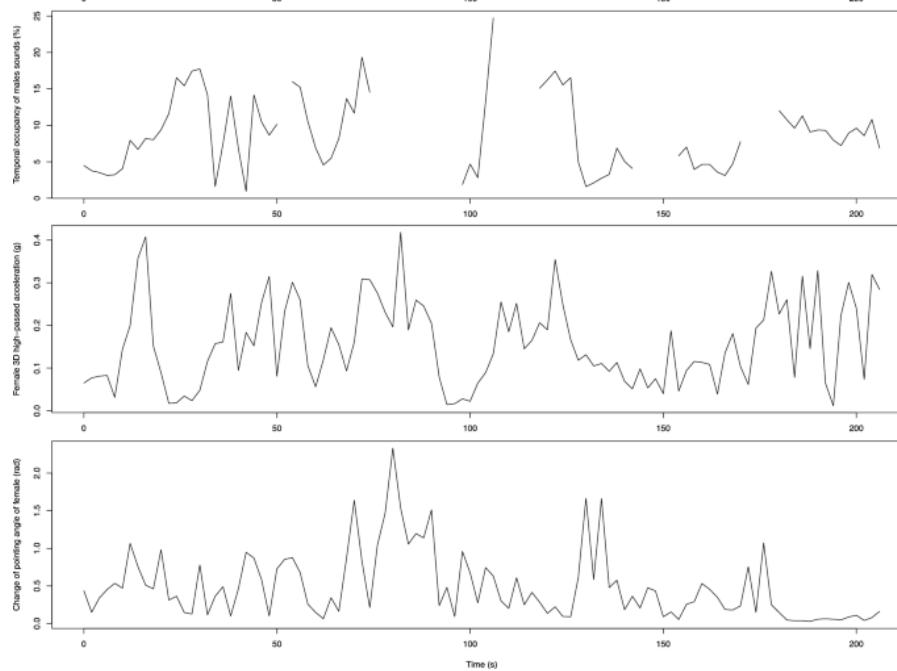
“Perfection is achieved not when there is nothing more to add, but when there is nothing left to take away”
– Antoine de Saint-Exupery

What makes a good visualisation

- Color Consistency
 - use same colors across multiple charts for consistency
 - avoid using colors with negligible contrast
 - avoid using too many colors
 - avoid using conventional colors to convey opposite meanings
 - pay heed to the needs of people who might be colorblind (check also in grayscale)
- Accurate Scaling

What makes a good visualisation

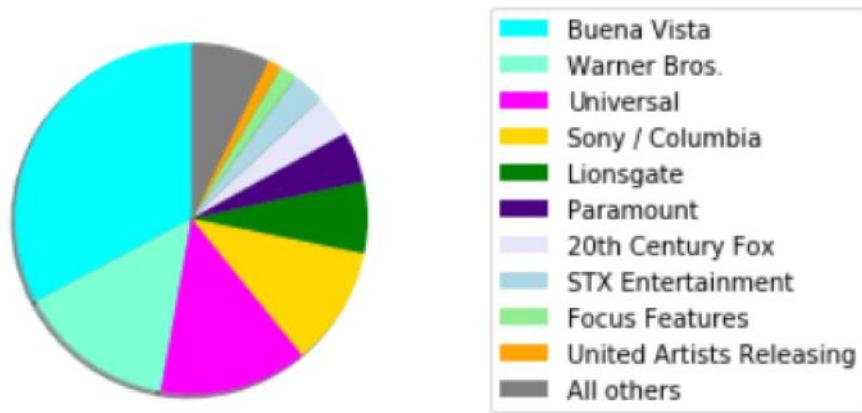
- identify & explain/infer from missing data



What makes a good visualisation

- labelling
 - label the axis correctly and consistently across all your charts.
 - avoid using acronyms that are not widely understood.
 - make the chart title as concise and descriptive as possible.
 - whenever possible, label the lines in your line chart directly rather than using a legend.
 - be consistent in formatting; if you are working with currency symbols, percentage signs and the decimal values, retain them across all your charts.

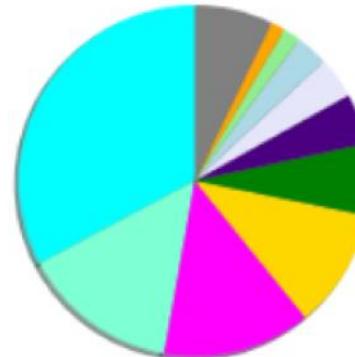
EXAMPLE



Market Share of Film Studios

PIE CHART

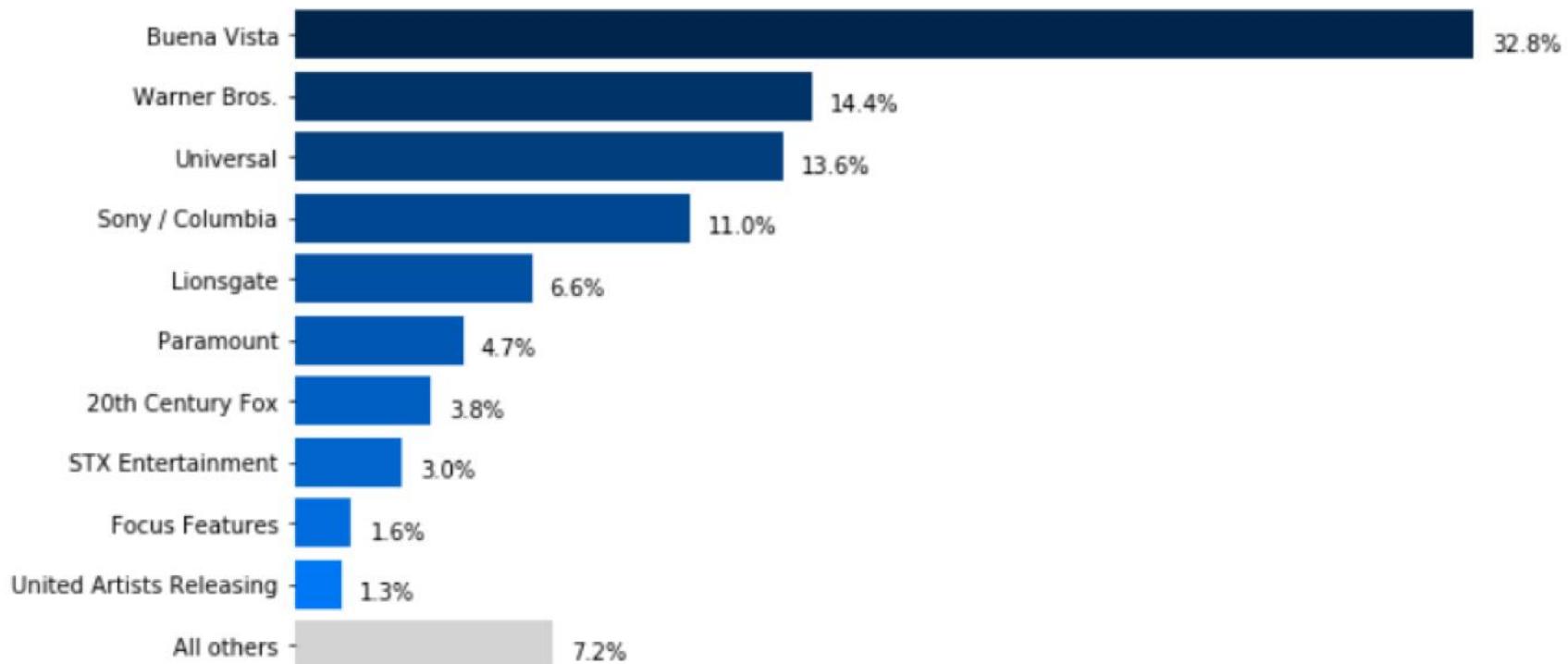
Not comprehensible!



Buena Vista
Warner Bros.
Universal
Sony / Columbia
Lionsgate
Paramount
20th Century Fox
STX Entertainment
Focus Features
United Artists Releasing
All others

BAR CHART

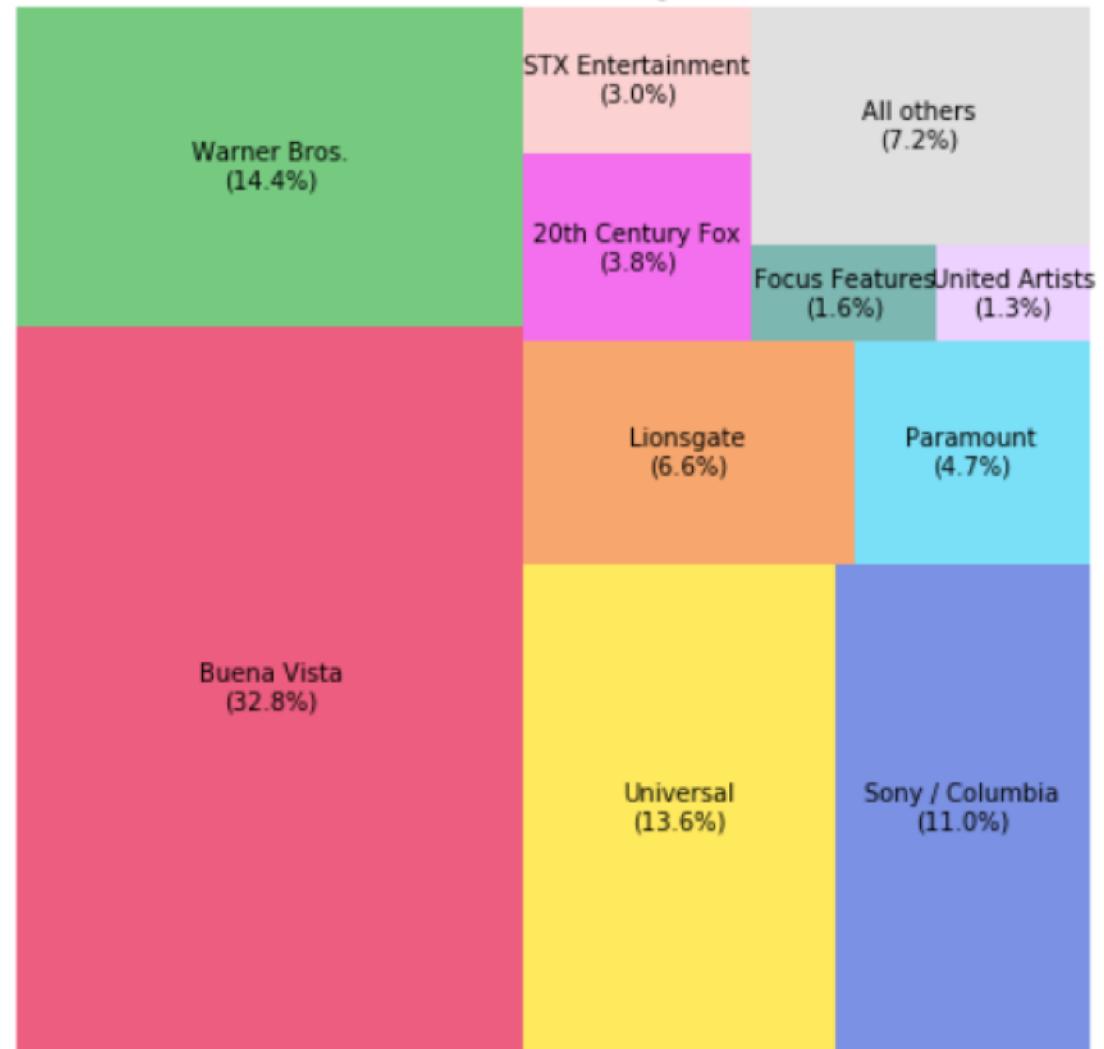
Market Share for Films Studios (Jan 1 - Oct 6, 2019)



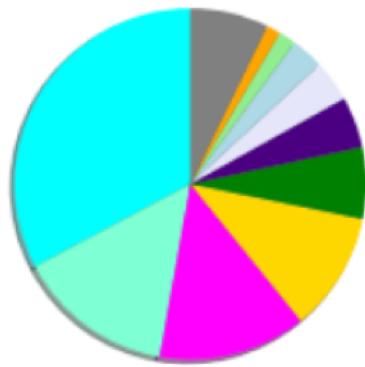
AREA PLOTS: TREE MAP



Market Share for Films Studios (Jan 1 - Oct 6, 2019)



AREA PLOTS: WAFFLE CHART

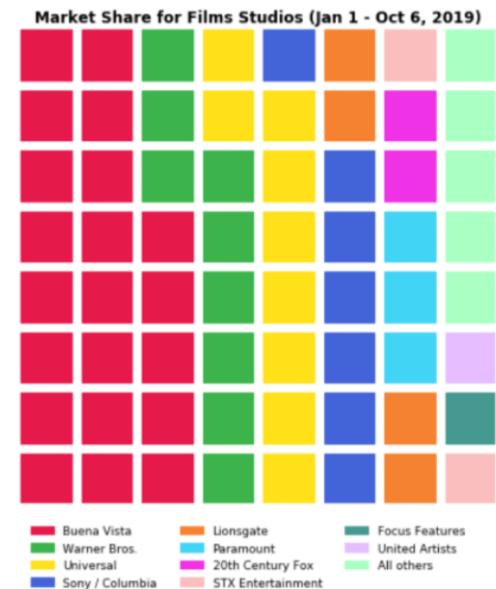
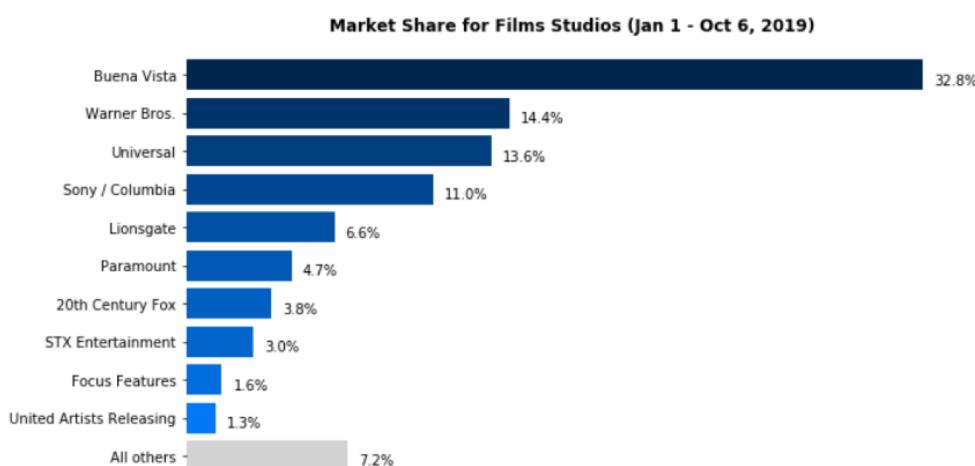
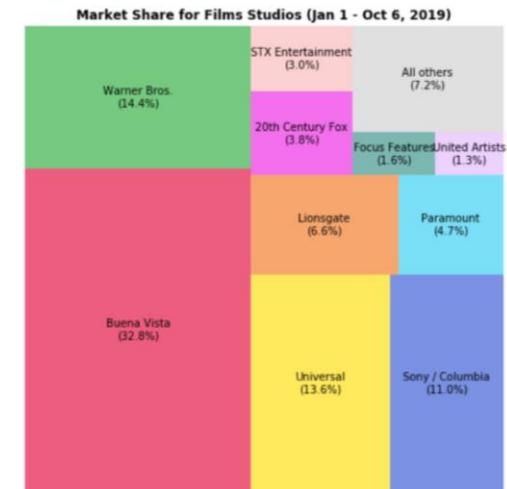
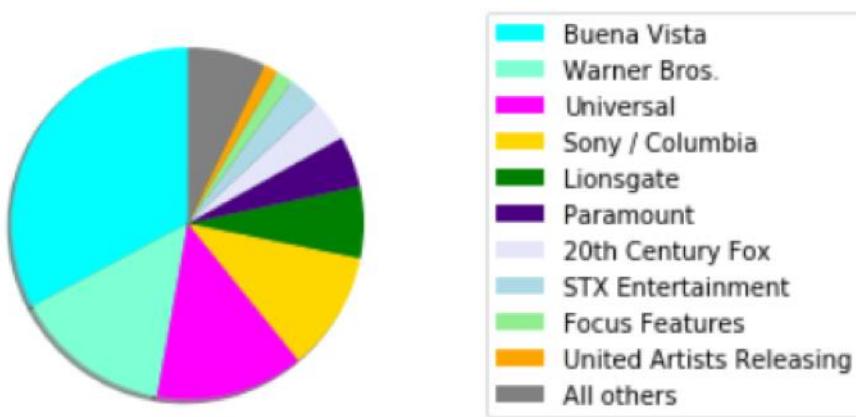


Market Share for Films Studios (Jan 1 - Oct 6, 2019)



Buena Vista	Lionsgate	Focus Features
Warner Bros.	Paramount	United Artists
Universal	20th Century Fox	All others
Sony / Columbia	STX Entertainment	

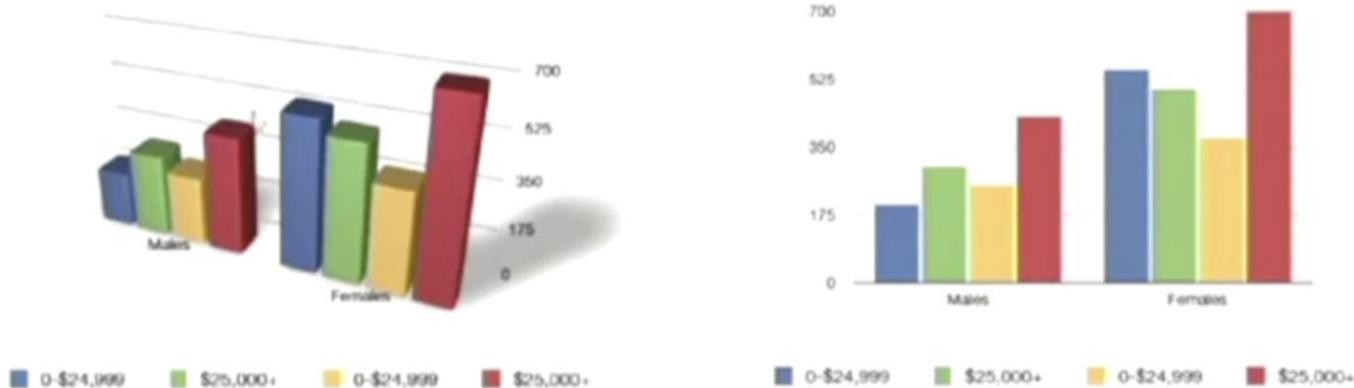
So which visualisation was best?



Tufte's Graphical Theory

- minimize data-to-ink ratio
- minimise lie factor (or increase graphical integrity)
- minimise chart junk
- use proper scales and labelling

$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$



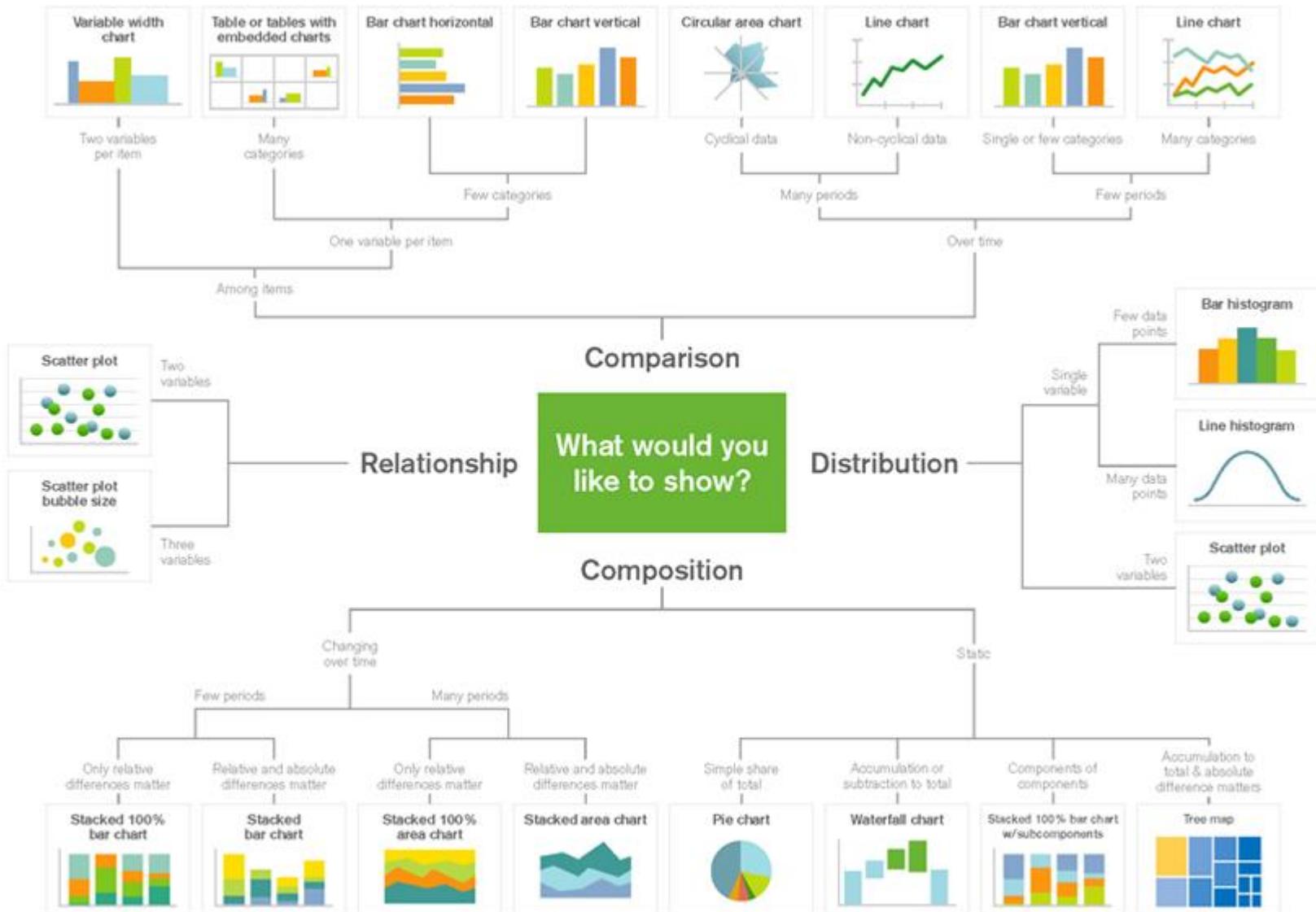
The good, the bad, & the ugly
~~mislading~~



Outline

- **Visualization**
 - why we visualise
 - **how to pick a plot**
 - **initial data vs final results visualization (some examples)**
 - bad designs and misleading graphs
- **Summarization**
 - measures of central tendency & dispersion
 - which measure to pick

How to choose the right plot?



How to choose the right plot?

- **distributions & compositions**
 - proportions
 - data distributions
- **comparisons**
 - group differences
- **associations**
 - relationships between variables
 - geographical data
- **variable types**

Initial Data vs Final Result Visualization

HISTOGRAMS

BOX-PLOT

SCATTER PLOT

PIE CHARTS & BAR CHARTS

MOSAIC PLOT

VIOLIN PLOT

RAIN-DROP

FUNNEL PLOTS

SPIDER PLOT / RADAR CHART

RADIAL HEAT MAP

CIRCOS PLOT

STREAMGRAPH

Not an exhaustive list!

Some plots used for both!



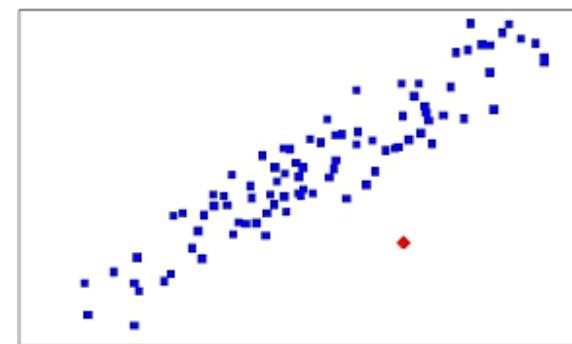
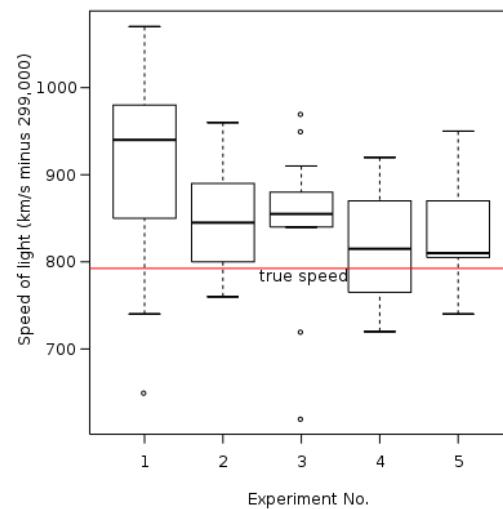
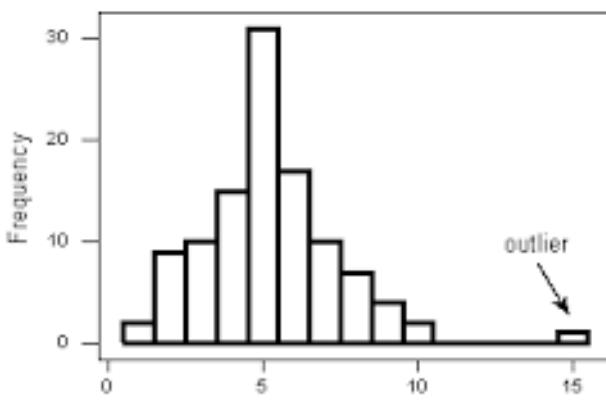
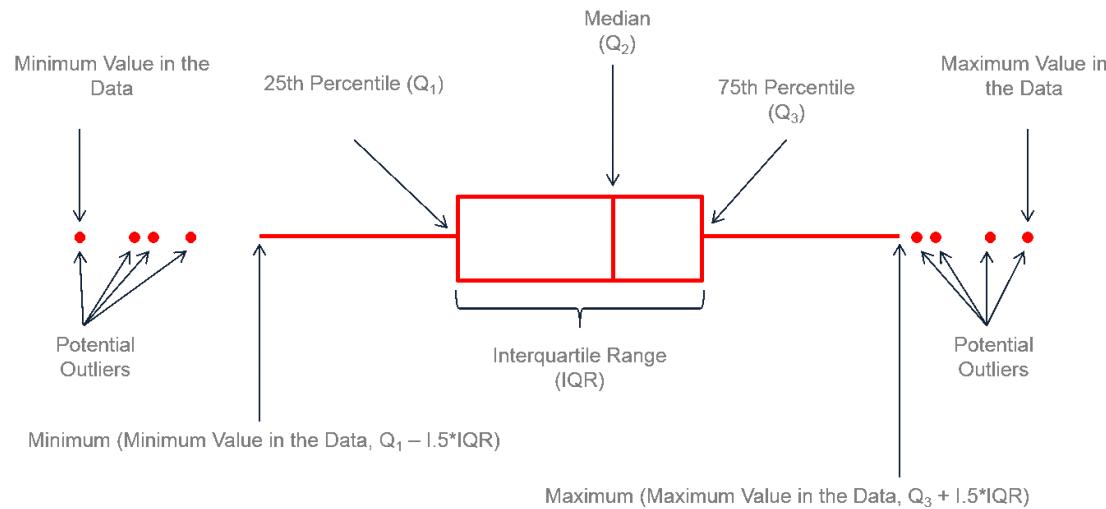
Data Visualization: What Info?

- allows for initial guesses of data distribution (normality) and direction of effect
 - ex: histograms (bin-width dependency), box-plots, scatter plots, pie charts, bar plots (already seen), etc

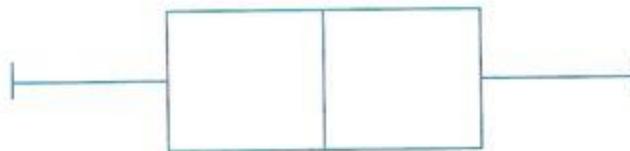
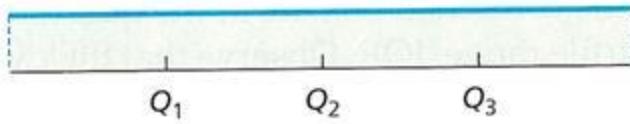


Data Visualisation

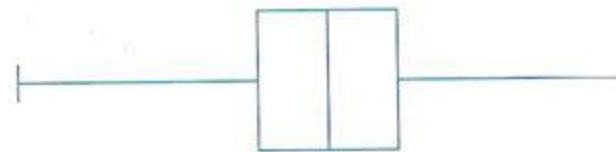
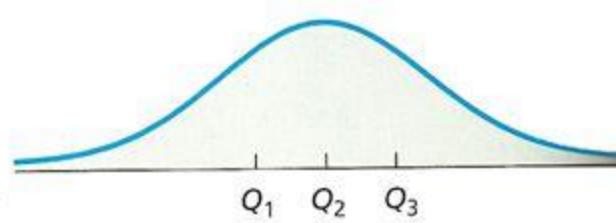
HISTOGRAM, BOXPLOT, SCATTER PLOT



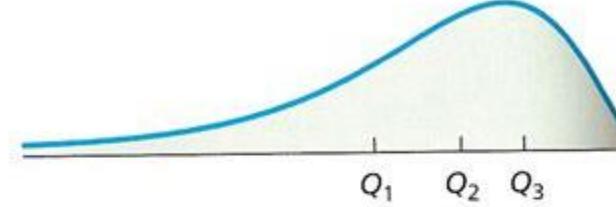
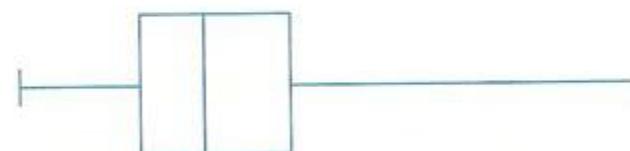
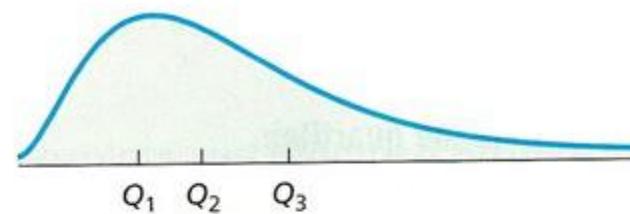
Boxplot

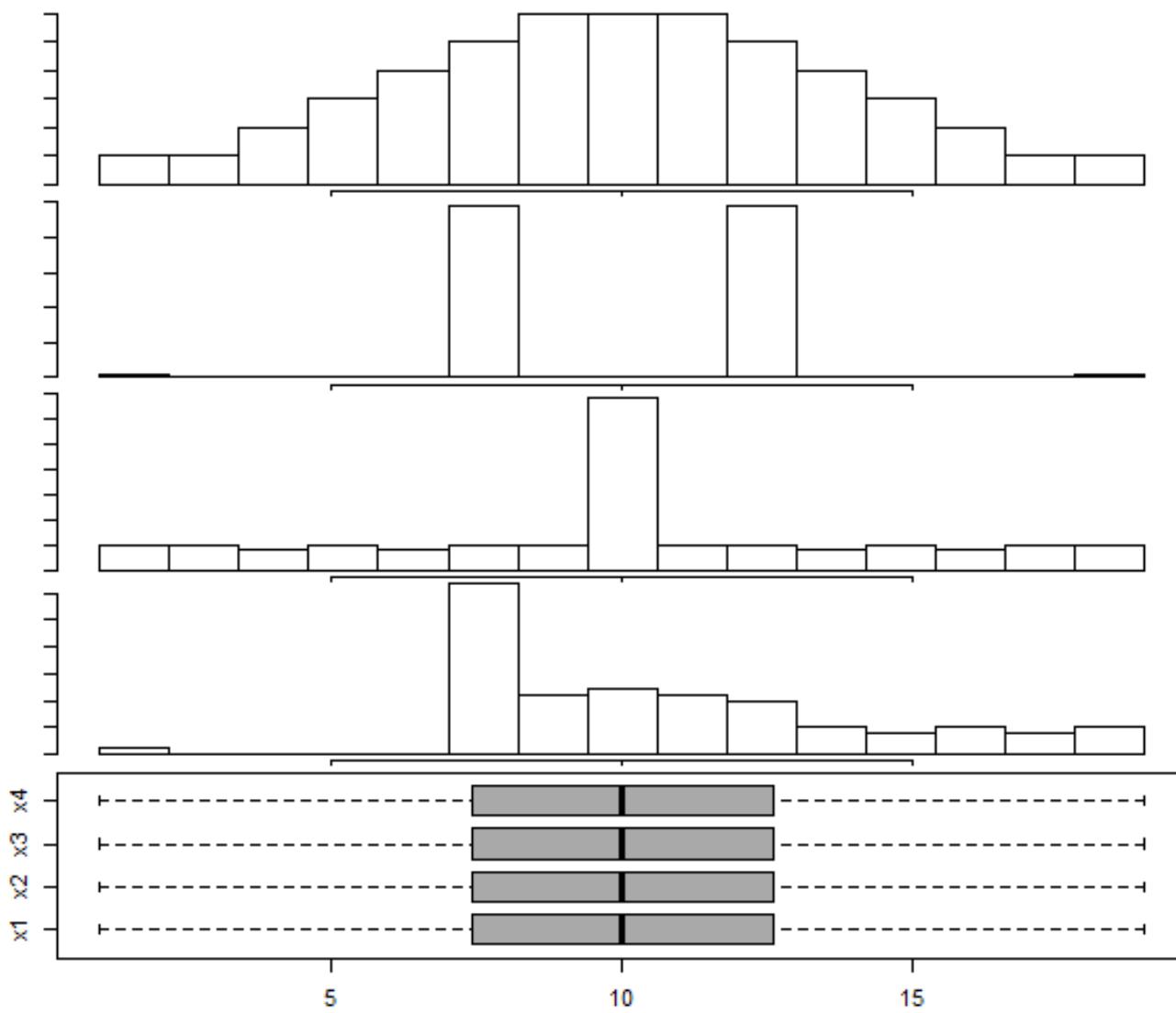


(a) Uniform



(b) Bell shaped





PIE CHART

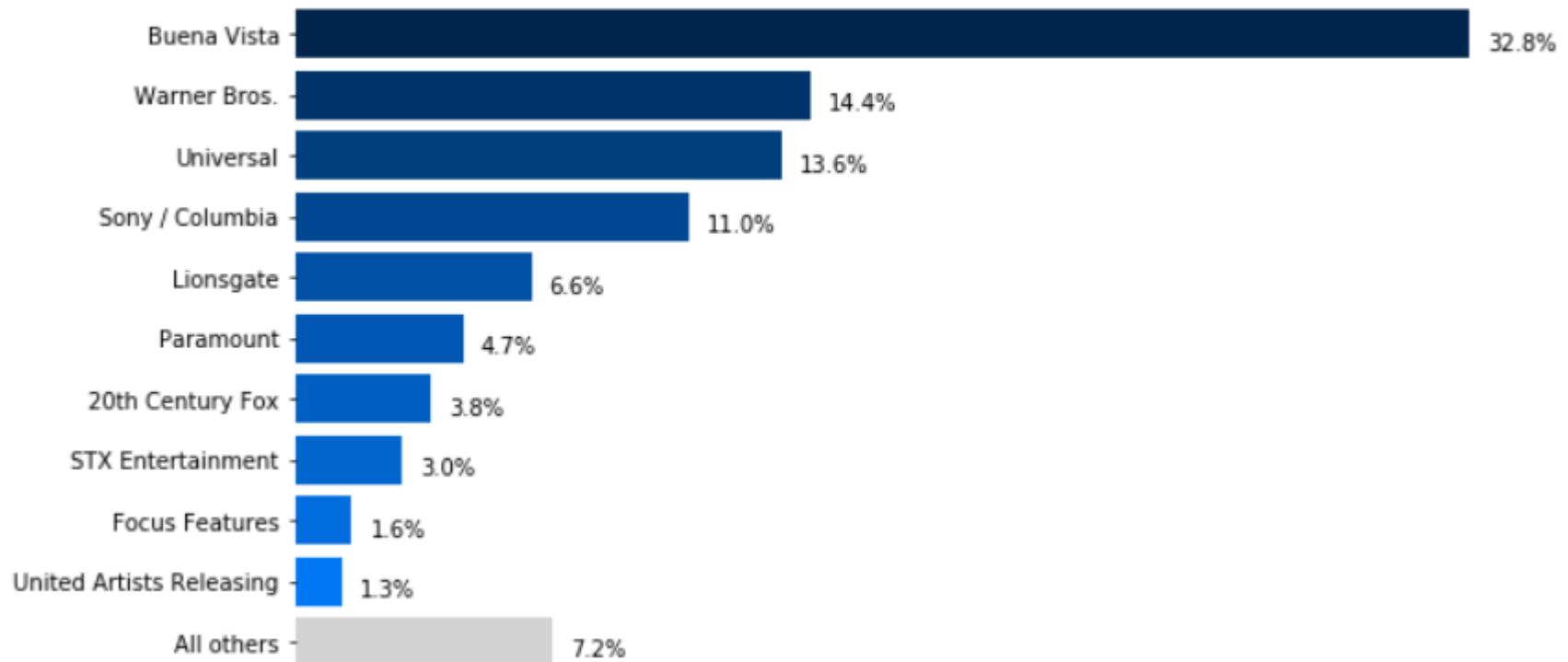
Not comprehensible!

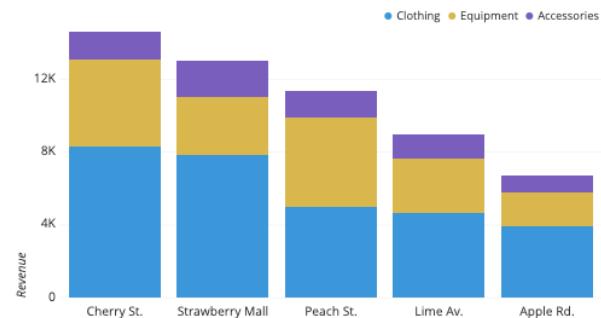


Buena Vista
Warner Bros.
Universal
Sony / Columbia
Lionsgate
Paramount
20th Century Fox
STX Entertainment
Focus Features
United Artists Releasing
All others

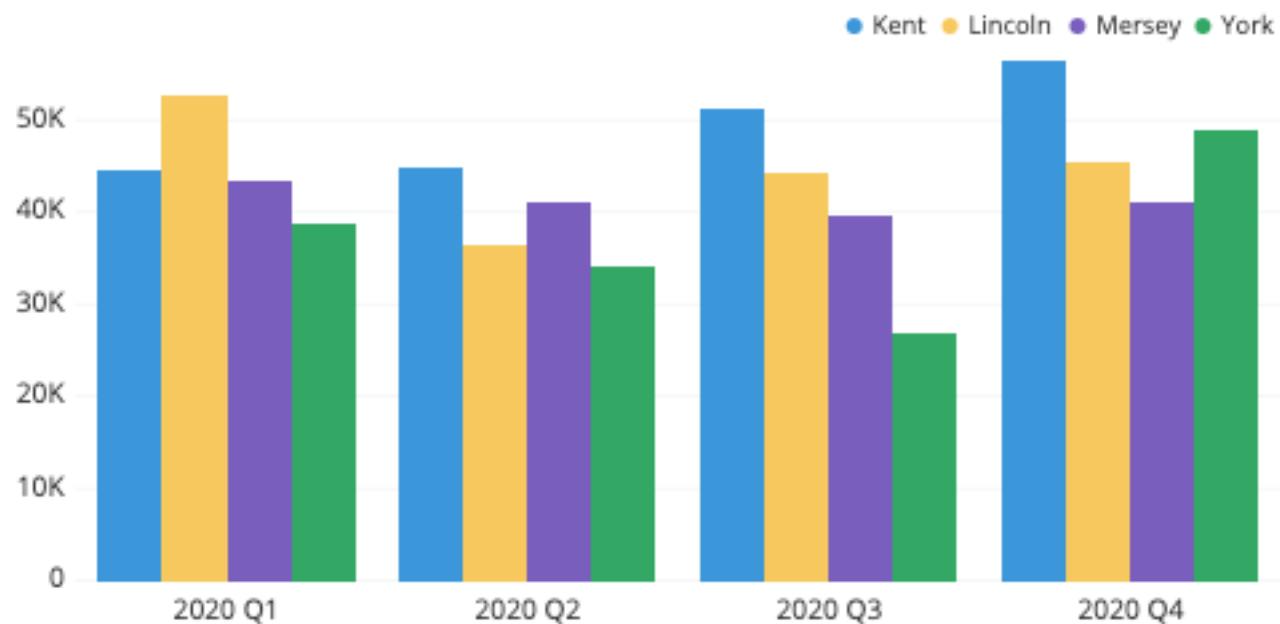
BAR CHART

Market Share for Films Studios (Jan 1 - Oct 6, 2019)





New Revenue





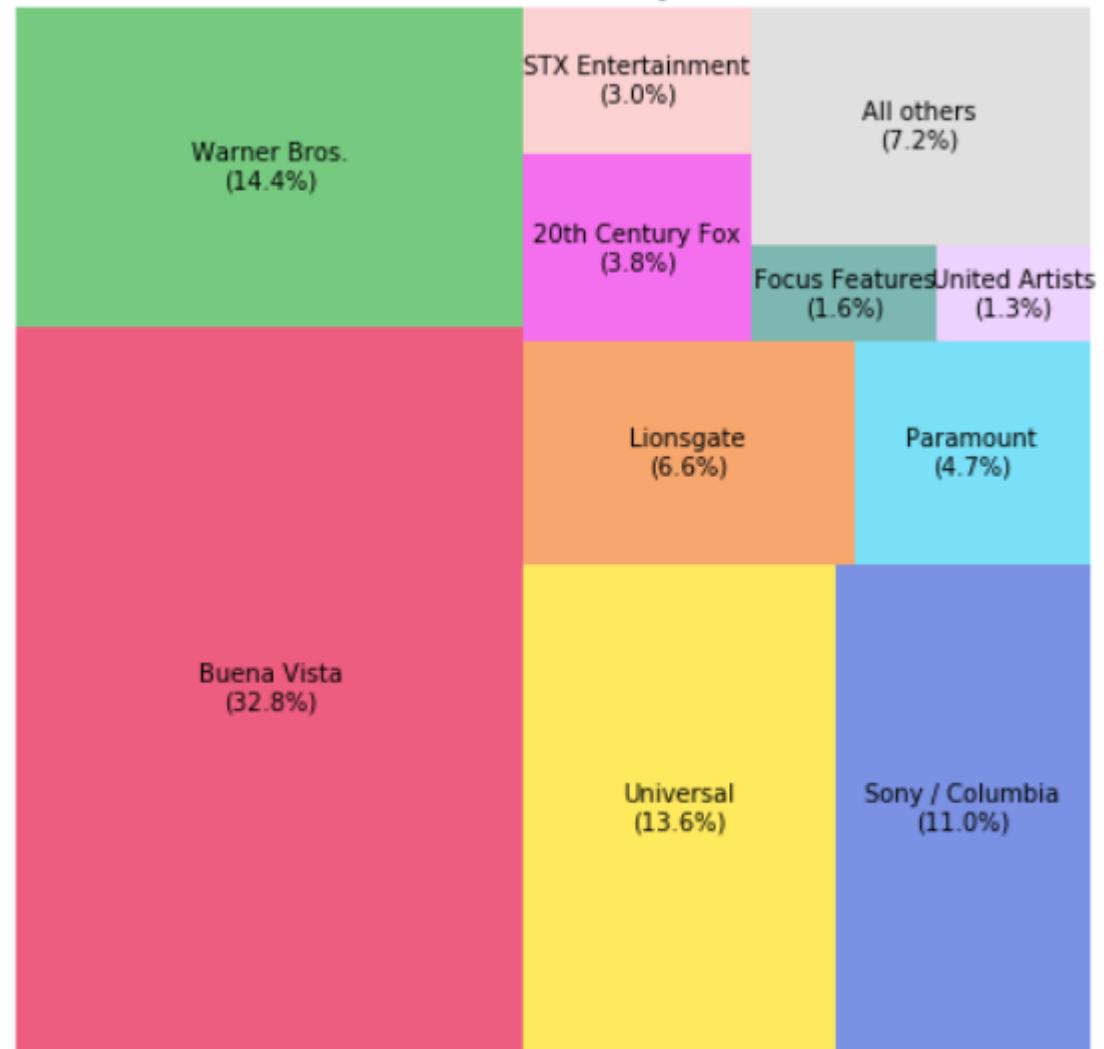
Pie vs Bar Charts

- use pie charts when
 - smaller no. of categories
 - readers can differentiate slices (unless you are making a point)
 - you don't need to rely on many colors or labels to explain the proportions
 - total adds up to 100%
- use bar charts when
 - have many categories (not too many)
 - need to compare numbers side-by-side (caution: more than two bars are hard for readers)

AREA PLOTS: TREE MAP



Market Share for Films Studios (Jan 1 - Oct 6, 2019)



AREA PLOTS: WAFFLE CHART



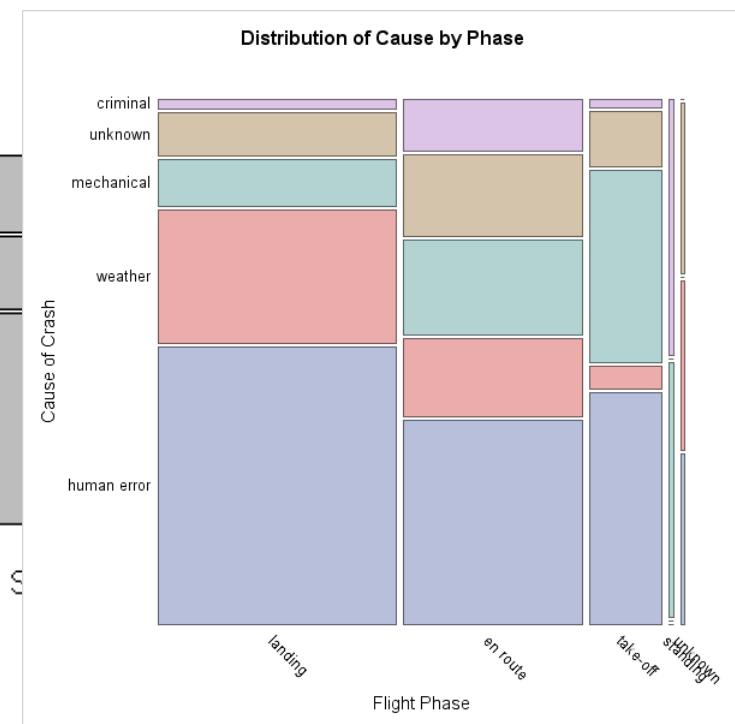
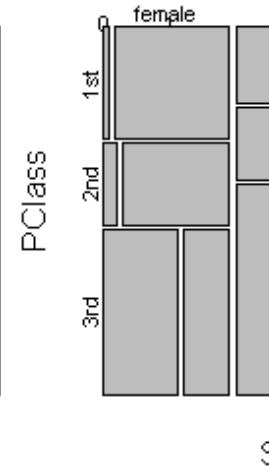
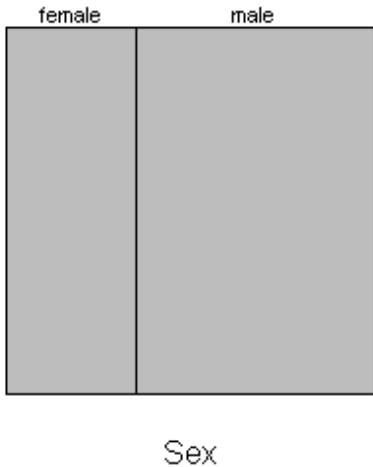
Market Share for Films Studios (Jan 1 - Oct 6, 2019)



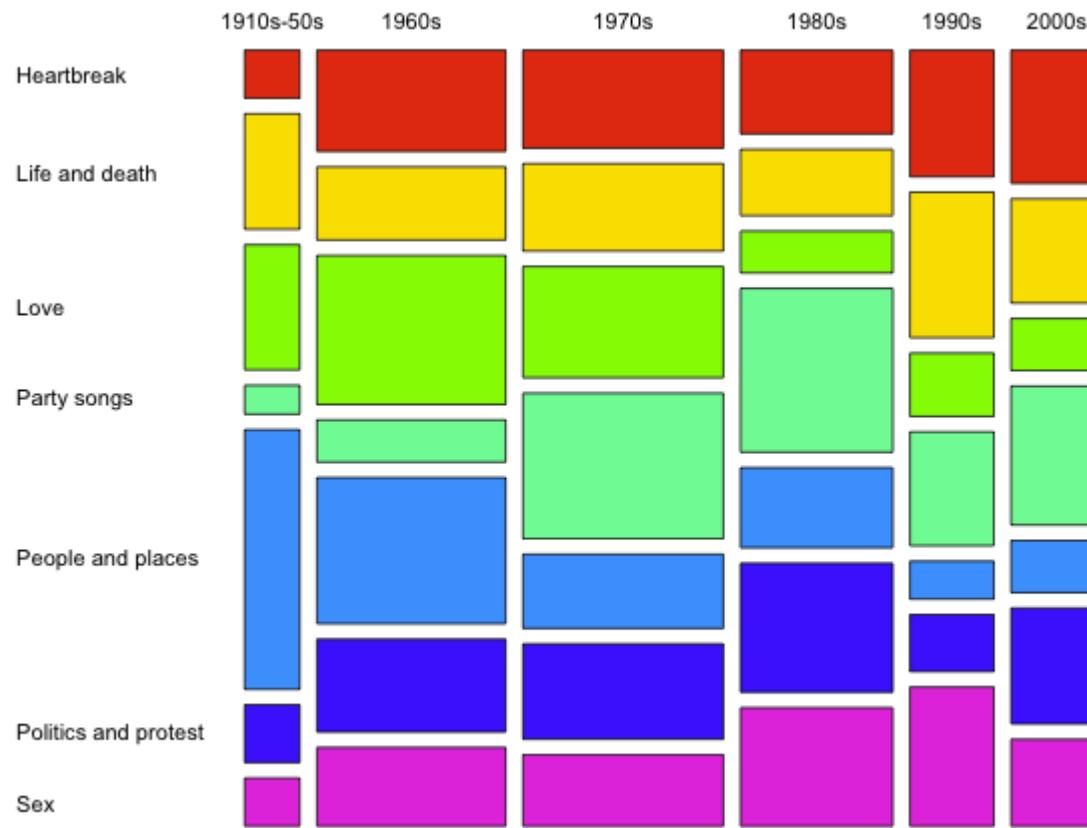


Data Visualisation

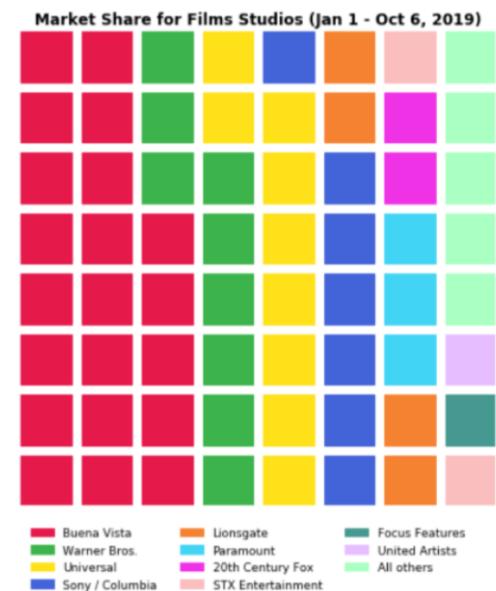
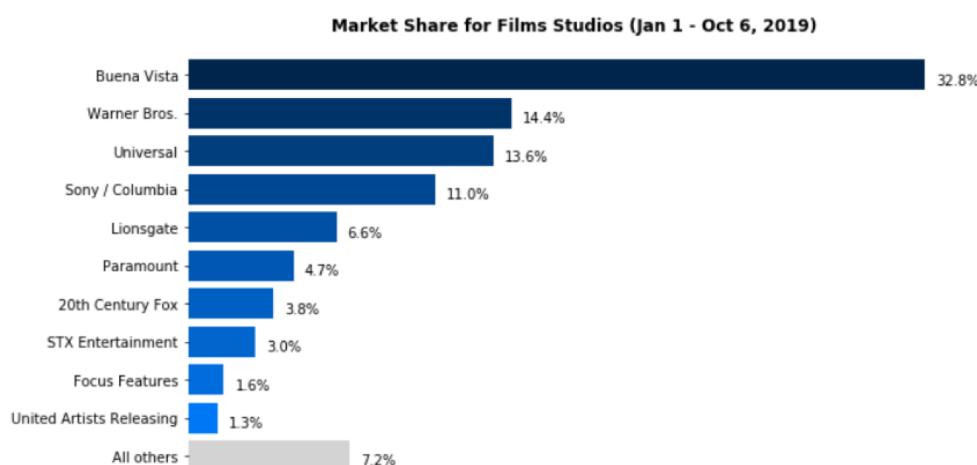
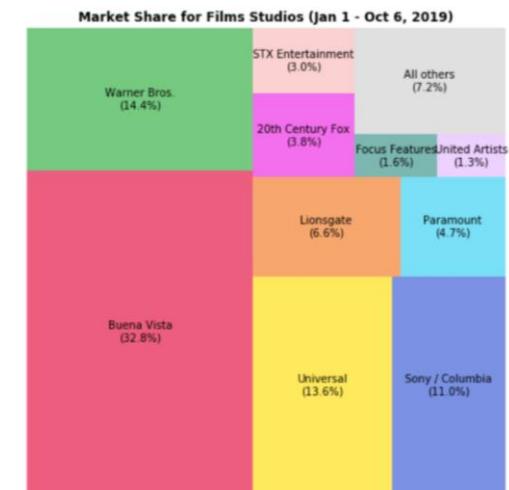
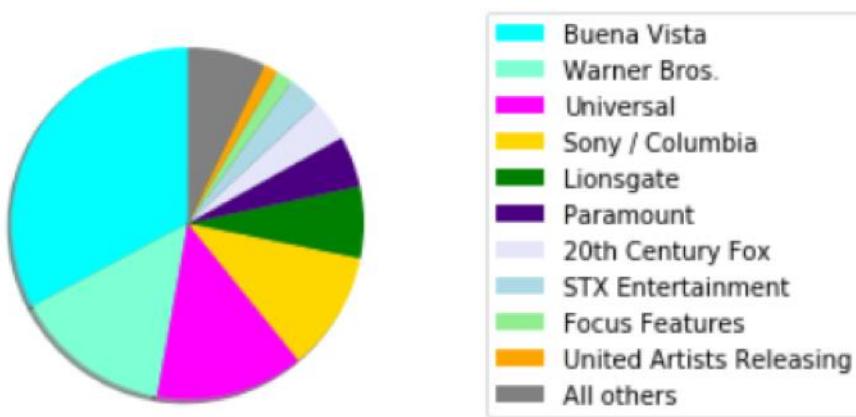
- mosaic plots
 - allows you to observe the relationship among two or more categorical variables



AREA PLOTS: MOSAIC PLOT

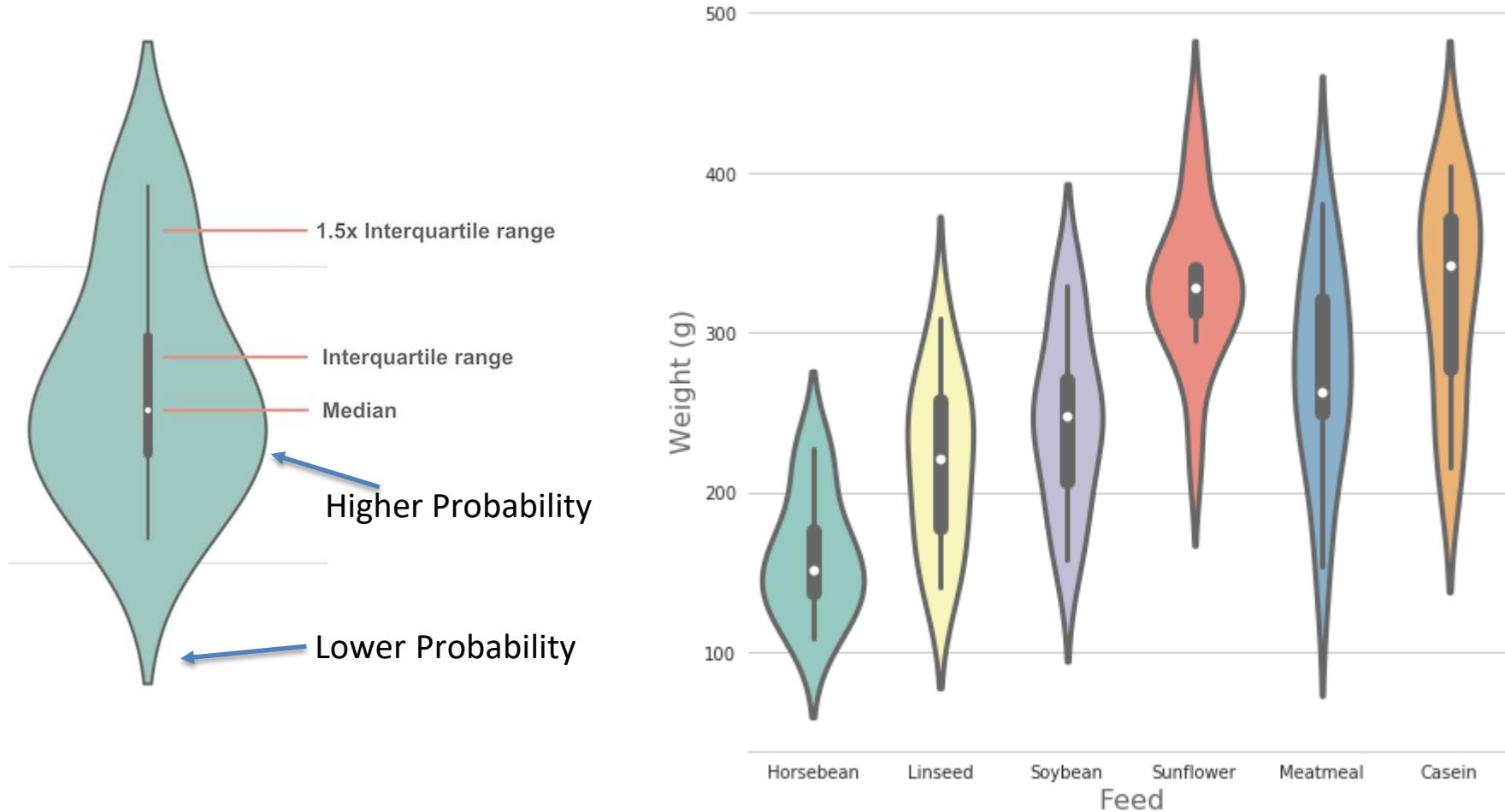


So which visualisation was best?

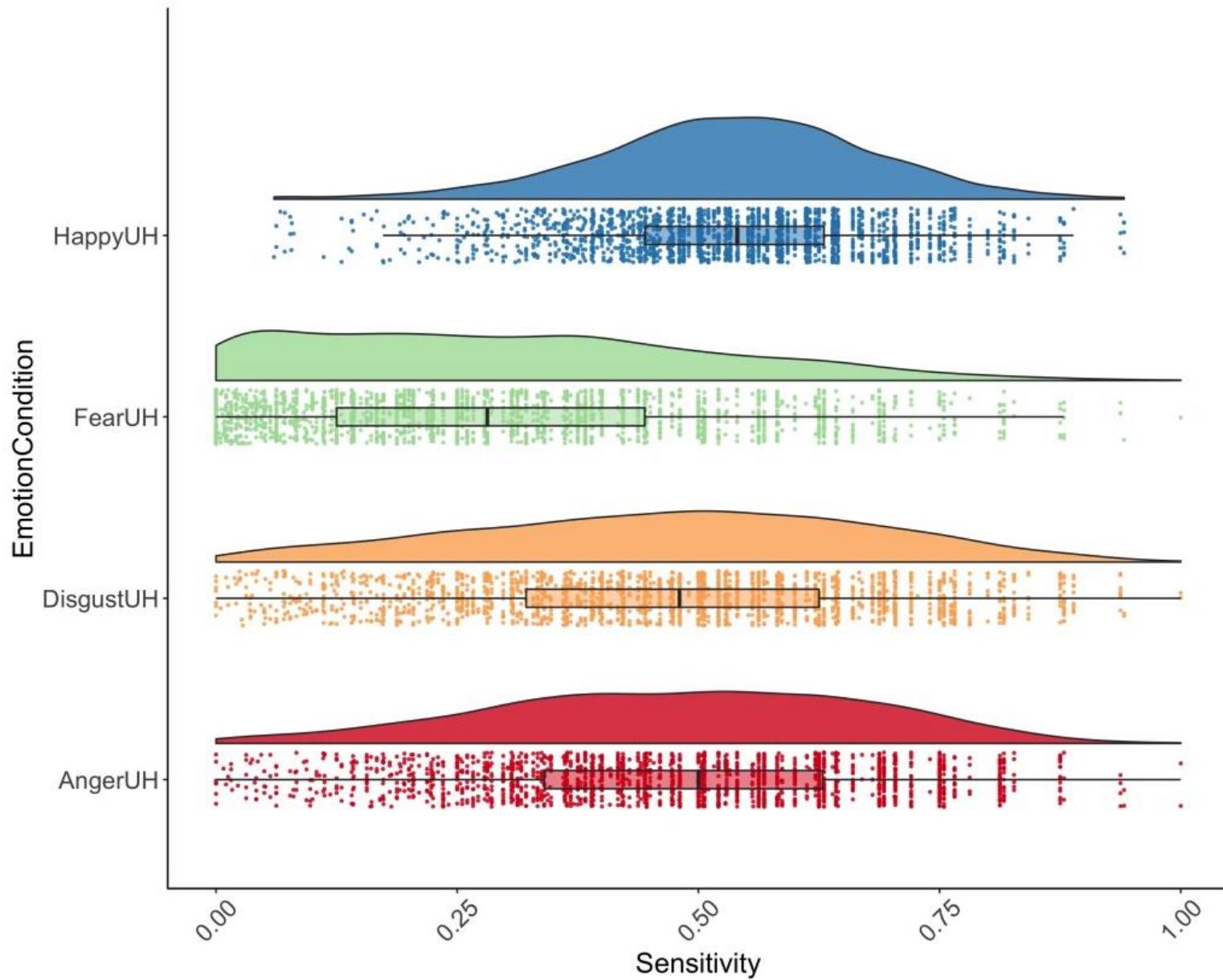


VIOLIN PLOT

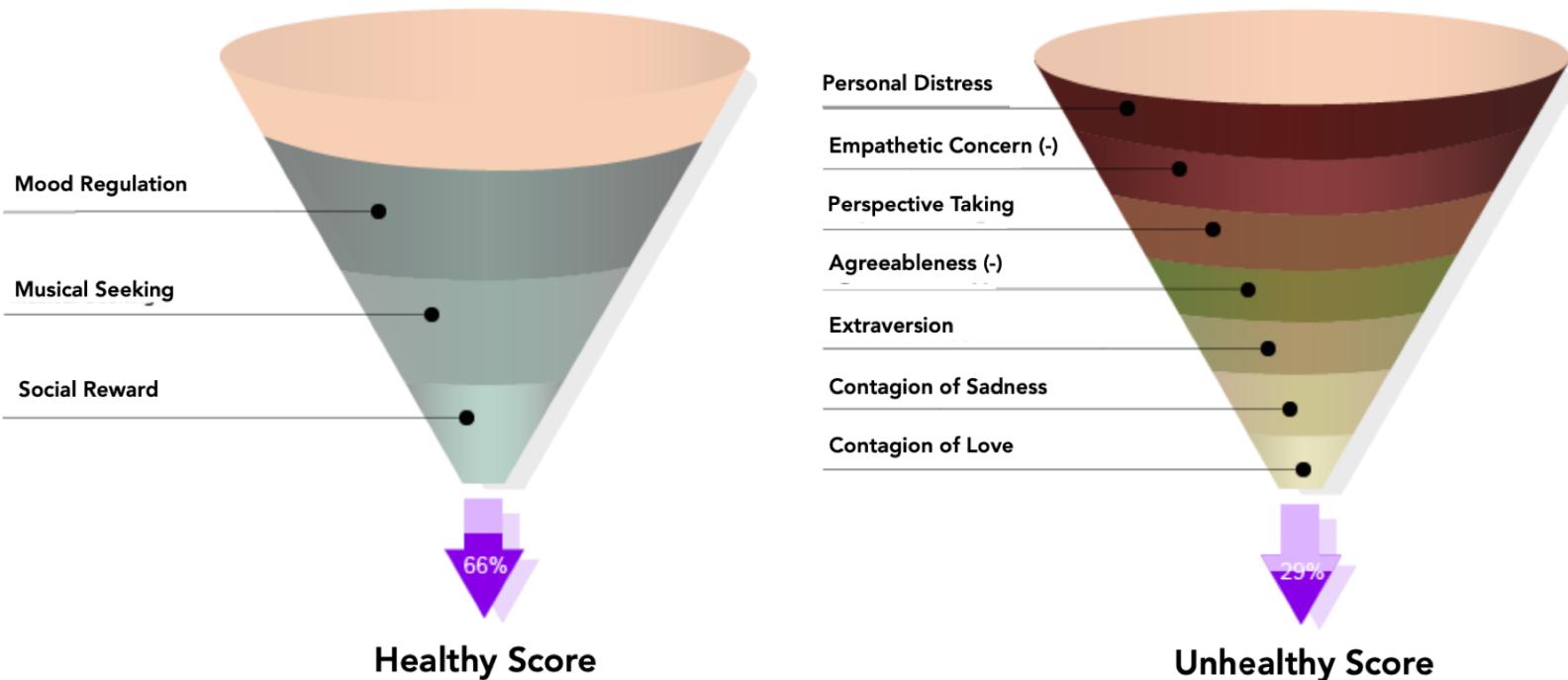
Chick weights by feed type



RAINDROP PLOT (Combo)

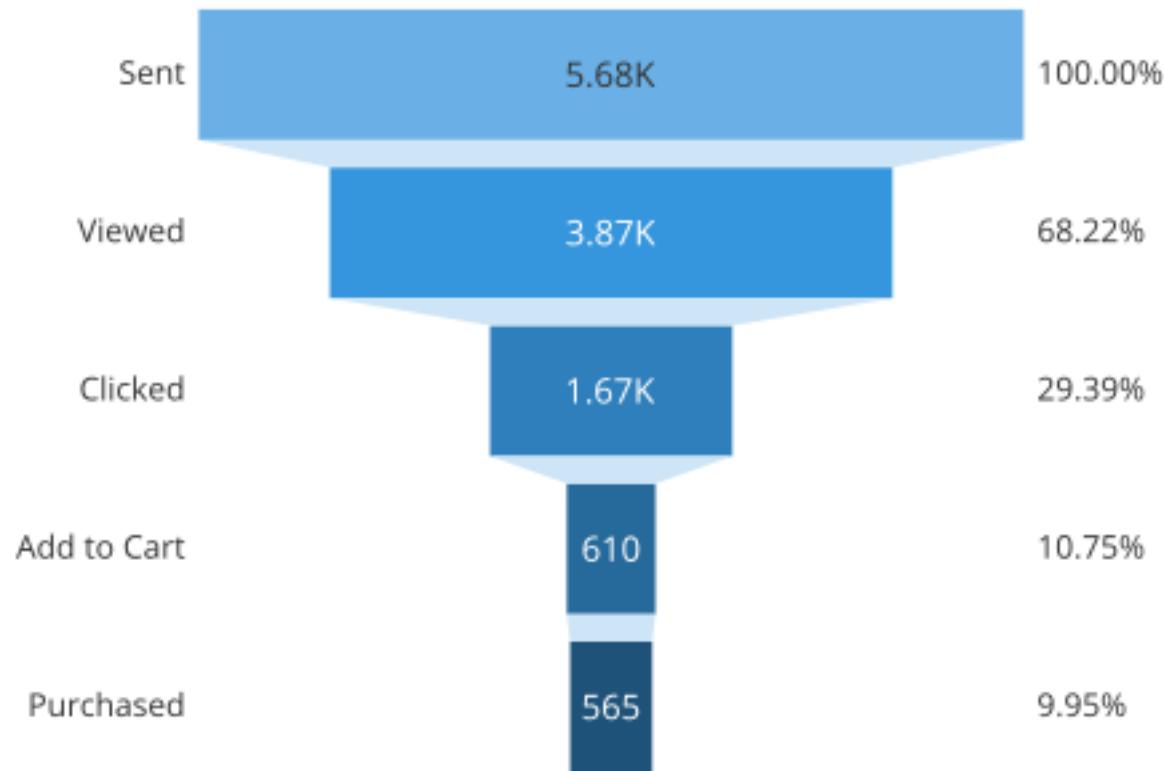


Visualizing Results



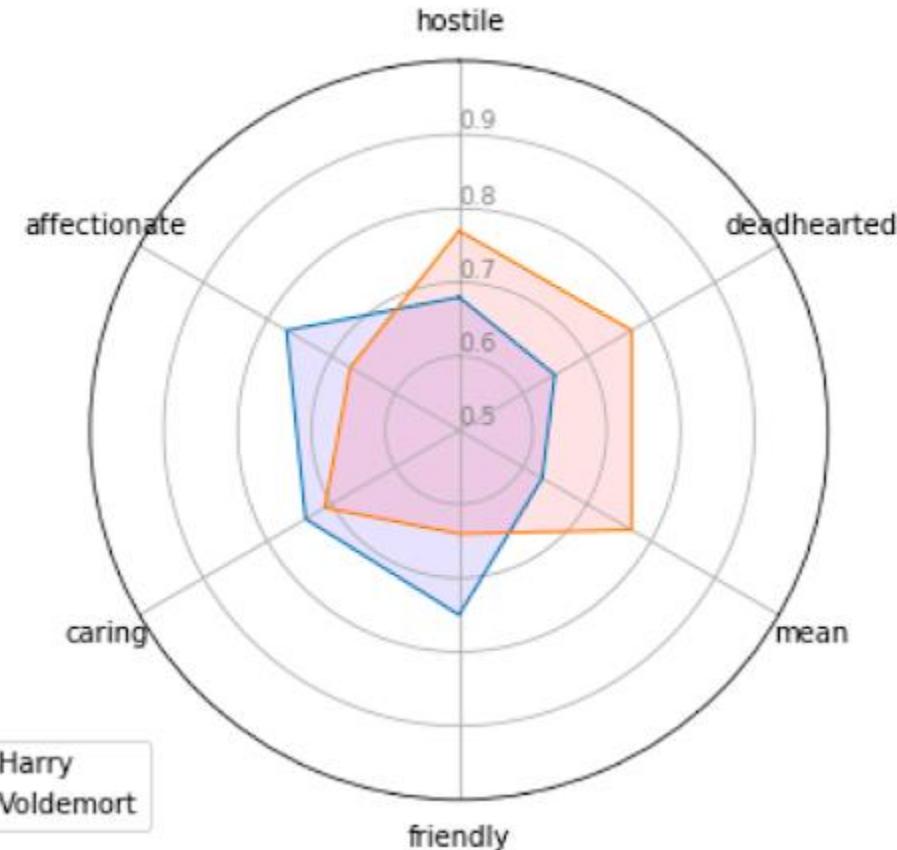
Funnel Plots
(Regression Results)

Describing Data

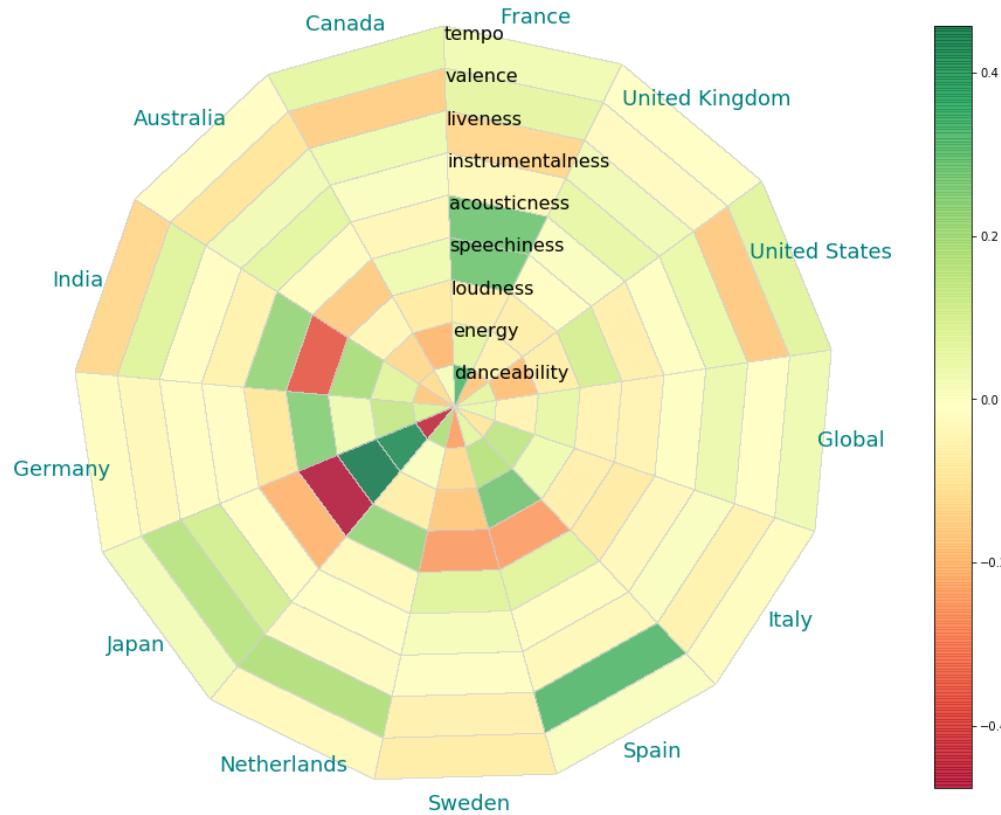


Funnel Plots

SPIDER PLOT / RADAR CHART



Describing Data



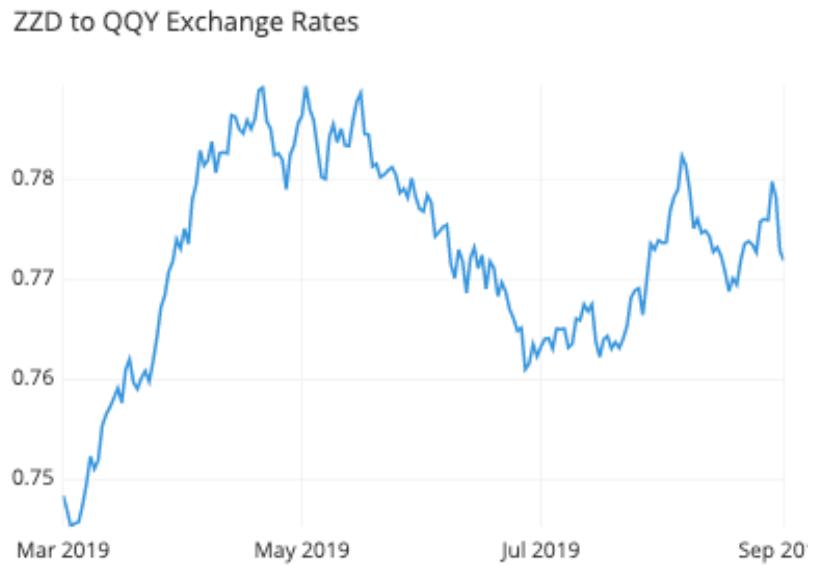
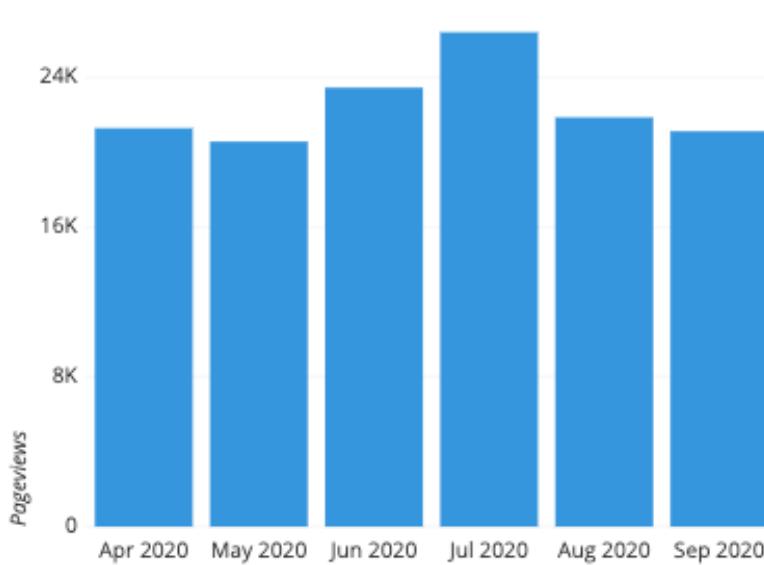
Radial Heat Map

How to choose the right plot?

- temporal changes
- proportions
- data distributions
- group differences
- relationships between variables
- geographical data

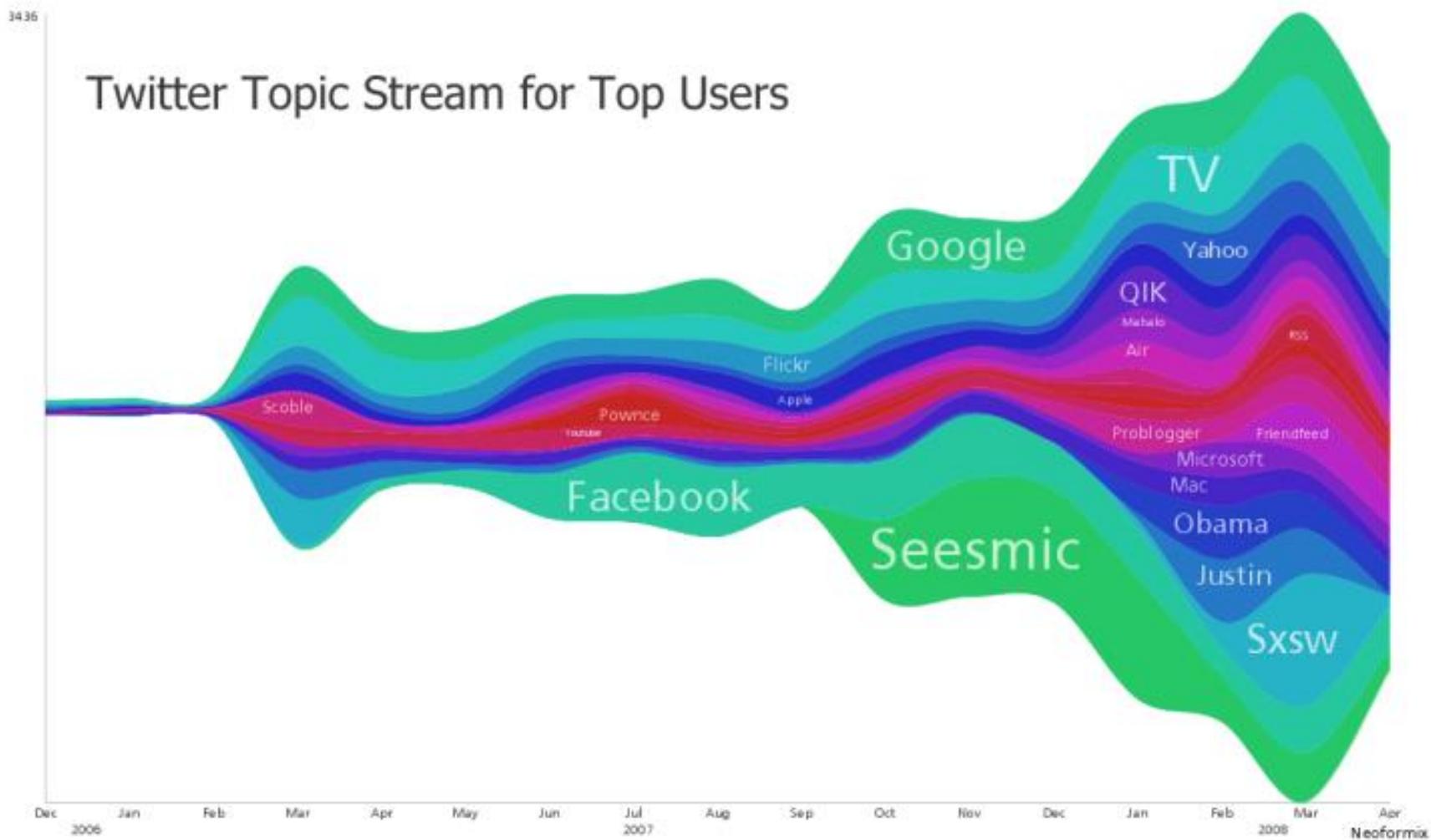
Temporal

- showing change over time



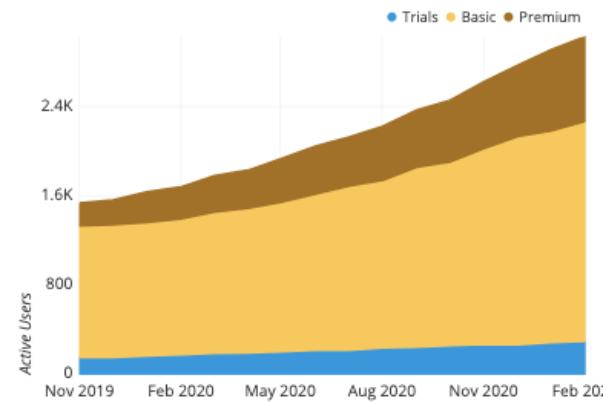
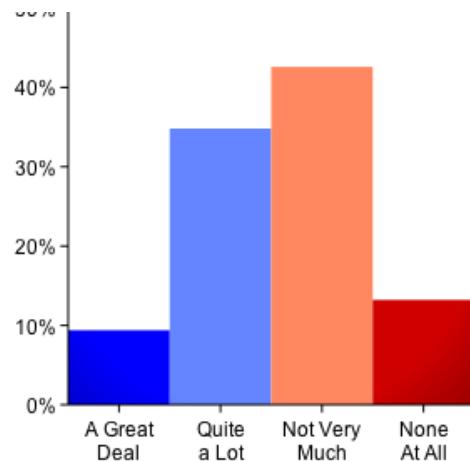
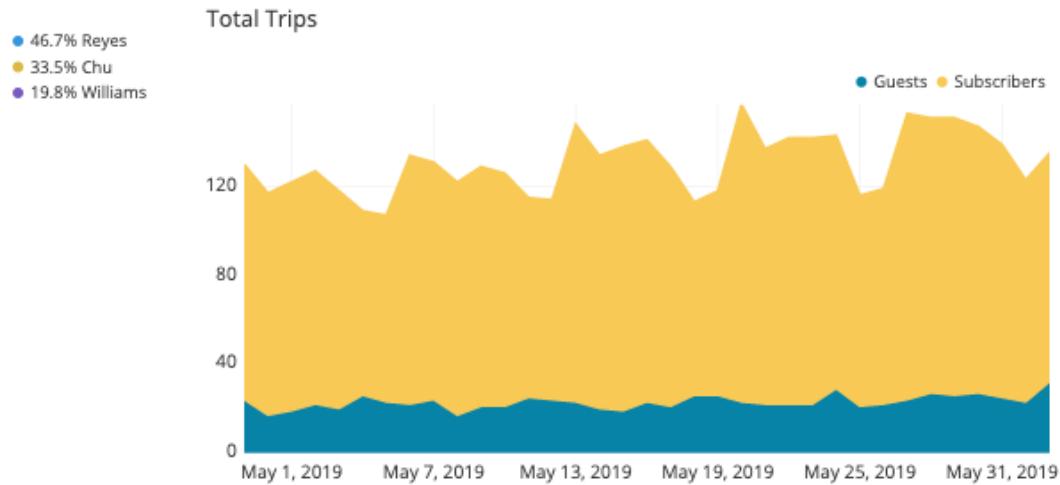
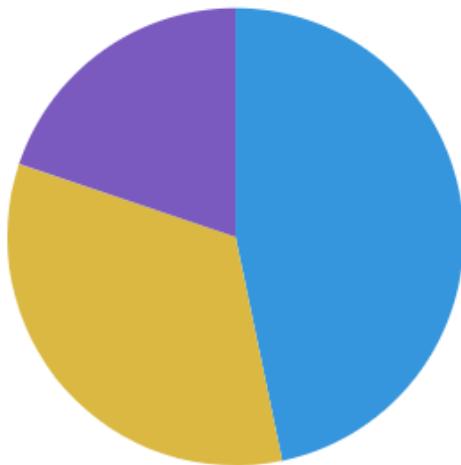
Temporal

- showing change over time

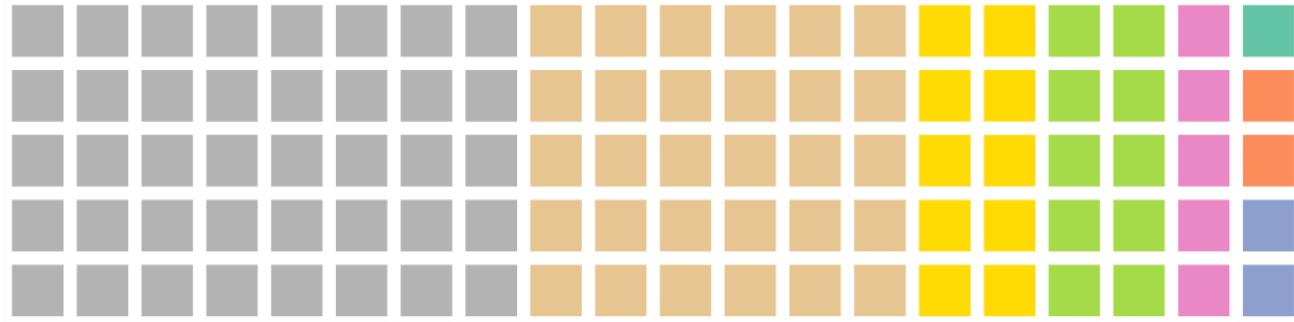


Proportions

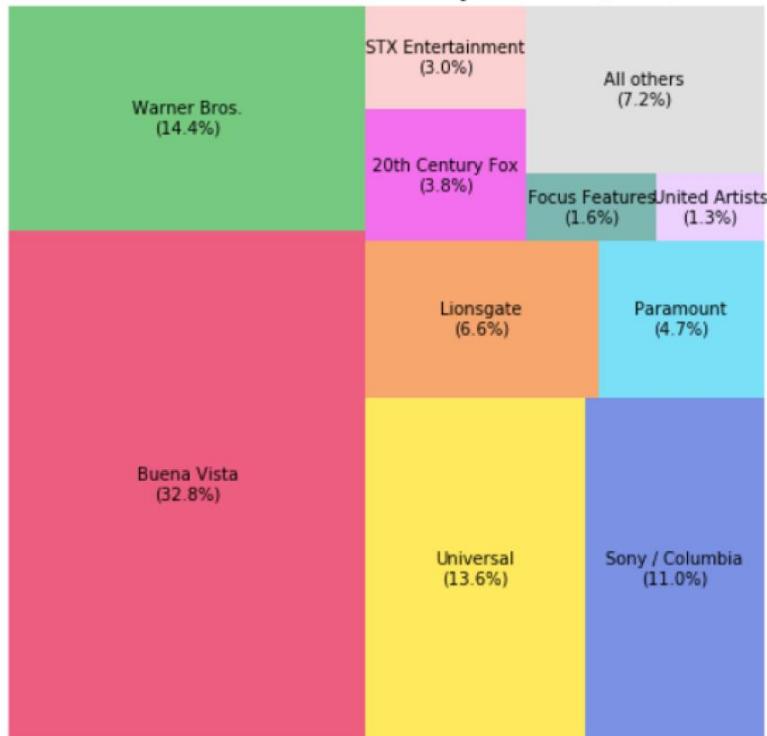
- showing a part-to-whole composition



Proportions

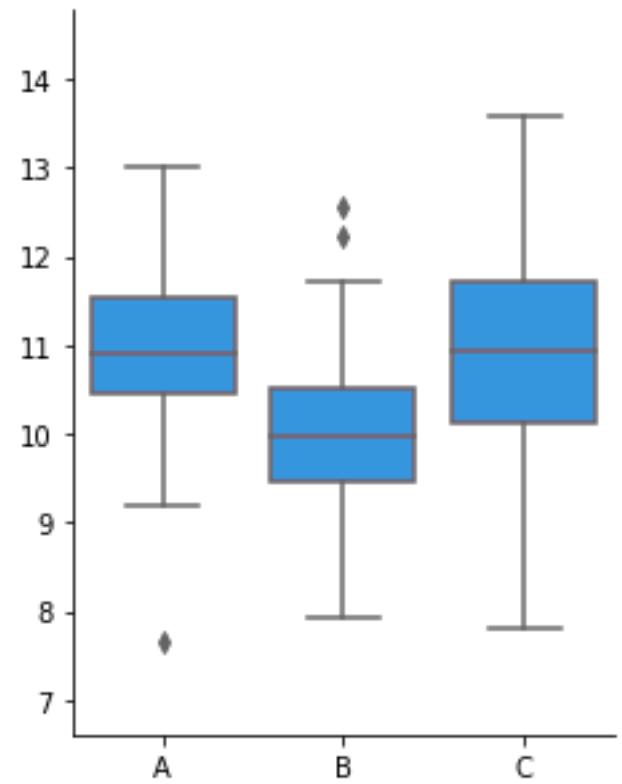
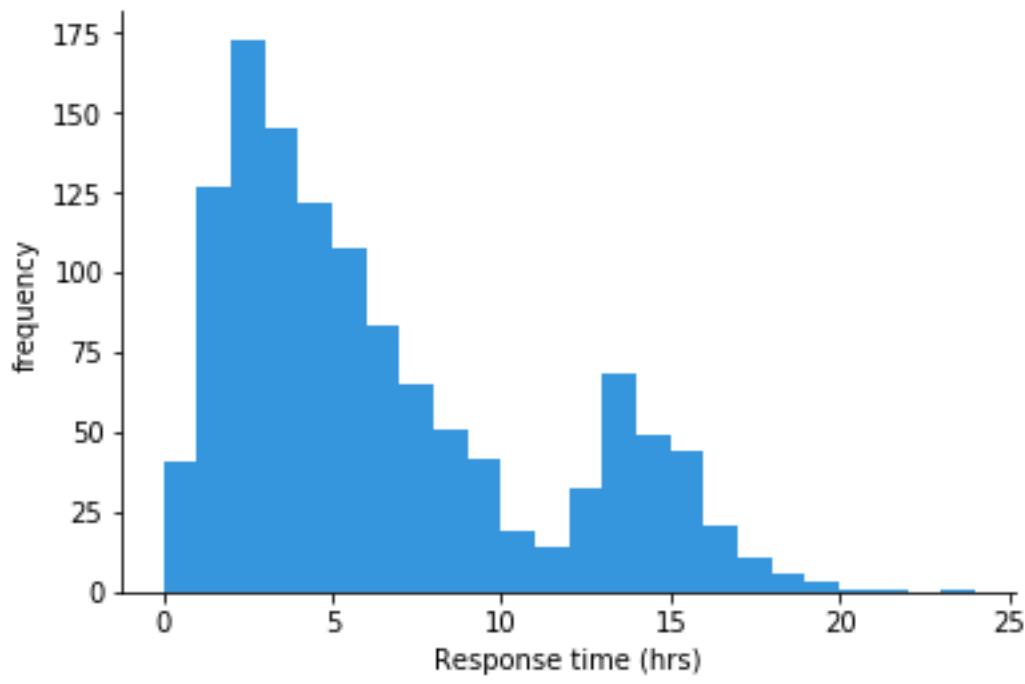


Market Share for Films Studios (Jan 1 - Oct 6, 2019)



Area plots

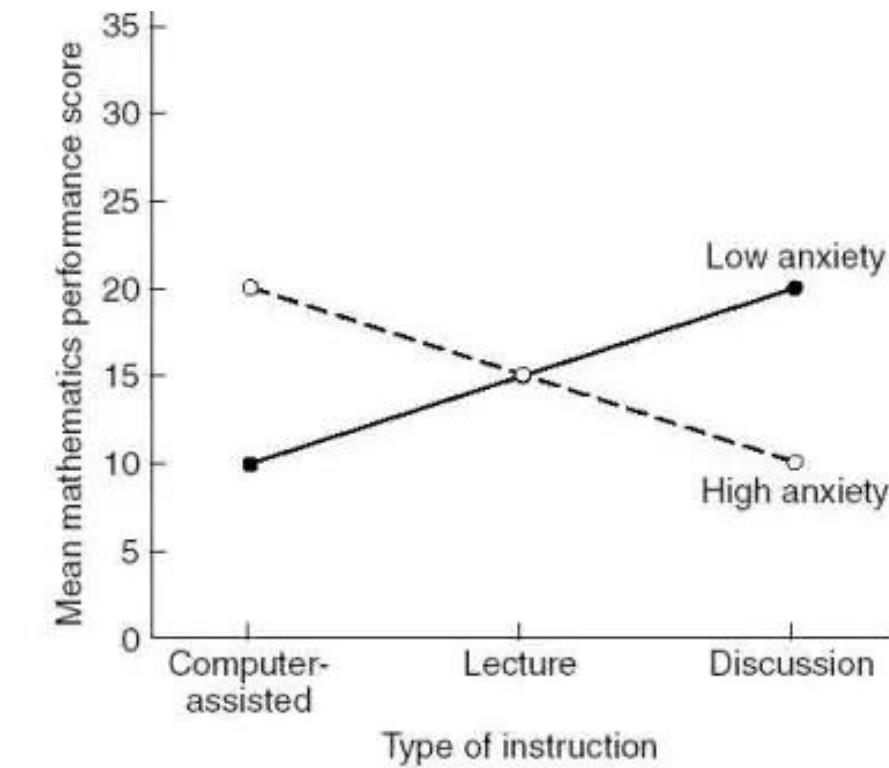
Data Distribution



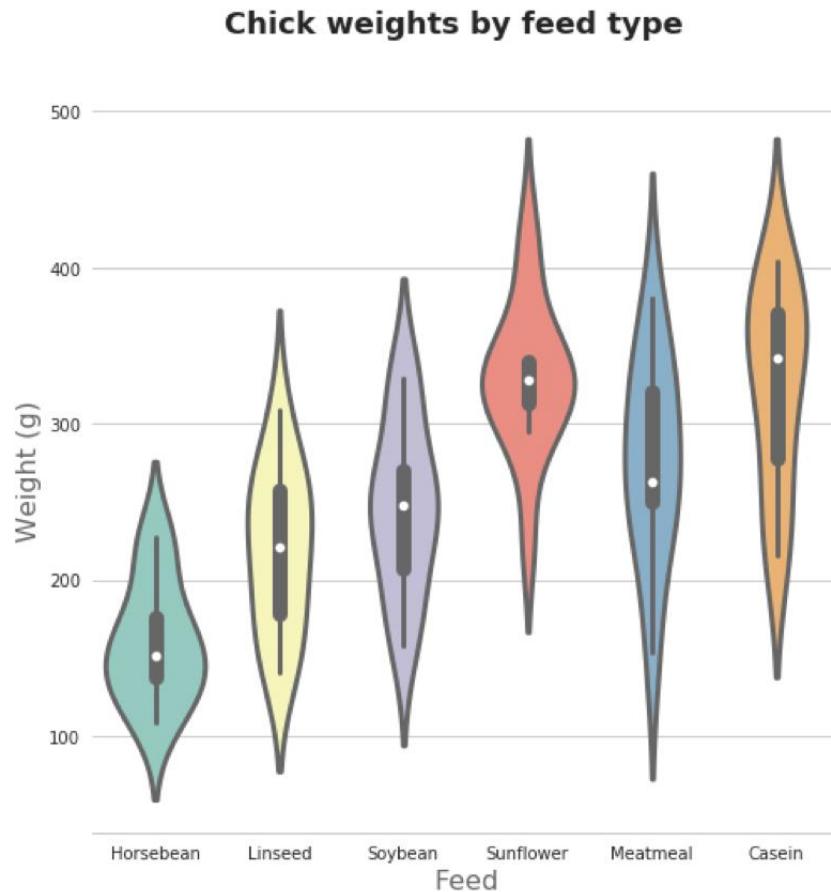
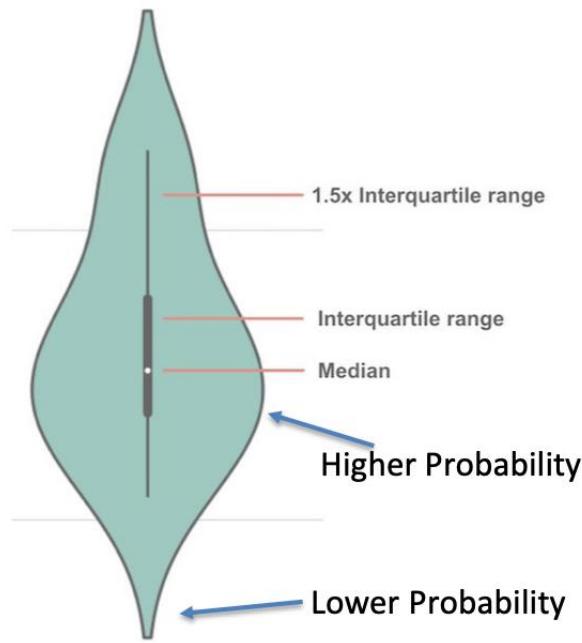
indicative of potential groups or group differences

Group Differences

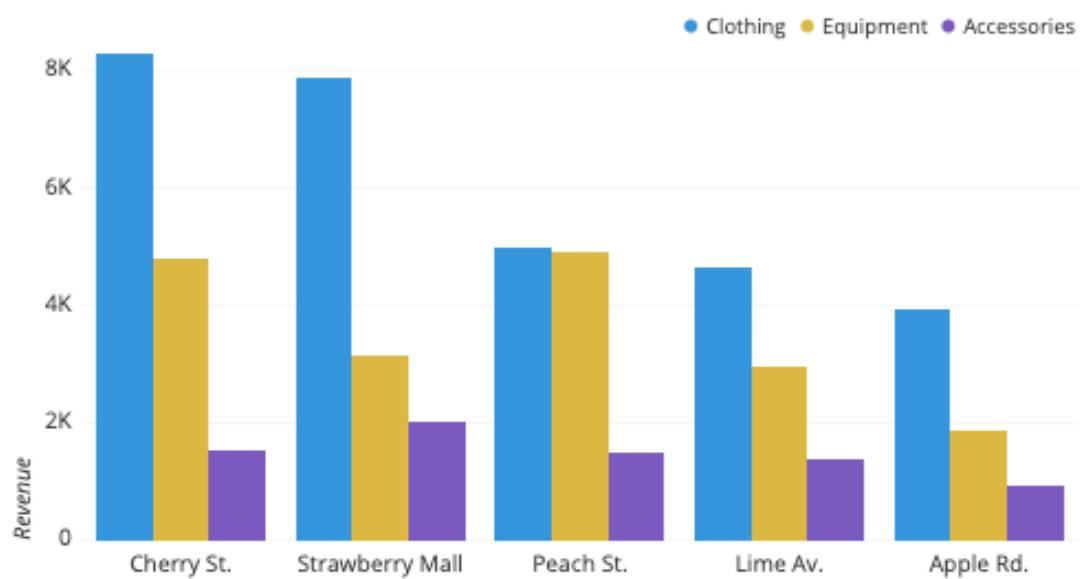
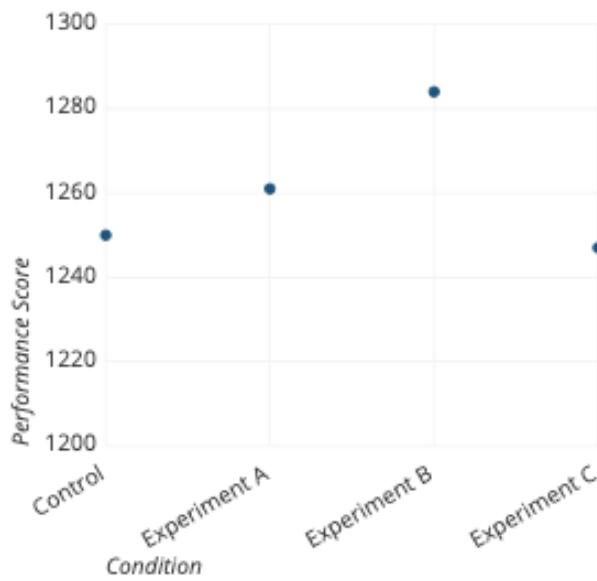
- main effects and interaction plots



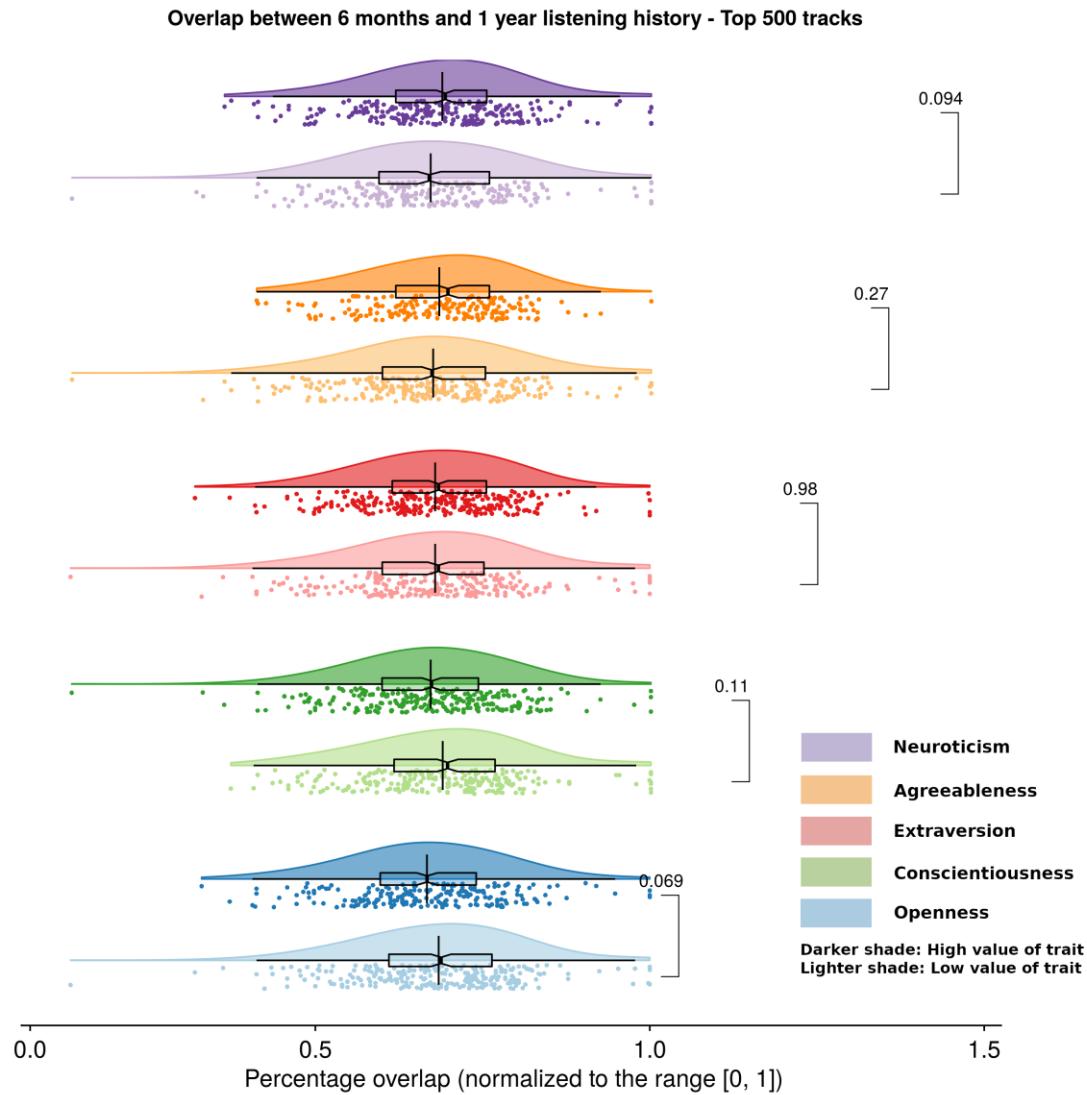
Data Distribution & Group Differences



Group differences

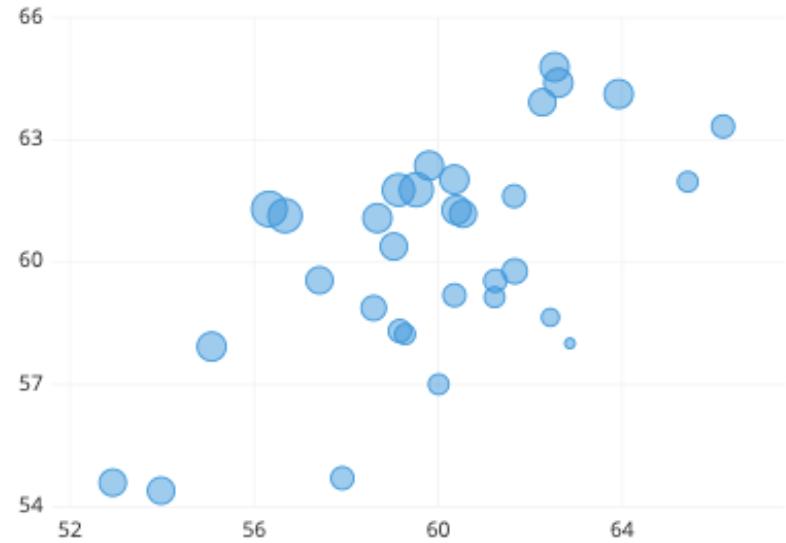
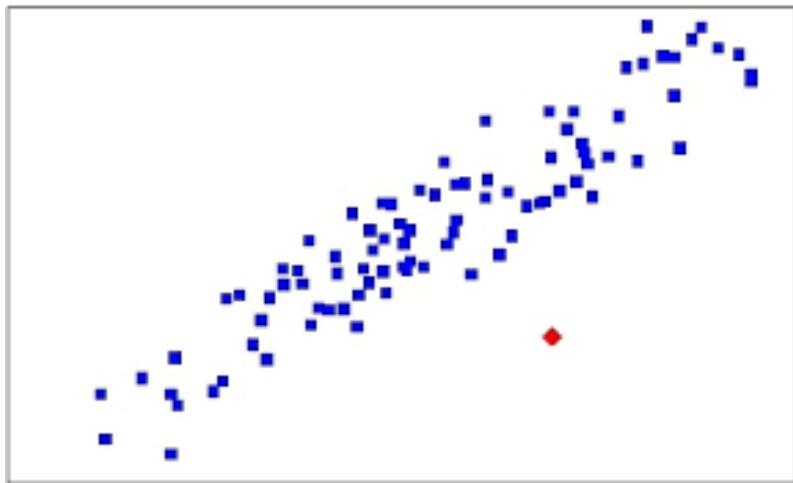


Describing Data + Group Differences



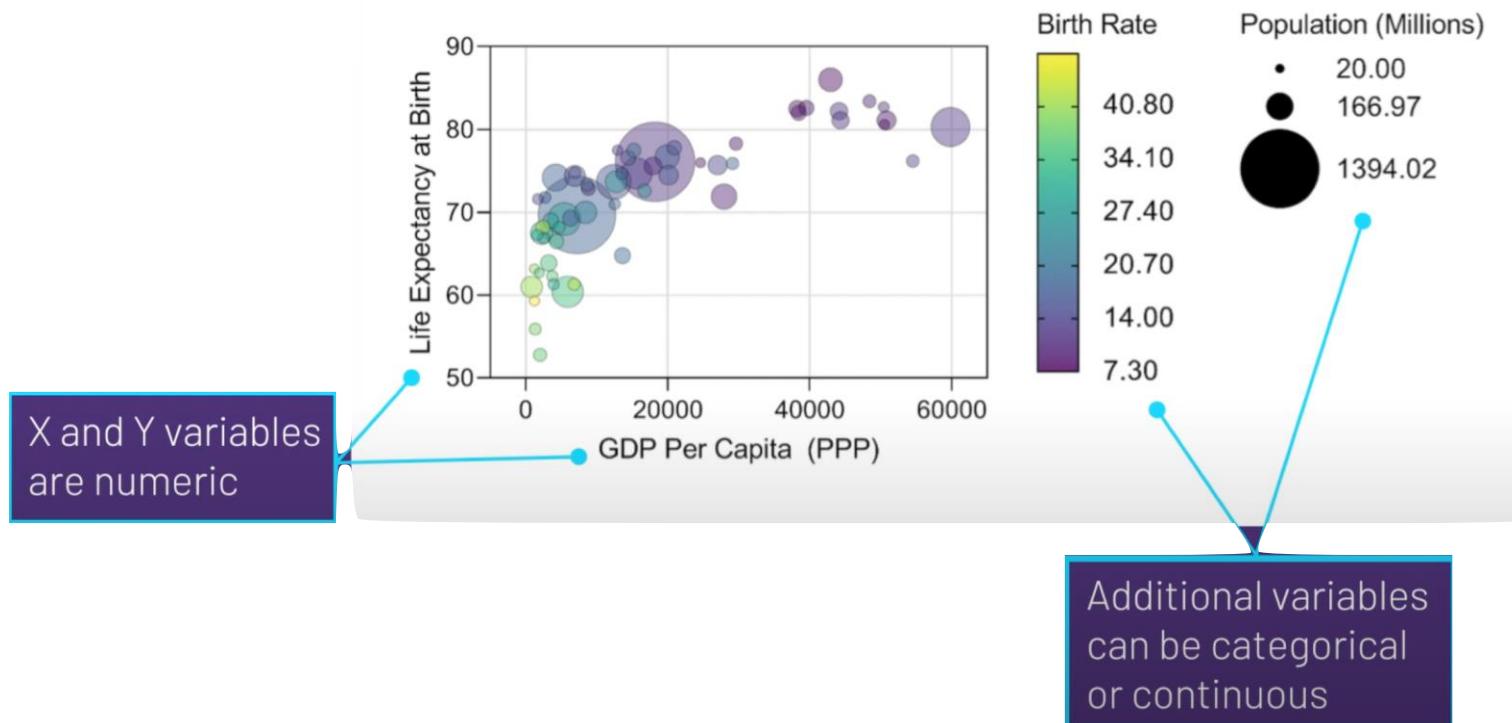
Association between variables

- scatter/bubble plots
 - allows you to observe the relationship between variables



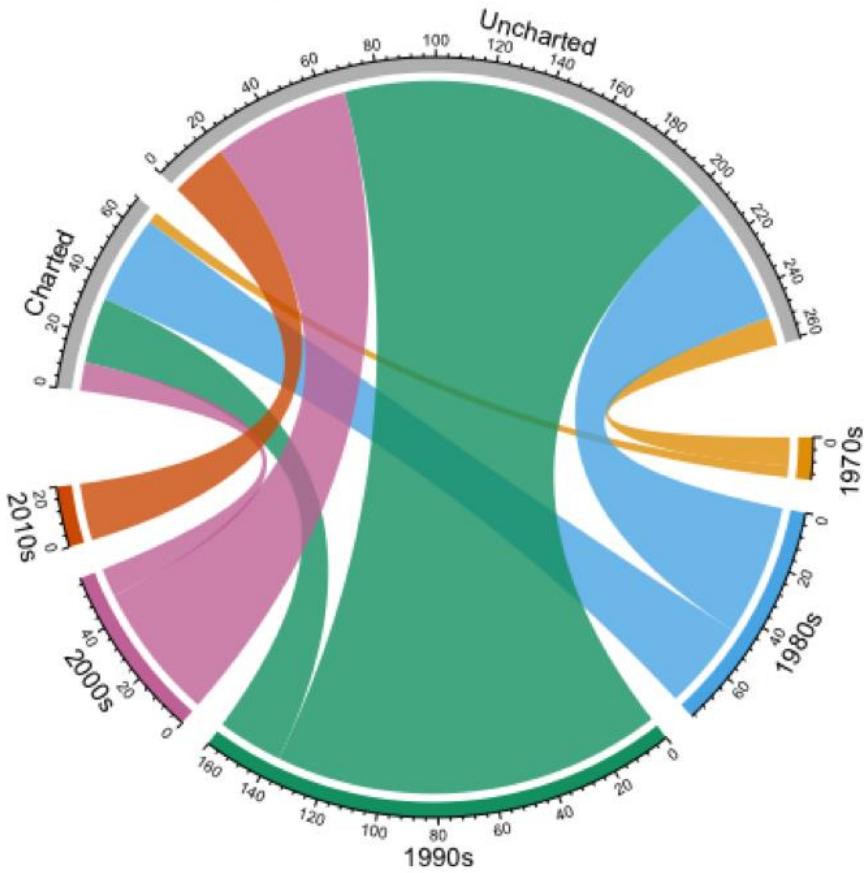
Association between variables

- bubble plots

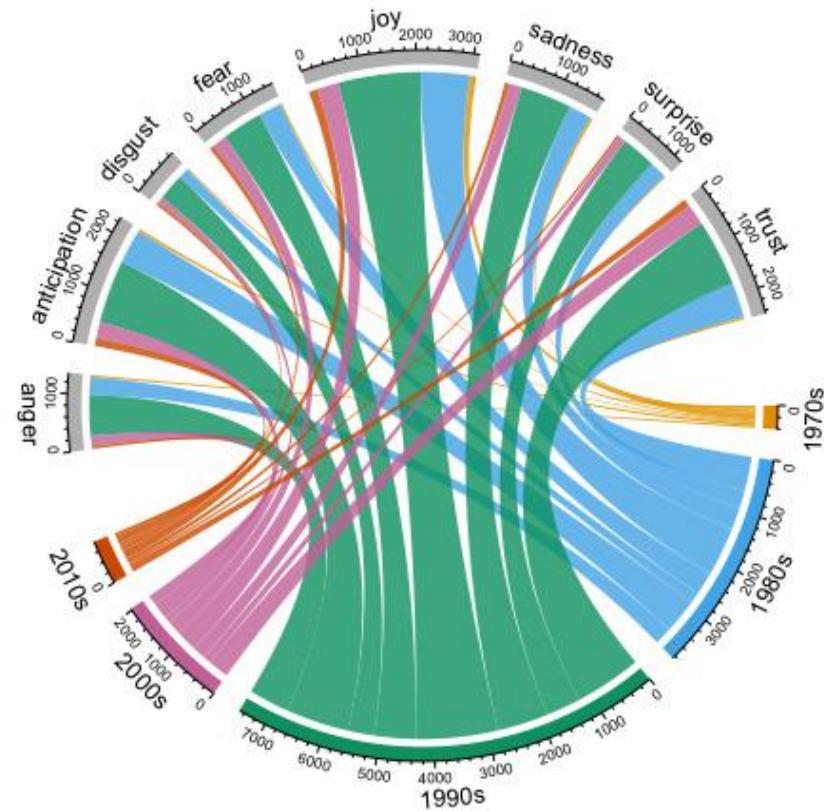


Association between variables

Relationship Between Chart and Decade



Relationship Between Mood and Decade



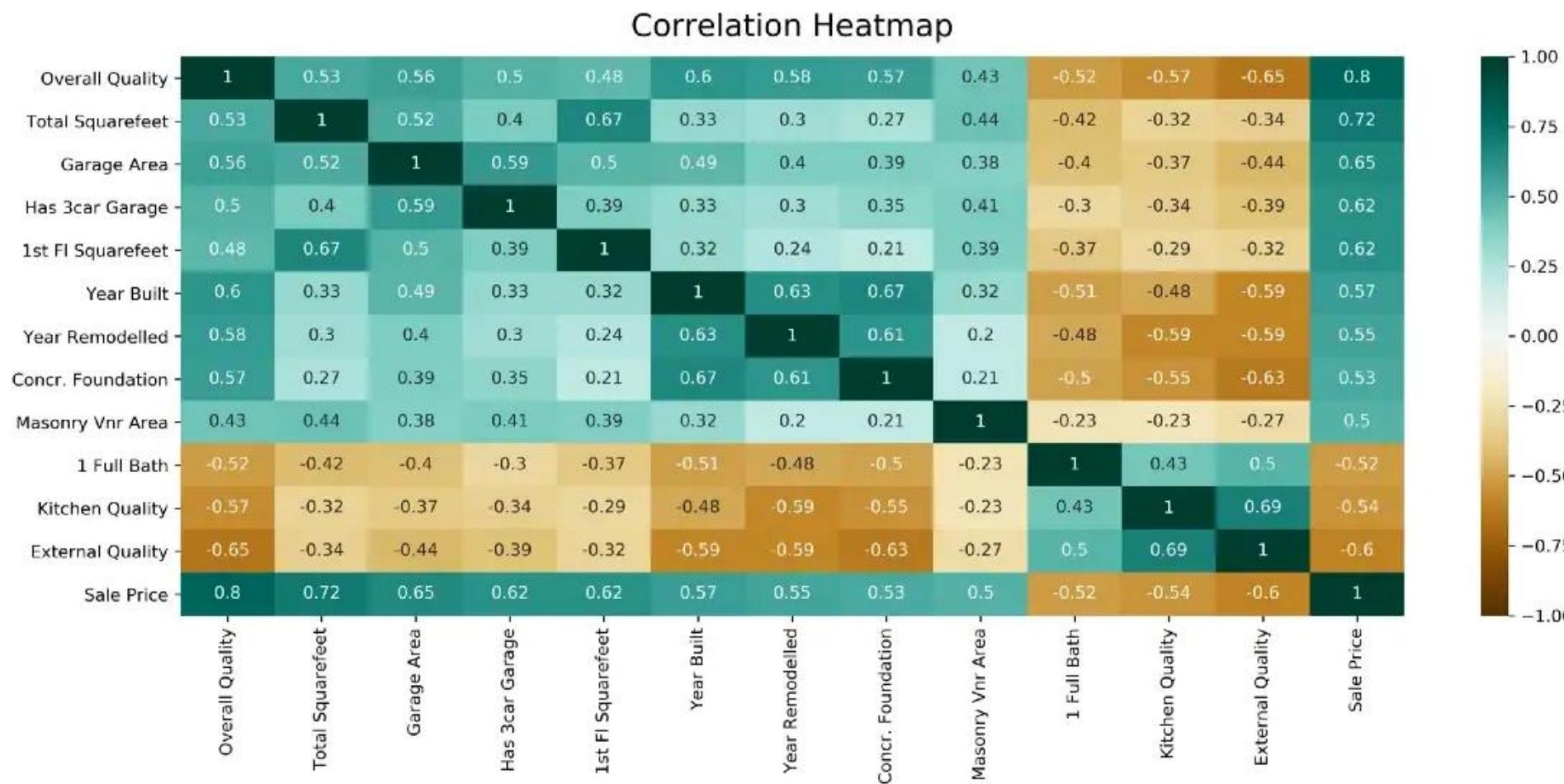
Association between variables

- heat maps depicting correlations

	Overall Qual	Total SF	Garage Area	Garage Cars_3.0	1st Flr SF	Year Built	Year Remod/Add	Foundation_PConc	Mas Vnr Area	Full Bath_1	Kitchen Qual_TA	Exter Qual_TA	SalePrice
Overall Qual	1.000000	0.534259	0.563904	0.502657	0.477136	0.602964	0.584654	0.571092	0.430041	-0.521553	-0.568011	-0.646351	0.800207
Total SF	0.534259	1.000000	0.524145	0.399740	0.668871	0.331811	0.300193	0.270644	0.441001	-0.418993	-0.316613	-0.341000	0.716714
Garage Area	0.563904	0.524145	1.000000	0.589214	0.498690	0.488023	0.397731	0.393544	0.380563	-0.402050	-0.365930	-0.435269	0.649897
Garage Cars_3.0	0.502657	0.399740	0.589214	1.000000	0.391699	0.333050	0.303772	0.349473	0.405799	-0.295060	-0.336226	-0.394001	0.619110
1st Flr SF	0.477136	0.668871	0.498690	0.391699	1.000000	0.323315	0.244190	0.212511	0.386482	-0.369359	-0.293941	-0.318021	0.618486
Year Built	0.602964	0.331811	0.488023	0.333050	0.323315	1.000000	0.629116	0.666546	0.320780	-0.509293	-0.478751	-0.591403	0.571849
Year Remod/Add	0.584654	0.300193	0.397731	0.303772	0.244190	0.629116	1.000000	0.608503	0.204234	-0.483858	-0.585228	-0.590271	0.550370
Foundation_PConc	0.571092	0.270644	0.393544	0.349473	0.212511	0.666546	0.608503	1.000000	0.208299	-0.500180	-0.550170	-0.626157	0.529047
Mas Vnr Area	0.430041	0.441001	0.380563	0.405799	0.386482	0.320780	0.204234	0.208299	1.000000	-0.229672	-0.226351	-0.269285	0.503579
Full Bath_1	-0.521553	-0.418993	-0.402050	-0.295060	-0.369359	-0.509293	-0.483858	-0.500180	-0.229672	1.000000	0.425653	0.496703	-0.520016
Kitchen Qual_TA	-0.568011	-0.316613	-0.365930	-0.336226	-0.293941	-0.478751	-0.585228	-0.550170	-0.226351	0.425653	1.000000	0.690116	-0.540860
Exter Qual_TA	-0.646351	-0.341000	-0.435269	-0.394001	-0.318021	-0.591403	-0.590271	-0.626157	-0.269285	0.496703	0.690116	1.000000	-0.600362
SalePrice	0.800207	0.716714	0.649897	0.619110	0.618486	0.571849	0.550370	0.529047	0.503579	-0.520016	-0.540860	-0.600362	1.000000

Association between variables

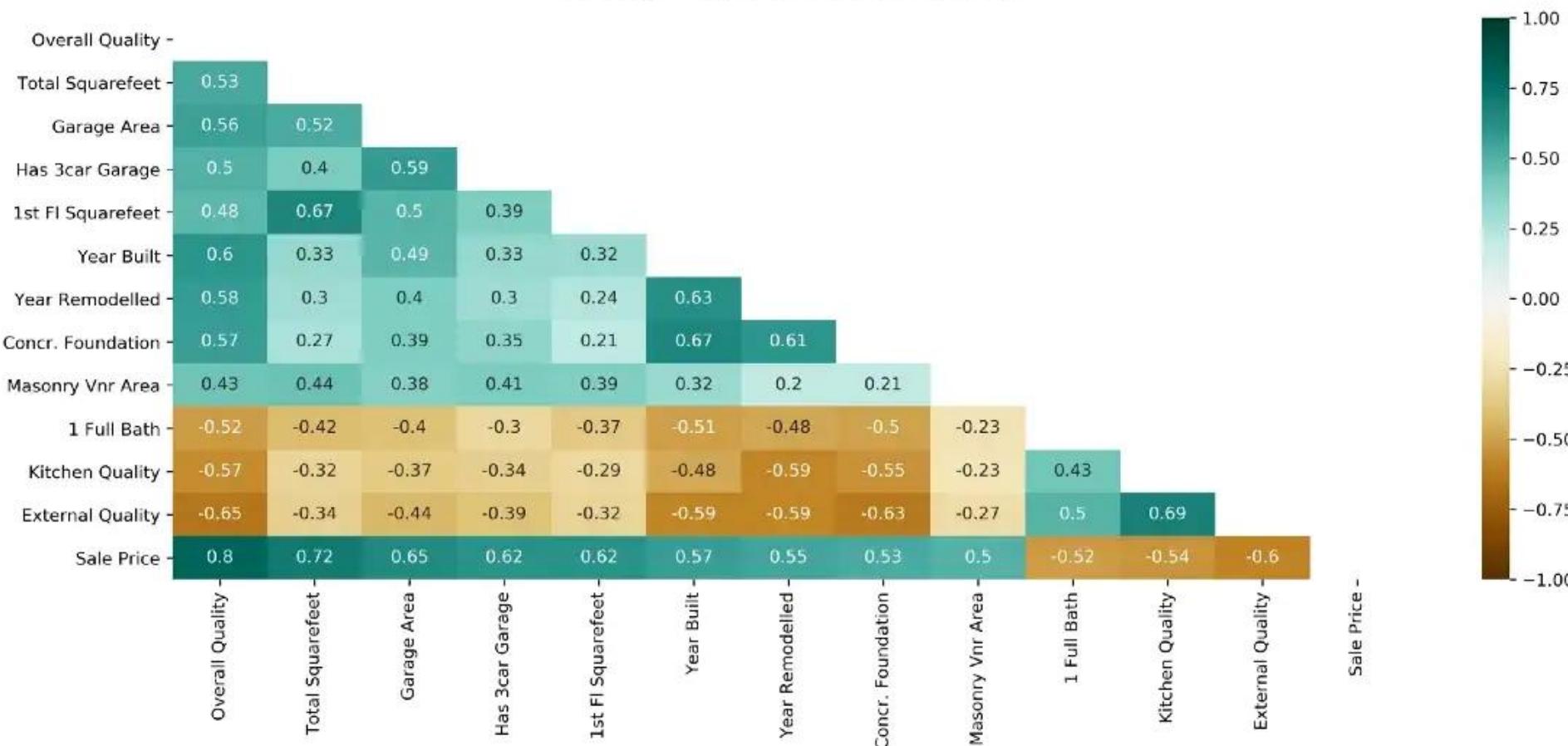
- heat maps depicting correlations



Association between variables

- heat maps depicting correlations

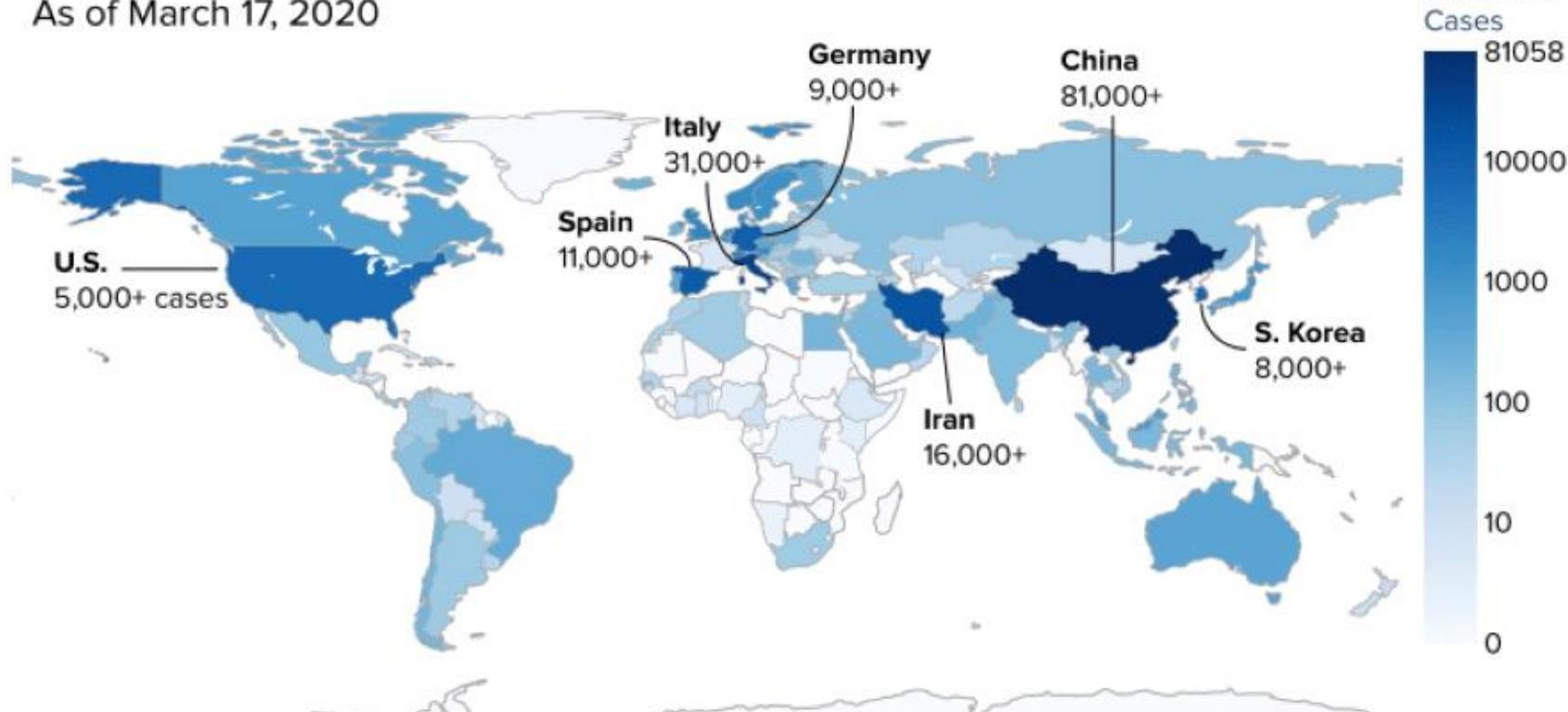
Triangle Correlation Heatmap



Geographical maps

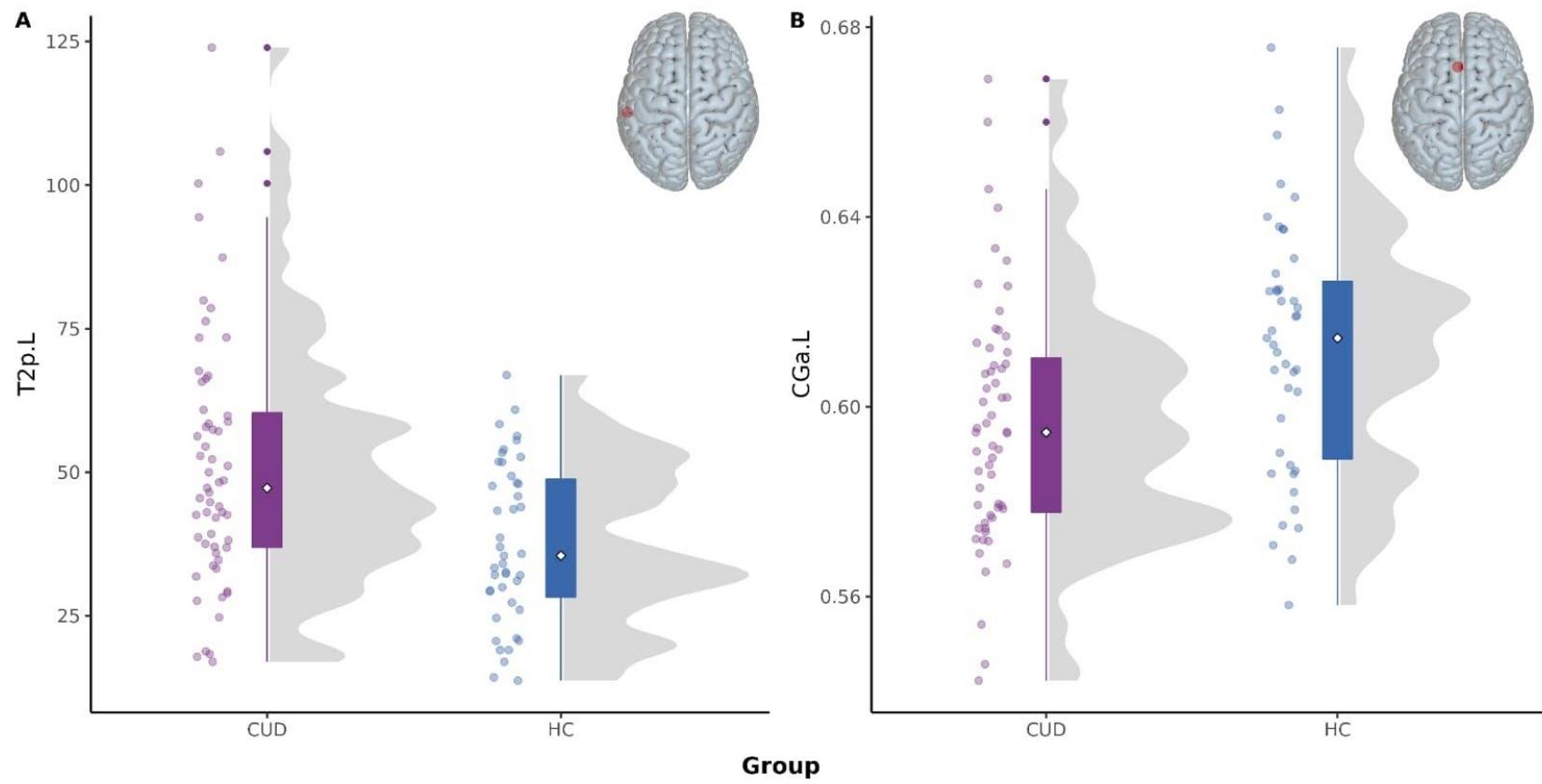
Reported coronavirus cases worldwide

As of March 17, 2020



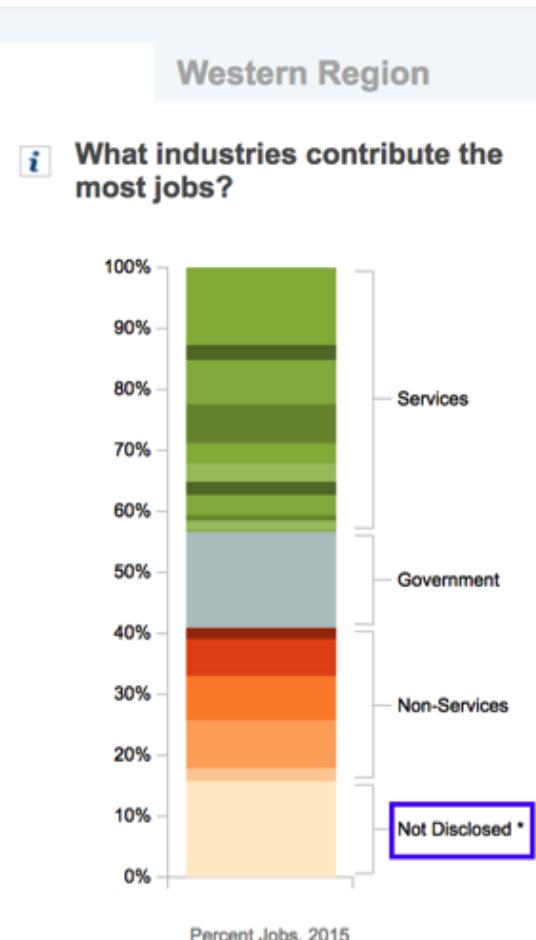
SOURCE: Johns Hopkins University. Data as of March 17, 2020 at 6 p.m. ET

Creative Combinations



What makes a good visualisation

- Storytelling
- Reduce Cognitive Load
- Less is more
- Missing data
- Color Consistency
- Labelling



closure of potentially confidential information. Categories where val
conomic Profile System (EPS) to create detailed reports.

To do or not to do

- Provide necessary Context around Visuals
- Ensure Simplicity and Clarity of Information
- Ensure Brevity and Avoid Unnecessary Information
- Use Simple and Easy to Understand Color Palettes
- Pay attention to Graphics in order to make sure that they are Visually Appealing
- Where possible, bring in Originality by relating, seemingly Unrelated data and subjects

To do or not to do

- Avoid using Too Many Variables within a single image which might result in distracting the viewers
- Be extremely careful of not visualizing data through an Unsuitable or Incorrect visualization format
- While using Scales in Data Visualization in order to depict differences between data points, it is important to ensure that the scale is consistent
- Poor Choice of Colors is another significant issue which should be avoided at all costs. Thus, it is important to:
 - avoid using colors with negligible contrast
 - avoid using too many colors
 - avoid using conventional colors to convey opposite meanings
 - pay heed to the needs of people who might be colorblind (check also in grayscale)

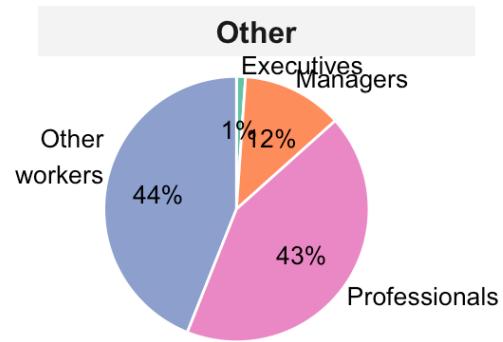
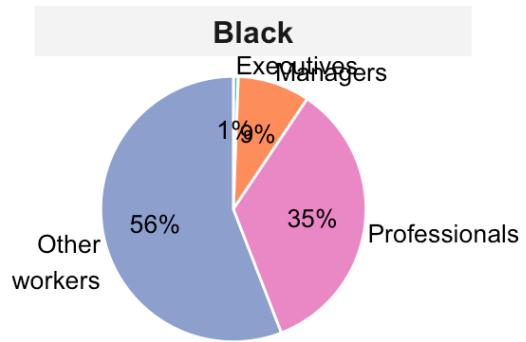
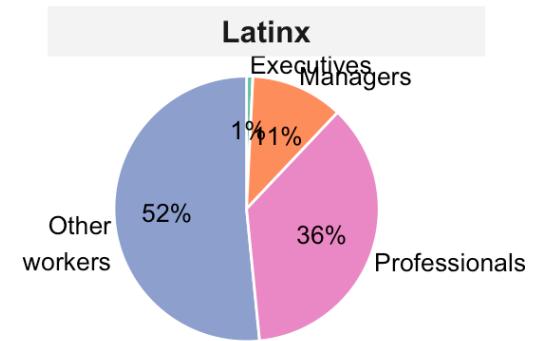
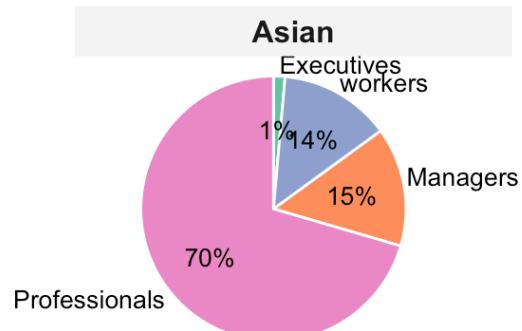
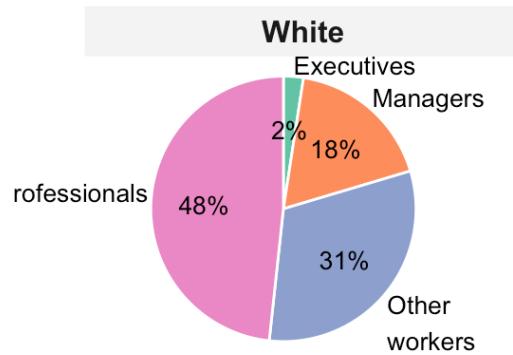


Outline

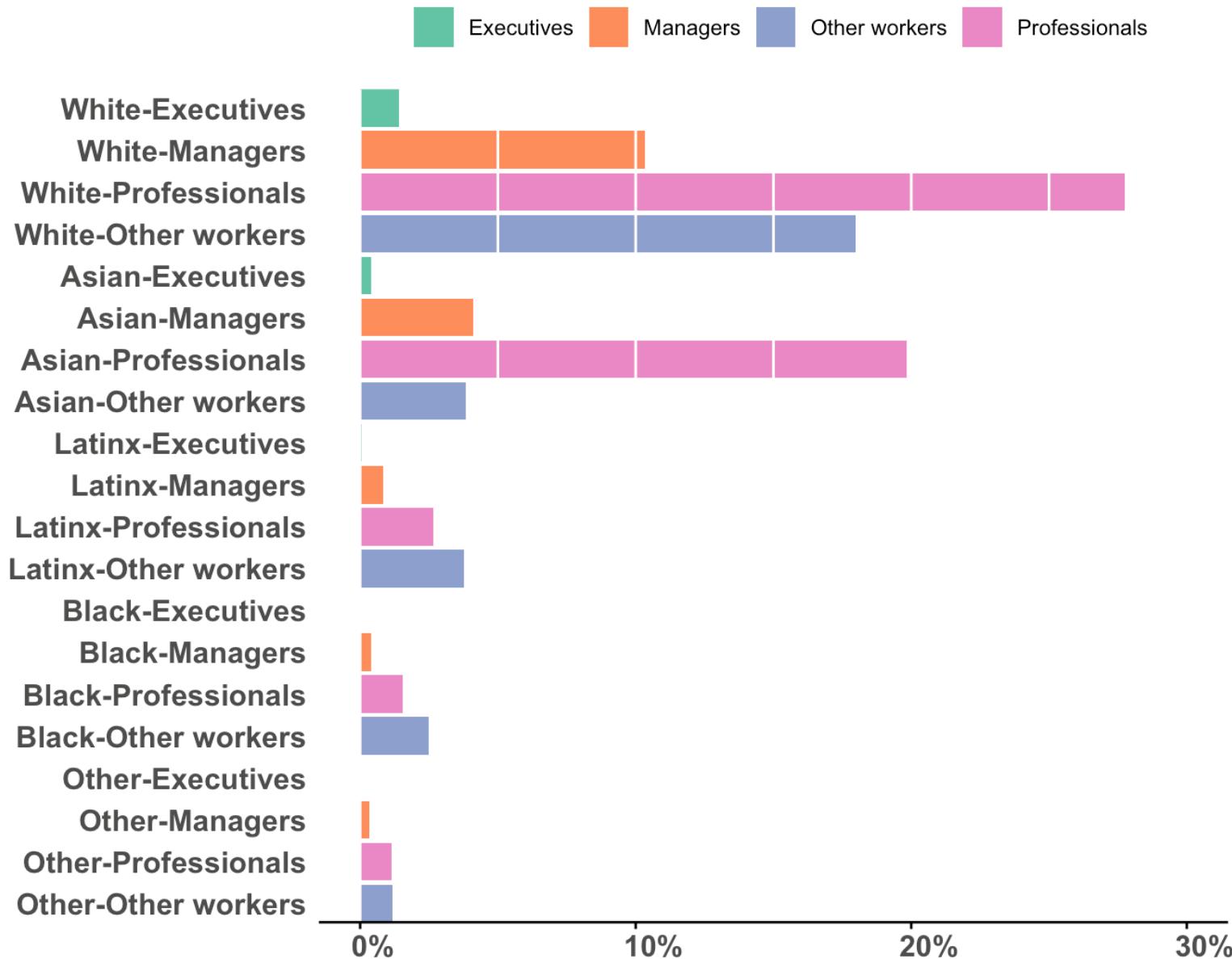
- **Visualization**
 - why we visualise
 - how to pick a plot
 - initial data vs final results visualization (some examples)
 - **bad designs and misleading graphs**
- **Summarization**
 - measures of central tendency & dispersion
 - which measure to pick

Bad Designs & Improvements

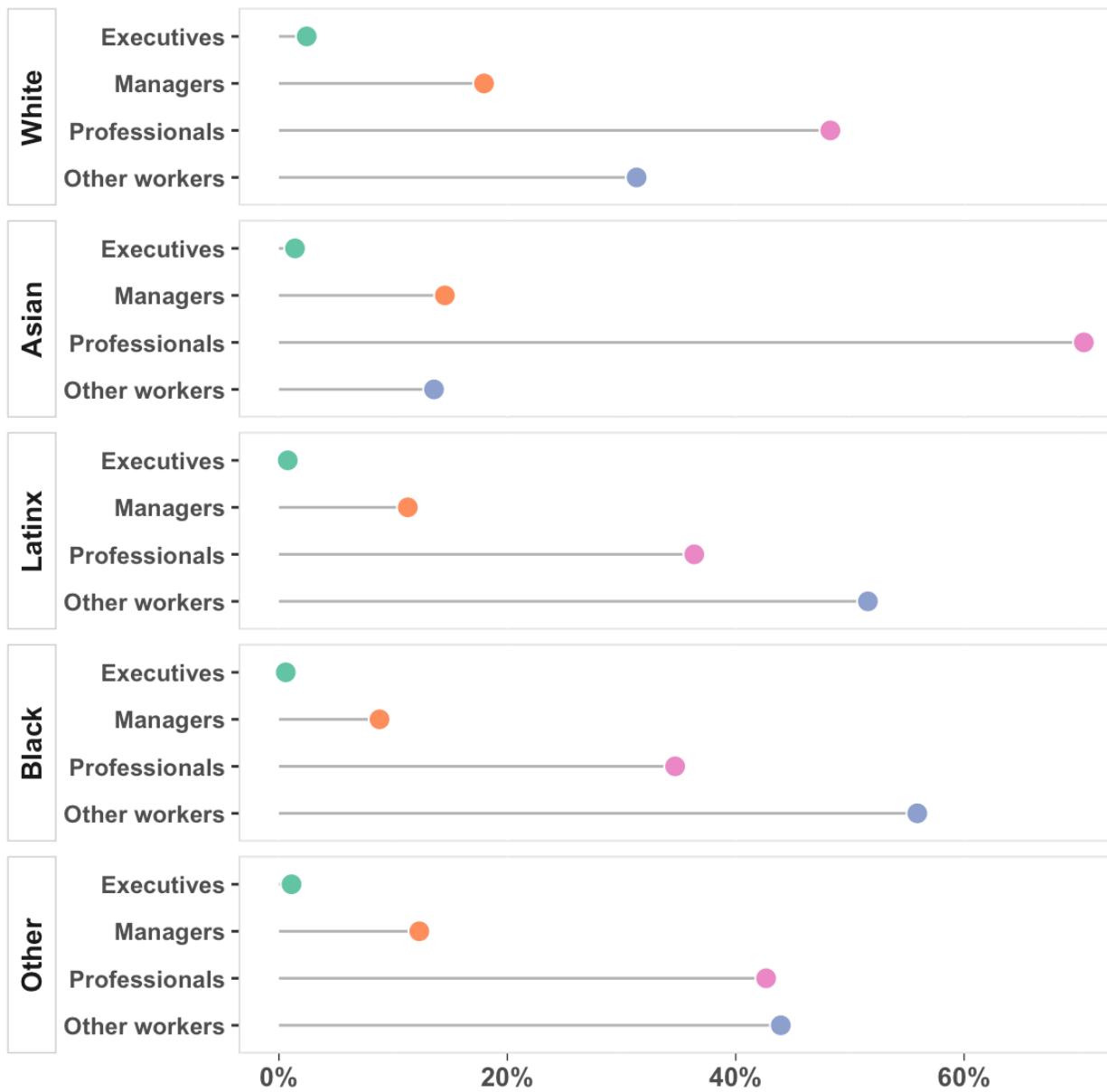
<https://nandeshwar.info/data-visualization/pie-chart-vs-bar-chart/>



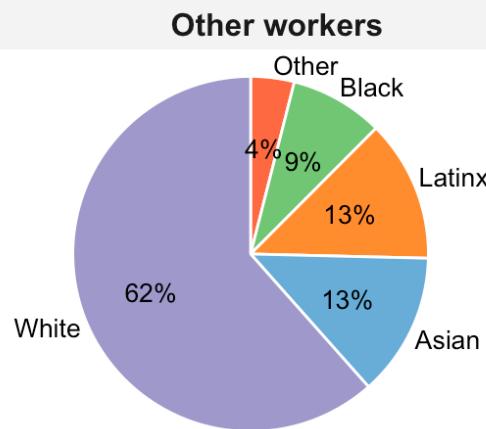
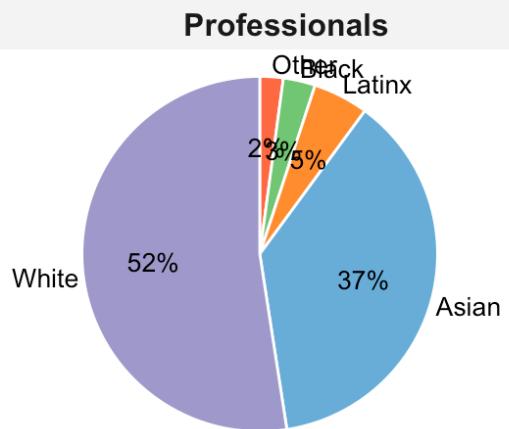
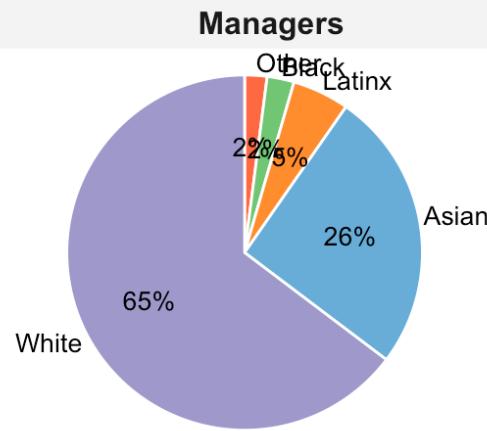
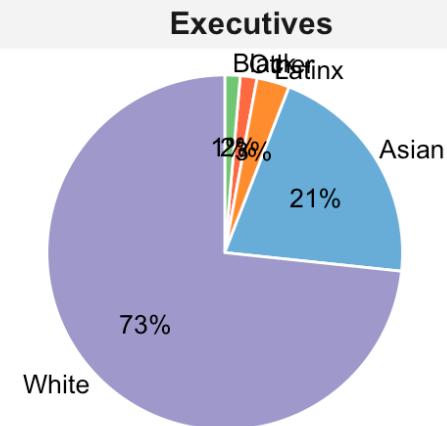
Source: Reveal, <https://www.revealnews.org/topic/silicon-valley-diversity/>



Source: Reveal, <https://www.revealnews.org/topic/silicon-valley-diversity/>



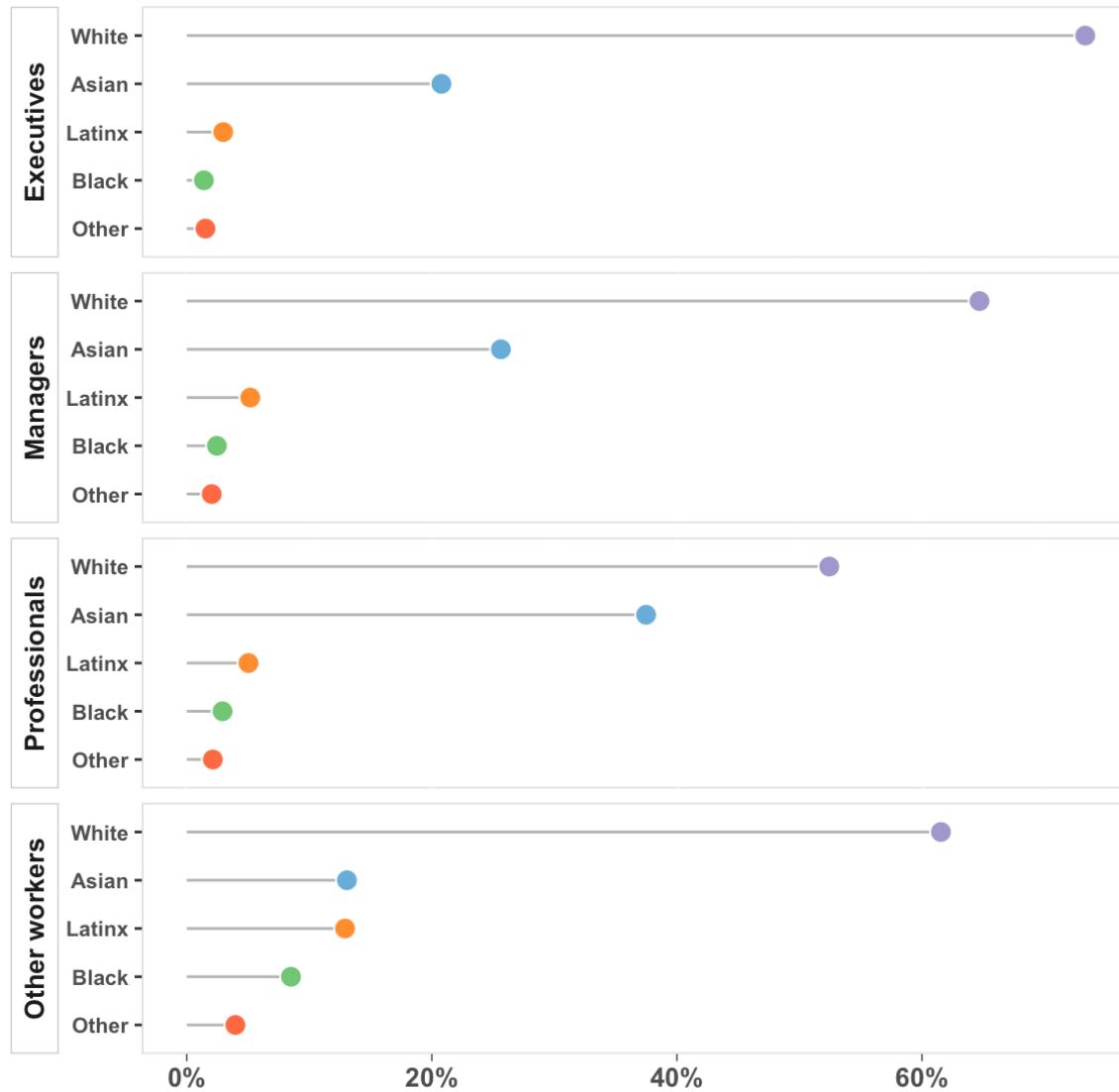
Source: Reveal, <https://www.revealnews.org/topic/silicon-valley-diversity/>



Source: Reveal, <https://www.revealnews.org/topic/silicon-valley-diversity/>



Source: Reveal, <https://www.revealnews.org/topic/silicon-valley-diversity/>



Source: Reveal, <https://www.revealnews.org/topic/silicon-valley-diversity/>

What if we want to compare genders
within the job categories and
ethnicities/races?

Job categories and ethnicity/race distribution by gender

□ Female □ Male

Executives



Managers



Of all female executives,
Black females are about
2% of them, and of all
male executives, Black males
are about 1% of them

Professionals



Other workers



0% 5% 10% 30% 50% 70%

0% 5% 10% 30% 50% 70%

Note: The x-axis is transformed using the square root function to see smaller values. Source: Reveal, <https://www.revealnews.org/topic/silicon-valley-diversity/>

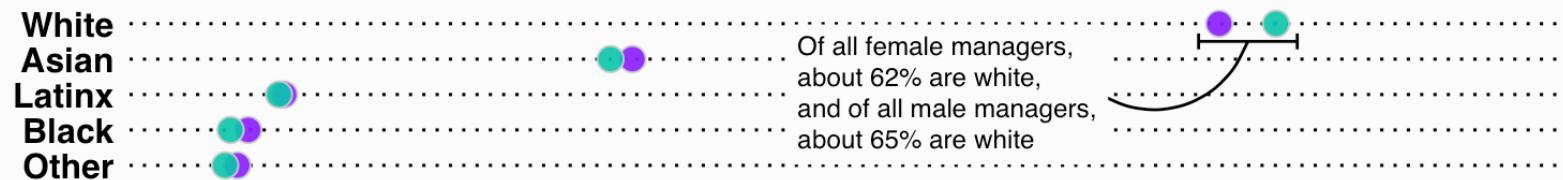
Job categories and ethnicity/race distribution by gender

○ Female ○ Male

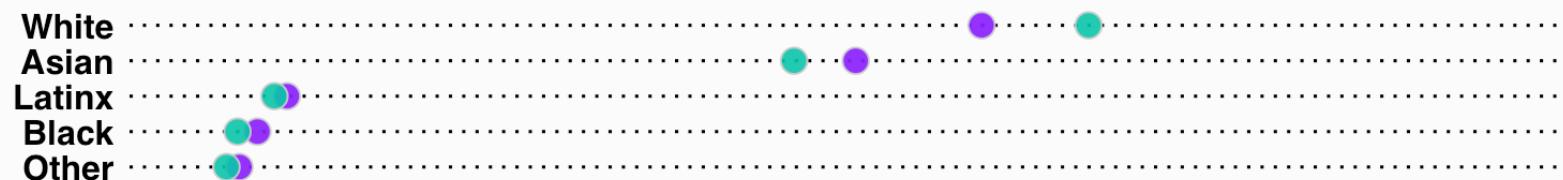
Executives



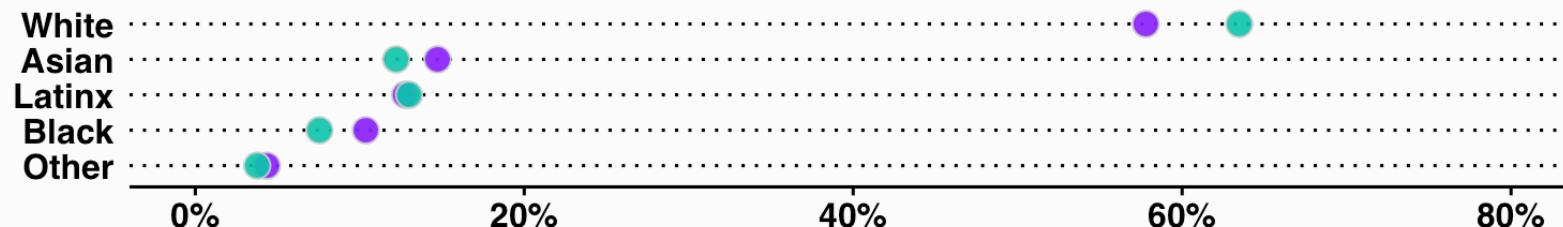
Managers



Professionals



Other workers



Job categories and ethnicity/race distribution by gender

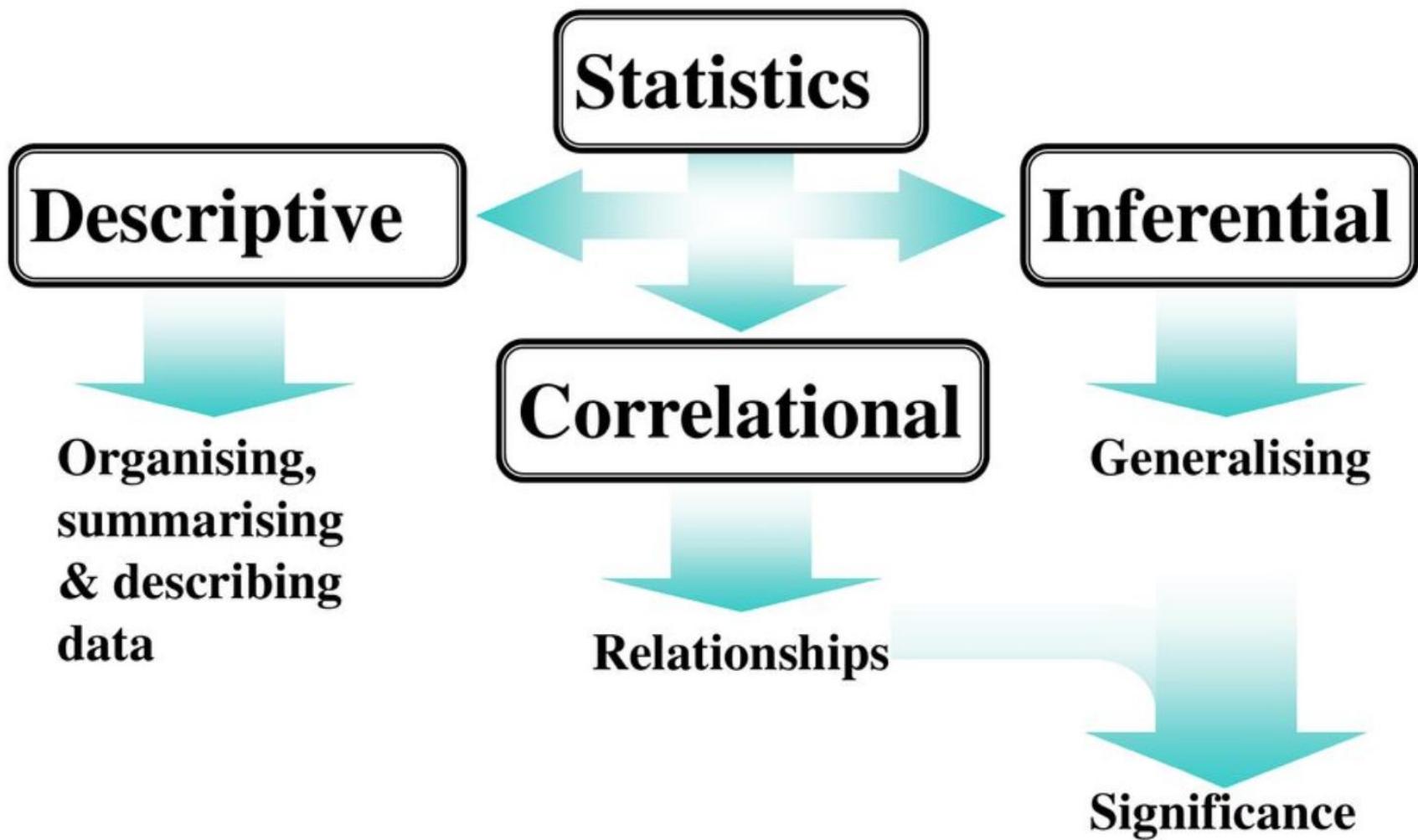
○ Female ○ Male





Outline

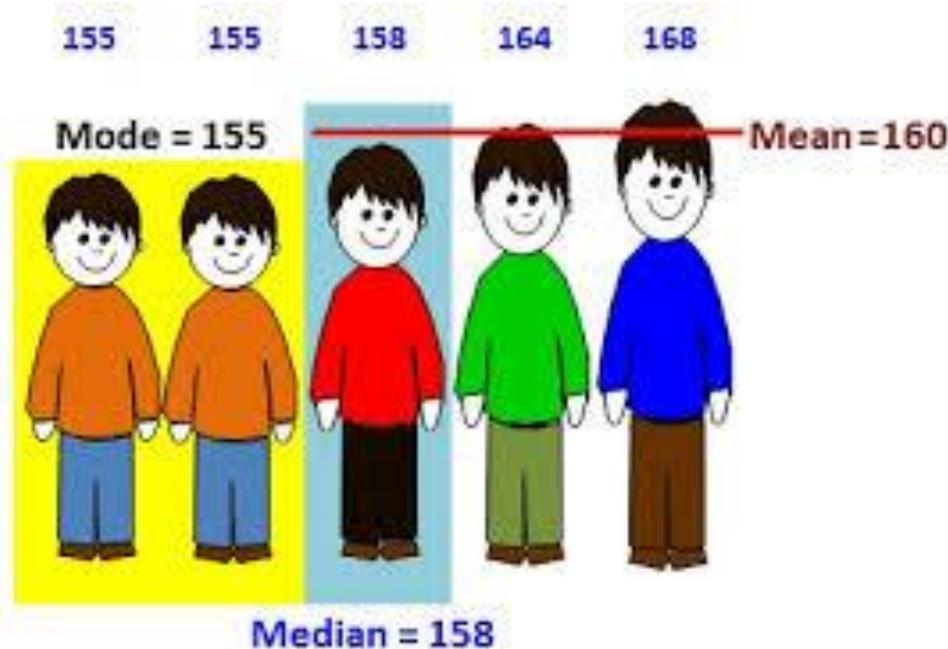
- **Visualization**
 - why we visualise
 - how to pick a plot
 - initial data vs final results visualization (some examples)
 - bad designs and misleading graphs
- **Summarization**
 - **measures of central tendency & dispersion**
 - **which measure to pick**



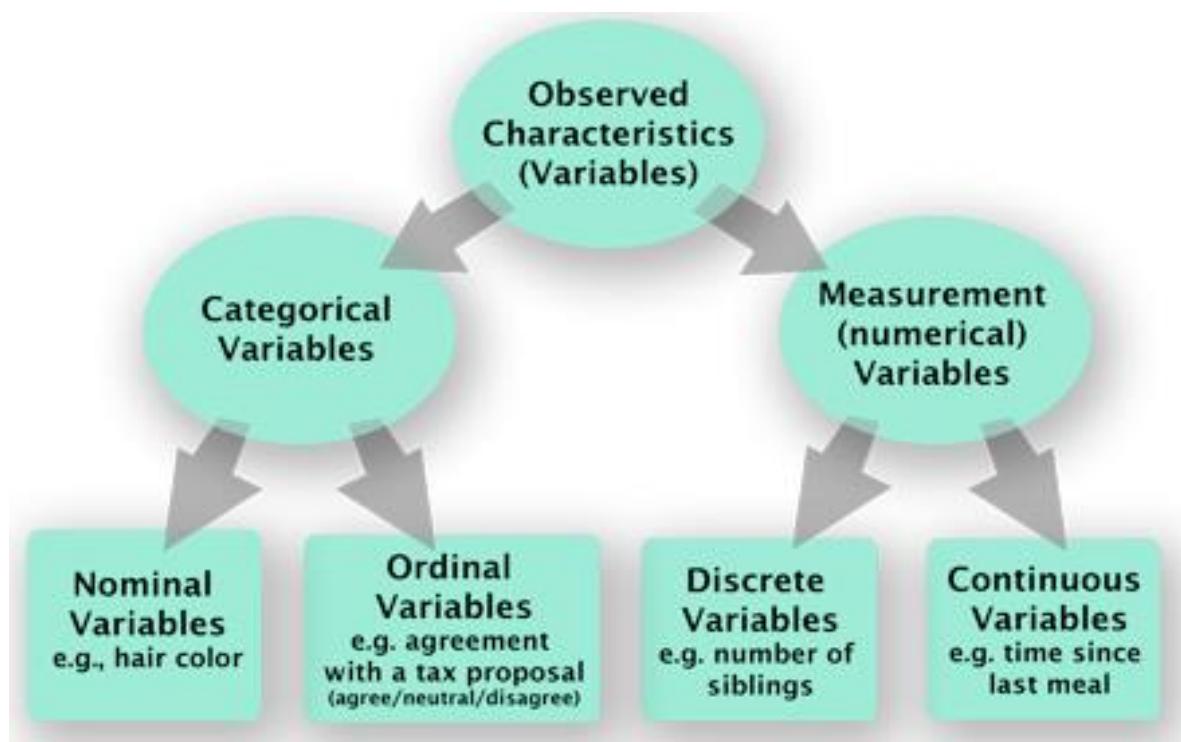
Descriptive Statistics

- Common descriptive statistics are:
 - Measure of **central tendency**
 - the most typical value of a given group of values
 - Measure of **dispersion**
 - how much all the other values in the group vary around the typical value

Measures of central tendency



Central Tendency for Variable Types



MODE

MEDIAN

MEAN MEAN

MODE
MEDIAN

Measures of central tendency

Advantages

Disadvantages

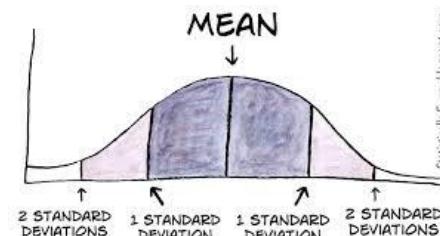
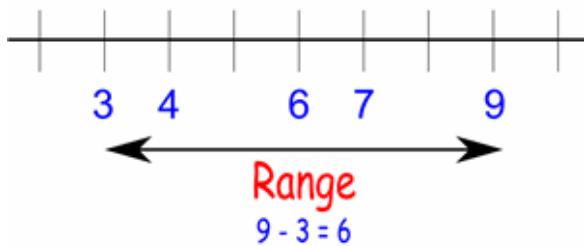
MEAN

:tion

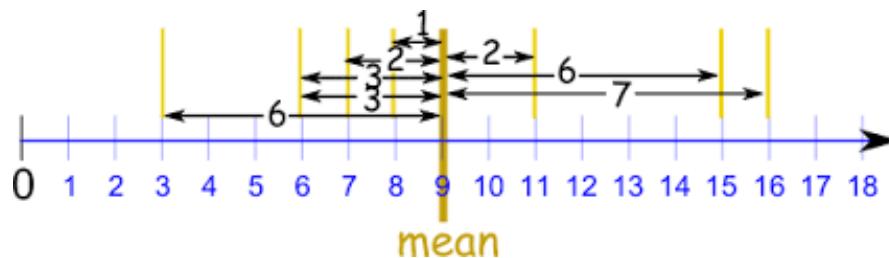
MEDIAN

MODE

Measures of dispersion/spread

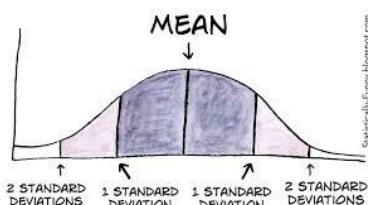
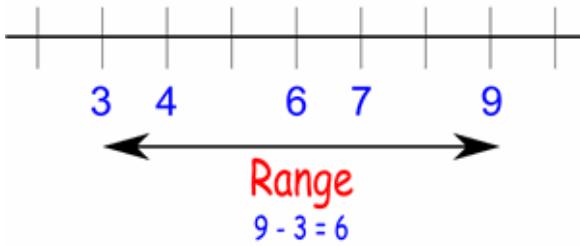


$$SD = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$



$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

Measures of dispersion/spread



$$SD = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

Advantages

--

Disadvantages

distorted by extreme values
no indication of grouping
around the mean

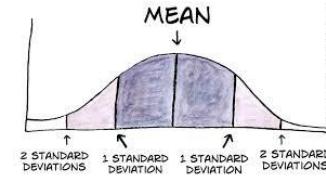
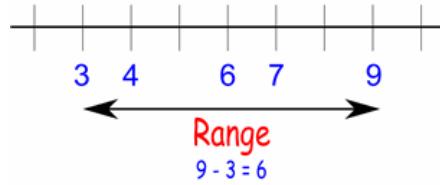
- Fundamental to significance testing, and forms basis of Analysis of Variance (ANOVA)
- Enables population parameters to be estimated from a sample of people

--

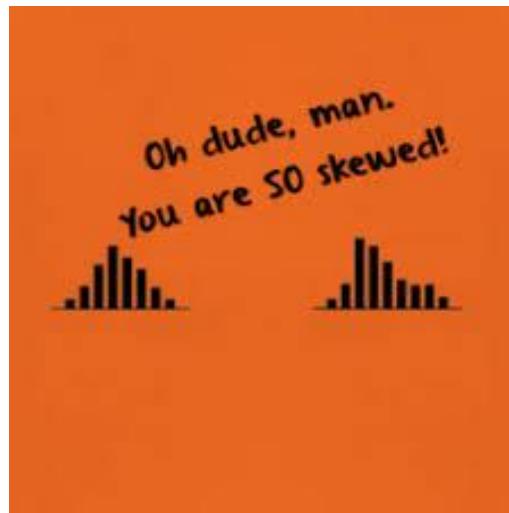
MEAN

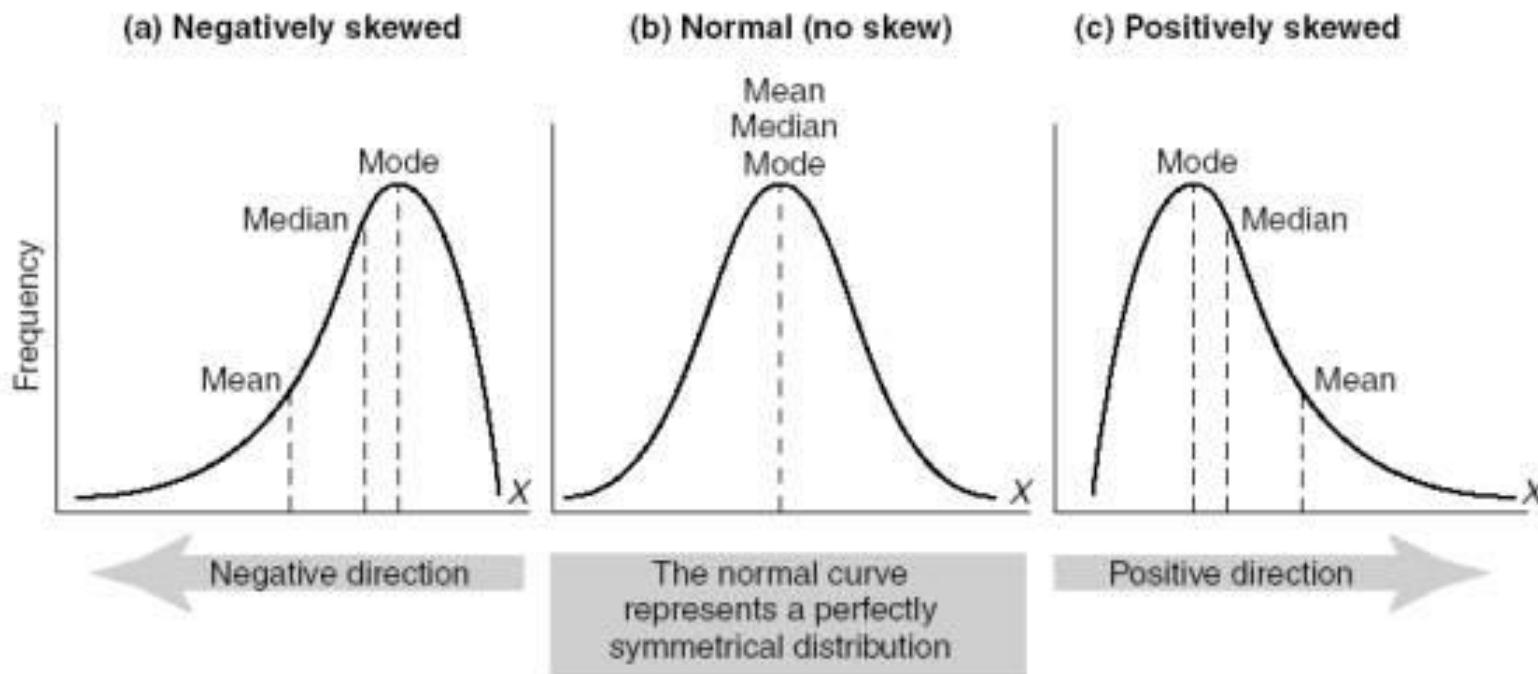
?

MODE MEDIAN

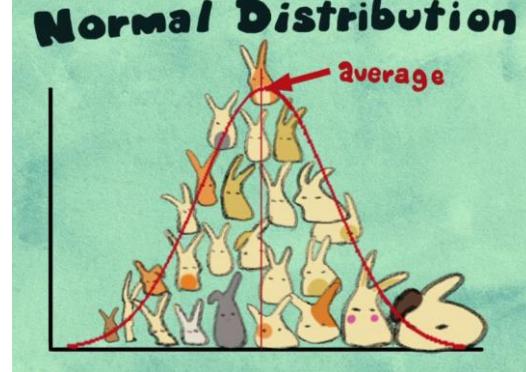


When do these measures fail to be representative ????



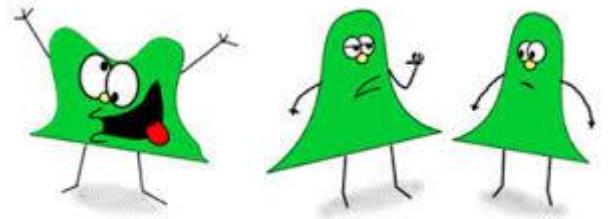


Normal Distribution



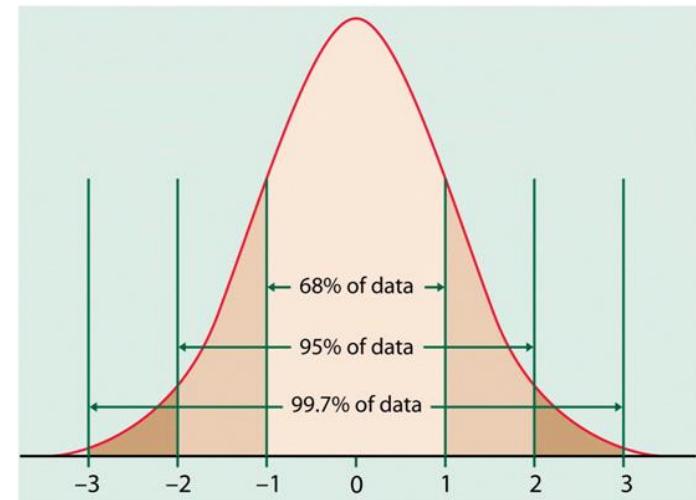
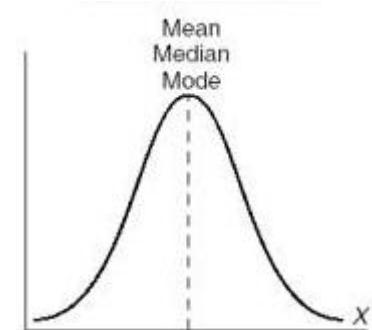
- A bell-shaped mathematical curve describing how values are distributed
- Data taken from a sample is **assumed** to be ‘normally distributed’, and must approximate this shape in order to use parametric tests of significance
- *Inferential statistics* (eg: t-tests, F-tests, regression analyses) require in some sense that the numeric variables are approximately normally distributed
- **Note:** it does not fit all populations

Normal Distribution

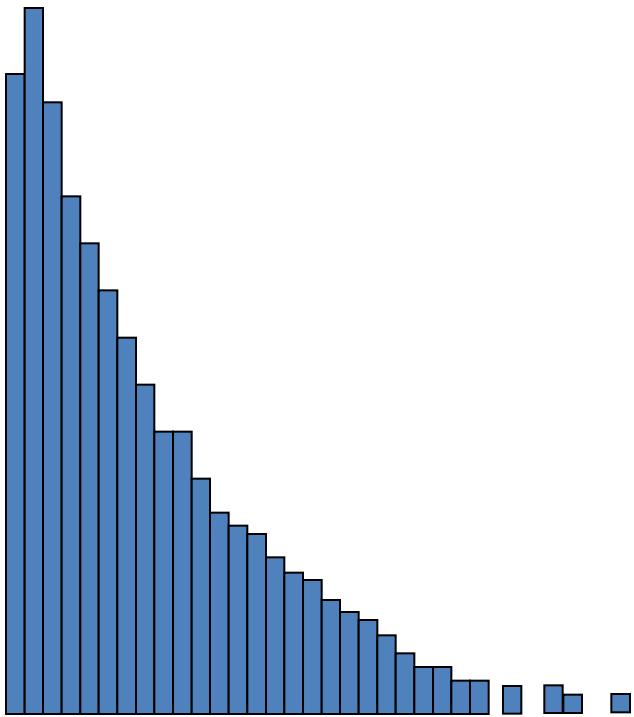


"KEEP YOUR EYE ON THAT GUY, TOM. HE'S NOT, YOU KNOW...NORMAL!"

- symmetrical about the horizontal axis midpoint
- mean, median, and mode all fall on the midpoint
- No matter what μ and σ are, the area between
 - $\mu-\sigma$ and $\mu+\sigma$ is about 68%;
 - $\mu-2\sigma$ and $\mu+2\sigma$ is about 95%;
 - $\mu-3\sigma$ and $\mu+3\sigma$ is about 99.7%
- Almost all values fall within 3 standard deviations

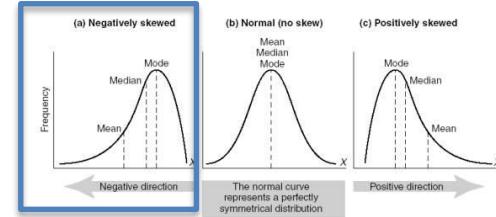


Skewed Distribution

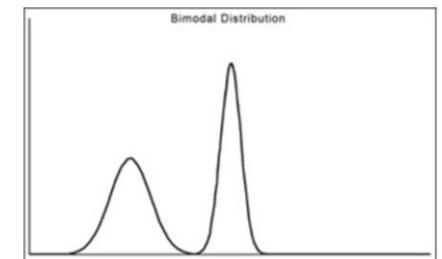


- Resembles an exponential distribution
- Lots of extreme values far from mean or mode
- Not straightforward to do useful statistical tests with this type of distribution

Skewed Distribution



- **Negative skew**
 - Result from relatively easy tasks, due to a ceiling effect
- **Positive skew**
 - Results from tasks which are hard to improve upon, due to a floor effect (such as RT —reaction time)
- **Bimodal**
 - Two distinct peaks
 - probable indicator of groups
 - ex: completion time of marathon runners



BRSM

Descriptive Statistics, Correlation

Vinoo Alluri & Bapi Raju

Descriptive

Organising,
summarising
& describing
data

Statistics

Inferential

Correlational

Relationships

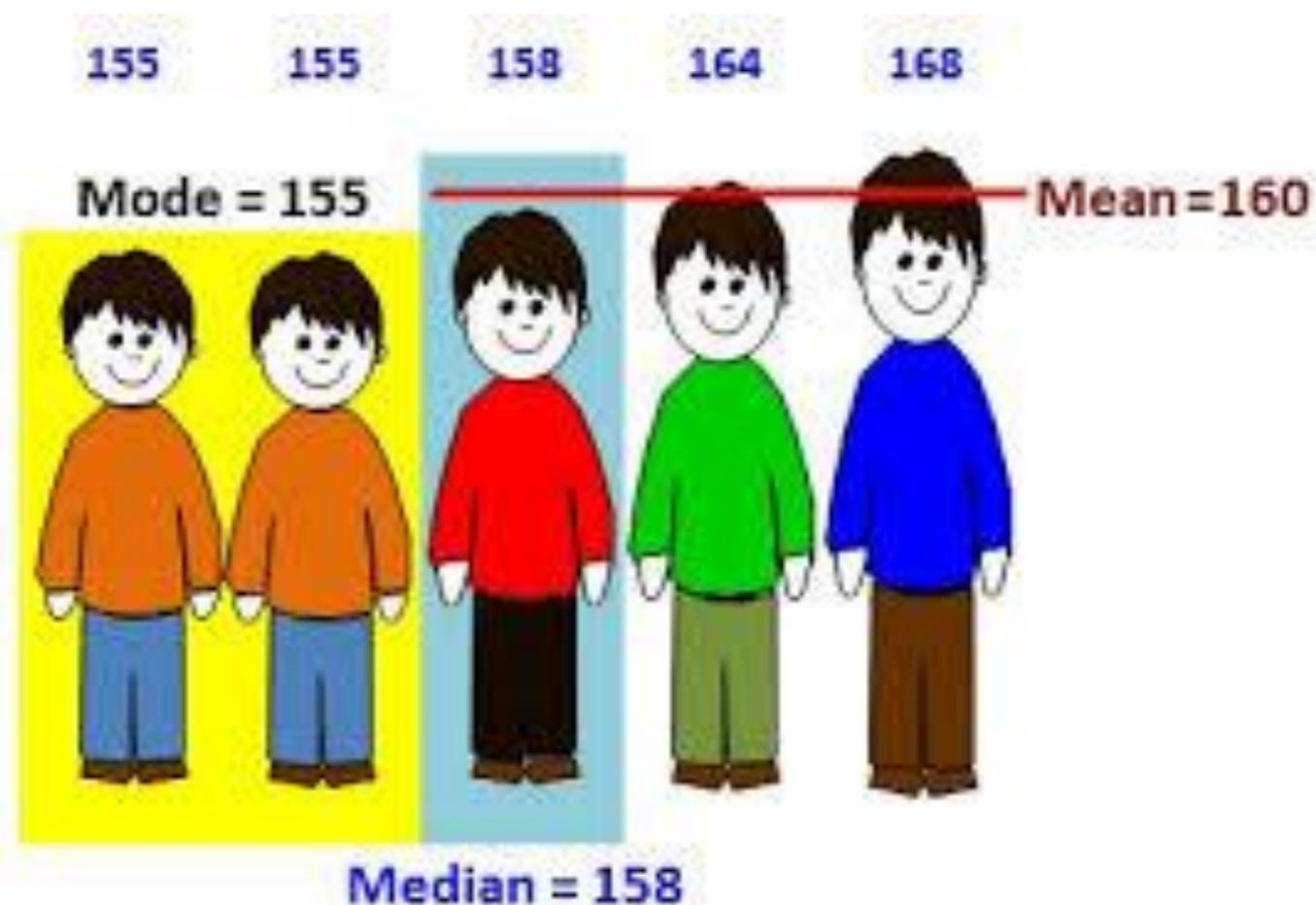
Generalising

Significance

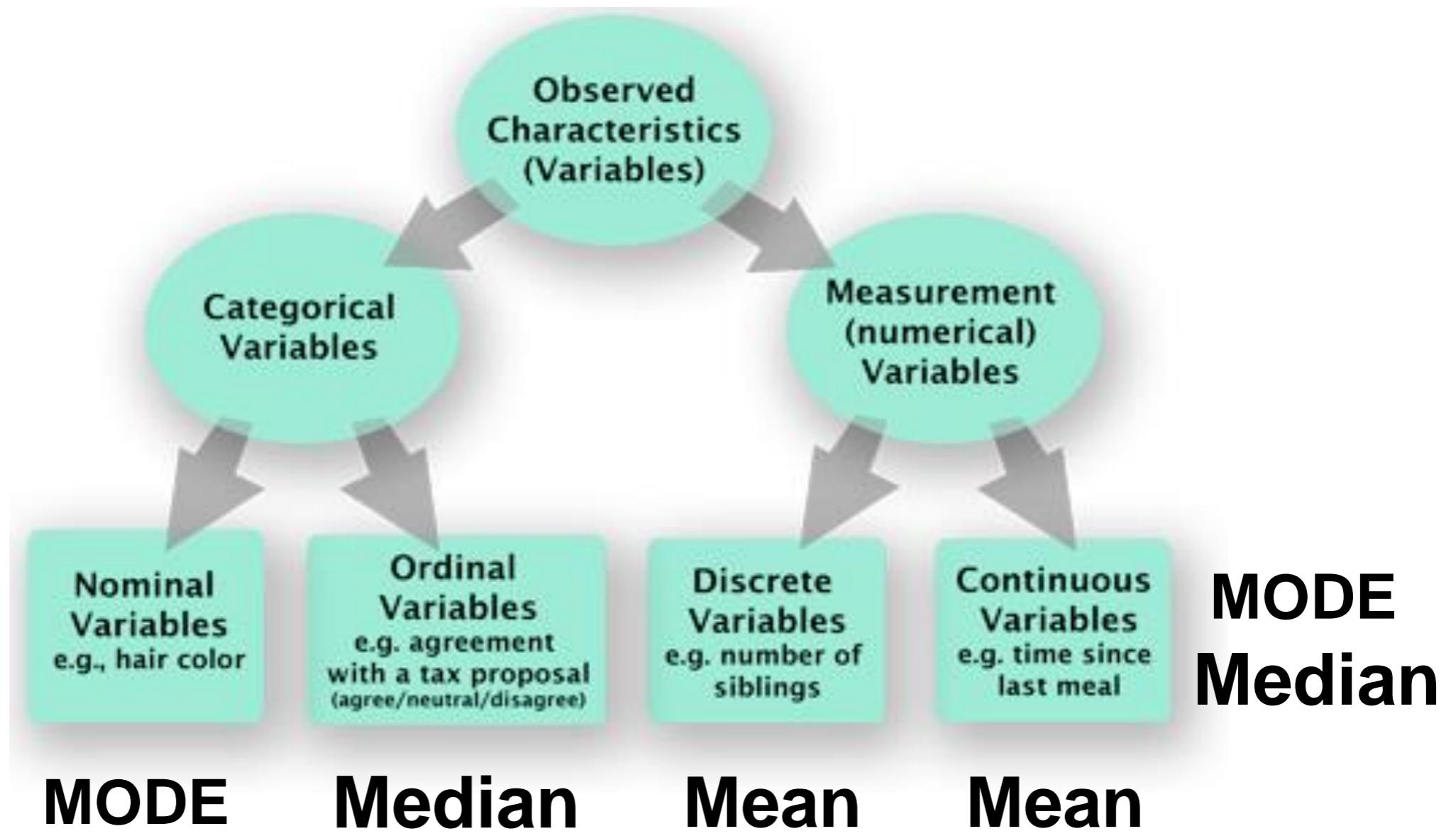
Descriptive Statistics

- Common descriptive statistics are:
 - Measure of **central tendency**
 - the most typical value of a given group of values
 - Measure of **dispersion**
 - how much all the other values in the group vary around the typical value

Measures of central tendency



Central Tendency for Variable Types



Measures of central tendency

Advantages

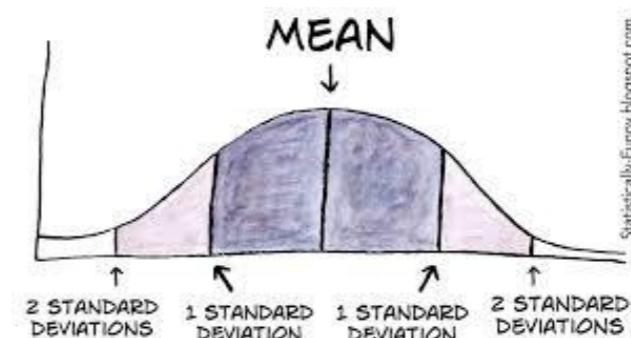
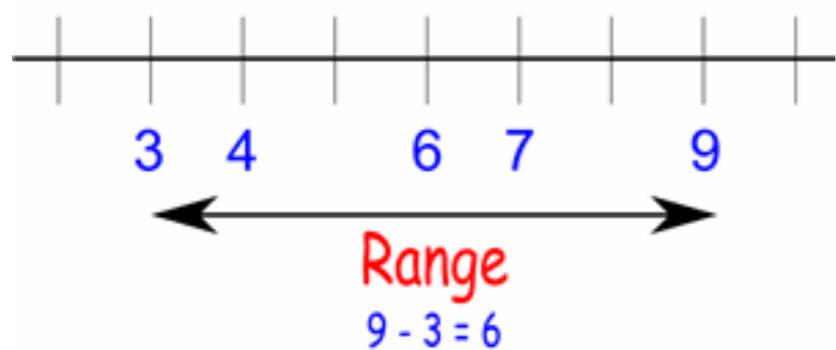
Mean

Disadvantages

Median

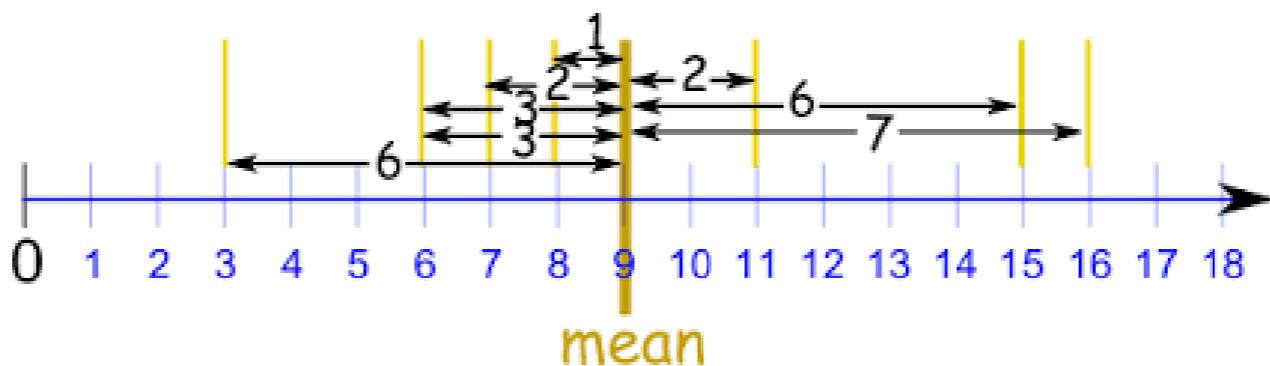
MODE

Measures of dispersion/spread



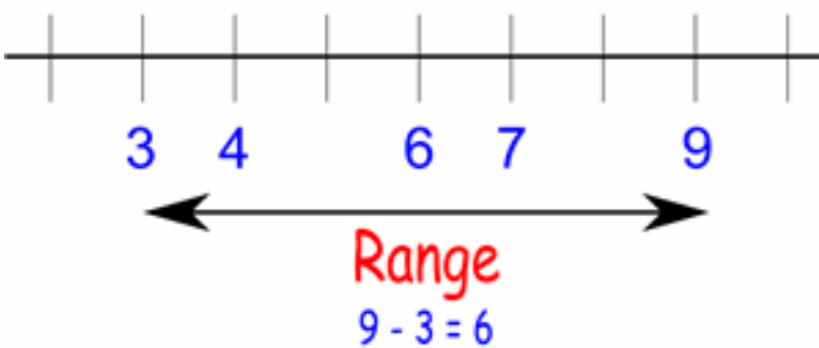
StatisticallyFunny.blogspot.com

$$SD = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$



$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

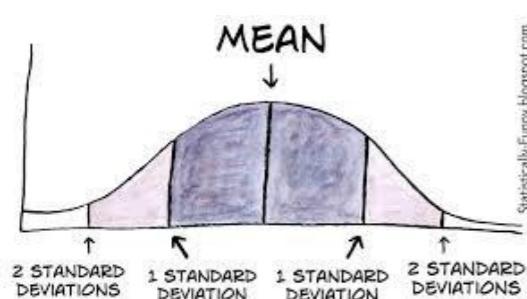
Measures of dispersion/spread



Advantages

Disadvantages

distorted by extreme values
no indication of grouping around
the mean



$$SD = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

- Fundamental to significance testing, and forms basis of Analysis of Variance (ANOVA)
- Enables population parameters to be estimated from a sample of people

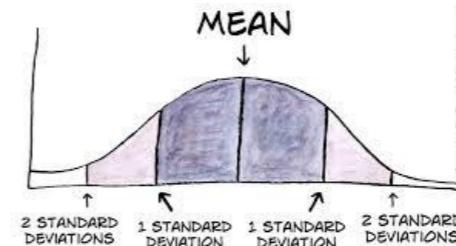
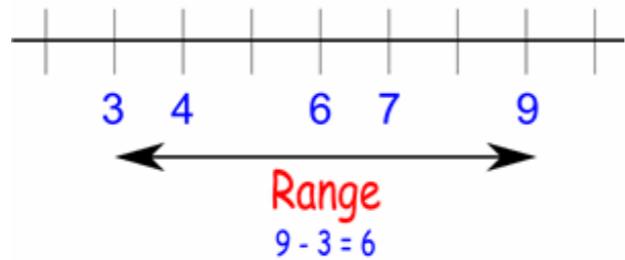
—

MEAN

?

MODE

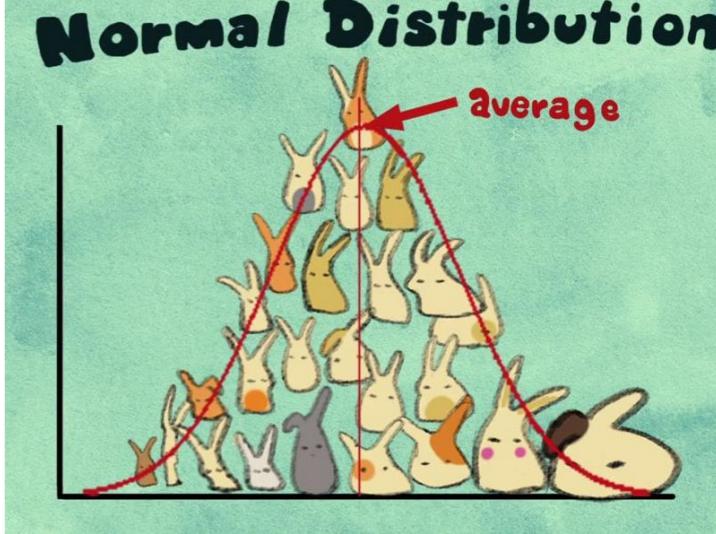
MEDIAN



When do these measures fail to be representative ????

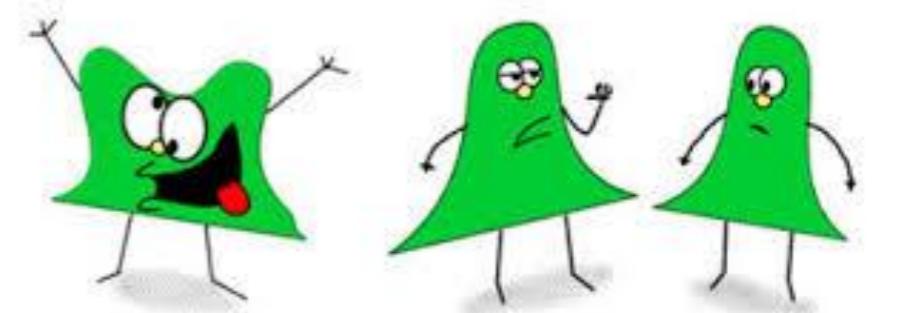


Normal Distribution



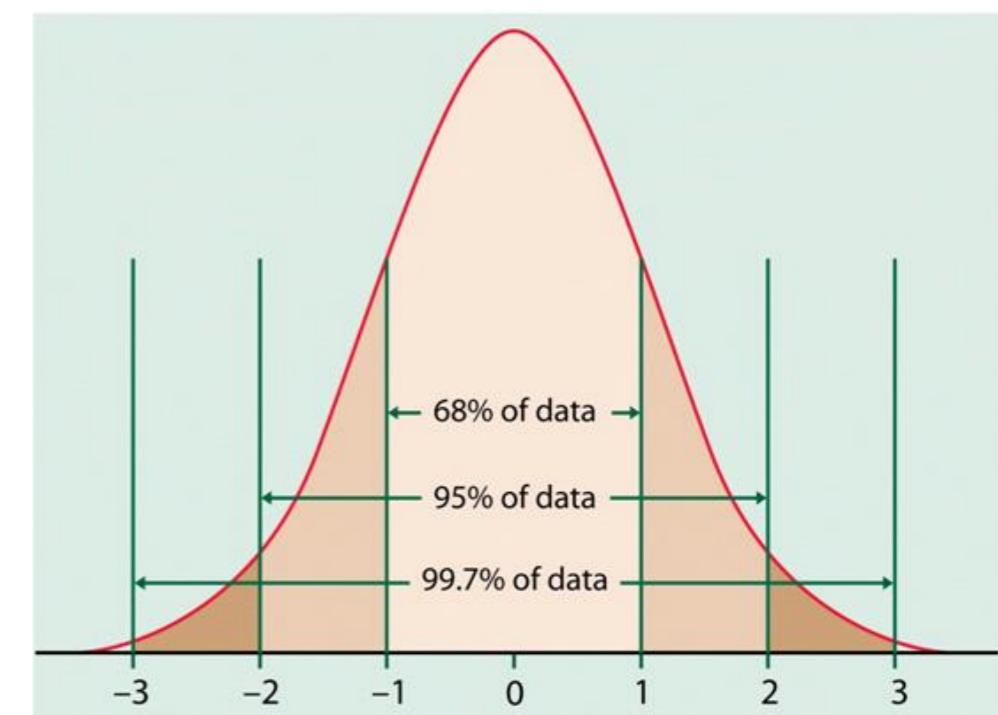
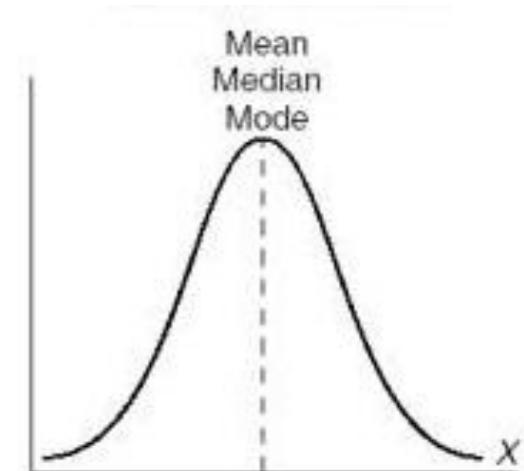
- A bell-shaped mathematical curve describing how values are distributed
- Data taken from a sample is **assumed** to be ‘normally distributed’, and must approximate this shape in order to use parametric tests of significance
- *Inferential statistics* (eg: t-tests, F-tests, regression analyses) require in some sense that the numeric variables are approximately normally distributed
- *Note:* it does not fit all populations

Normal Distribution

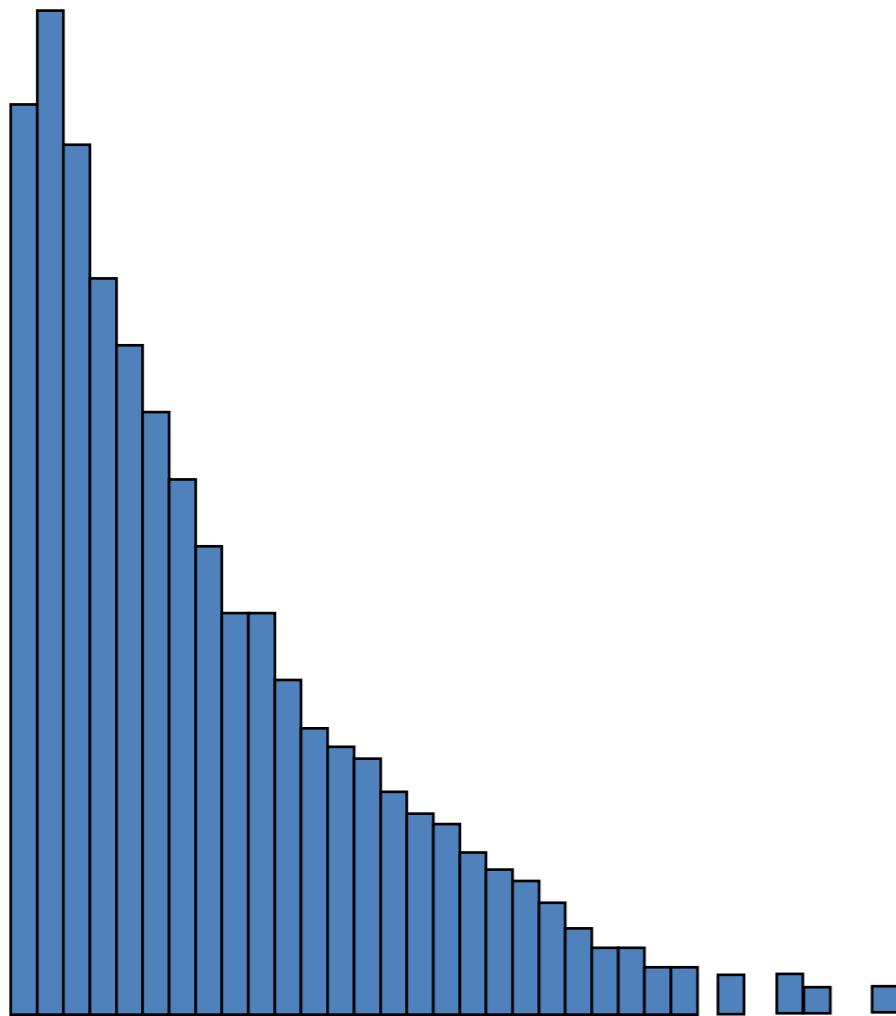


"KEEP YOUR EYE ON THAT GUY, TOM. HES NOT, YOU KNOW..NORMAL!"

- symmetrical about the horizontal axis midpoint
- mean, median, and mode all fall on the midpoint
- No matter what μ and σ are, the area between
 - $\mu-\sigma$ and $\mu+\sigma$ is about 68%;
 - $\mu-2\sigma$ and $\mu+2\sigma$ is about 95%;
 - $\mu-3\sigma$ and $\mu+3\sigma$ is about 99.7%
- Almost all values fall within 3 standard deviations

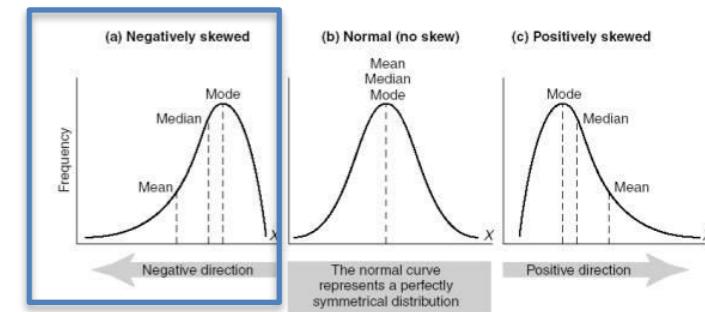


Skewed Distribution

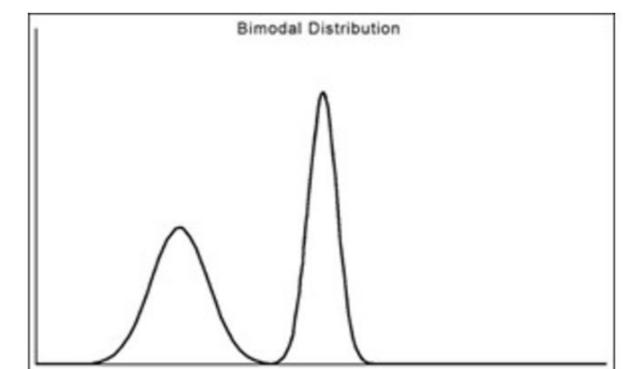


- Resembles an exponential distribution
- Lots of extreme values far from mean or mode
- Not straightforward to do useful statistical tests with this type of distribution

Skewed Distribution

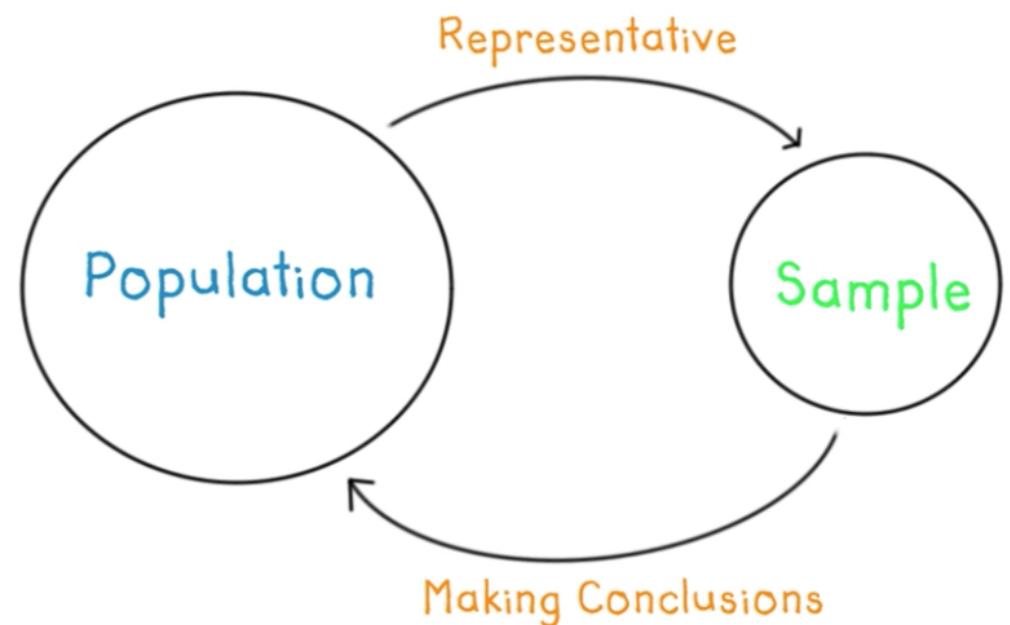


- **Negative skew**
 - Result from relatively easy tasks, due to a ceiling effect
- **Positive skew**
 - Results from tasks which are hard to improve upon, due to a floor effect (such as RT —reaction time)
- **Bimodal**
 - Two distinct peaks
 - probable indicator of groups
 - ex: completion time of marathon runners

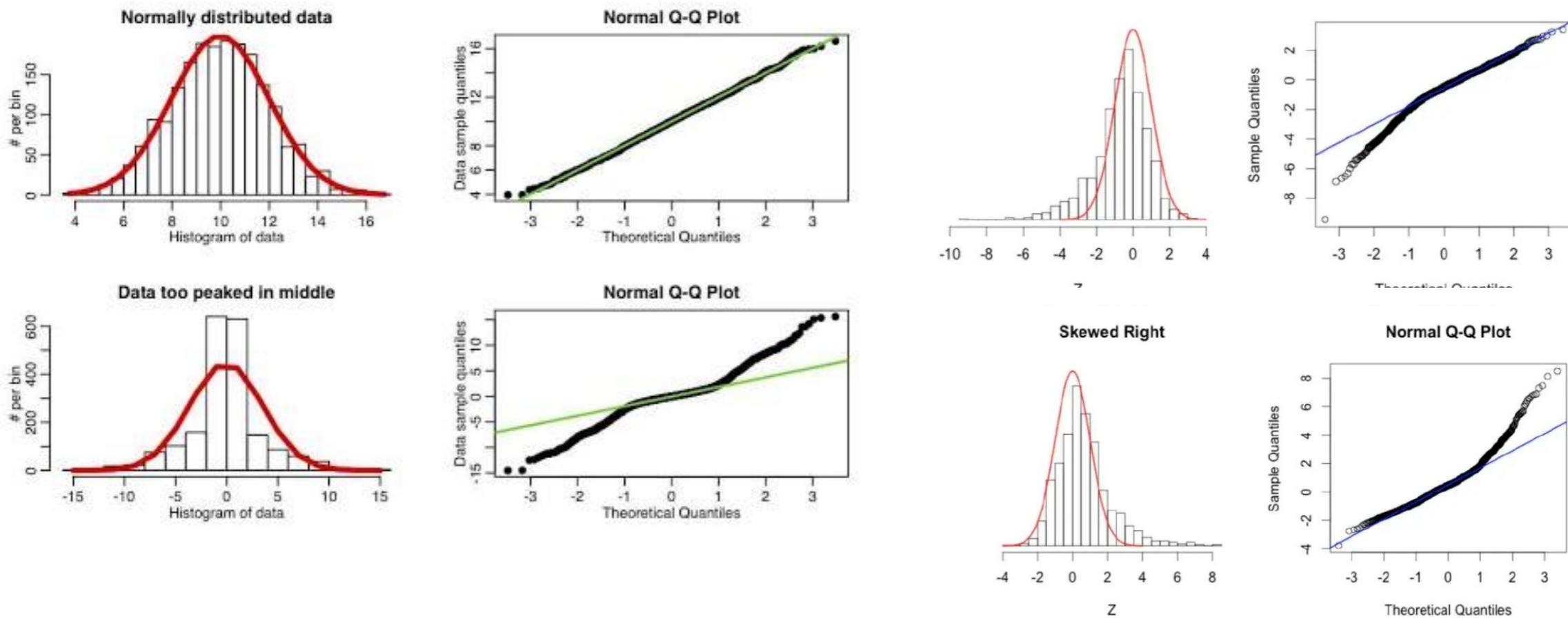


Normality in Real-World Data

- real-world data is usually skewed
- parametric tests assume that we are sampling from a normally distributed population



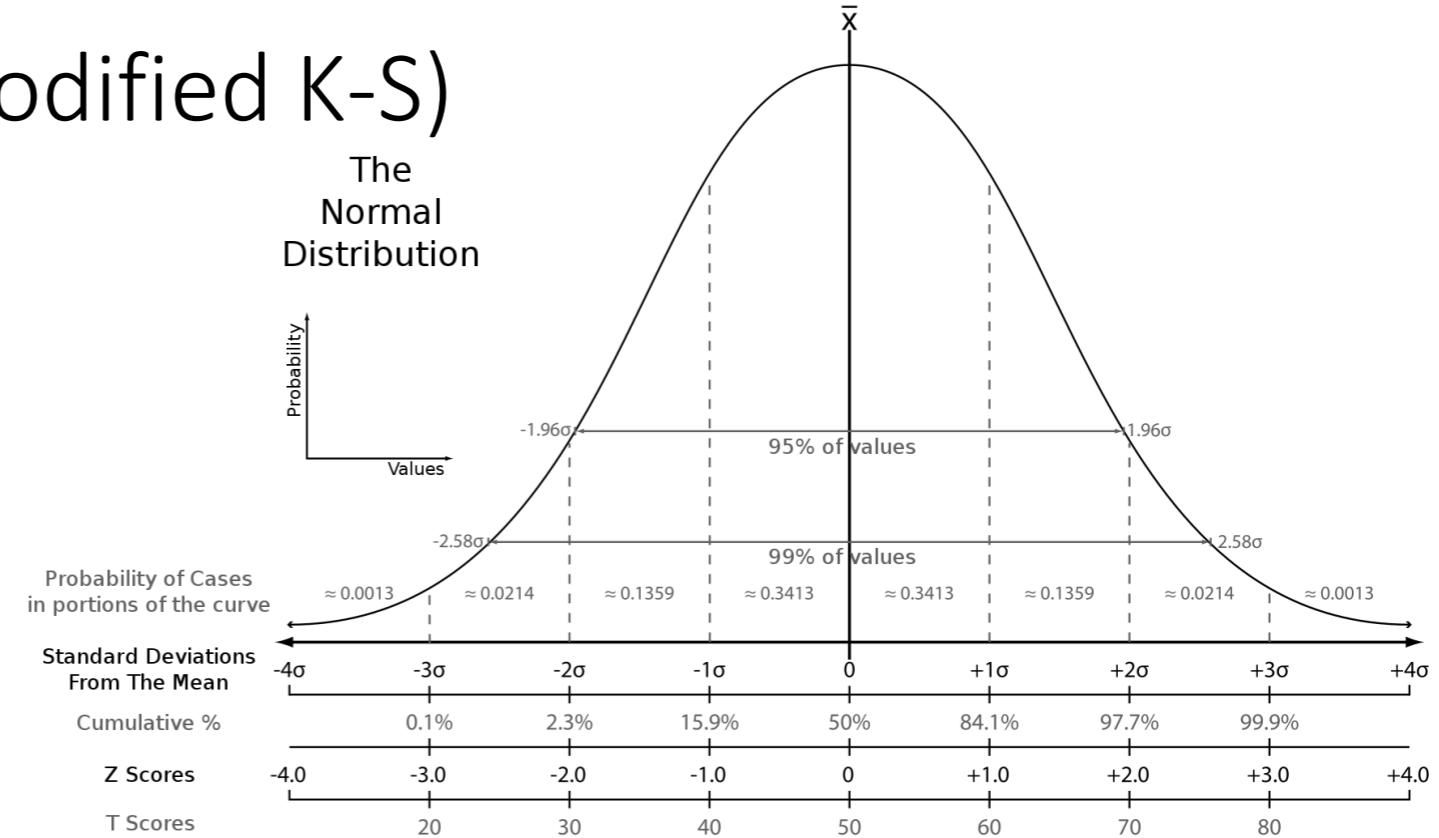
Testing Normality



- Q-Q plot: graphical technique (can also use it to test any theoretical distribution)
- theoretical quantiles plotted on x-axis and sample quantiles plotted on y-axis

Testing Normality

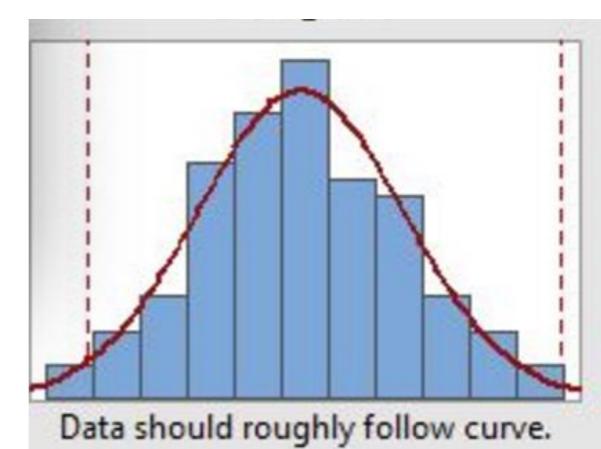
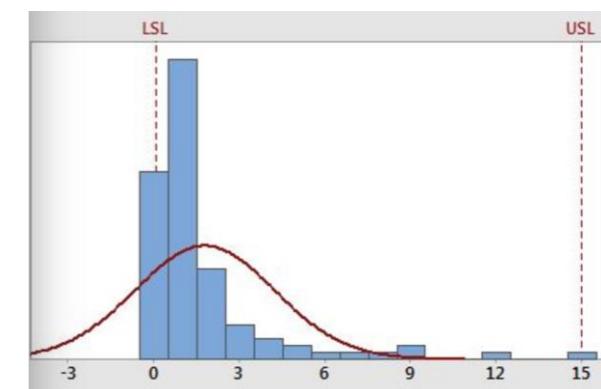
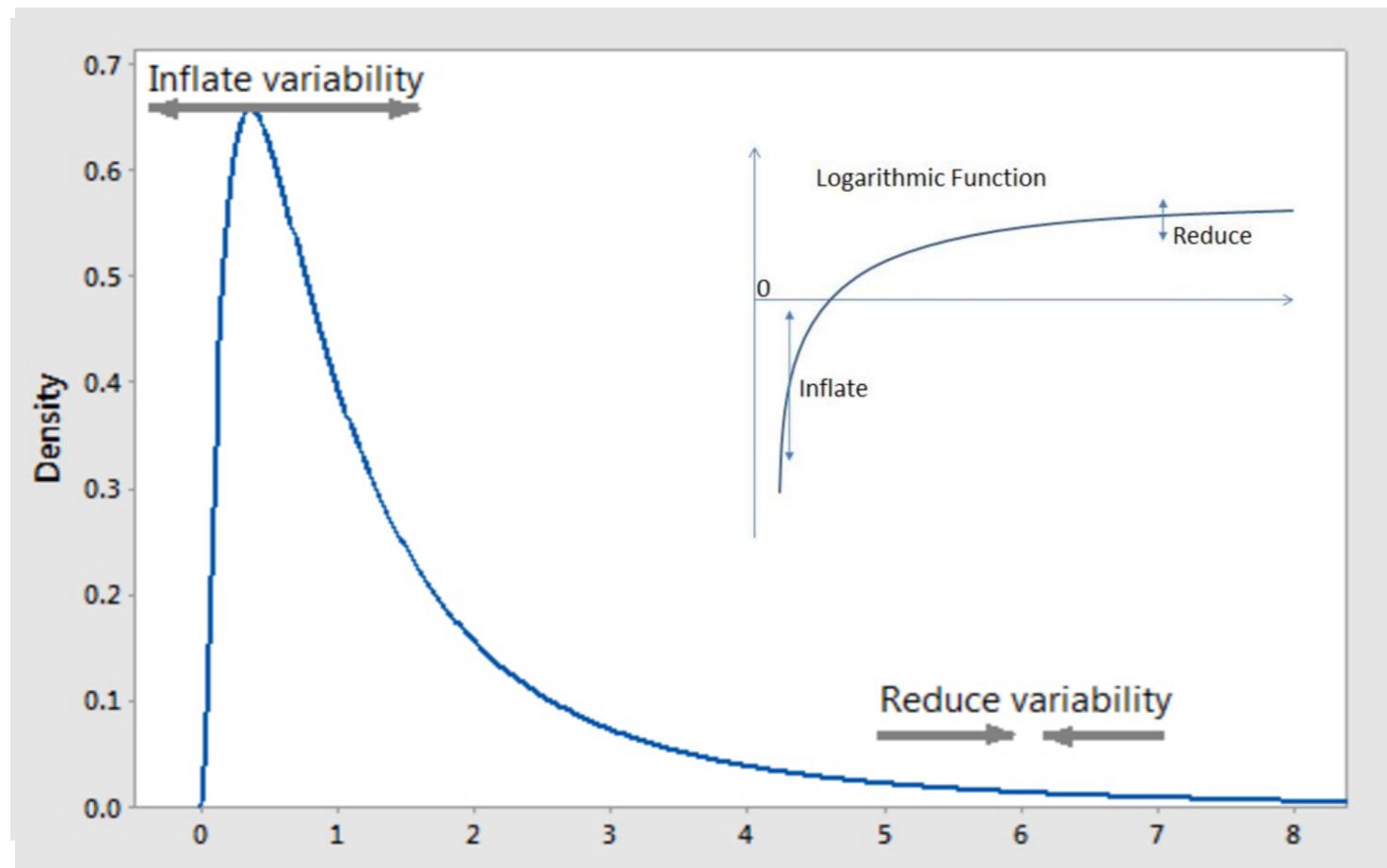
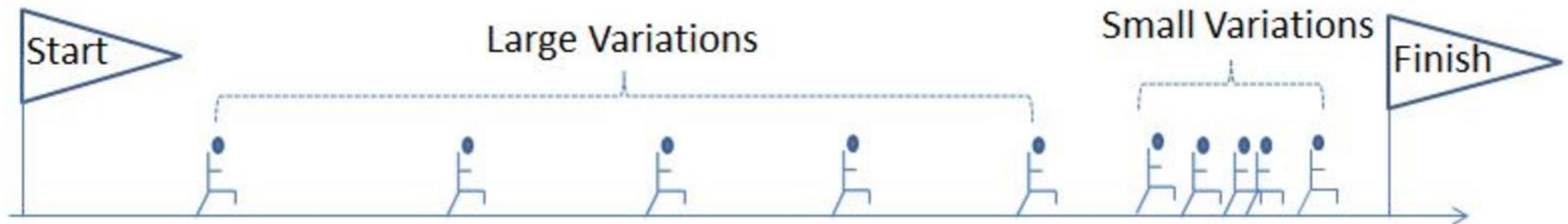
- Tests to assess normality (null hypothesis: data are sampled from a population that follows a normal distribution)
 - Kolmogorov-Smirnov (≥ 50)
 - Shapiro-Wilk (for smaller sample size, i.e. < 50)
 - Anderson-Darling (modified K-S)
 - Lilliefors test
 - Cramer-von Mises
 - etc..



Testing Normality

- For non-normal data
 - transform to normal distribution (eg: sqrt, log)
 - if it works - use parametric tests
 - if still not normal - use non-parametric tests
 - if you have groups of data, you **MUST** test each group for normality.

EXAMPLE



Normality Transforms

Moderately positive skewness	\sqrt{X}
Substantially positive skewness	$\log_{10}X$
Substantially positive skewness (with zero values)	$\log_{10}(X + C)$
Moderately negative skewness	$\sqrt{K-X}$
Substantially negative skewness	$\log_{10}(K-X)$

$C = \text{a constant added to each score so that the minimum score is 1}$

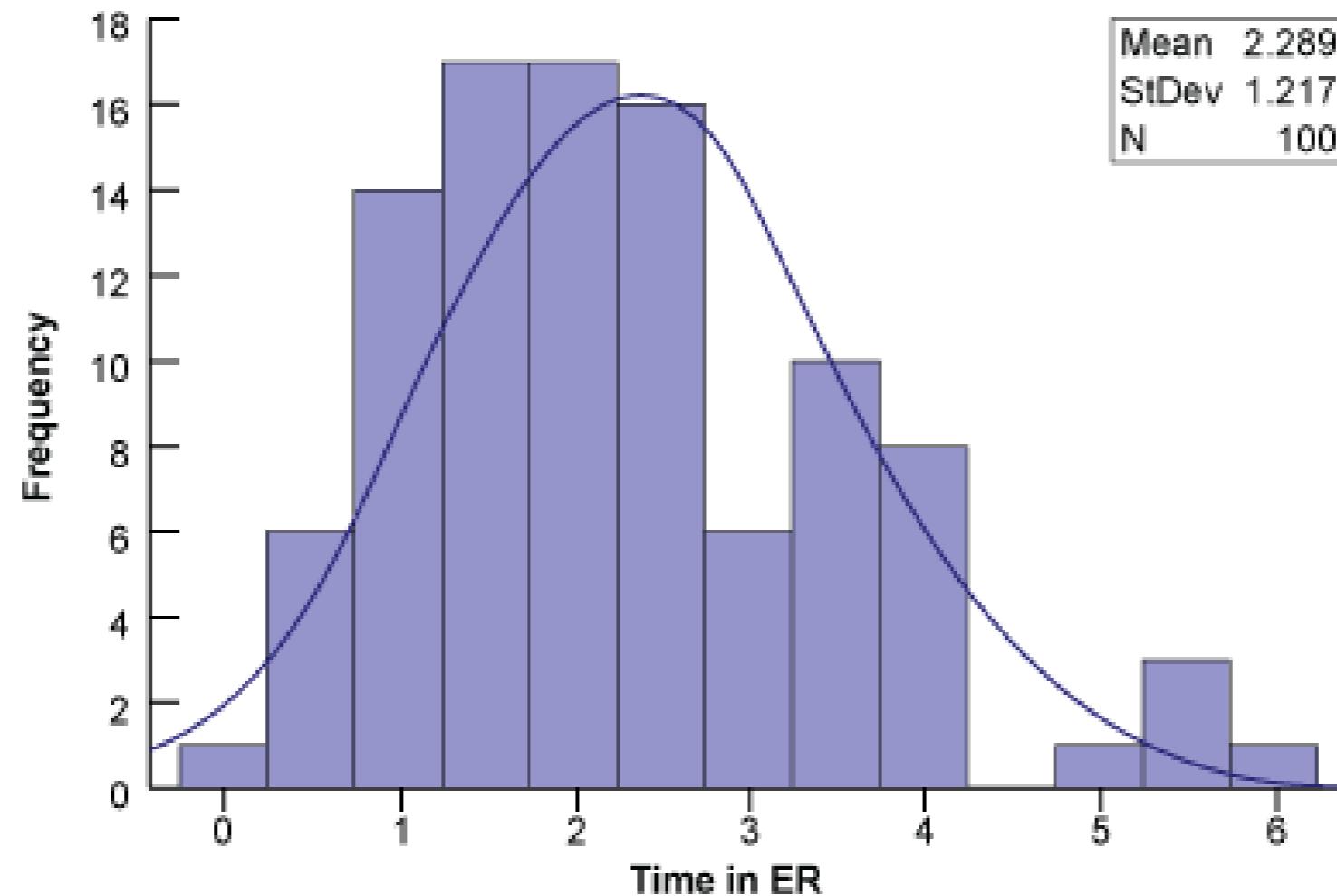
$K = \text{a constant from which each score is subtracted so that the minimum score is 1}$

Box-Cox transformation

- Box & Cox (1964) developed a procedure to identify an appropriate exponent ($\text{Lambda} = \lambda$) to use to **transform non-normal data into a “normal shape.”**
- power transformation
- increases the applicability and usefulness of statistical techniques based on the normality assumption
- is **not** a guarantee for normality
- only works if all the data is positive and greater than 0
(adding a constant (c) to all data)

EXAMPLE

hospital's target time for processing, diagnosing and treating patients entering the ER

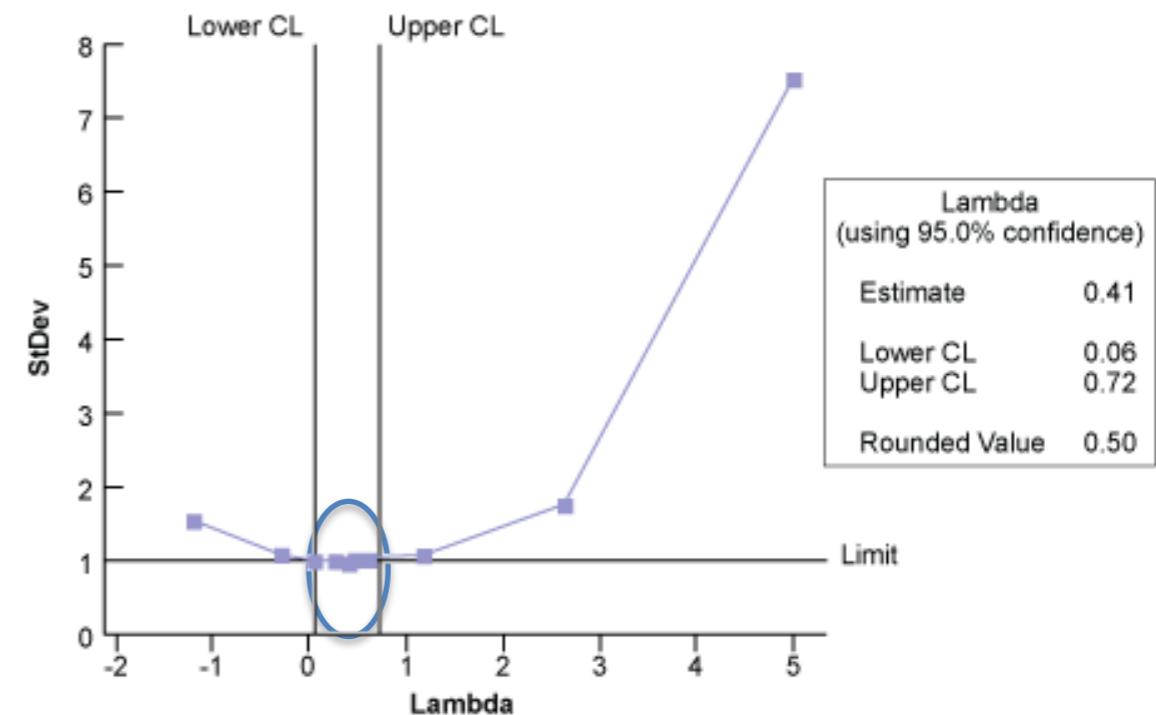
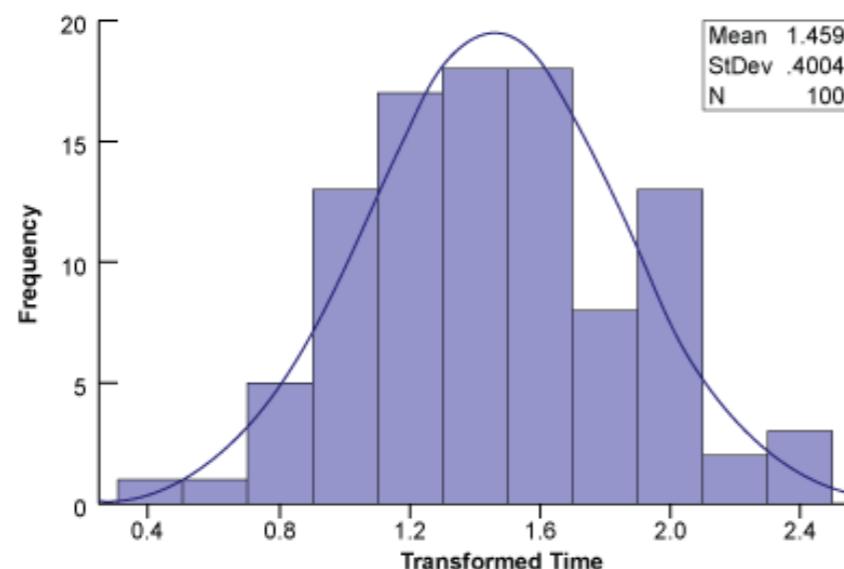
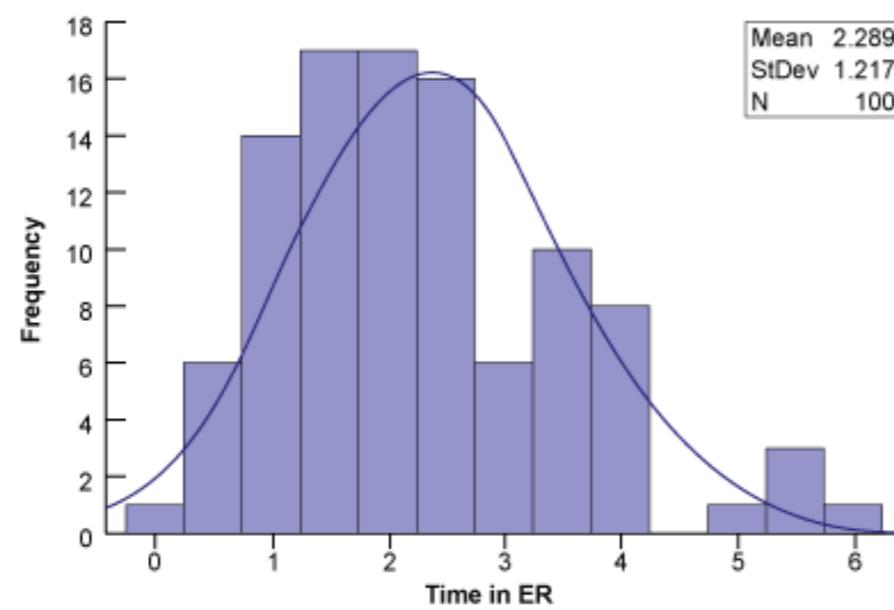


typically it is four hours or less

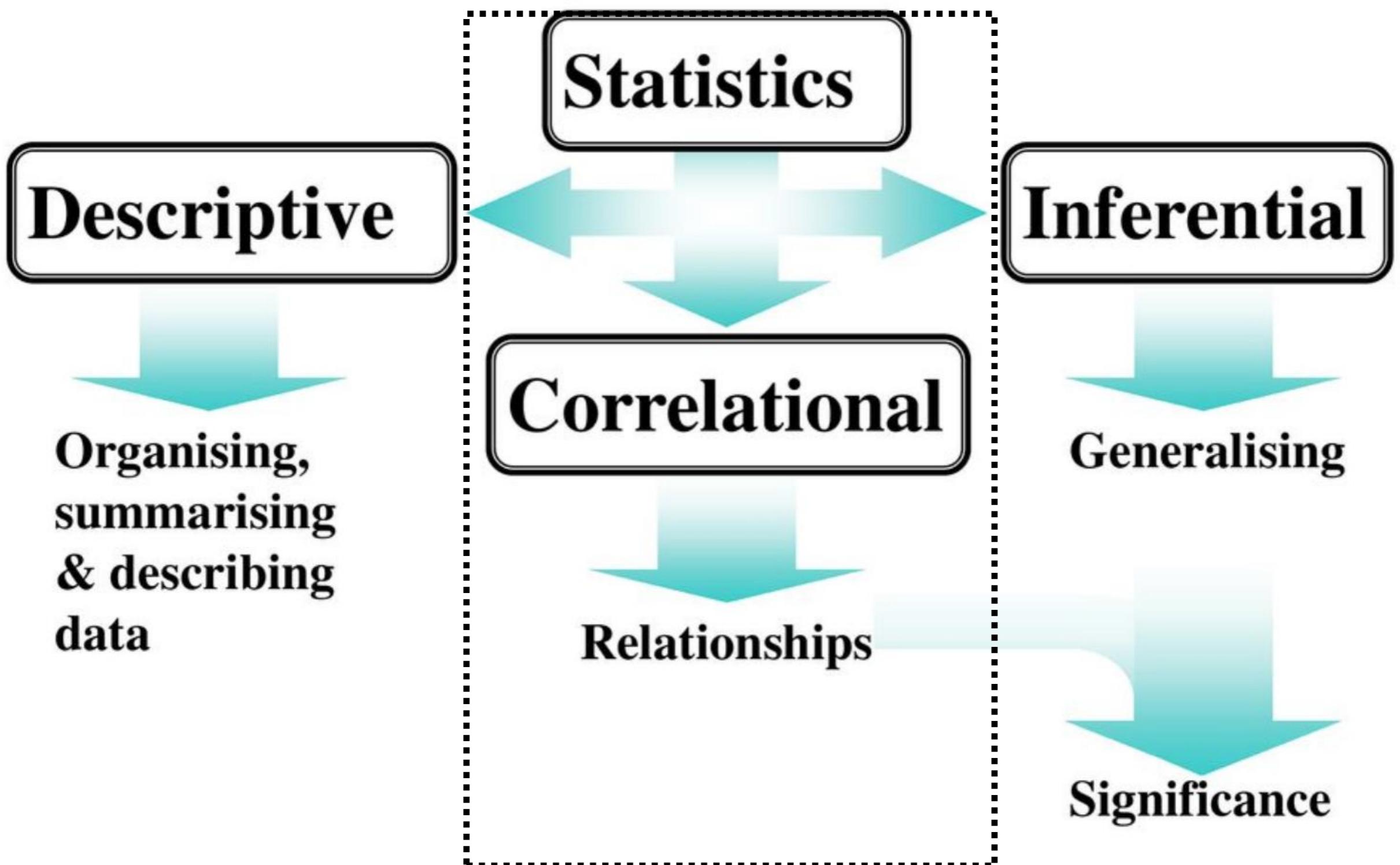
EXAMPLE

hospital's target time for processing, diagnosing and treating patients entering the ER

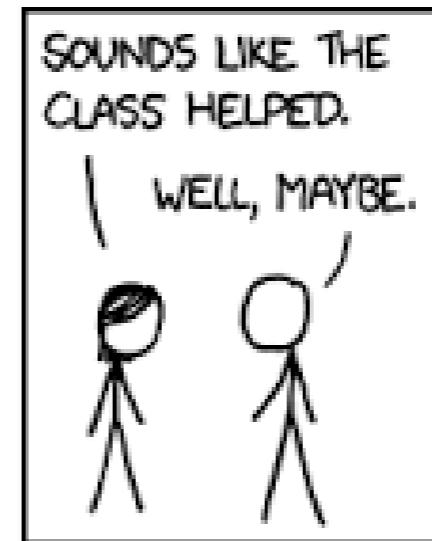
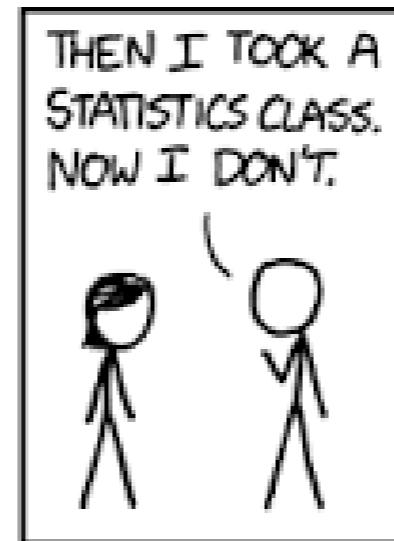
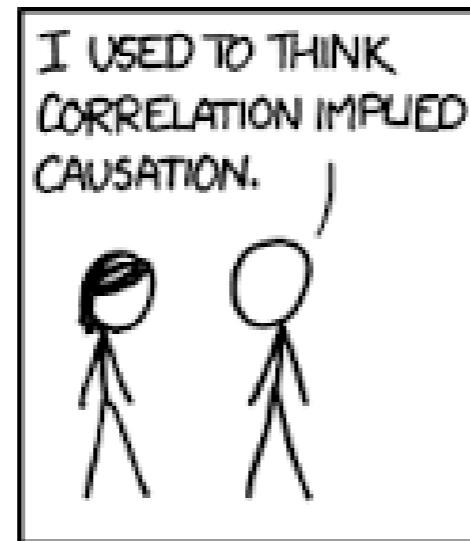
the “optimal value” is the one which results in the best approximation of a normal distribution curve



$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ (y_t^\lambda - 1)/\lambda & \text{otherwise.} \end{cases}$$

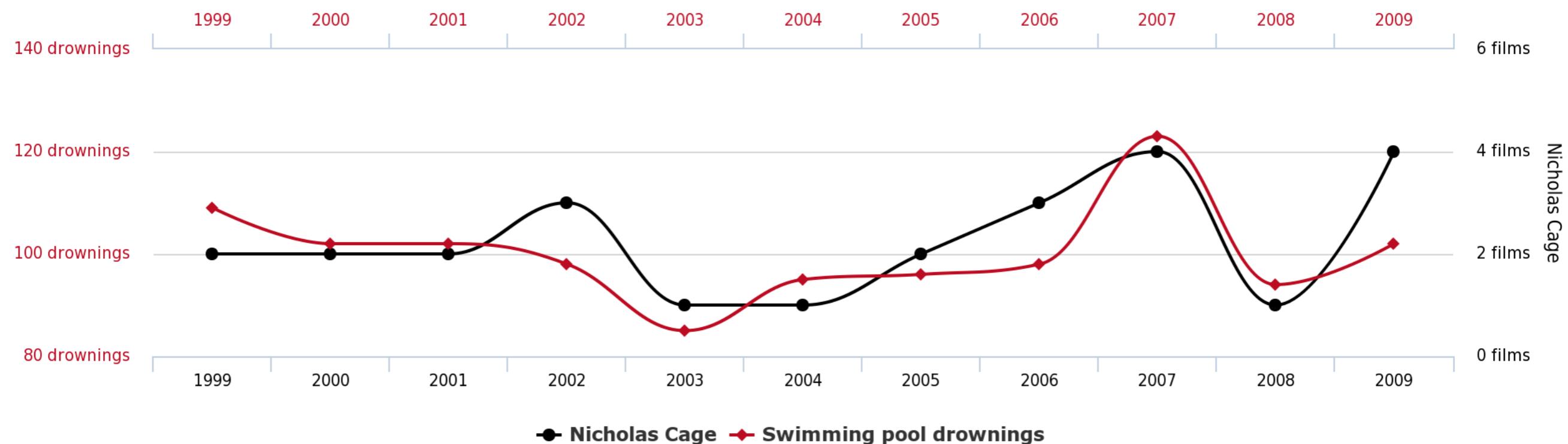


Correlation

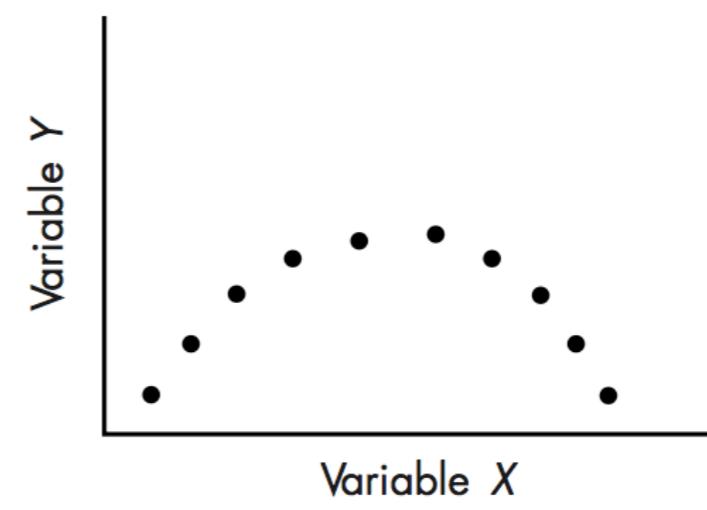
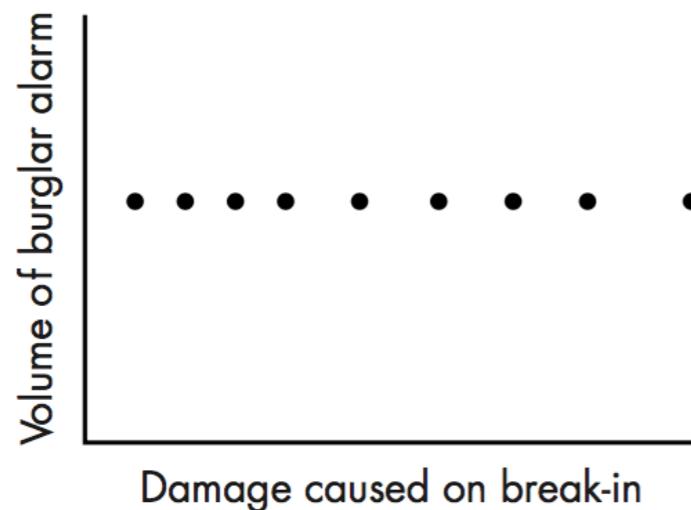
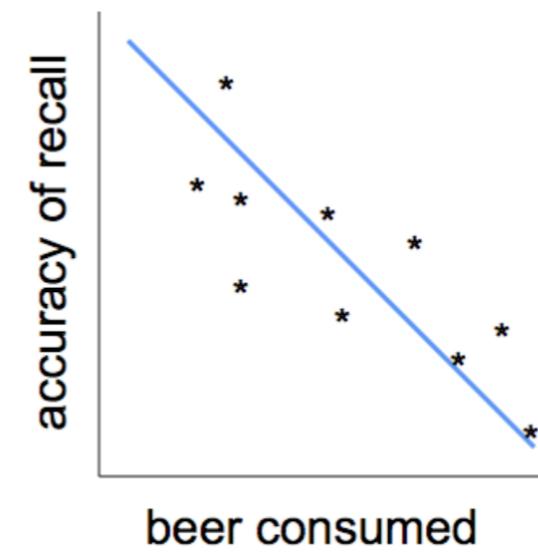
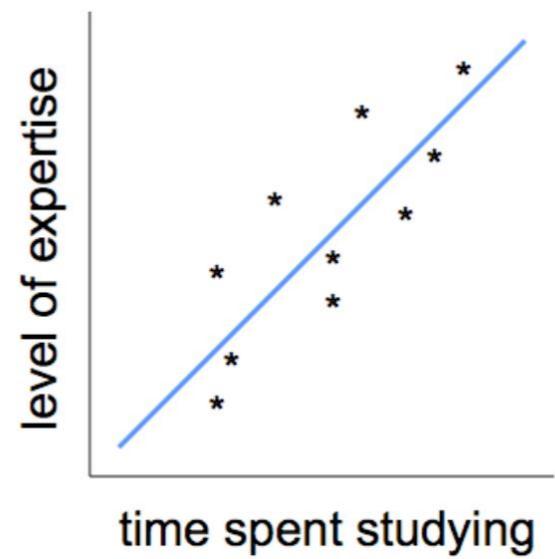


Not Causality

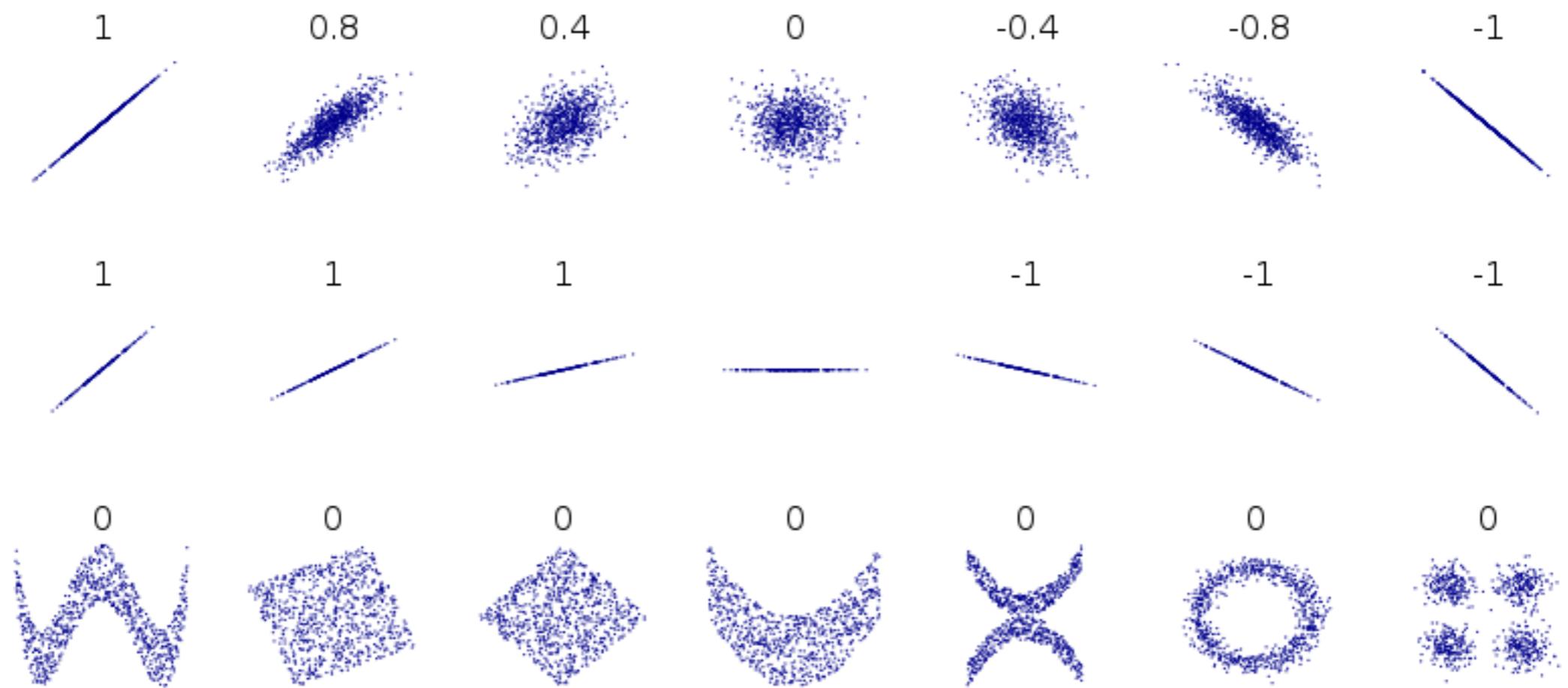
Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in



Correlation



Pearson's r



$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Correlation

- calculation of correlation between two variables is a descriptive measure of the association
- testing the correlation for significance is an inferential procedure

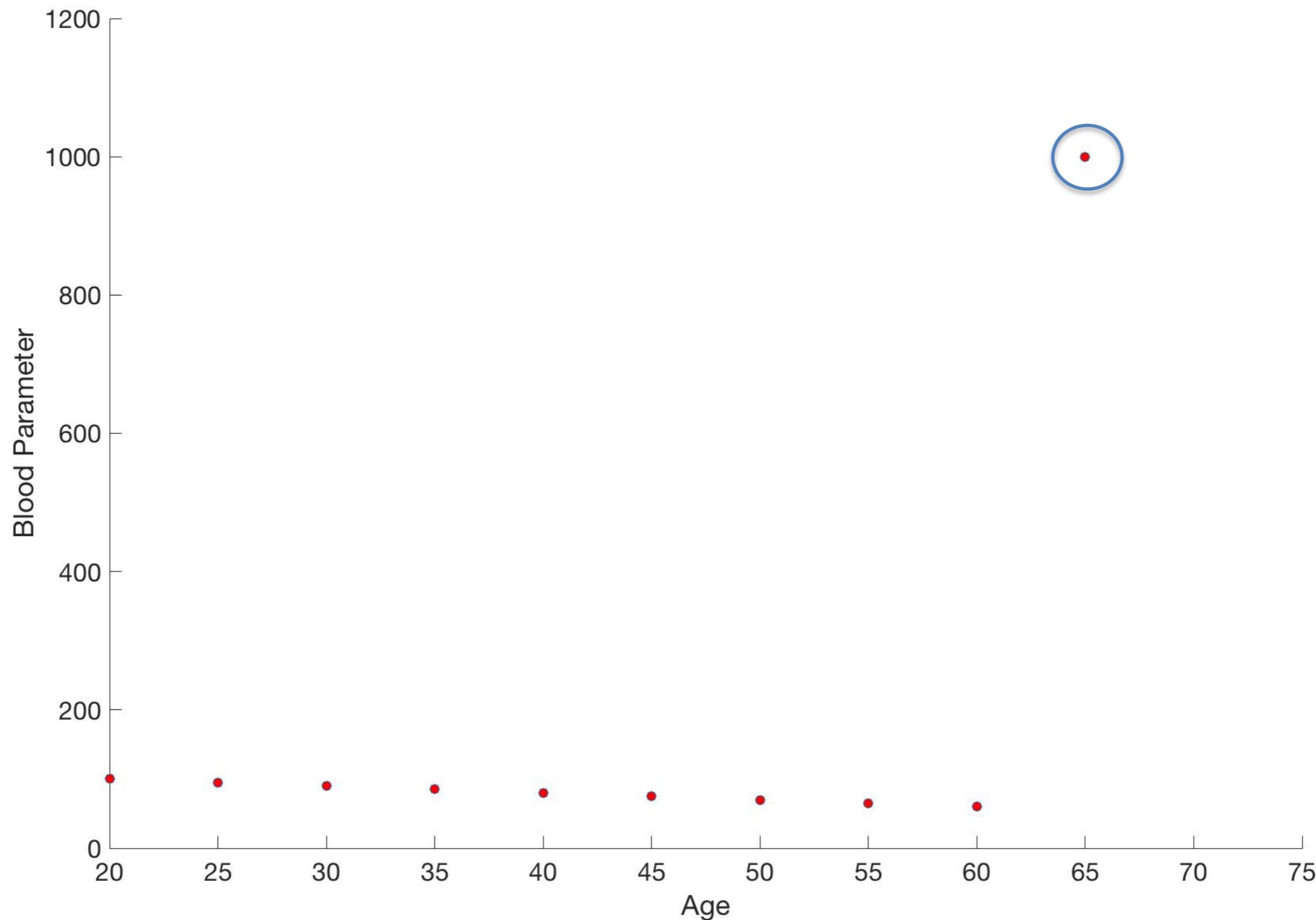
Variable Y\X	Quantitative X	Ordinal X	Nominal X
Quantitative Y	Pearson r	Biserial r_b	Point Biserial r_{pb}
Ordinal Y	Biserial r_b	Spearman rho/Tetrachoric r_{tet}	Rank Biserial r_{rb}
Nominal Y	Point Biserial r_{pb}	Rank Biserial r_{rb}	Phi, L, C, Lambda

r = correlation coefficient

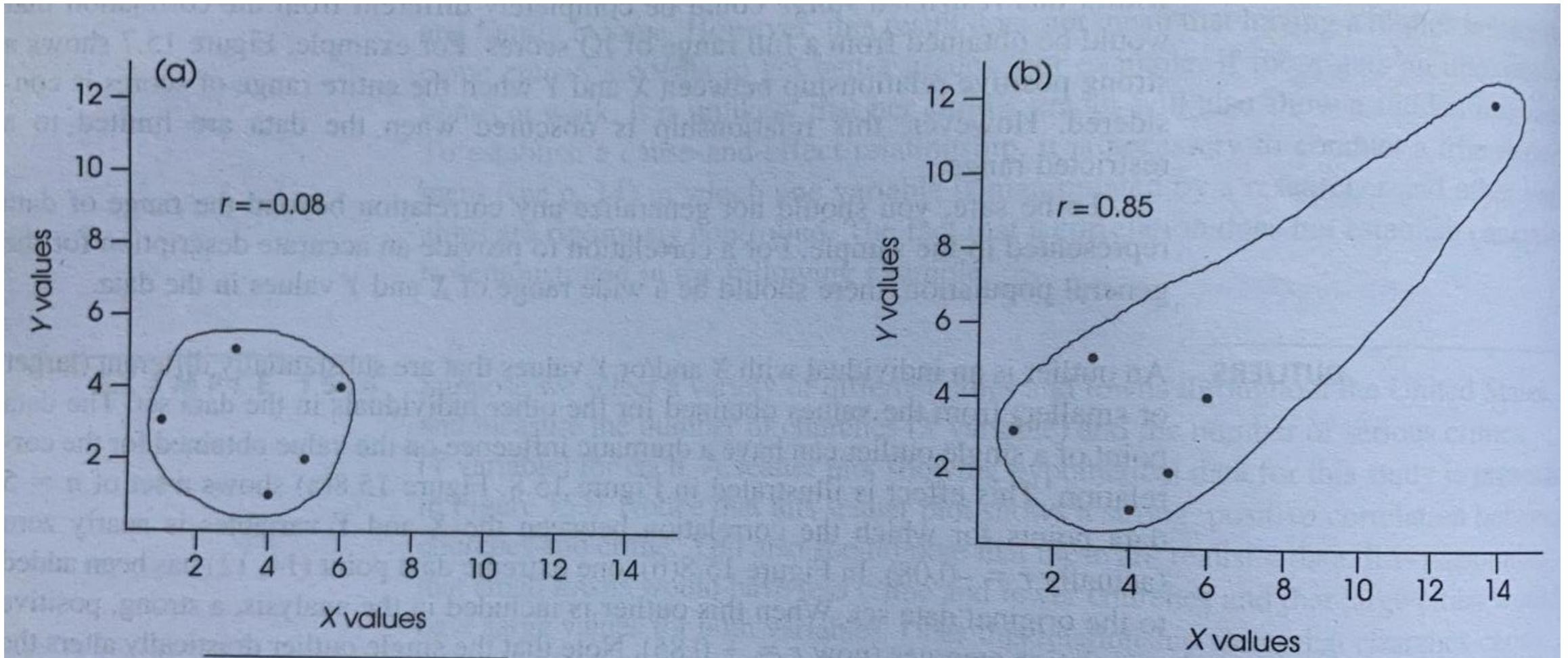
r^2 = coefficient of determination

***r* = ?**

Pearson's ***r* = .48**

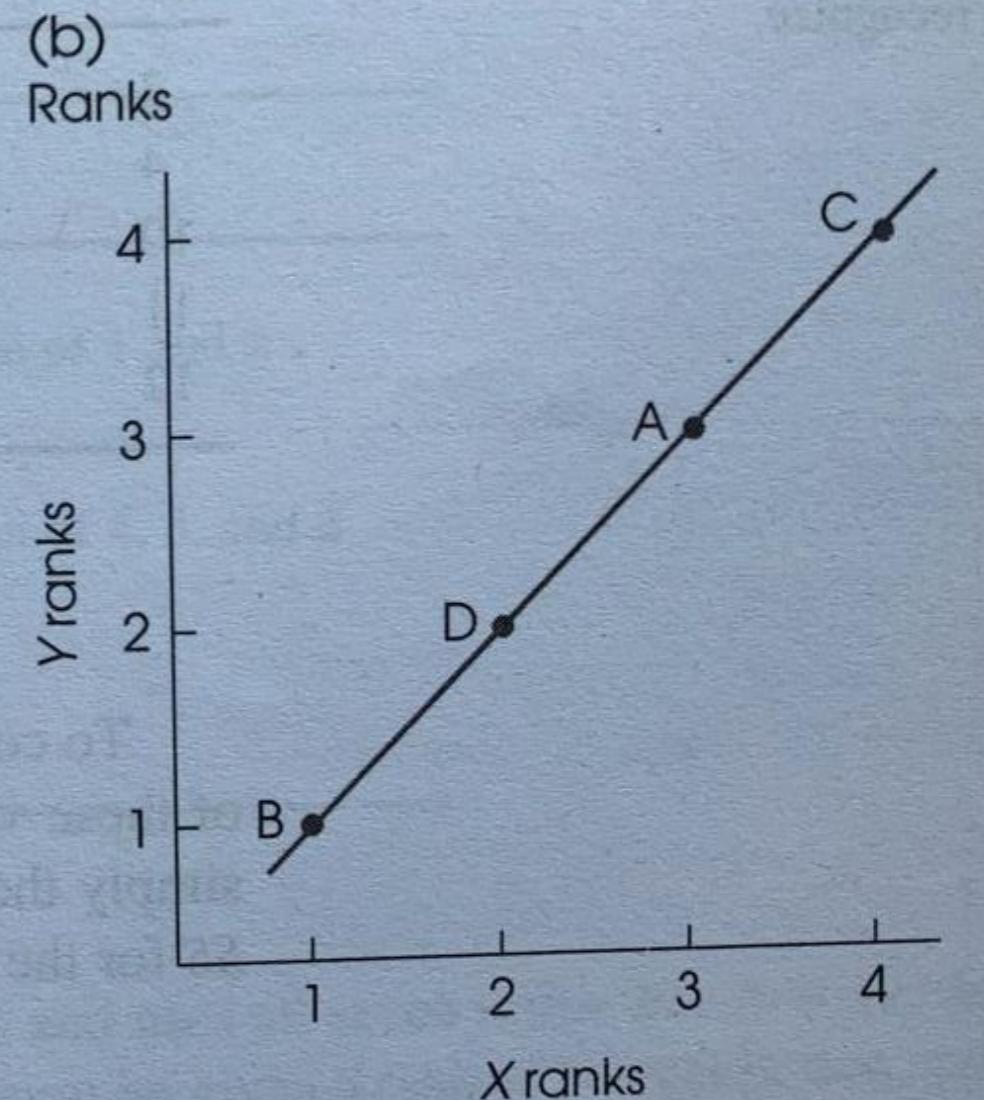
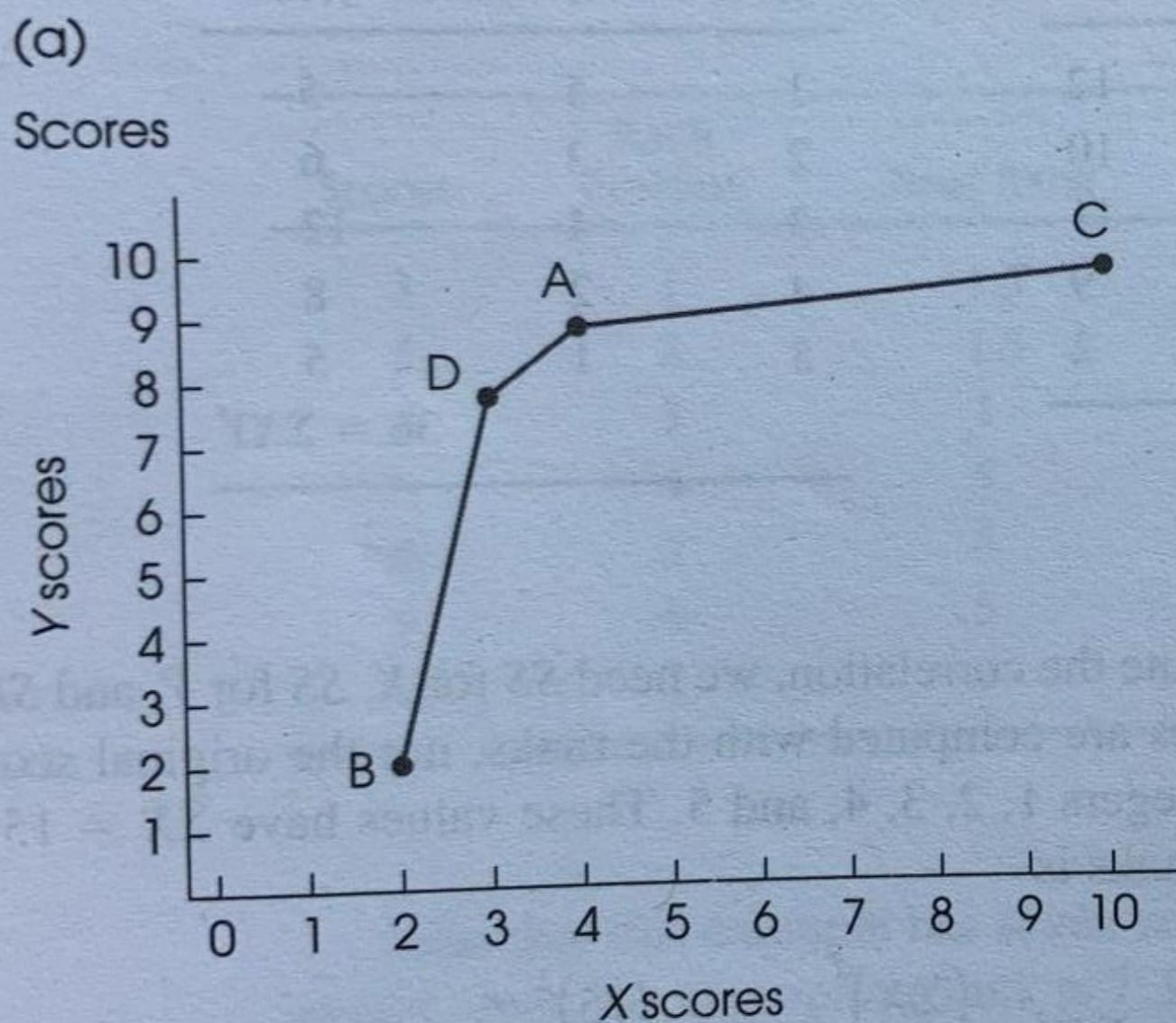


Pearson's r



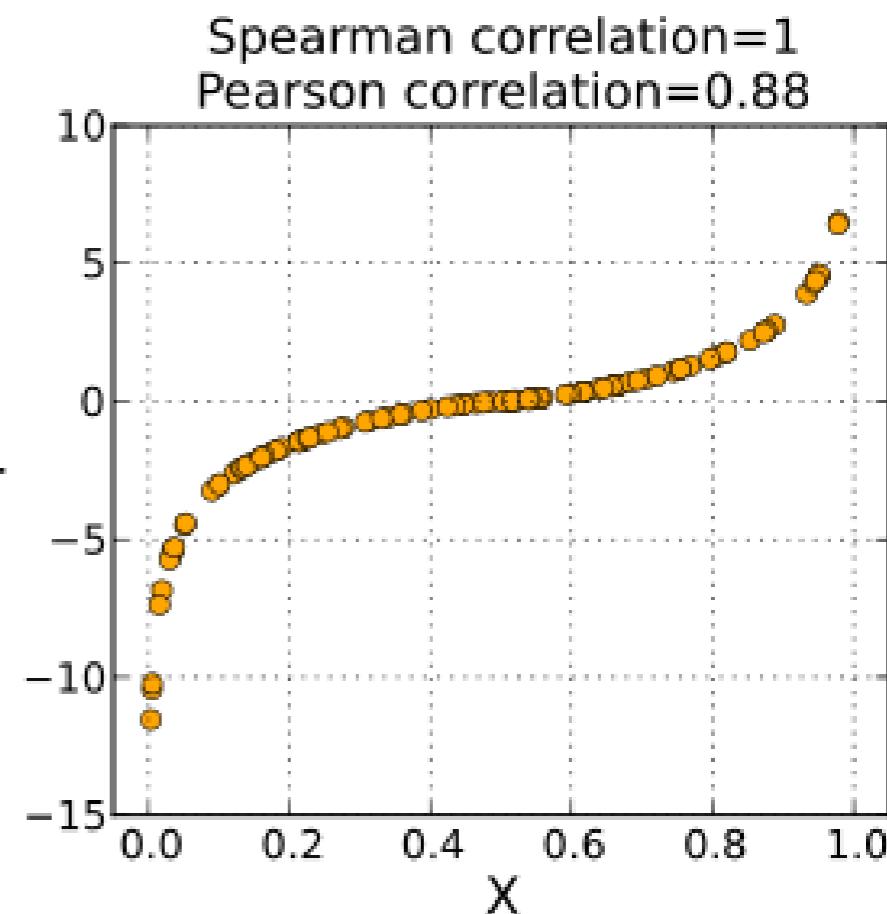
sensitive to outliers

Spearman's *rho*



Spearman's rho

- Pearson's correlation coefficient on the ranks of the data
- deals with ordinal data
- If there are no repeated values, a perfect Spearman's correlation occurs when each of the variables is a perfect monotone function of the other



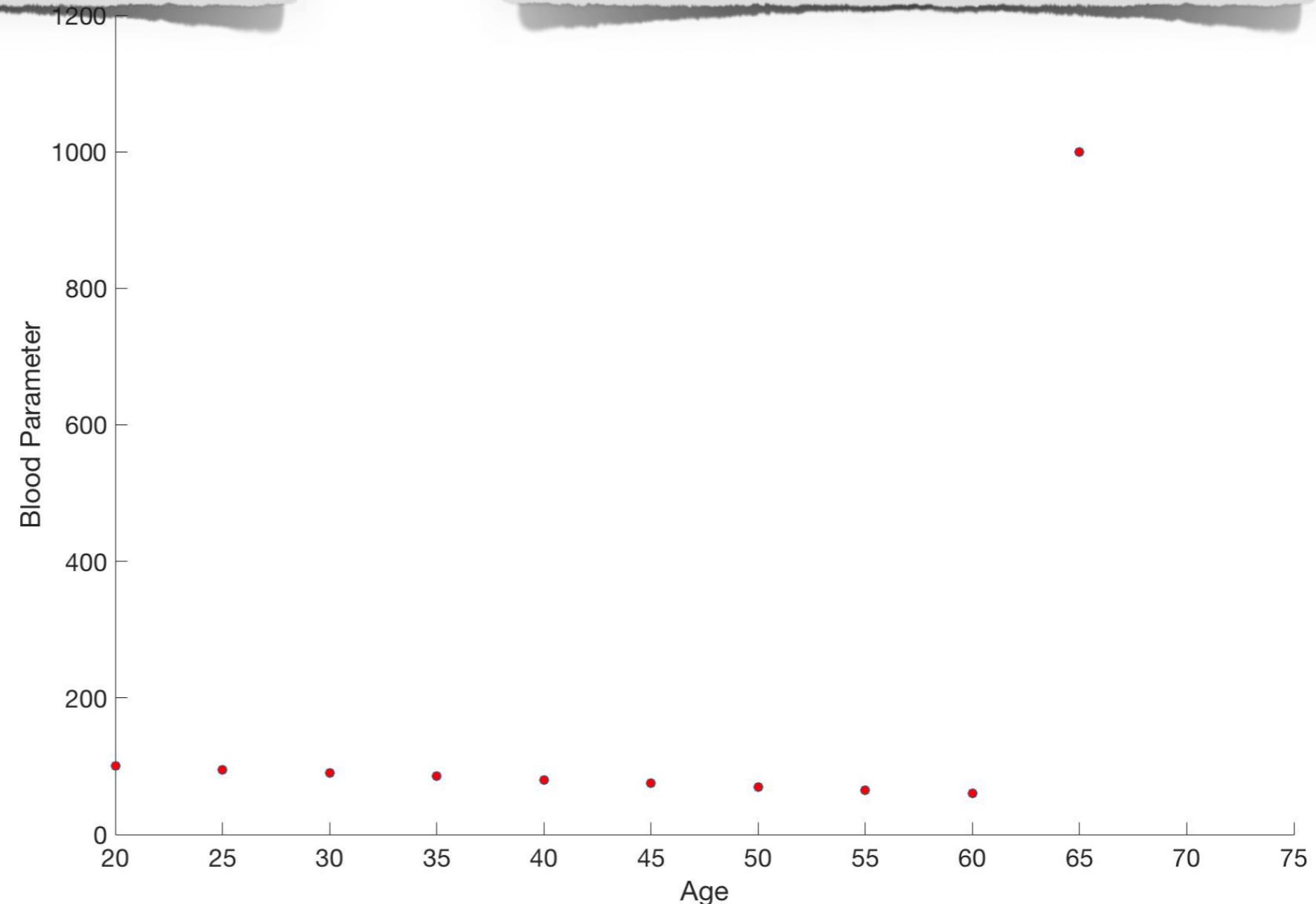
Pearson's r vs Spearman's ρ

- Pearson's sensitive to outliers

Pearson's $r = .48$

$r = ?$

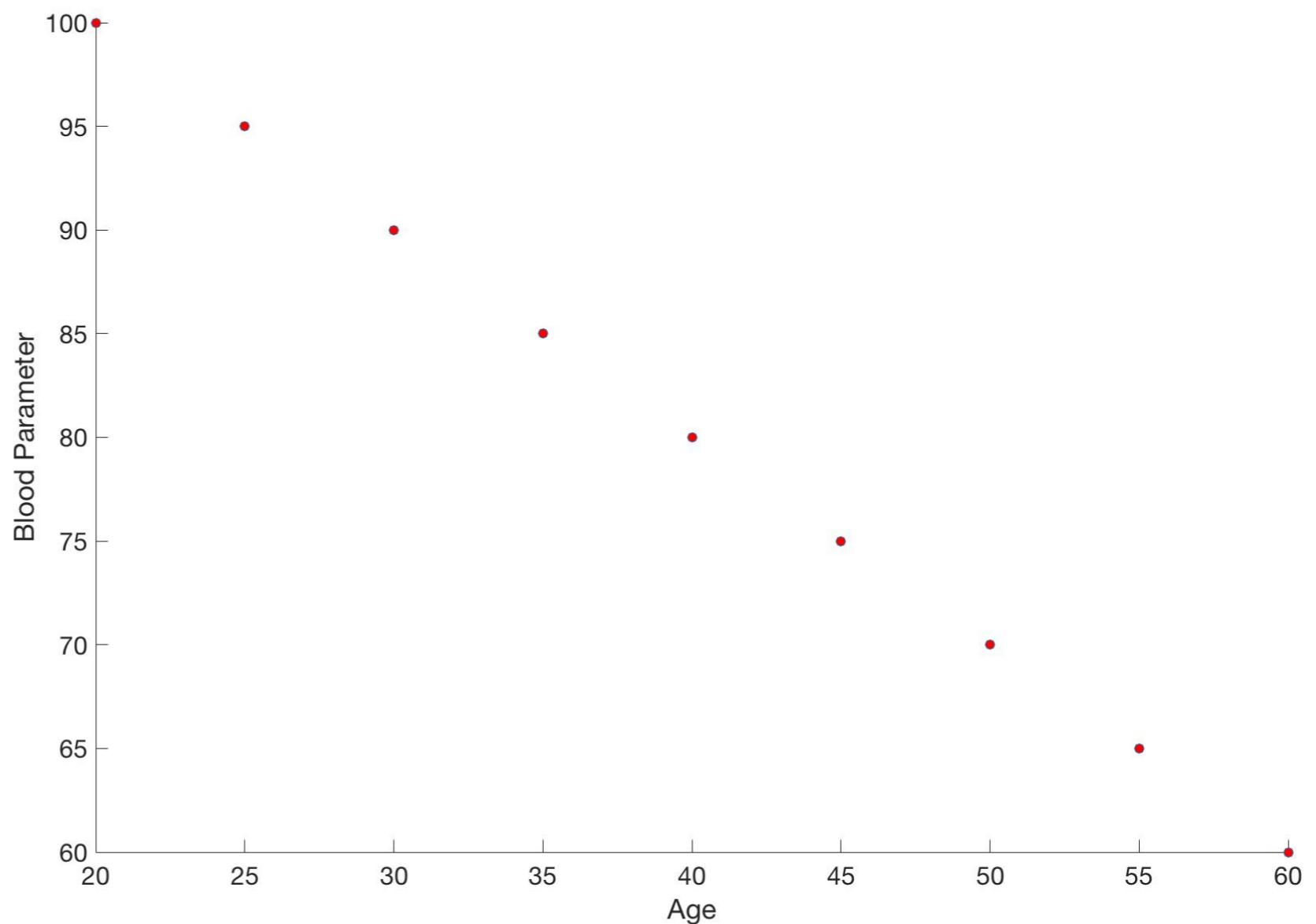
Spearman's $r = -.45$



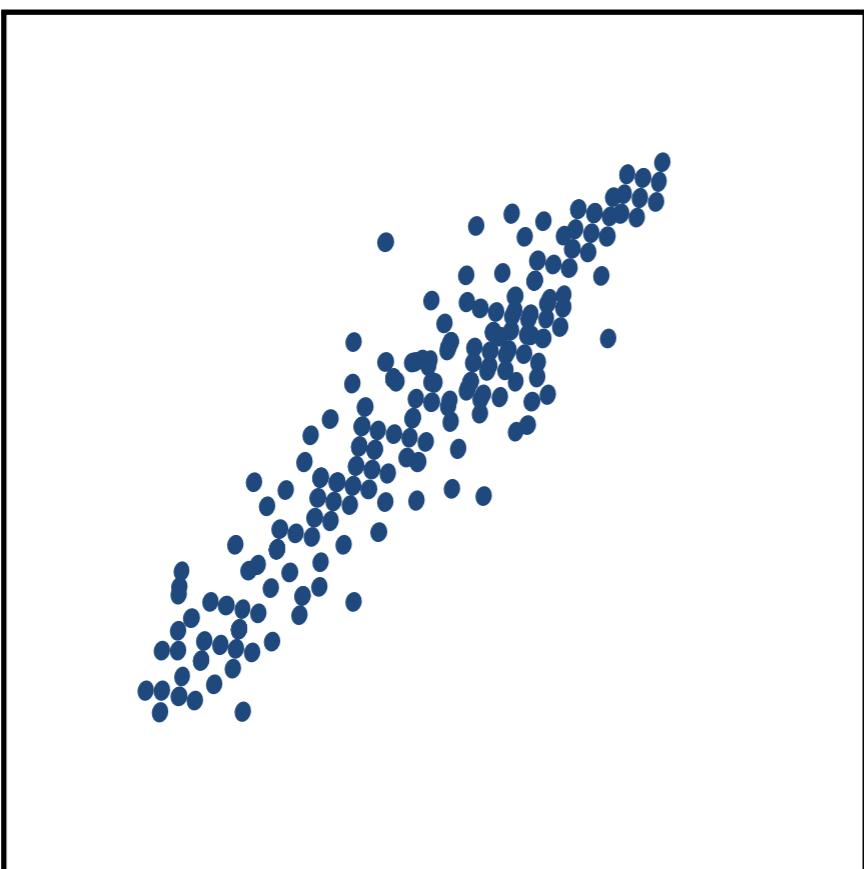
Pearson's r vs Spearman's ρ

Pearson's $r = -1$

Spearman's $r = -1$

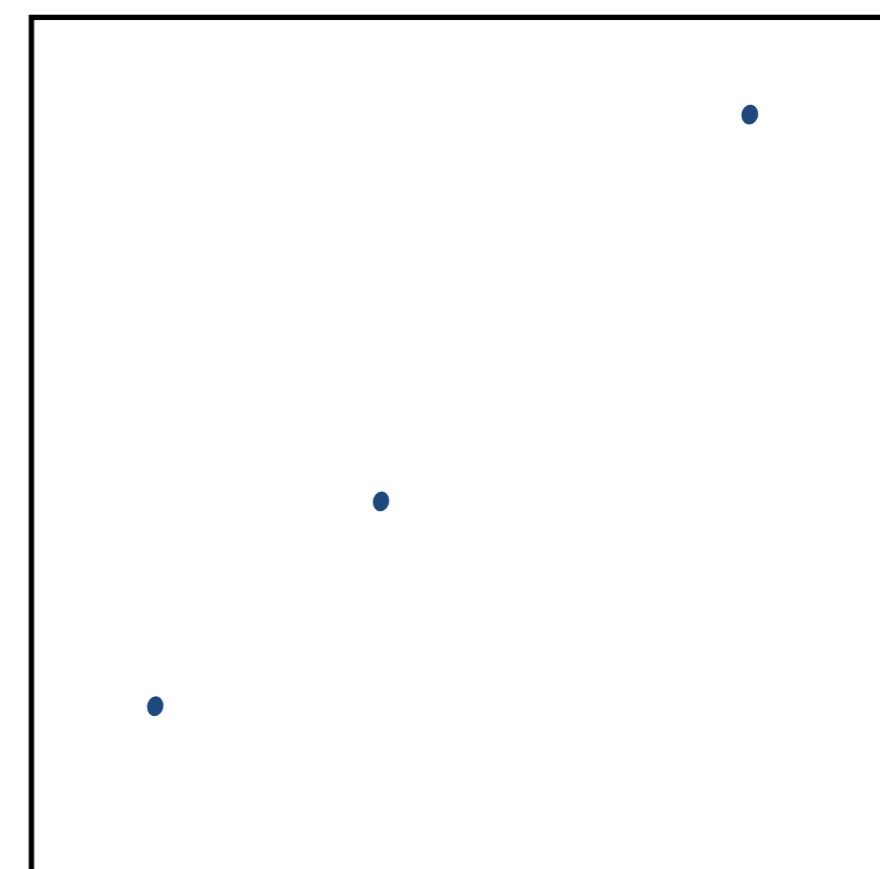


Significance of Correlation



$$r = 0.85$$

Is this significant?

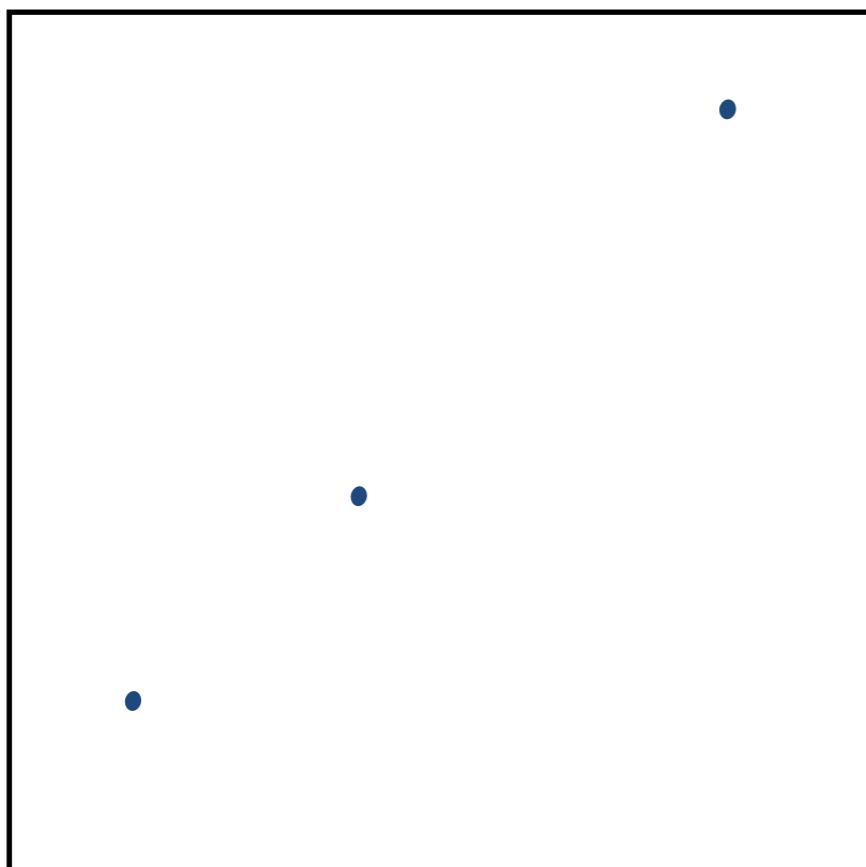


$$r = 0.99$$

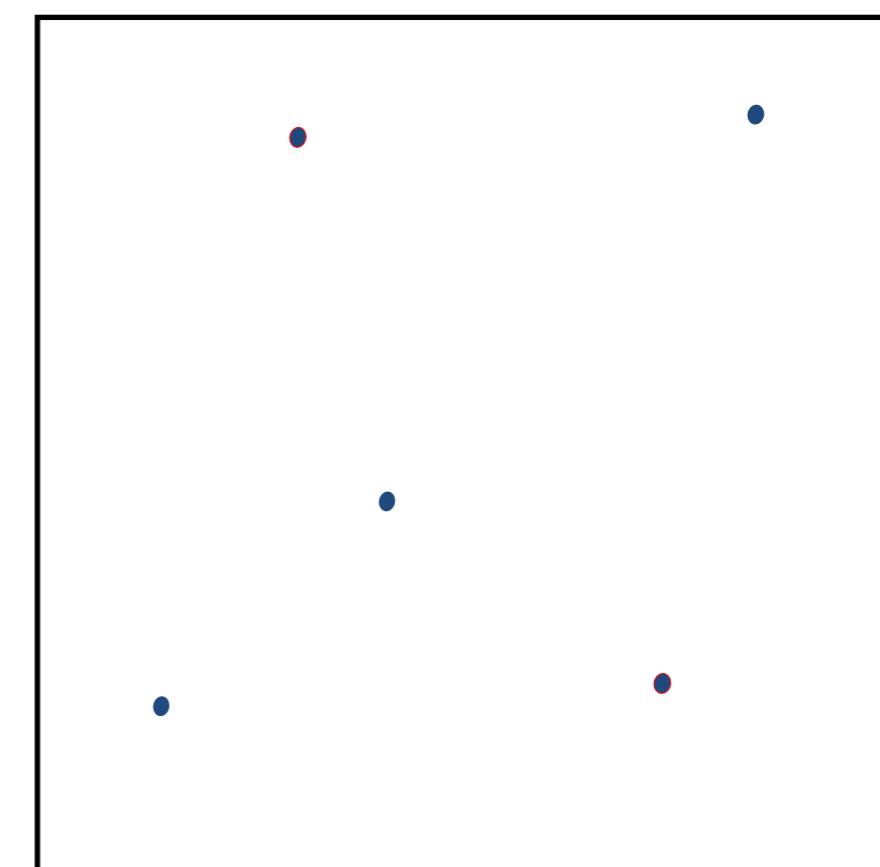
Is this significant?

Significance of Correlation

Add 2 more points to the plot



$$r = 0.99$$

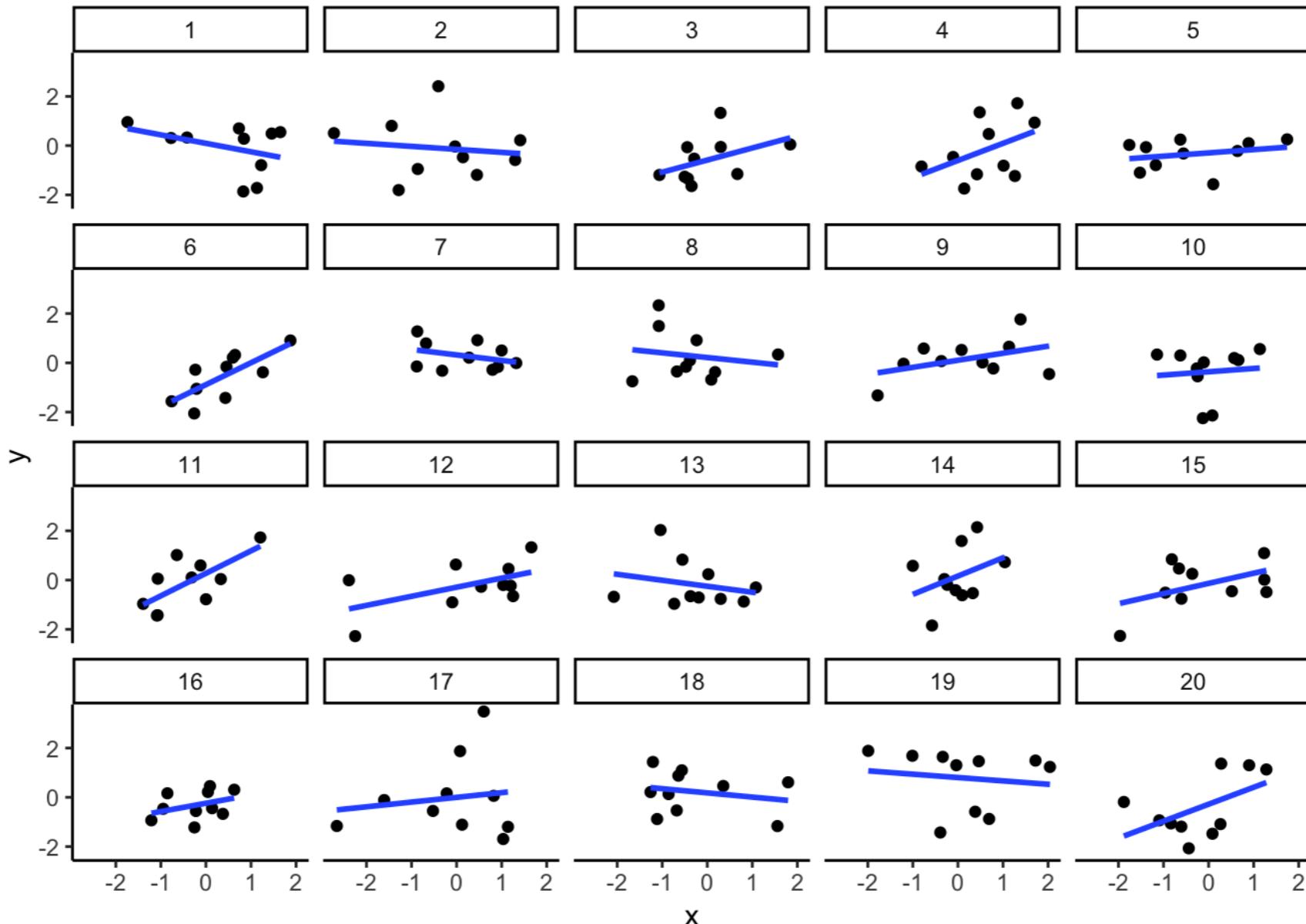


$$r = 0.05$$

Strength & Significance

- Strong relationship shown by correlation coefficient close to +/-1
 - apparently ‘strong’ relationships may not be statistically significant
 - e.g., sample size - when n is low, the odds are high that a ‘good’ correlation will occur by chance

Let's Simulate

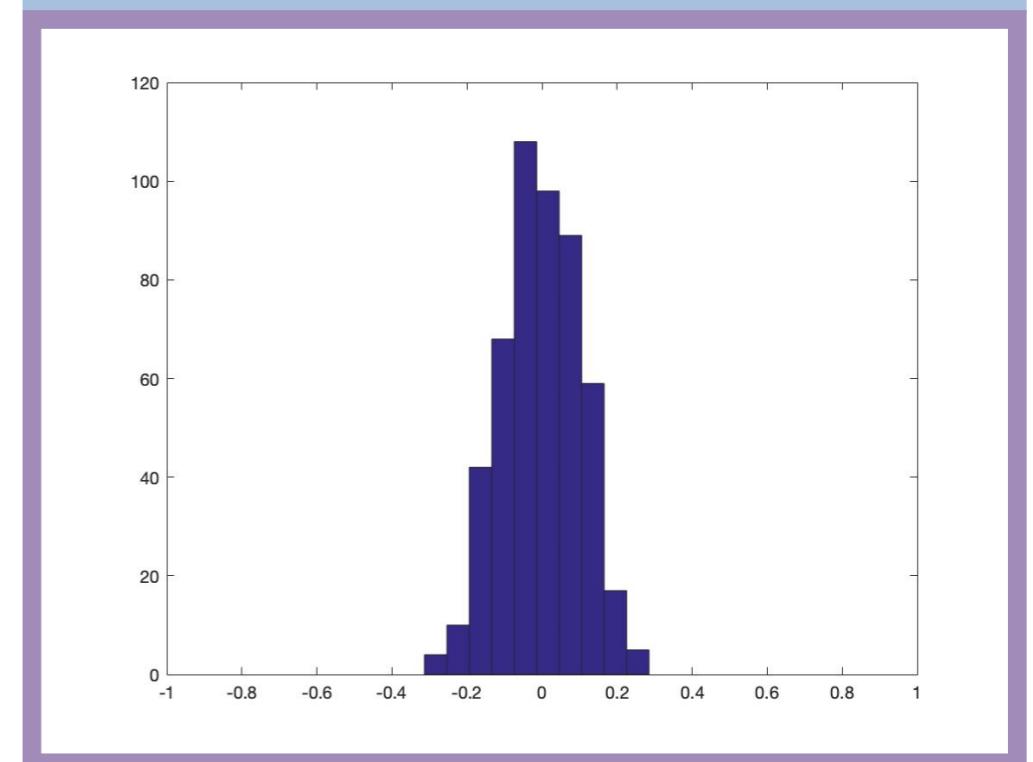
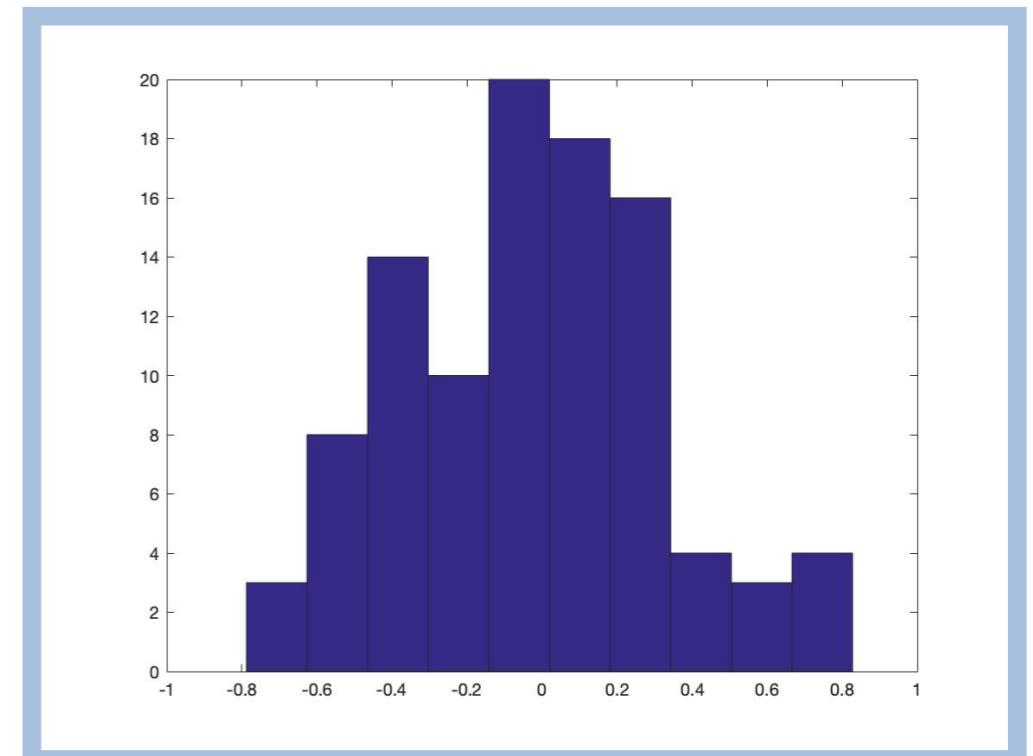
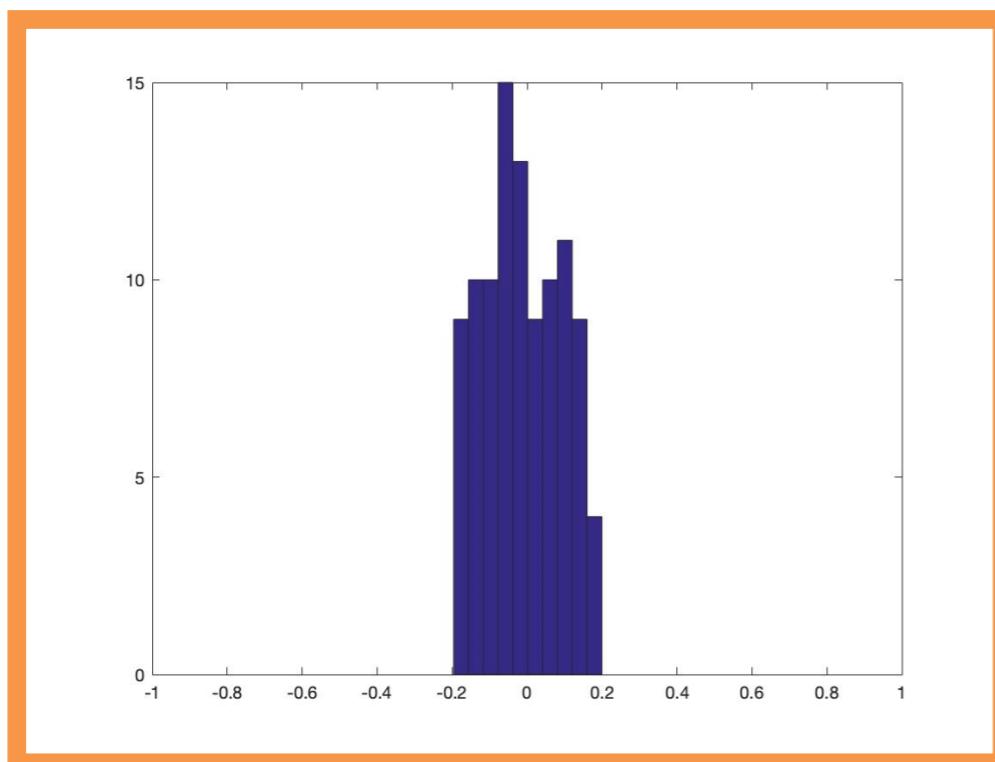


Let's make fake data: 20 draws/iterations of random numbers for two variables
For each, sample size will be 10 and scatter plot them.

Let's Simulate

How would the distributions of r look like for the following:

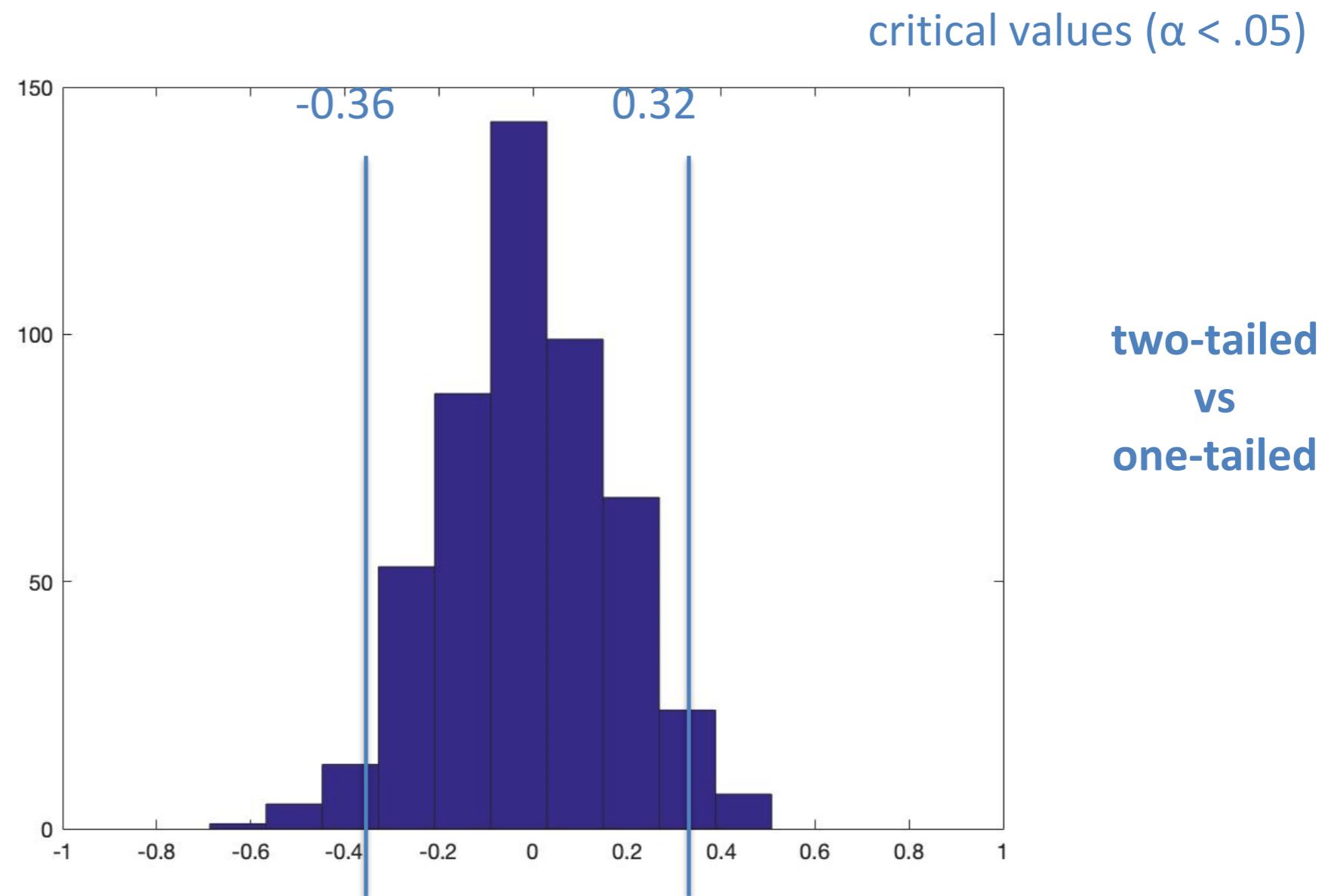
- i) sample size = 10, iterations = 100
- ii) sample size = 100, iterations = 100
- iii) sample size = 100, iterations = 500



Let's Simulate

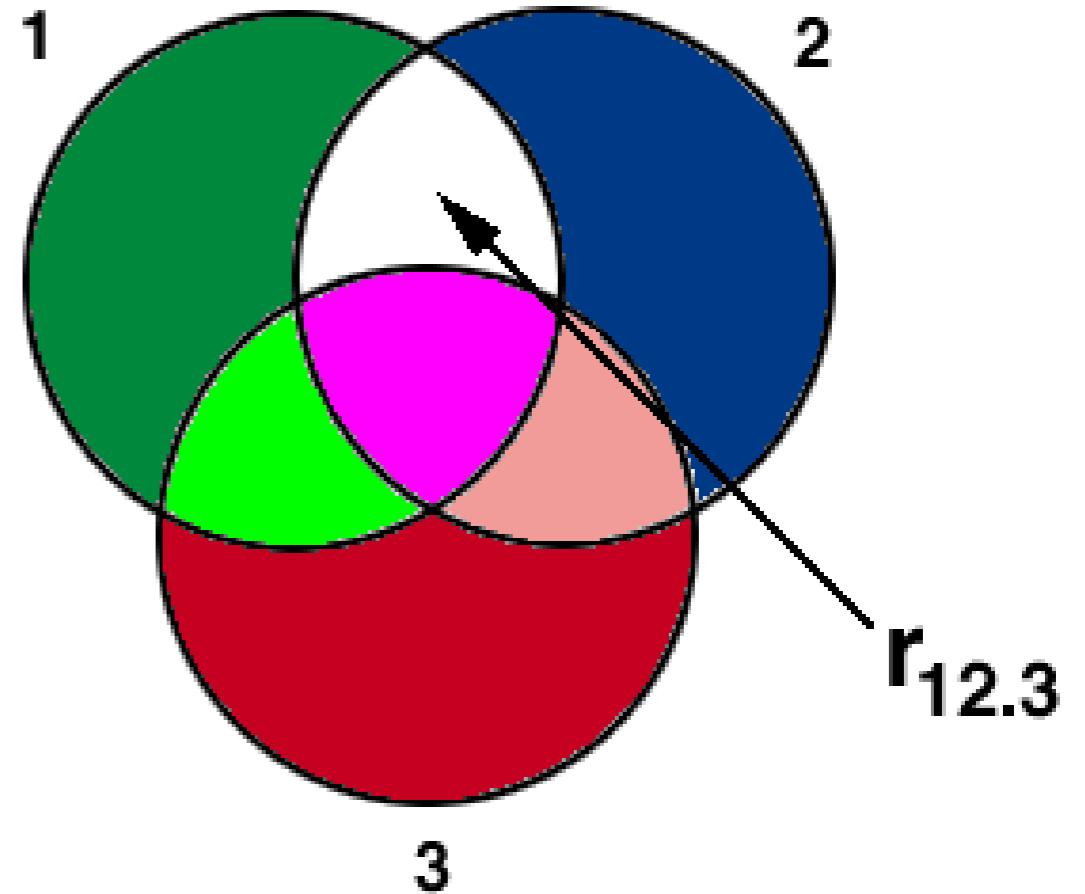
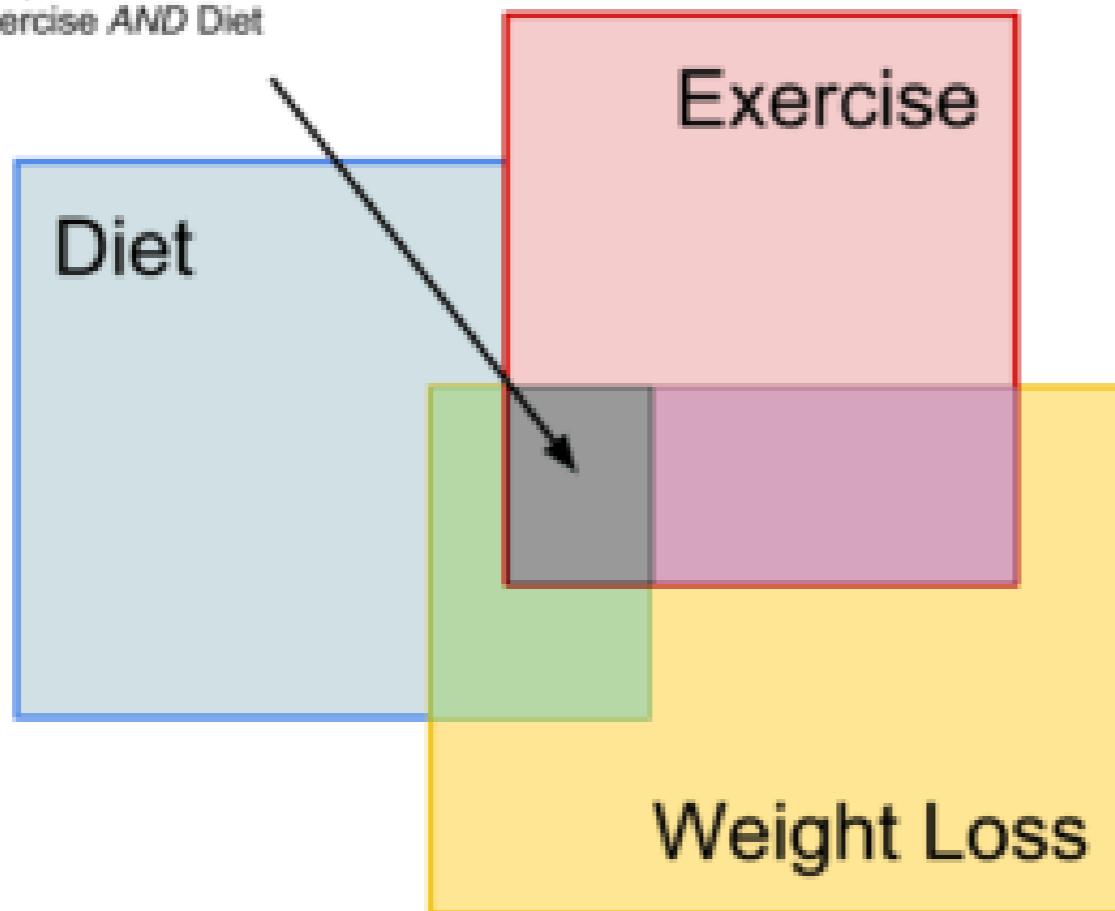
What would the critical r values be
for a sample size of 30?

i) $n = 30$, iterations = 500



Partial Correlation

Unique correlation of
Exercise AND Diet



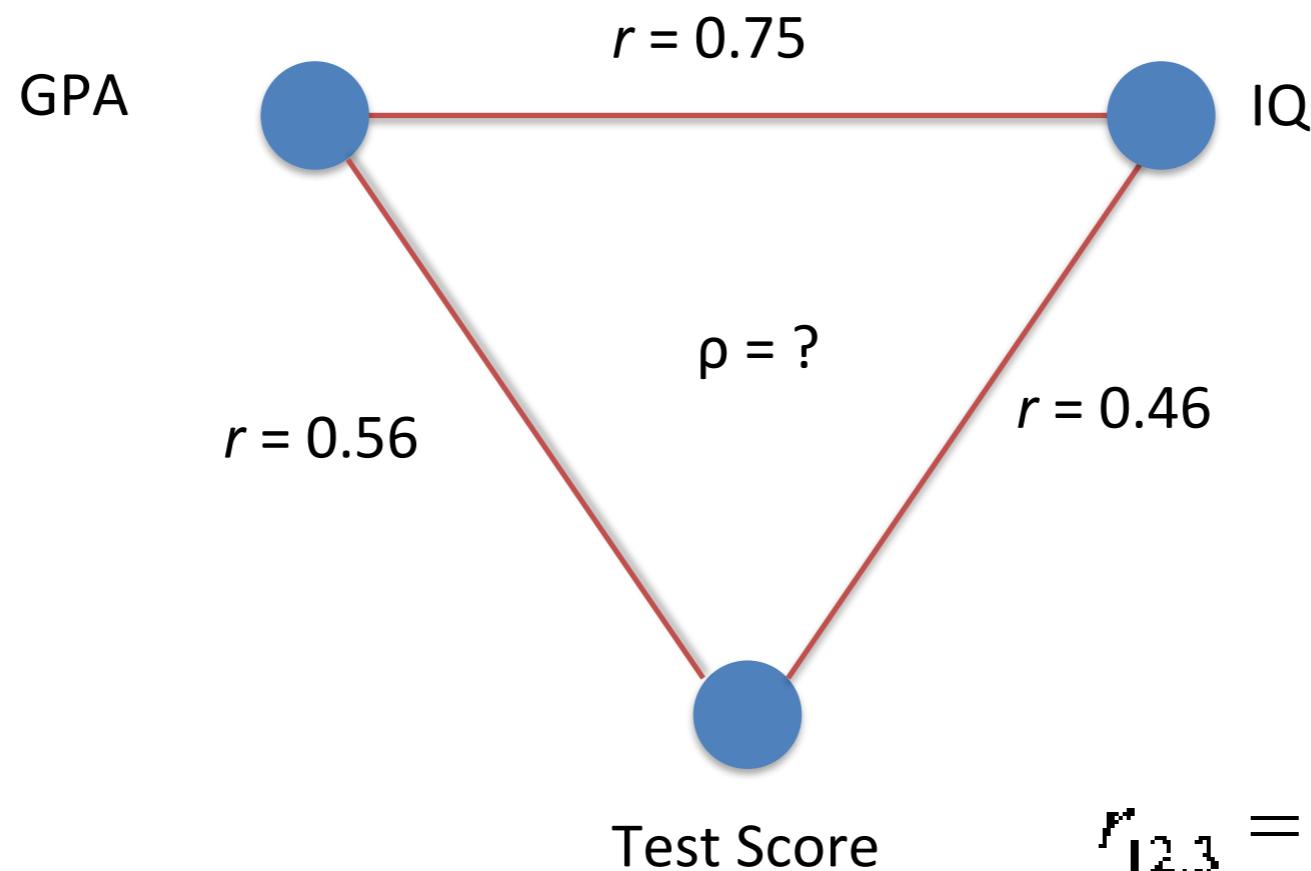
Partial Correlation

- measure of association between two variables, while controlling or adjusting the effect of one or more additional variables
 - What is the relationship between test scores and IQ scores after controlling for no. of hours of study?

Partial Correlation

- assumptions (Pearson)
 - all pairs of variables have a linear relationship
 - points are independent of each other
 - pairs of variables are bivariate normal
(typically each variable is normally distributed)
 - non-parametric version for non-linear and or non-normal data

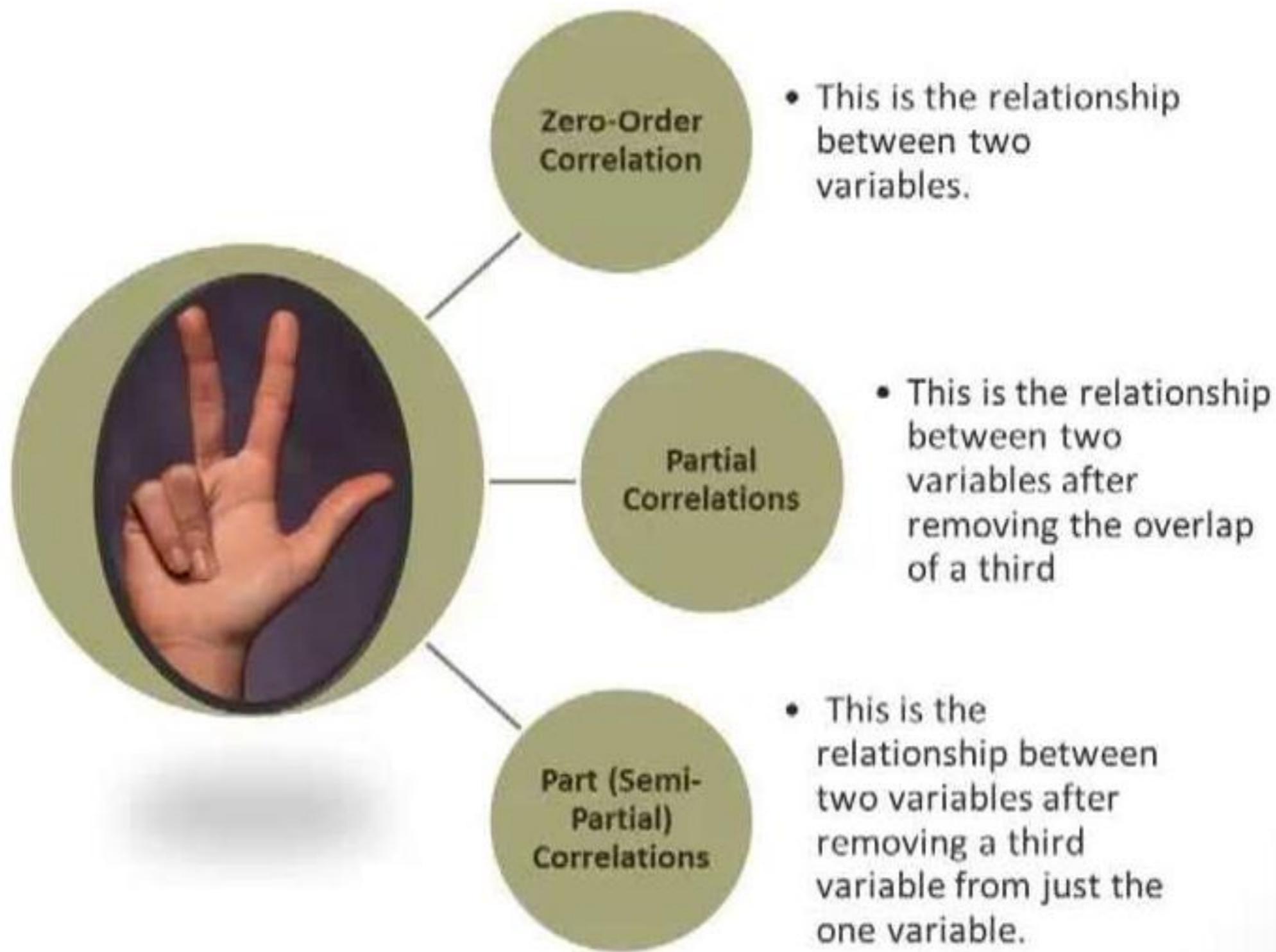
Activity/Assignment: Partial Correlation



$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Semi-Partial Correlation

- measure of association between two variables, while controlling or adjusting the effect of one or more additional variables **only on one of the two variables**
 - eg: you are interested in understanding the relationship between study time, tutoring, and exam scores while considering the potential confounding effect of study time on the relationship between tutoring and exam scores
 - how would you proceed?



$$r_{1(2.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{23}^2}} \text{ and } r_{2(1.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}}$$

BRSM Reliability & Outliers

Vinoo Alluri & Bapi Raju

Reliability

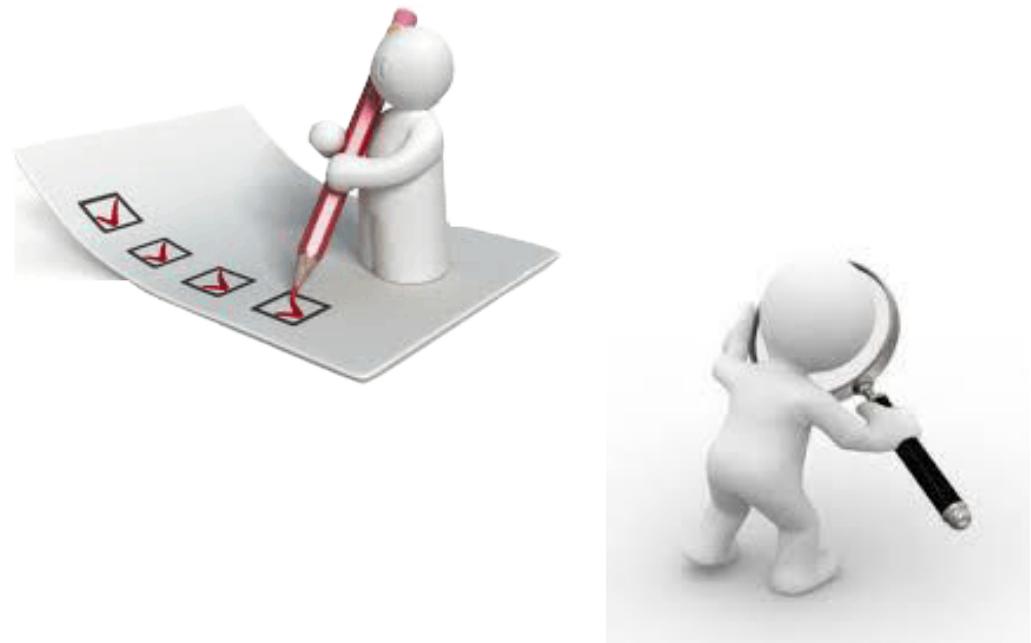
Reliability
/re-ly-a-bi-li-ti/

1. To be able to produce good results time after time. 2. How much a person can be depended on.

- **consistency** and **stability** of a research instrument
(ex: measure or score or person)
- any measure we use in research should be reliable, otherwise it's useless
- **repeatability** of a method/test or research findings

Kinds of Reliability

- Tools/methods or measuring device
- People



Threats to Reliability

- **measurement error:** equipment malfunction, human error, or ambiguous wording in survey questions
- **instrumentation changes:** measurement instruments are not consistent across repeated measurements, changes in the instrument itself can introduce variability and affect reliability.
- **practice effects:** Participants might improve their performance in a task due to practice or learning effects, leading to different results on subsequent administrations
- **sampling variability:** In experiments involving small sample sizes, random fluctuations in the characteristics of the participants can lead to unreliable results.

Threats to Reliability

- **participant error**: any factor which adversely alters the way in which the participant responds
 - ex: interview at 11 am vs 6 pm
- **participant bias**: any factor which produces a false/biased response
 - ex: mental health questionnaire in a company
- **researcher error**: any factor which alters the researcher's interpretation
 - ex: fatigue effects if interview all day
- **researcher bias**: any factor which induces bias in the researcher's recording of responses
 - ex: subjective interpretation (to get the “result” you expect)

Kinds of Reliability

stability and degree of agreement
between **people** during measurements

stability and consistency of
method/tool/apparatus
over time/repeated
measurements

Intra-Rater
Inter-Rater
Reliability

Test-Retest
Reliability

Internal
Consistency

Parallel
Alternate
Form

coherence of attributes constituting the
method/tool/apparatus

equivalence of two versions of
the method/tool/apparatus to
compare results

Kinds of Reliability

Cohen's Kappa (nominal; 2 raters)

Fleiss' Kappa(nominal; >2 raters)

Kendall's coefficient of concordance (ordinal)

Krippendorff's Alpha (all measurement levels)

Intra-Rater
Inter-Rater
Reliability

Test-Retest
Reliability

Internal
Consistency

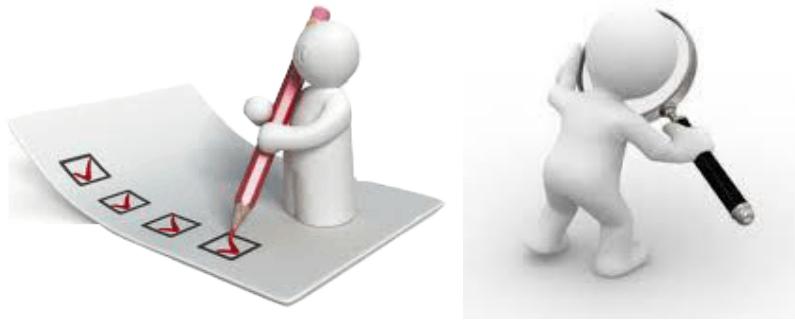
Parallel
Alternate
Form

Cronbach Alpha

Split-Half

Kuder Richardson-20/21

Pearson's correlation



Reliability

- For people (reliability of participants)
 - Inter-rater or Inter-observer Reliability - degree of agreement between two participants or observers simultaneous recorded measurements
 - ▶ correlation, helps in outlier detection
 - Intra-observer Reliability - degree of agreement within the same observer's measurements on repeated occasions



Reliability

EXAMPLE

- For people (reliability of participants)
 - Inter-rater or Inter-observer Reliability

Does this specimen have a chin?

	Kevin	Mayla
1. <i>Pan troglodytes</i>	No	No
2. <i>Australopithecus afarensis</i>	No	No
3. <i>Paranthropus aethiopicus</i>	No	Yes
4. <i>Homo erectus</i>	No	No
5. <i>Homo sapiens</i>	Yes	Yes

<http://www.passbiology.co.nz/biology-level-3/human-evolution>

1, 2, and 4 probably don't; 5 probably does.

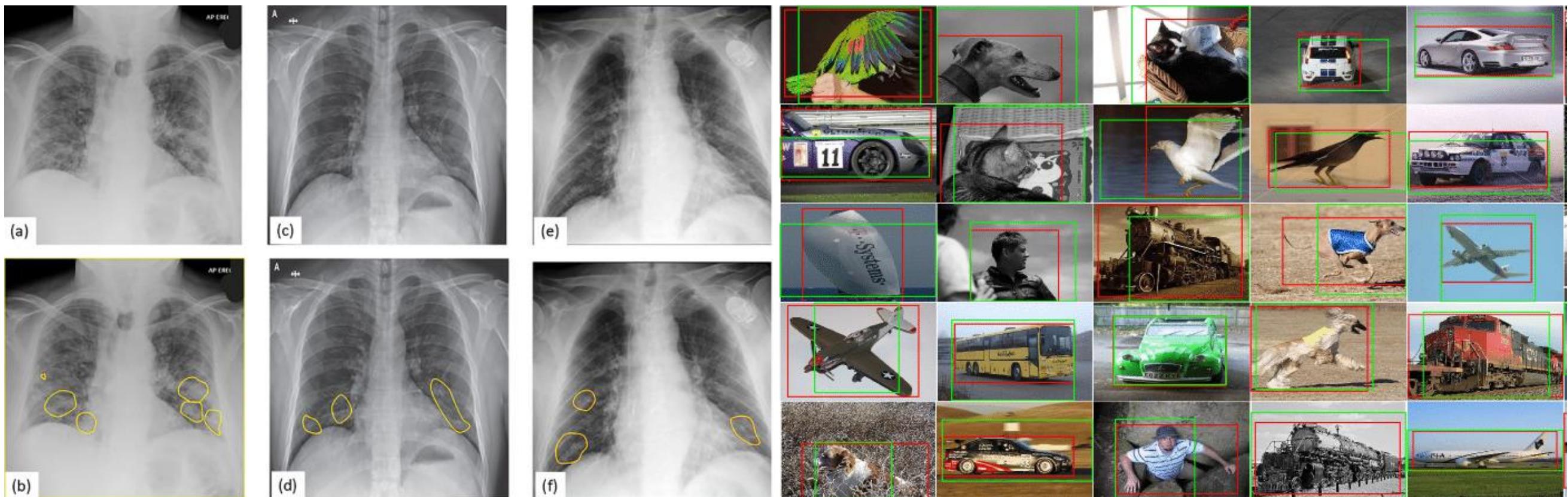
https://www.youtube.com/watch?v=fq_LNTPgVF8&app=desktop



Reliability

EXAMPLE

- For people (reliability of participants)
 - Inter-rater or Inter-observer Reliability

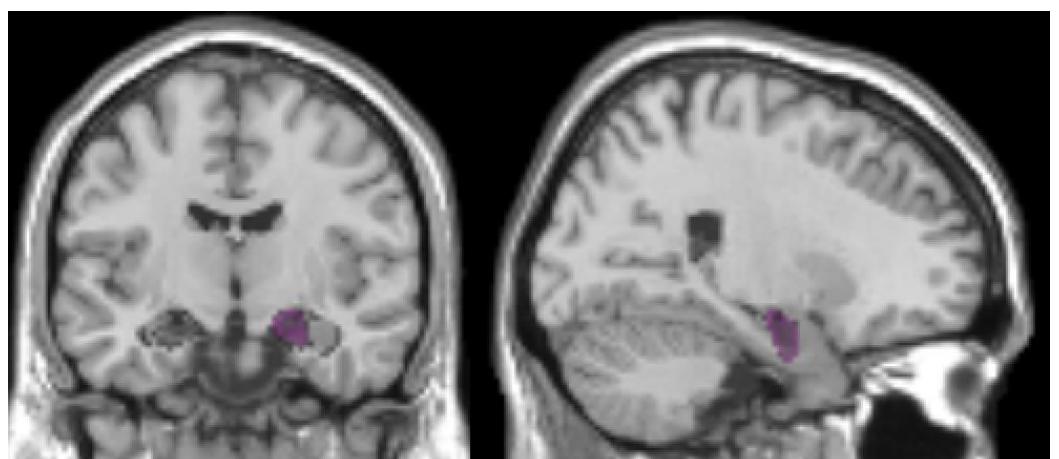


How many annotators per dataset?



Reliability

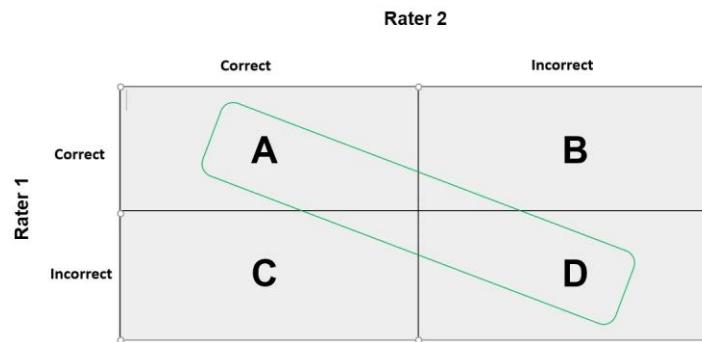
- For people (reliability of researchers)
 - Similar to participants
 - not common
 - can be assessed in qualitative research when you have more than one PI
 - ex: qualitative thematic analysis





Reliability

- ▶ *Cohen's kappa*: a quantitative measure of reliability for two raters that are rating the same thing, correcting for how often the raters may agree by chance
- ▶ can be used to check consistency of the same rater at two different time points
- ▶ used when the variable is nominal



50 images rated by 2 raters

	Yes2	No2
Yes1	20	30
No1	35	15



Reliability

- ▶ *Cohen's kappa*: a quantitative measure of reliability for two raters that are rating the same thing, correcting for how often the raters may agree by chance

r1=['yes','no','yes','no','yes','no','yes','no','yes']

r2=['yes','yes','yes','no','no','yes','yes','yes']

Agreement= sum of agreements /
total number of instances = $(4+2)/9 = 0.66$

	Yes2	No2
Yes1	4	1
No1	2	2



Reliability

- Internal consistency: Is the measurement device consistently measuring what you want it to measure?
 - ▶ Average inter-item correlation finds the average of all correlations between pairs of questions
 - ▶ Split Half Reliability: all items that measure the same thing are randomly split into two. The two halves of the test are given to a group of people and find the correlation between the two. The split-half reliability is the correlation between the two sets of scores.
 - ▶ Kuder-Richardson 20: average correlation for all the possible split half combinations in a test.



Reliability

- Internal consistency: Is the measurement device consistently measuring what you want it to measure?
 - ▶ *Cronbach's alpha:*
 - ▶ was developed in 1951 by Cronbach Lee to meet the need of finding an objective way of measuring the internal consistency reliability of an instrument used in a research work
 - ▶ mostly used when the research being carried out has multiple-item measures of a concept
 - ▶ typically used in questionnaires/surveys (self-reported)



Reliability

- Internal consistency: Is the measurement device consistently measuring what you want it to measure?
 - ▶ *Cronbach's alpha:*

$$\alpha = \frac{k\bar{r}}{(1+(k-1)\bar{r})}$$

- ▶ \bar{r} = mean inter-indicator correlation
- ▶ k=number of indicators or number of items



Reliability

EXAMPLE

- Internal consistency:
 - we have a 5 item scale showing data collected from 100 respondents

0 = Never 1 = Almost Never 2 = Sometimes 3 = Fairly Often 4 = Very Often

1. In the last month, how often have you been upset because of something that happened unexpectedly? 0 1 2 3 4
2. In the last month, how often have you felt that you were unable to control the important things in your life? 0 1 2 3 4
3. In the last month, how often have you felt nervous and “stressed”? 0 1 2 3 4
4. In the last month, how often have you felt confident about your ability to handle your personal problems? 0 1 2 3 4
5. In the last month, how often have you felt that things were going your way? 0 1 2 3 4



Reliability

EXAMPLE

- Internal consistency:
 - we have a 5 item scale showing data collected from 100 respondents
 - Correlate 100 responses x 5 items matrix

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	1.0				
Item 2	.35	1.0			
Item 3	.42	.31	1.0		
Item 4	.25	.38	.41	1.0	
Item 5	.21	.36	.46	.31	1.0

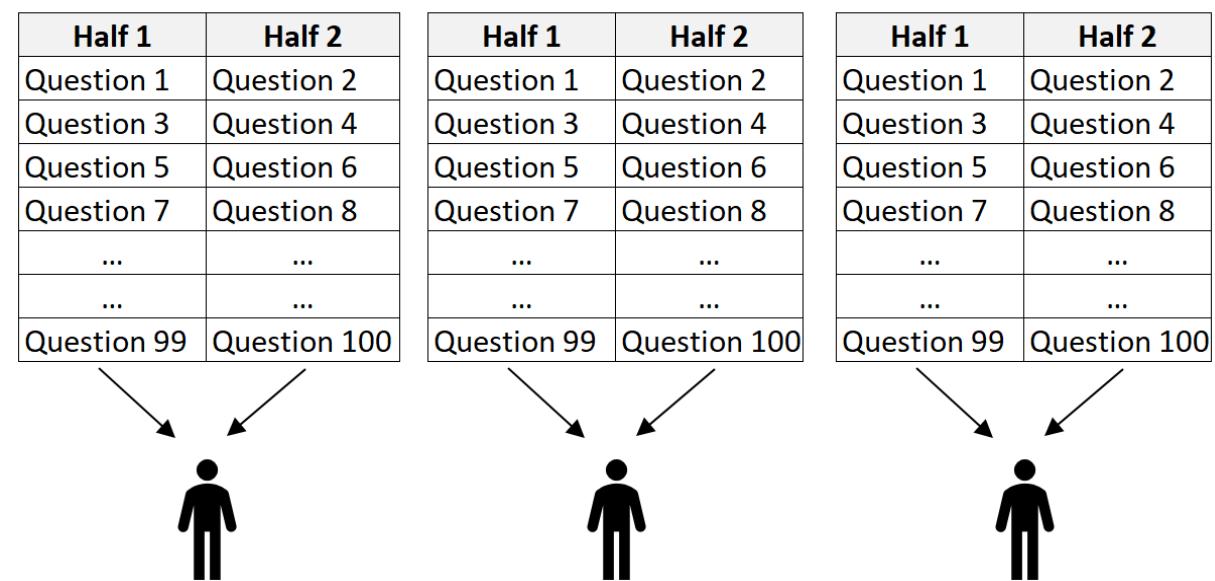
$$\alpha = \frac{k\bar{r}}{(1+(k-1)\bar{r})} = .73$$

Cronbach's alpha	Internal consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable



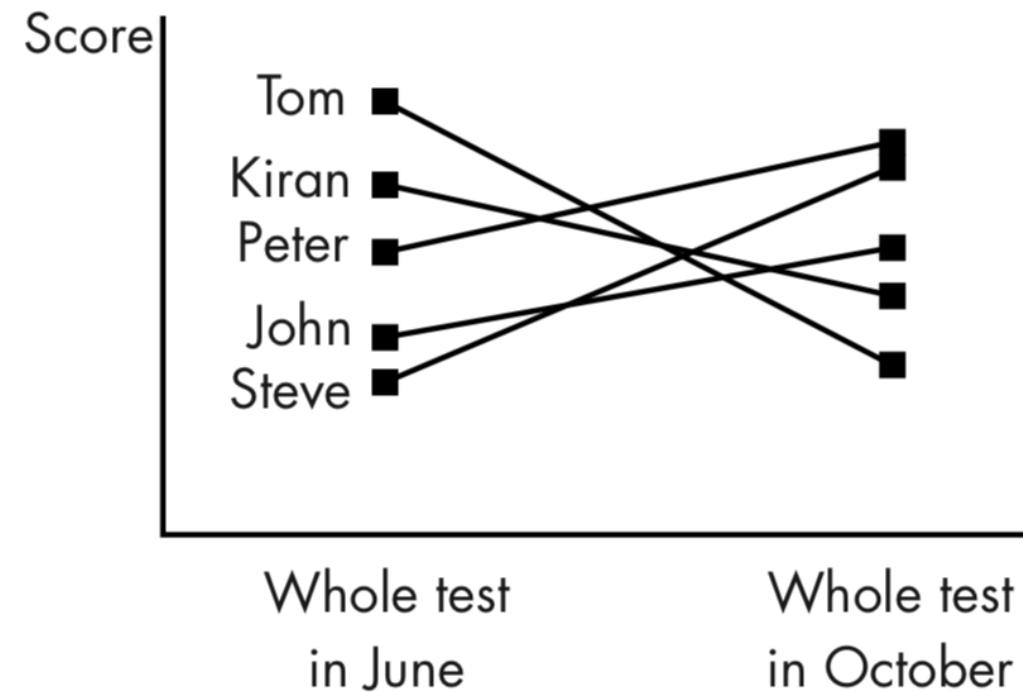
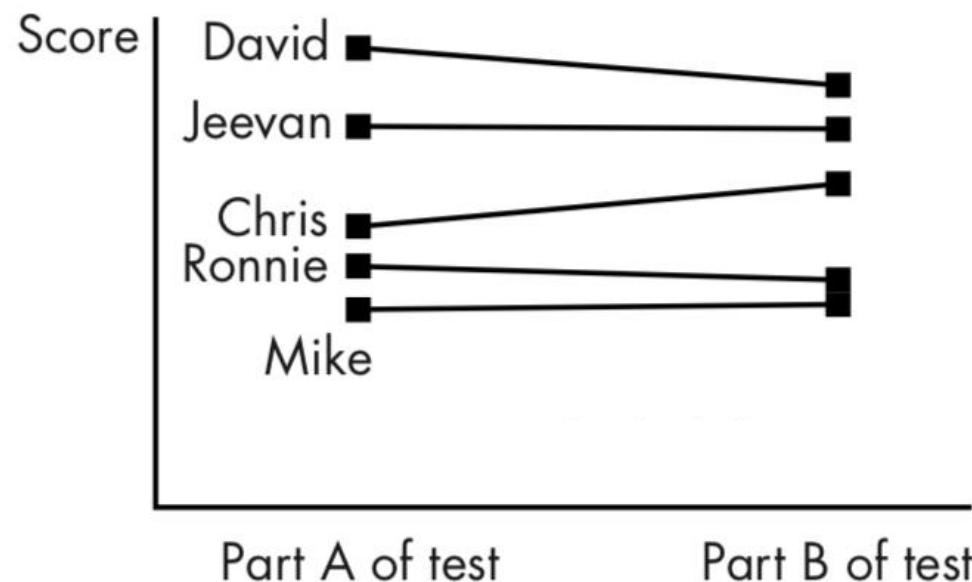
Reliability

- Internal consistency: Is the measurement device consistently measuring what you want it to measure?
 - ▶ *Split-half:*
 - ▶ uses only some of available correlations;
 - ▶ compare results of one half to the other half.
 - ▶ If the test is reliable then people's scores on each half should be similar





Reliability

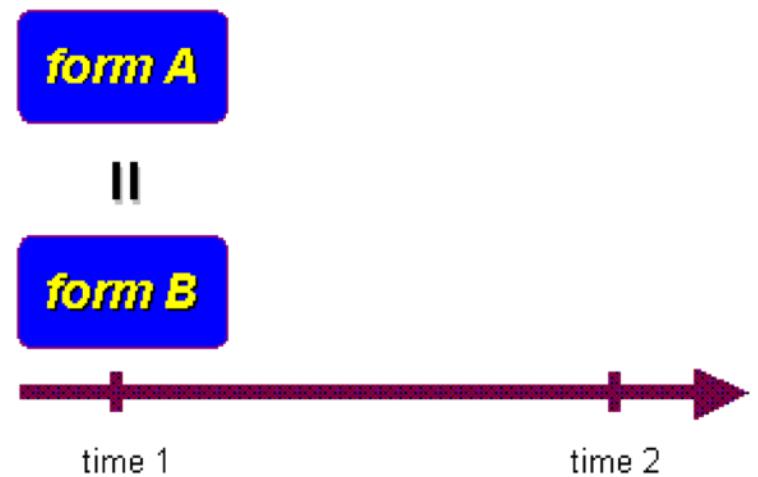


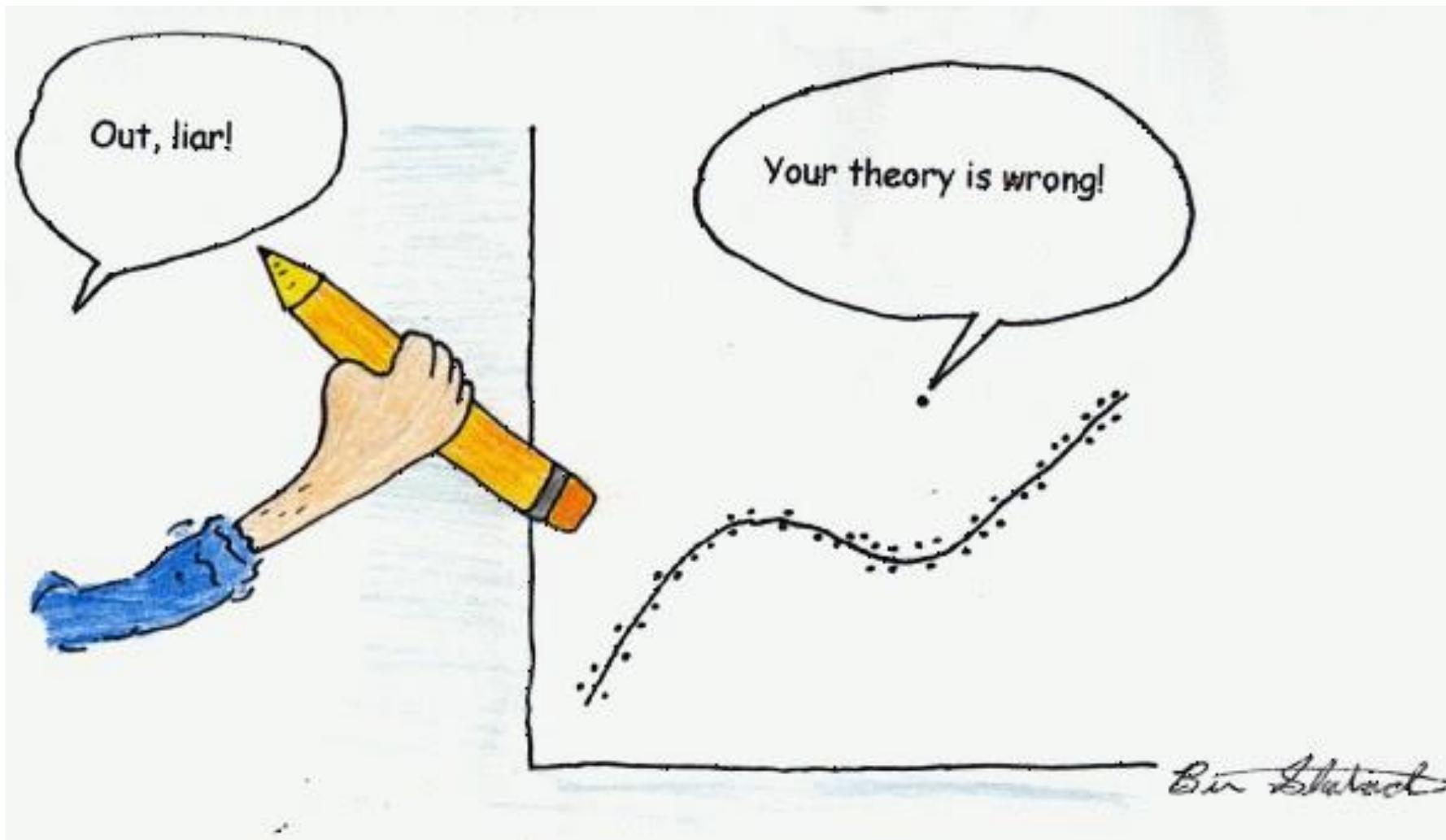
What kind of reliability and how good/bad is it?



Reliability

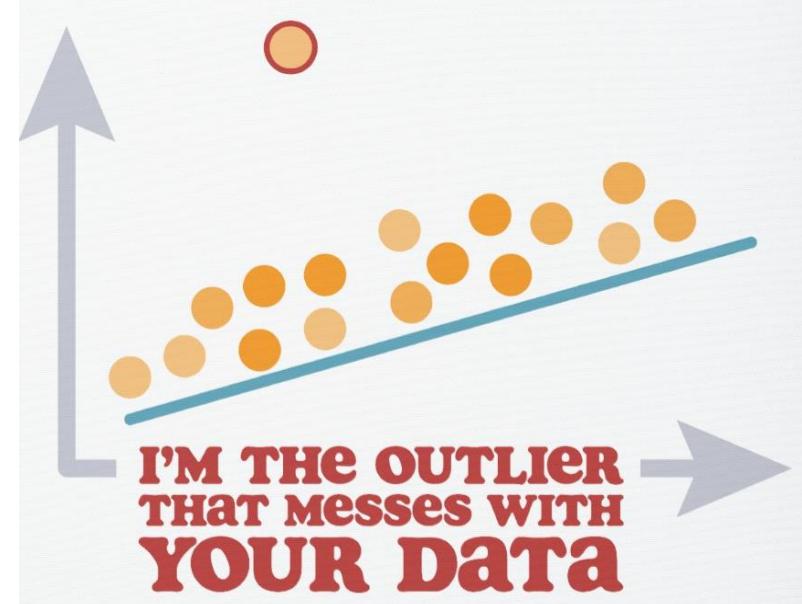
- parallel forms:
 - measure of reliability obtained by administering different versions of an assessment tool (both versions must contain items that probe the same construct, skill, knowledge base, etc.) to the same group of individuals
 - can avoid some problems inherent with test-retesting





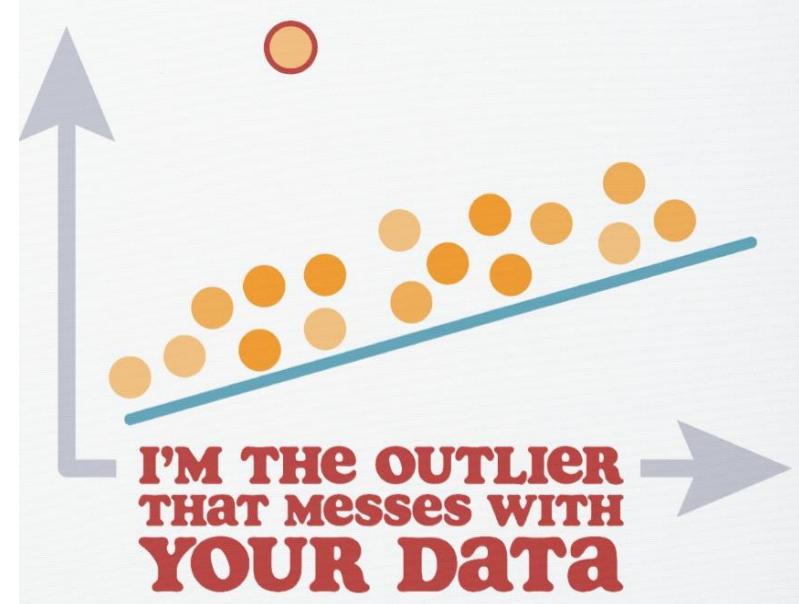
To have or not to have

Outliers



- detecting outliers is of major importance for almost any quantitative discipline (ie: Physics, Economy, Finance, Machine Learning, Cyber Security, Cognitive Science)
- not as common when sample size is low
 - ex: neuroimaging, qualitative studies involving interviews
- individual vs item/scale/stimulus

Outliers

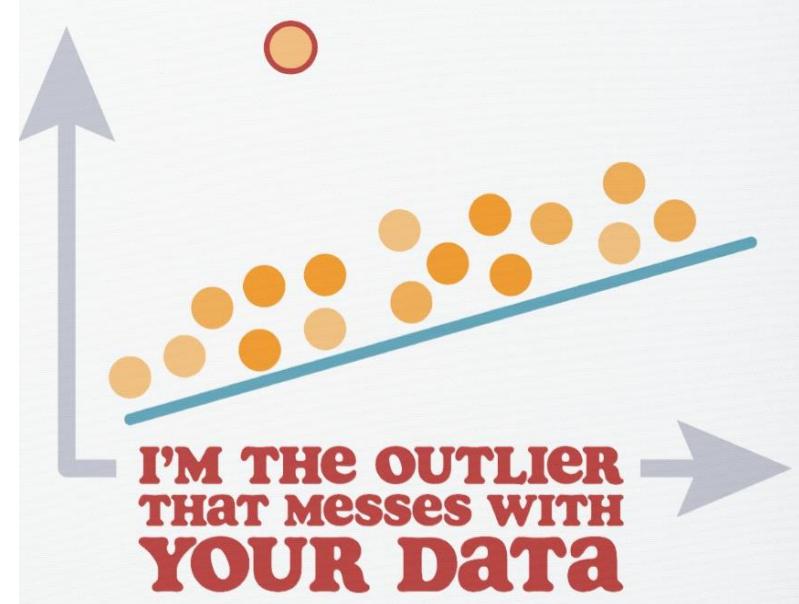


- probable causes?
 - measurement/execution errors (instrument errors/data extraction or experiment planning errors)
 - eg: improper scanner handling
 - data entry errors, missing data (human errors)
 - eg: entering 999 for missing values and using it for analysis

Dealing with Outliers

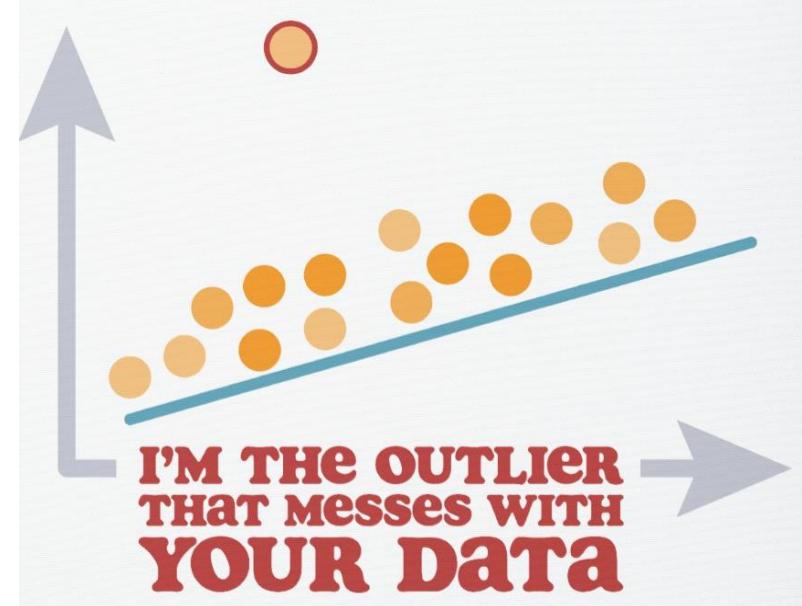
- omit
- replace (ex: with mean)
- using different analysis methods (ex: non-parametric tests)
- valuing the outliers
- data transformation

Outliers



- probable causes?
 - measurement/execution errors (instrument errors/data extraction or experiment planning errors)
 - eg: improper scanner handling
 - data entry errors, missing data (human errors)
 - eg: entering 999 for missing values and using it for analysis
 - data processing errors (data manipulation or data set unintended mutations)
 - eg: multiplying interval data

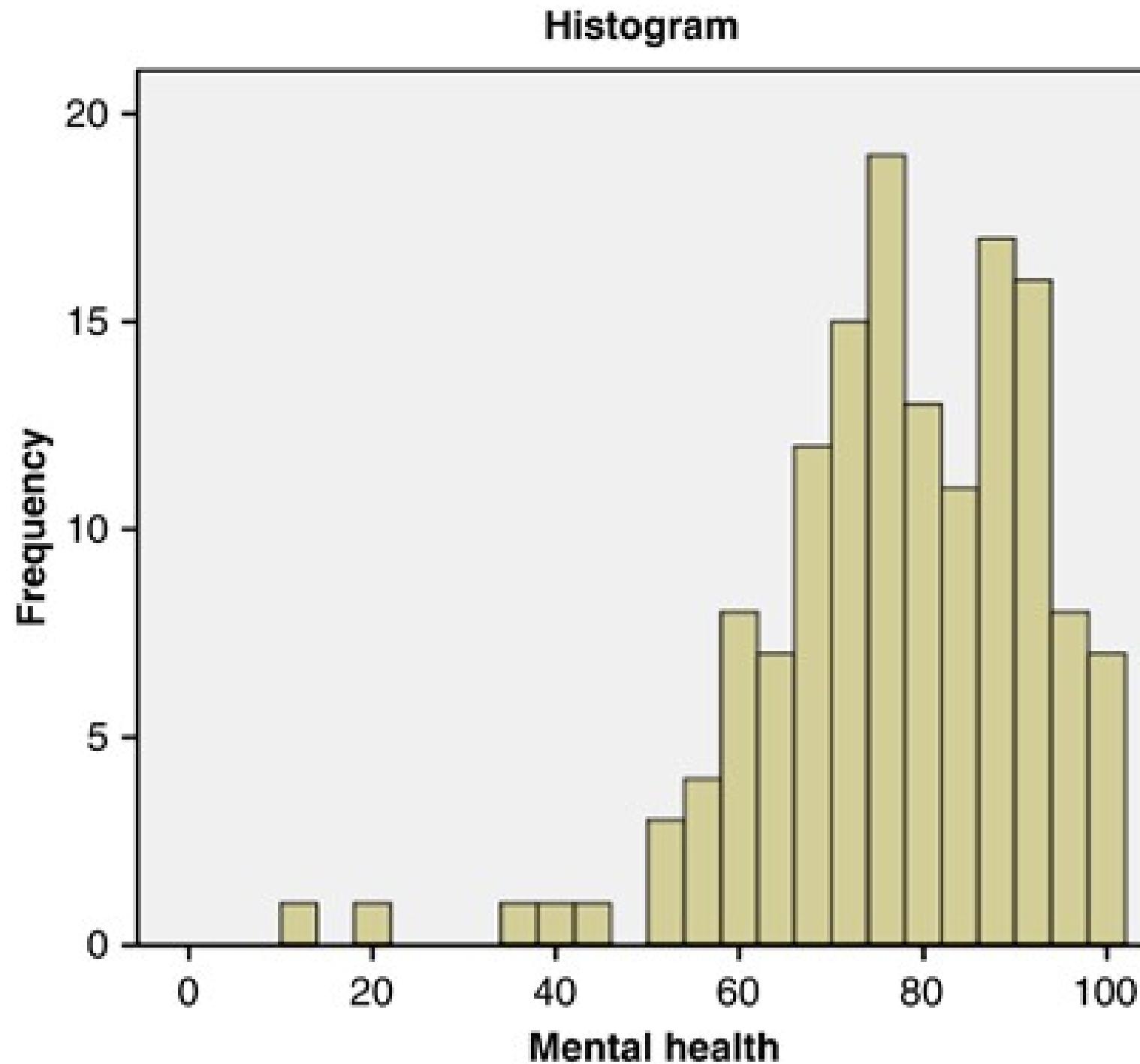
Outliers



- probable causes?
 - sampling errors (extracting or mixing data from wrong or various sources)
 - e.g: measure the weight of athletes but also include some wrestlers
 - natural (not an error, novelties in data or inherent data variability)

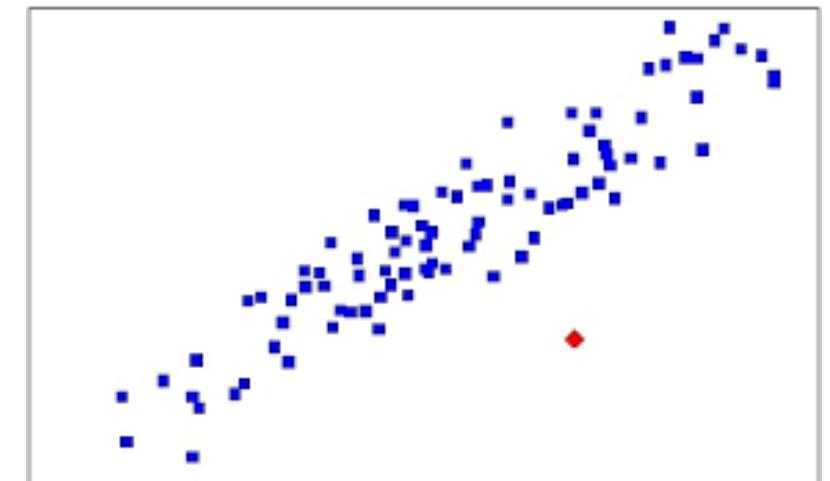
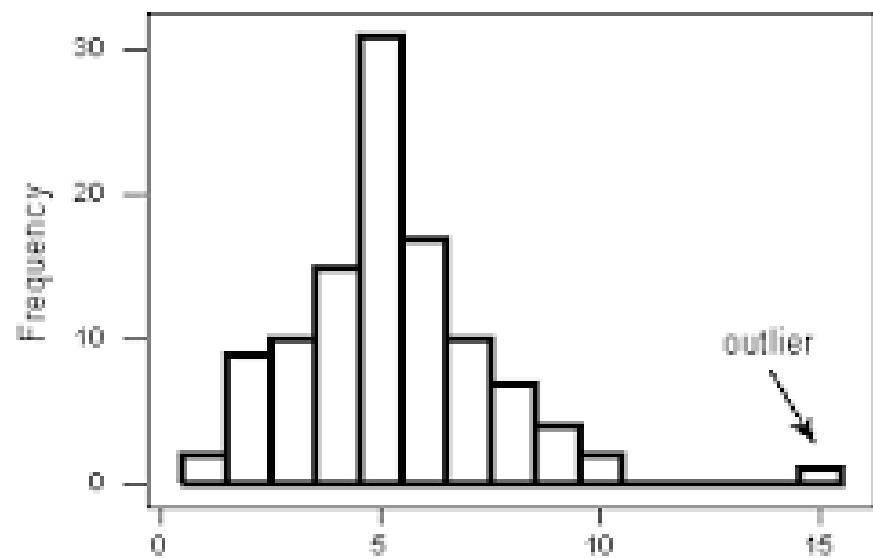
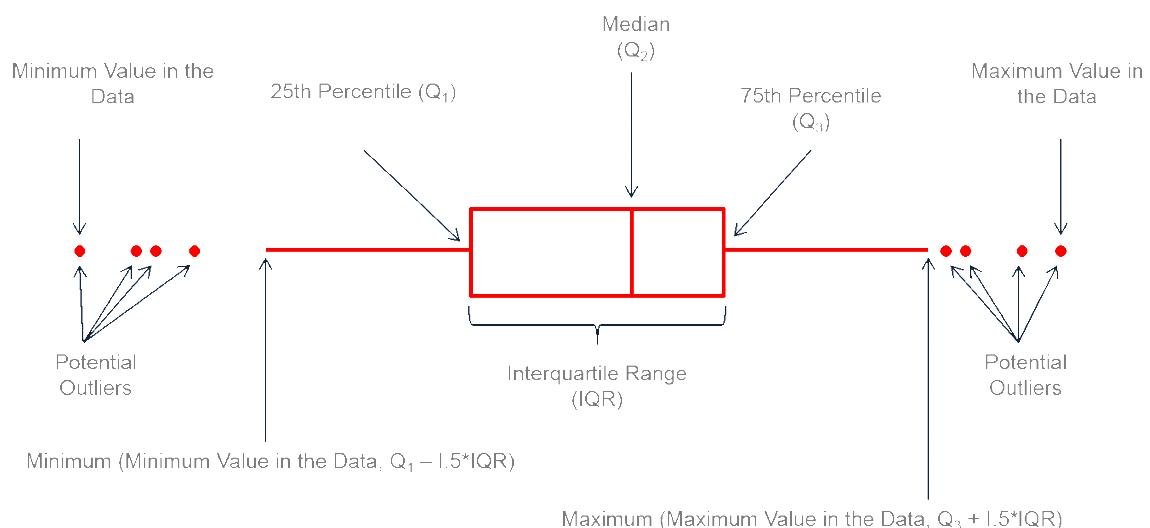
EXAMPLE

Natural Outliers



Outlier Detection

- graphical representations help (eg: scatter plot, box plot, histogram)



Outlier Detection

Intuitive way of detecting outliers (esp. in a perceptual experiment or survey)?

Outlier Detection

- graphical representations help (scatter plot, box plot, histogram)
- $>1.5 \times \text{InterQuartile Range}$
- 2/3 SDs from mean (depending on the nature of data)
- Grubbs' test (single), Tietjen-Moore test (multiple), etc..

EXAMPLE

Outlier (individual) Detection

- 2/3 SDs from mean (depending on the nature of data)
 - check individual 2SDs away from mean rating of each

37 participants

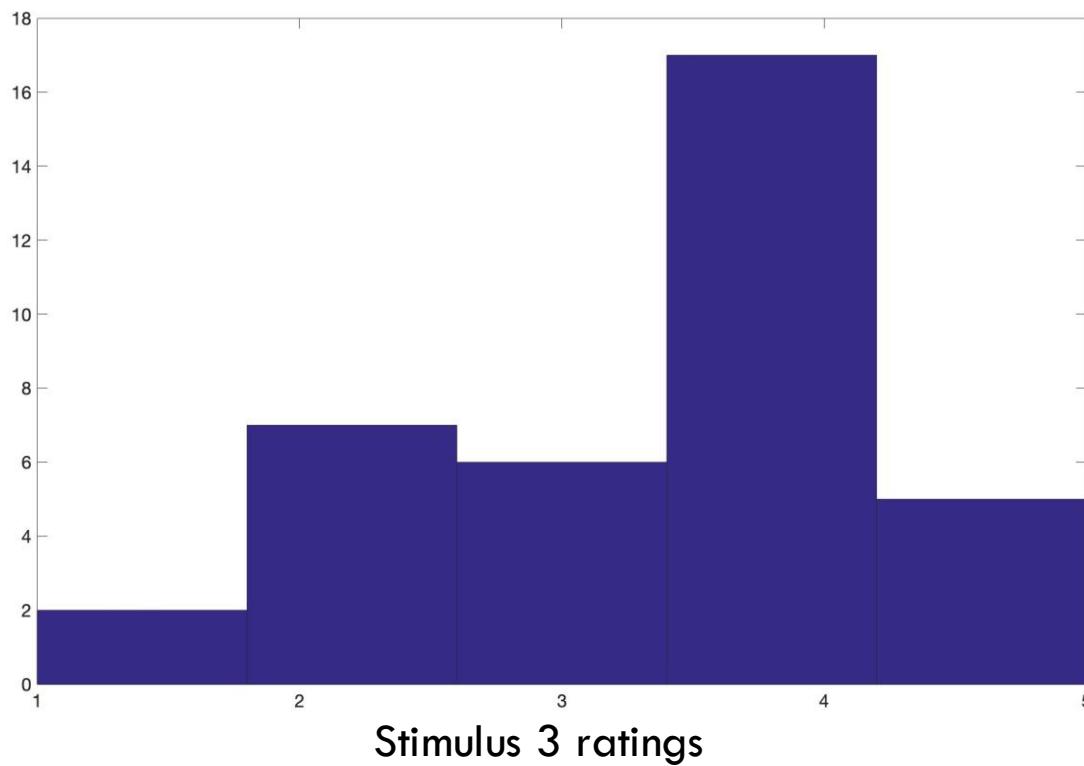
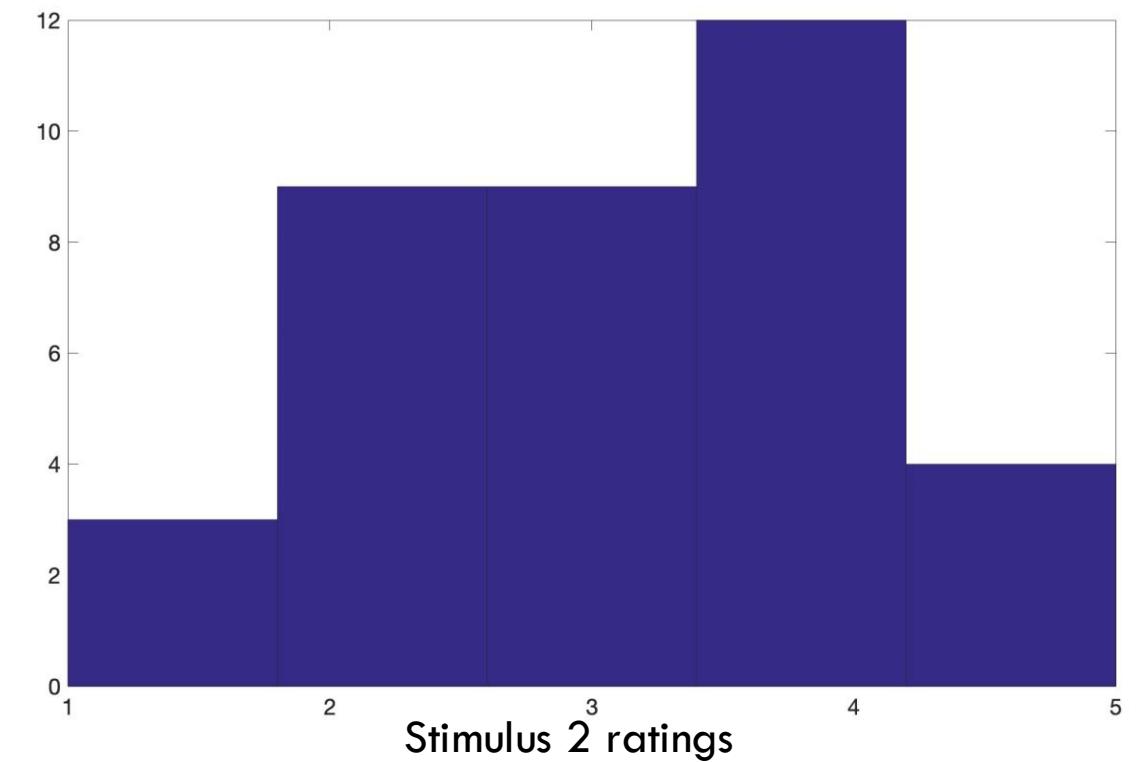
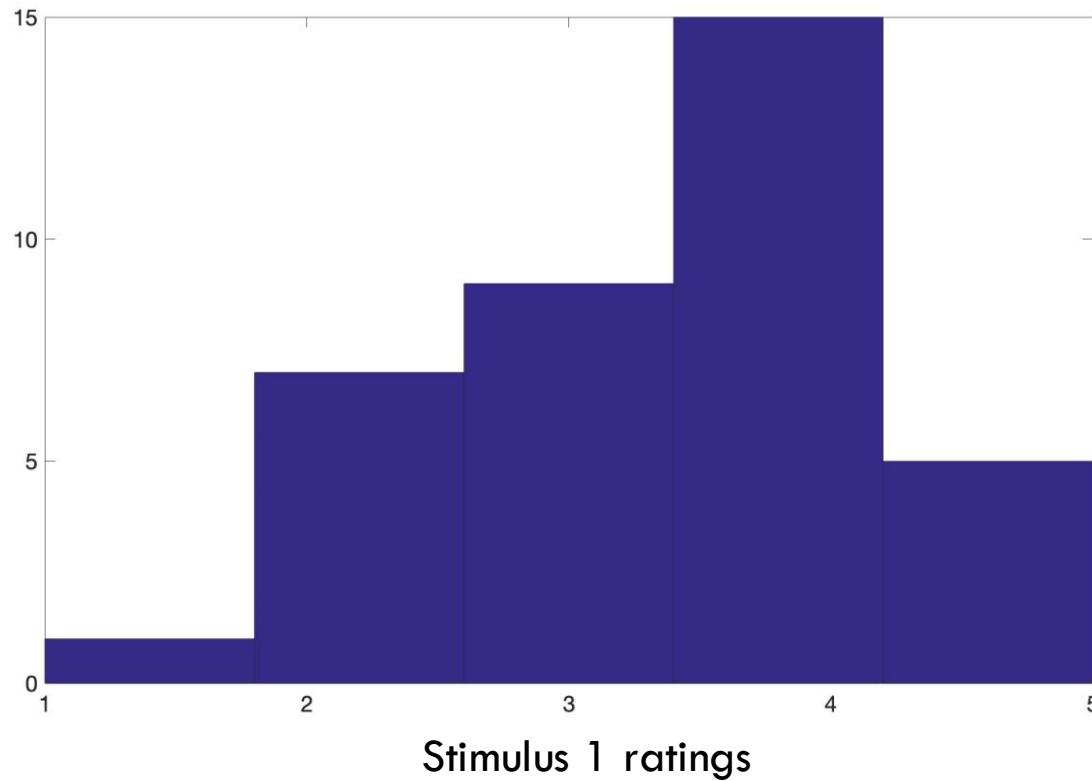


37 x 100 Arousal ratings

Rate Arousal (Energy) on a 5-point Likert scale
of
100 musical excerpts

EXAMPLE

Outlier Detection

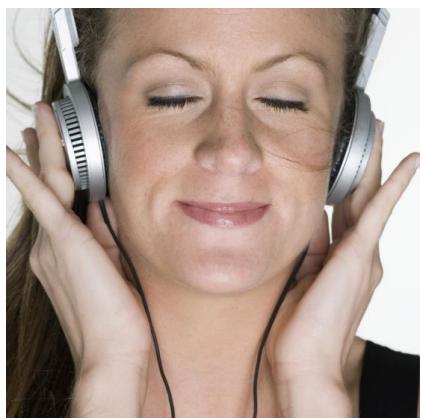


1 = low energy
5 = high energy

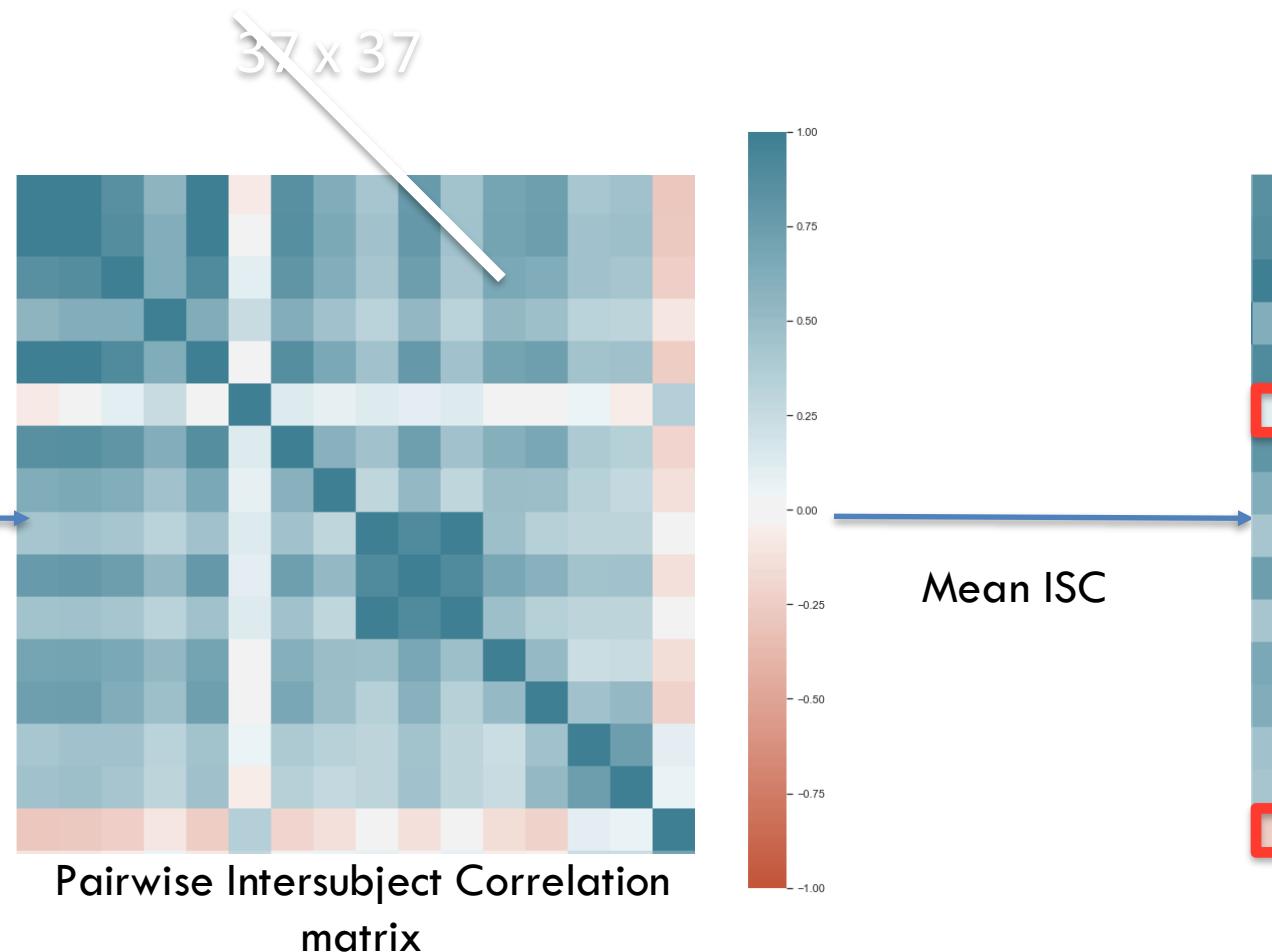
EXAMPLE

Outlier (individual) Detection

- 2SDs away from mean rating of each



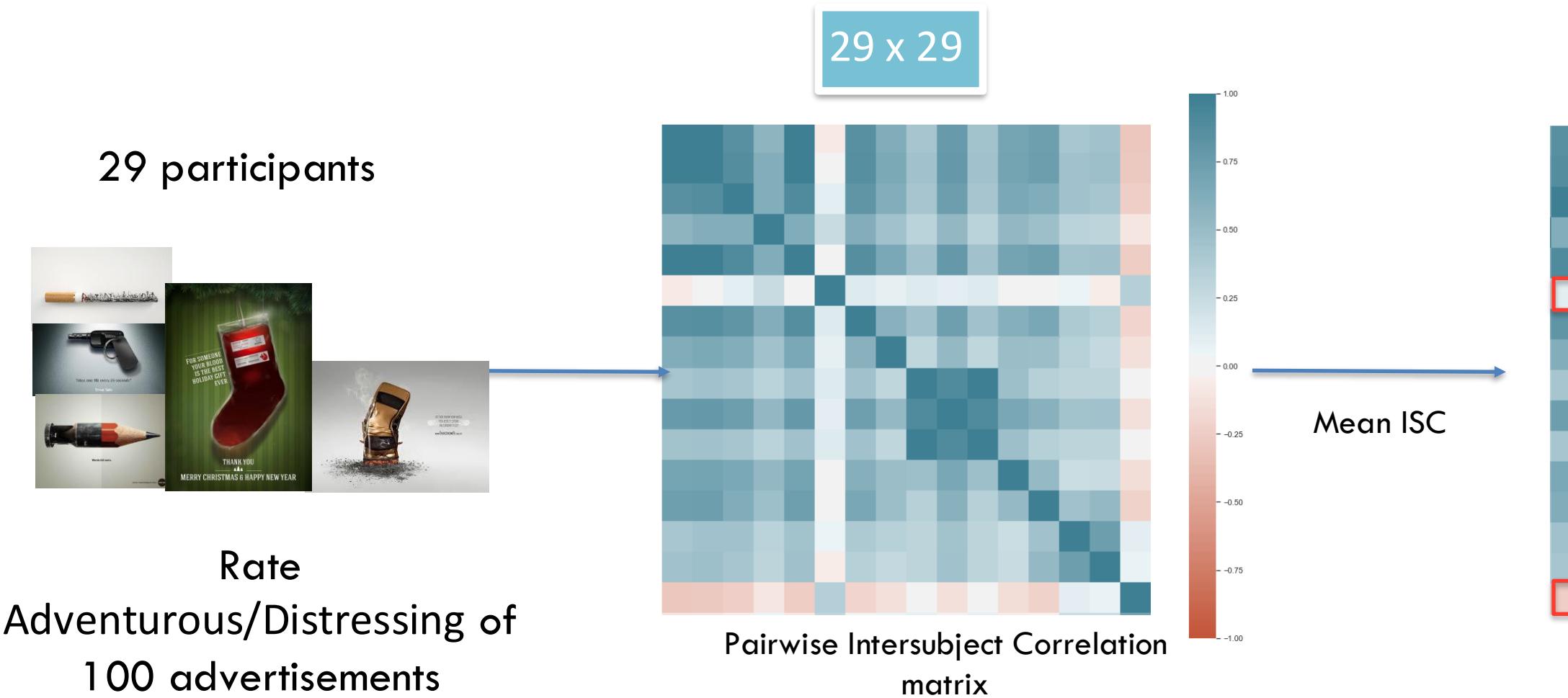
37 x 100 Arousal
ratings



EXAMPLE

Outlier (individual) Detection

- 2SDs away from mean rating of each



not always suitable (especially for subjective ratings)!

Dealing with Outliers

- omit
- replace (ex: with mean)
- using different analysis methods (ex: non-parametric tests)
- valuing the outliers
- data transformation

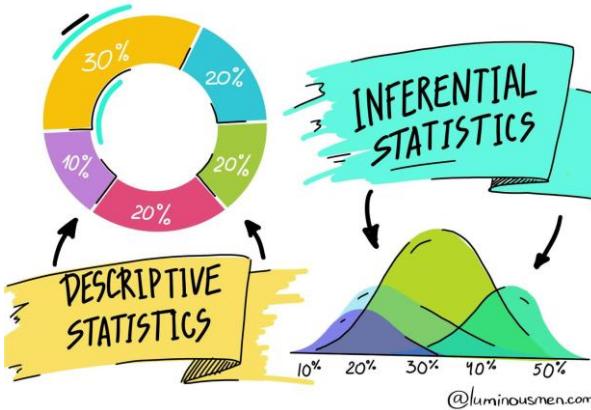
Activity: Missing Values

- Omit
- Replace by frequent value (Mode)
- Replace by Mean / Median

**Submit any 4 methods (names and 2-line description
for estimating missing values!**

Hypothesis Testing

Why do we need inferential statistics?



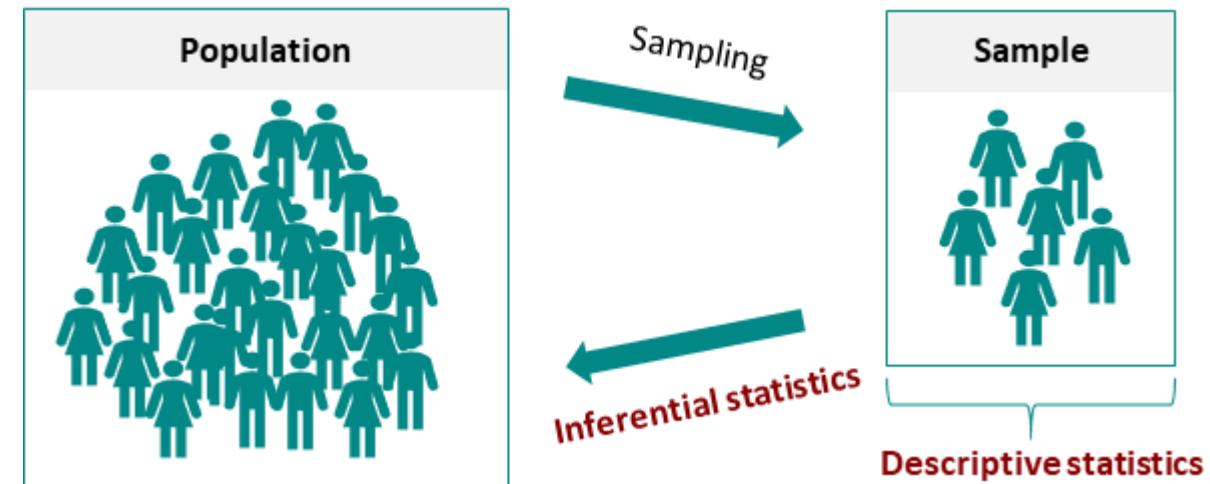
Descriptive Statistics

- Organise
- Summarise
- Simplify
- Describe and present data

Inferential Statistics

- Generalise from samples to populations
- Hypothesis testing
- Make predictions

Inferential statistics allow us to *infer* or generalize observations made with samples to the larger population from which they were selected.



What is a Hypothesis?

Research Question
(ideas)



A specific testable statement
(that guides an experiment and
statistical analysis)

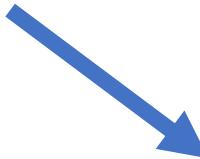
What Is a Real Hypothesis?

- A hypothesis is an educated guess, based on observation.
- Usually, a hypothesis can be supported or refuted through experimentation or more observation.
- A hypothesis can be disproven, but not proven to be true.



Research Question – Is online teaching effective?

Hypothesis Statement – Students taught offline perform better than students taught online



(ASSUMPTION) – based on previous studies, observations, experiences, etc.

Null Hypothesis and Alternative Hypothesis

H_0 VS H_1
or
 H_a

Students taught online vs offline
perform equally well on exams (no
difference/null)

You perform experiments to check if the H_0 holds true or not.
By disproving H_0 you accept the H_A

Students taught offline perform
better than students taught online

Students taught online perform
better than students taught offline

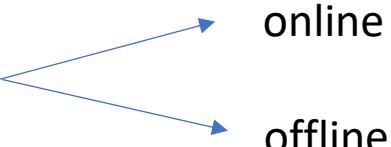
Variables in a hypothesis

Hypothesis Statement – Students taught offline perform better than students taught online

INDEPENDENT VARIABLE
DEPENDENT VARIABLE

not changed by the other variables you are trying to measure

Teaching method



online
offline

Value is changed or affected by the independent variable/s

Exam performance

Individuals with more years of education have higher income

H_0 – No relationship between years of education and income

H_1 - Individuals with more years of education have higher income

Leopards are stronger than Tigers

H_0 – Leopards and Tigers are equally strong, no difference

H_1 – Tigers are stronger than Leopards

H_2 – Leopards are stronger than Tigers

EXERCISE AND ANXIETY



Exercise effects on anxiety

H_0 - Exercise has no effect on anxiety

H_1 - Exercise lowers anxiety

H_2 – Exercise increases anxiety

IV – Exercise (exercising, not exercising)

DV – anxiety levels

Directionality in a hypothesis

This prediction is typically based on past research, accepted theory, extensive experience, or literature on the topic.

Else your statistical outcome can be misleading, by ignoring other outcomes
e.g. Does a technical degree impart technical skills?

High quality of engineering education leads to higher technical skills

Ho – Quality of engineering education has no effect on technical skills

H₁ - High quality of engineering education leads to higher technical skills

EXERCISE AND **ANXIETY**



Directionality?

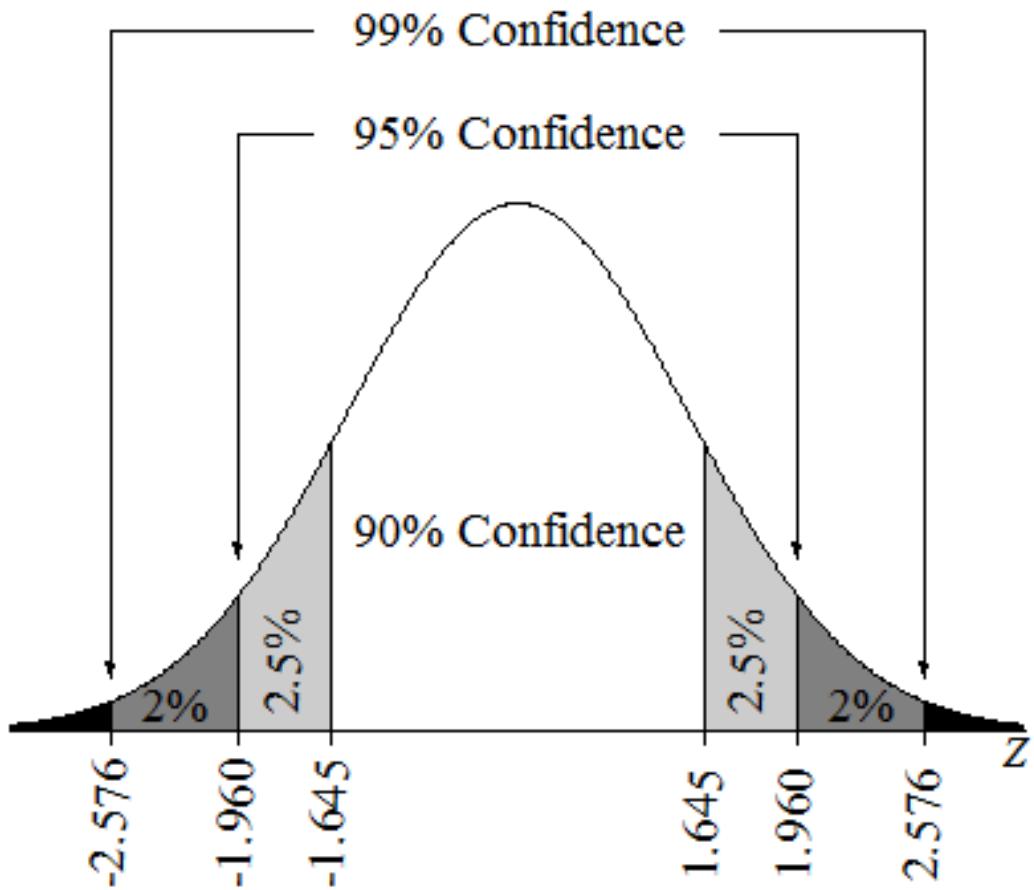
Exercise effects on anxiety

H_0 - Exercise has no effect on anxiety

H_1 - Exercise lowers anxiety

H_2 – Exercise increases anxiety

Confidence Intervals



Confidence Level	α (level of significance)	$Z_{\alpha/2}$
99%	1%	2.575
95%	5%	1.96
90%	10%	1.645

$$CI = \bar{x} \pm z \frac{s}{\sqrt{n}}$$

CI = confidence interval

\bar{x} = sample mean

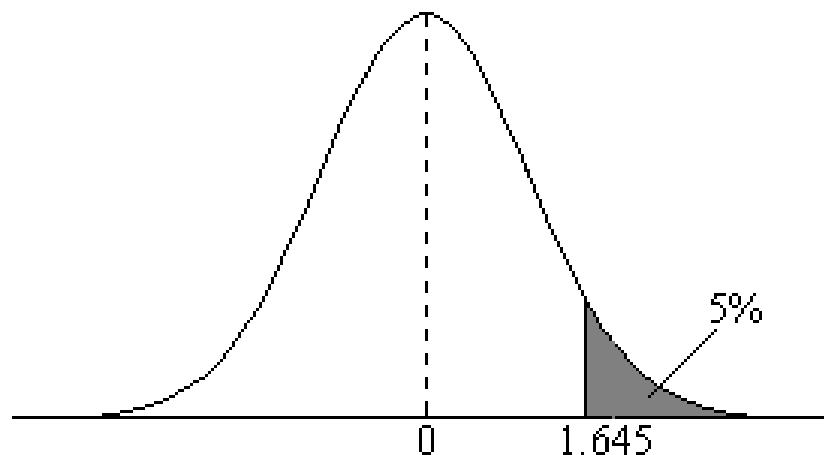
z = confidence level value

s = sample standard deviation

n = sample size

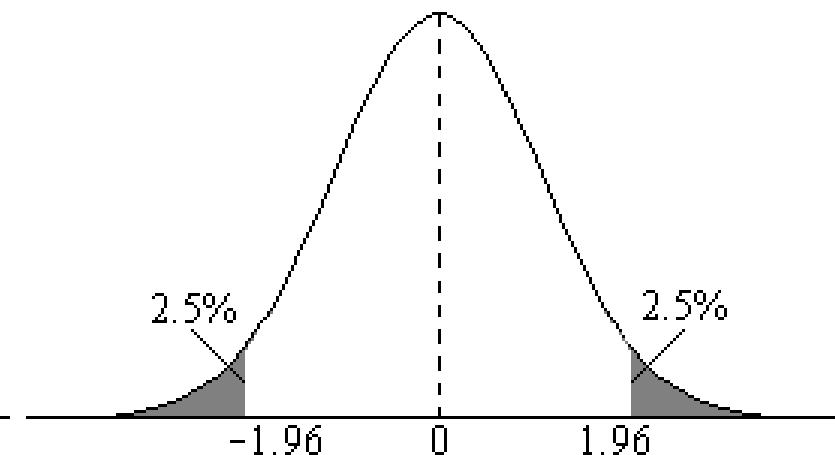
Hypothesis testing

Directional hypothesis



(a) One-tailed test

Non-Directional hypothesis



(b) Two-tailed test

General Rule: Use two-tailed test.

Only if direction is known from prior studies (justified reason), use one-tail test.

Criterion (α) for significance – 5% (0.05) for most behavioural studies (95 % CI)

If $p > 0.05 \rightarrow$ Accept the H_0

If $p \leq 0.05 \rightarrow$ Reject the H_0 & accept H_A

One-tailed vs two-tailed test

When is a one-tailed test NOT appropriate?

- Choosing a one-tailed test for the sole purpose of attaining significance is not appropriate.
- Choosing a one-tailed test after running a two-tailed test that failed to reject the null hypothesis is not appropriate, no matter how "close" to significant the two-tailed test was.
- Using statistical tests inappropriately can lead to invalid results that are not replicable and highly questionable—a steep price to pay to show significance in your results

Exercise effects on anxiety

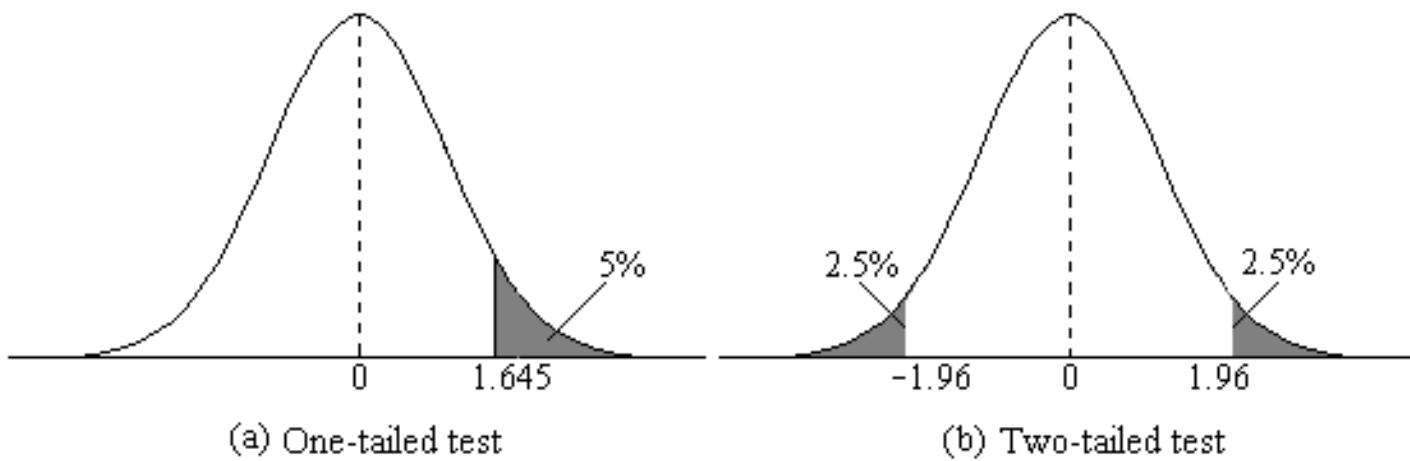
H_0 - Exercise has no effect on anxiety

H_1 - Exercise lowers anxiety

H_2 – Exercise increases anxiety?

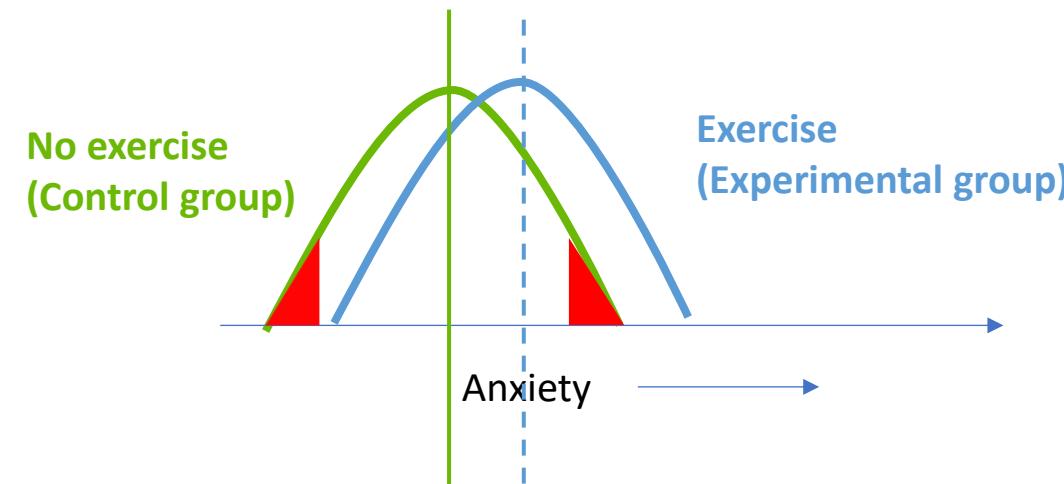
EXERCISE AND ANXIETY

Studies show that it is very effective at enhancing overall cognitive function.

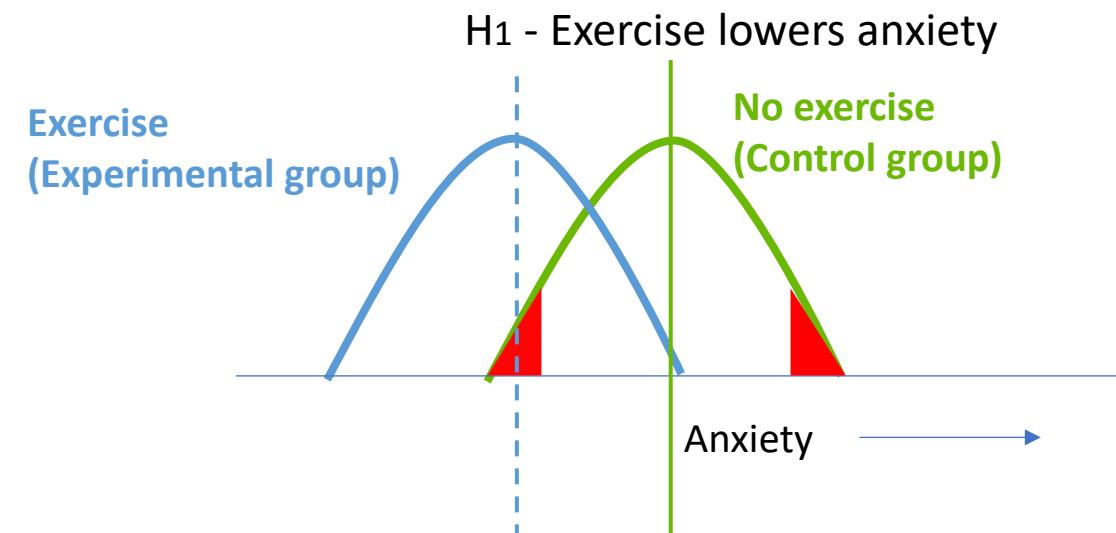


When $p > .05$, we retain the null hypothesis – there is less difference between the groups.

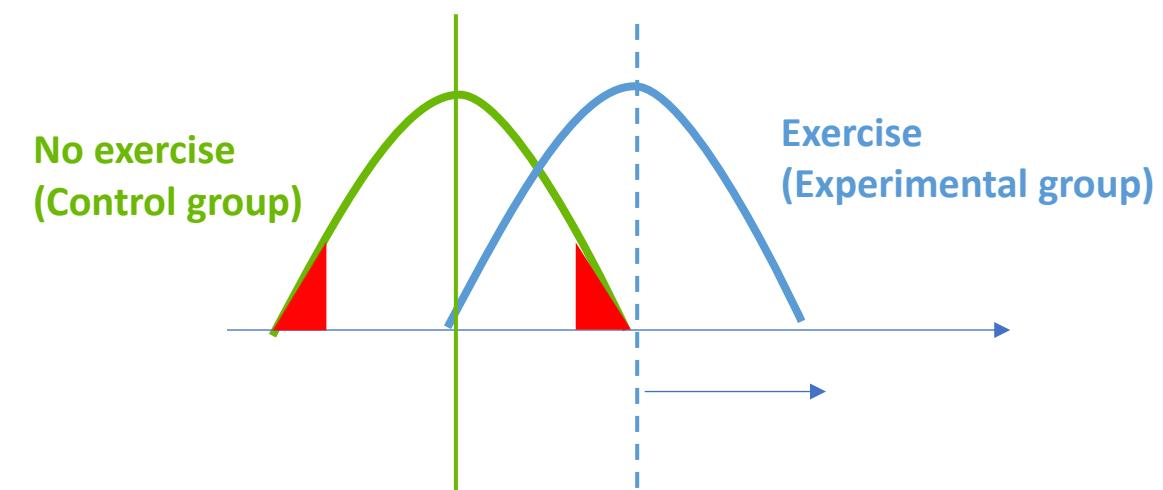
H_0 - Exercise has no effect on anxiety



When $p \leq .05$, we reject the null hypothesis - there is a 'significant' difference between the two groups.



H_2 - Exercise increases anxiety



Another Directional Hypothesis

You have a new drug to treat pain that is cheaper than the existing drug and you only want to confirm if the new drug is less effective than the existing drug

Whether the new drug is better than the existing drug does not matter.

H₀ - Null hypothesis – No difference between new drug and existing drug to treat pain

H₁ - Alternate hypothesis – Is the new drug less effective than the existing drug



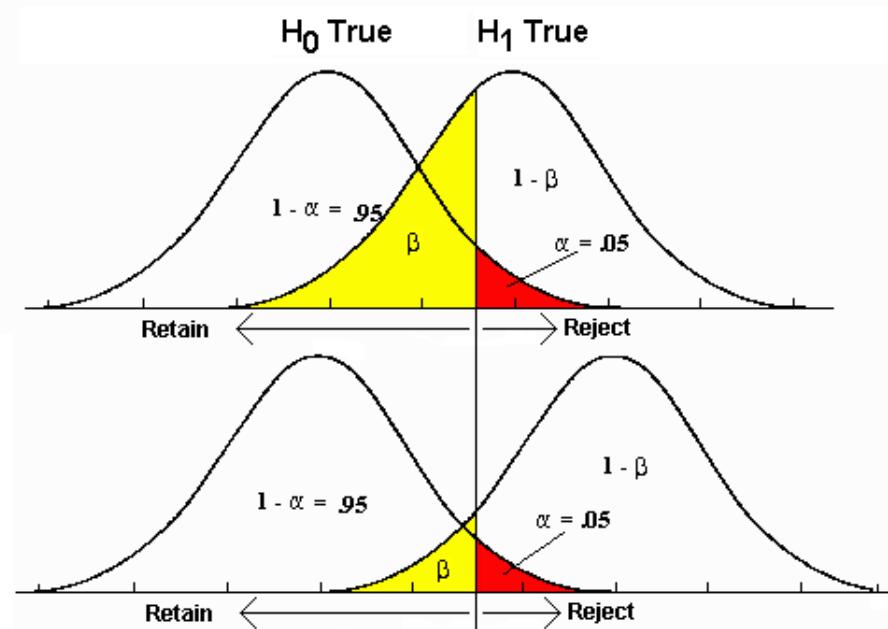
EXAMPLES

- H_0 , H_A , IV, DV, one or two tailed test?
- Smoking is injurious to the lungs
- Videogaming can lower attention span
- Does repetition in advertising improve sales?
- Air pollution is more fatal than COVID19
- Is there a difference in leadership style between men and women?

Ever wondered how and why statistics came into picture?

<https://nautil.us/how-eugenics-shaped-statistics-238014/>

Types of Errors in hypothesis testing



Decision from statistical tests	Reject Ho	Accept Ho
	Correct Decision Sensitivity/Power $1 - \beta$	Type 1 Error "False Positive" α
	Type 2 Error "False Negative" β	Correct Decision Specificity $1 - \alpha$

Fail to find a difference when there is one

Underreacting!

- Sample size is too small (high variability)
- Choosing one-tailed instead of two-tailed test
- Wrong statistical test

[Reality:]

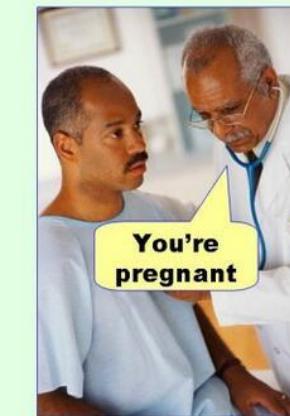
Ho False

Ho True

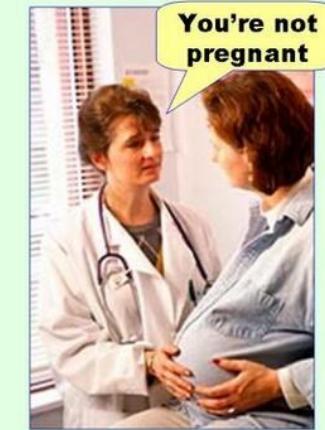
Observe difference when none exists

Overreacting!

Type I error
(false positive)

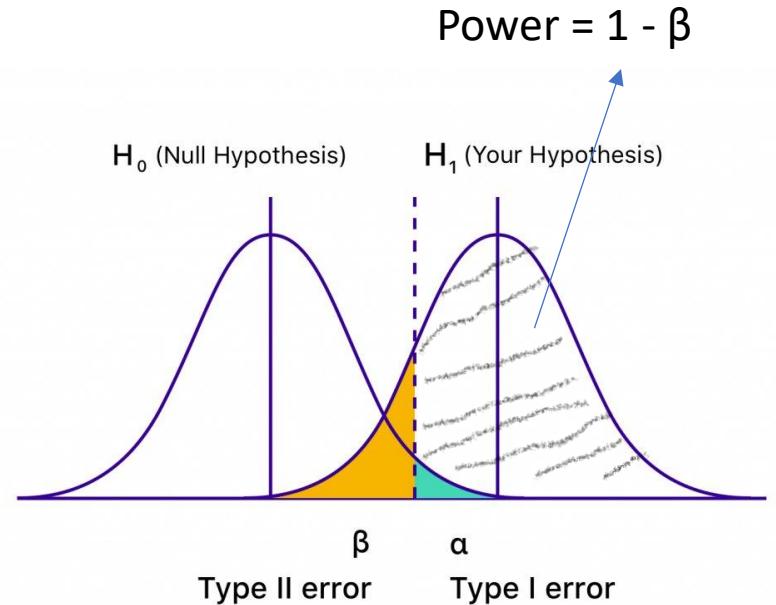


Type II error
(false negative)



Power

- Power - the probability that your test will find a statistically significant difference when such a difference actually exists.
- In other words, power is the probability that you will reject the null hypothesis when you should (and thus avoid a Type II error).
- It is generally accepted that power should be .8 or greater; that is, you should have an 80% or greater chance of finding a statistically significant difference when there is one.



Power

Power is calculated using statistical software. You need to know –

- What type of test you plan to use (e.g., independent t-test, paired t-test, ANOVA, correlation, regression, etc.)
- The alpha value or significance level you are using (usually 0.05 or 0.01)
- The expected effect size
- The sample size you are planning to use

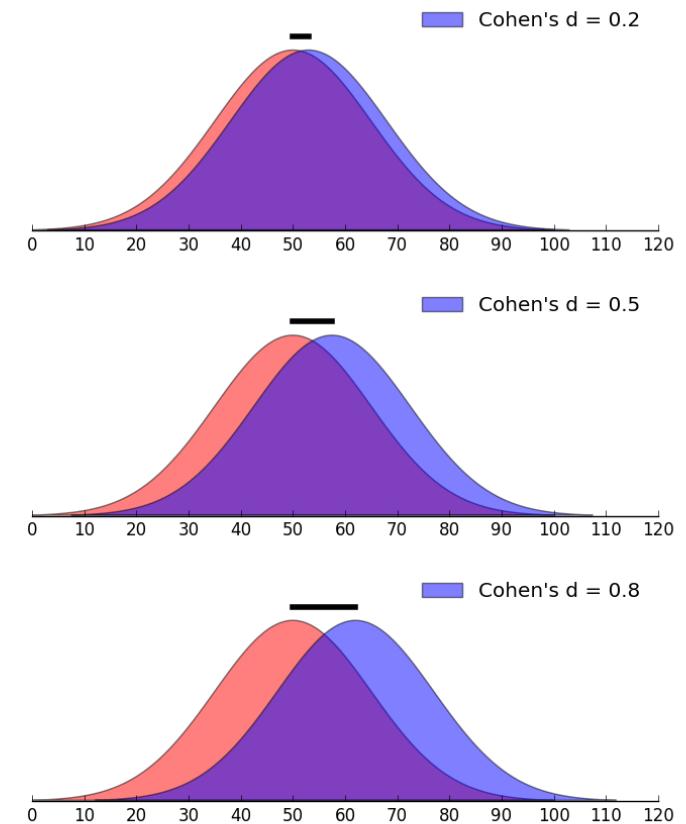
As your sample size increases, so does the power of your test.

- optimal sample means that you have collected more information -- which makes it easier to correctly reject the null hypothesis when you should.
- A power value is between 0 and 1.
- If the power is less than 0.8, you typically need to increase your sample size.

Effect Size

E.g. you evaluate the effect of a group discussion on student knowledge using pre and post tests on 1000 students. The mean score on the pre test was 83 out of 100 while the mean score on the post test was 84.

- What if you simply found a statistical difference by virtue of a large sample size (> 1000 or 10000)?
- If you calculate the effect size – you get a standard method to defining the importance of the statistical difference



Cohen's d effect size interpretation

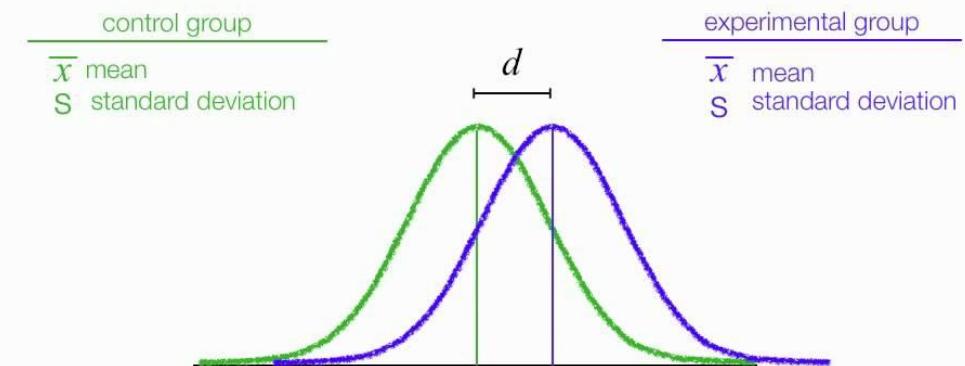
< 0.1 = trivial effect
0.1 - 0.3 = small effect
0.3 - 0.5 = moderate effect
> 0.5 = large difference effect

Effect Size

- Effect size is a quantitative measure of the *strength of a phenomenon*.
- Effect size emphasizes the **size** of the difference or relationship
- Examples:
 - the correlation between two variables (specifically r^2)
 - $r=.1$ weak, $r=.5$ moderate, $r=.7$ strong, $r=.9$ very strong
 - the regression coefficient in a regression (B_0, B_1, B_2)
 - Relative to model and field
 - the mean differences in t tests (use Cohen's D)
 - $d = .2$ is small; $r = .5$ is medium; $r = .8$ is large
 - The mean differences in ANOVA (use eta)
 - $.01$ is small, $.06$ medium, $.14$ large

$$\text{Cohen's Effect size} = \frac{\text{Mean}_{\text{treatment}} - \text{Mean}_{\text{control}}}{\text{Standard deviation}_{\text{pooled}}}$$

$$d = \frac{\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}}{s_{\text{pooled}}}$$



Basic formula for sample size - Continuous data

$$\text{Number of samples per group (n)} = \frac{2 \times (Z_{(1-\alpha/2)} + Z_\beta)^2 \times \sigma^2}{\Delta^2}$$

Where Δ = size of difference, minimal effect of interest

α = significance level (eg 0.05)

β = power, probability of detecting a significant result (typically 80%, 90%)

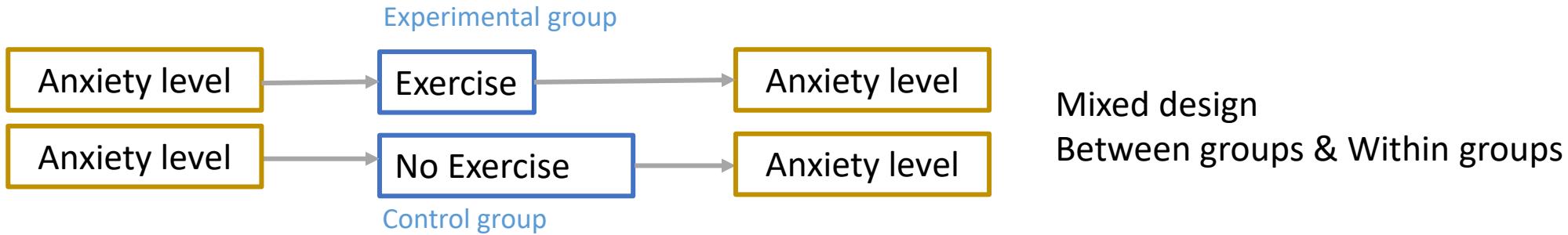
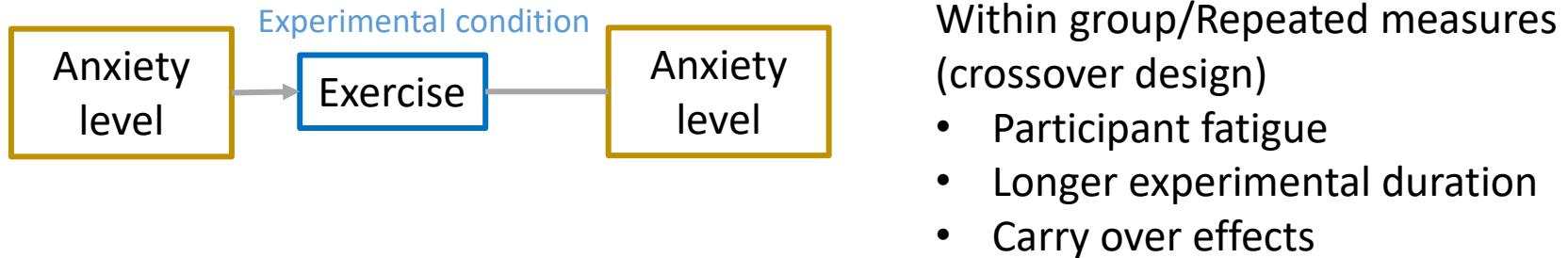
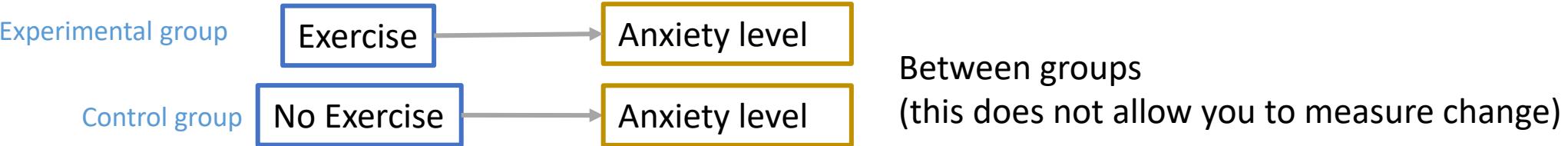
σ = SD of data

Z_p = points on normal distribution to give required power and significance

DV: Anxiety level

Do people who exercise have lower levels of anxiety?

IV: Exercise



Hypothesis Testing Practicals

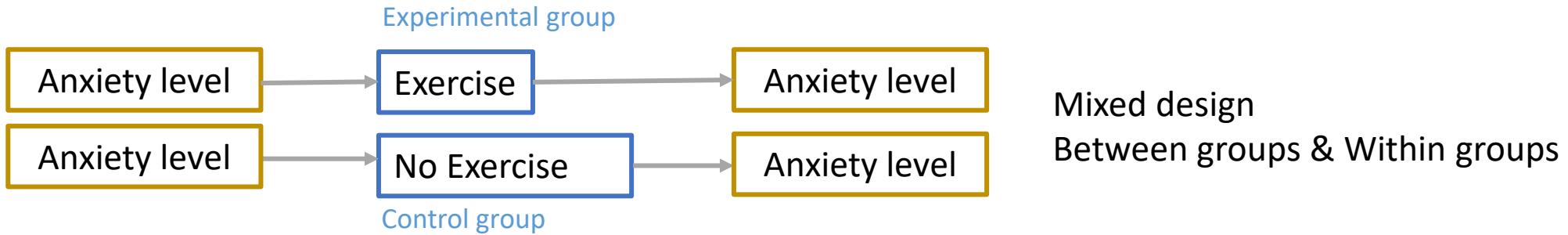
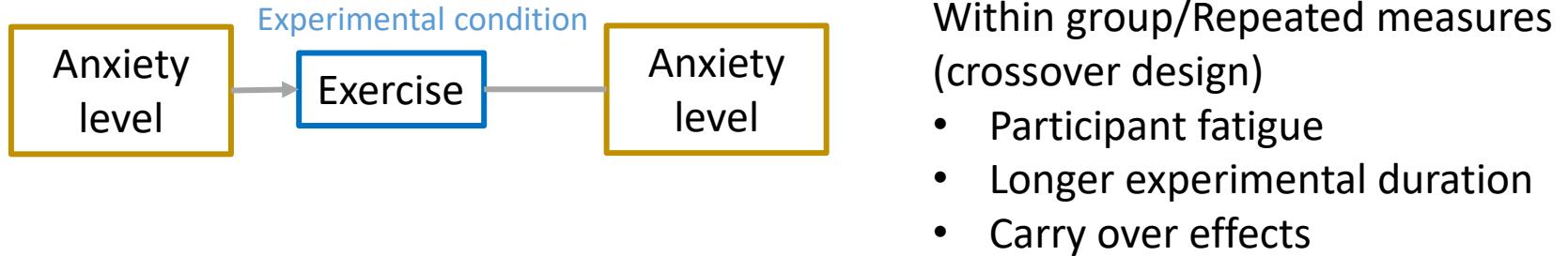
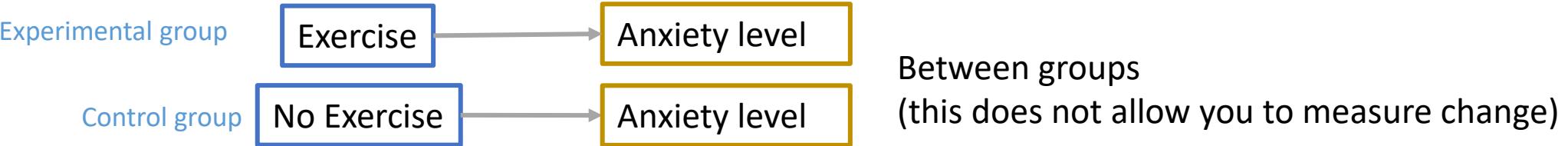
10/2/2022

Spring 2022

DV: Anxiety level

Do people who exercise have lower levels of anxiety?

IV: Exercise



DV: Anxiety level

IV: Exercise

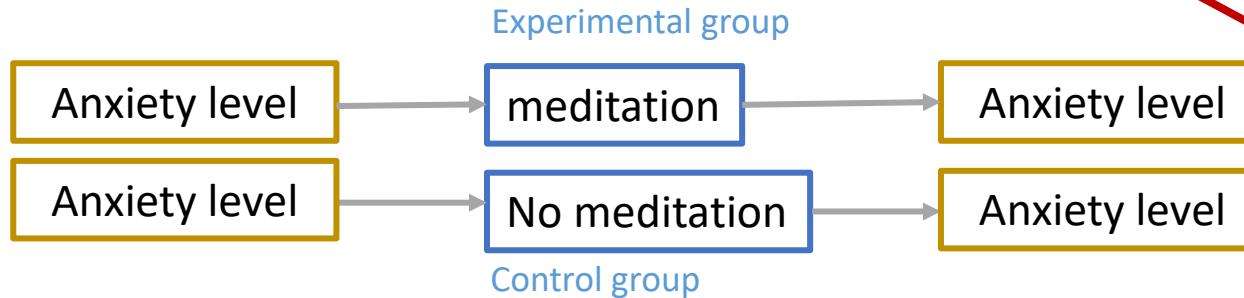
People who exercise have lower levels of anxiety



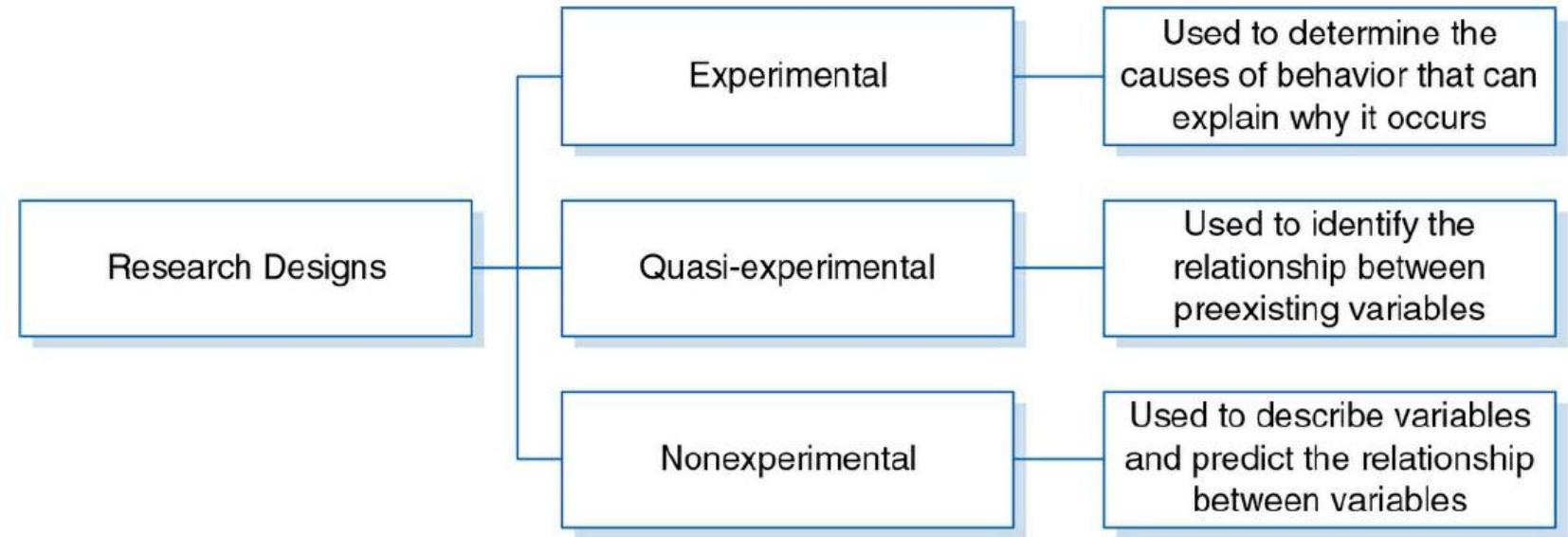
Within group/Repeated measures
(crossover design)

- Participant fatigue
- Longer experimental duration
- Carry over effects

Exercise lowers anxiety



Mixed design
Between groups & Within groups



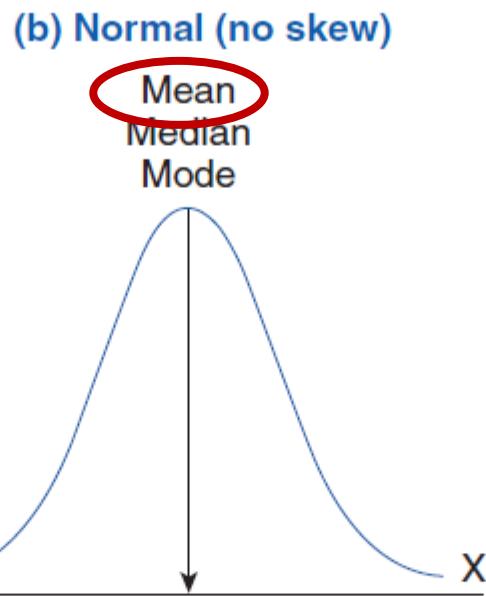
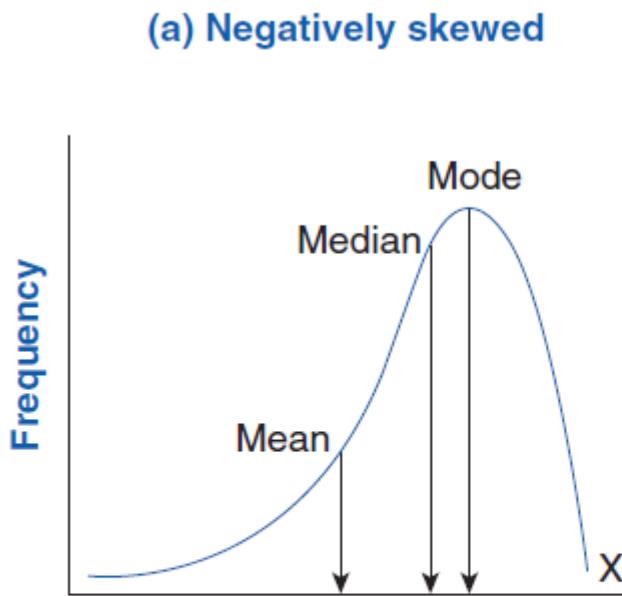
- Exercise lowers anxiety
- People who exercise have lower levels of anxiety
- A class teacher observes and records the behaviour of her students when they exercise and when they don't exercise

Anxiety levels

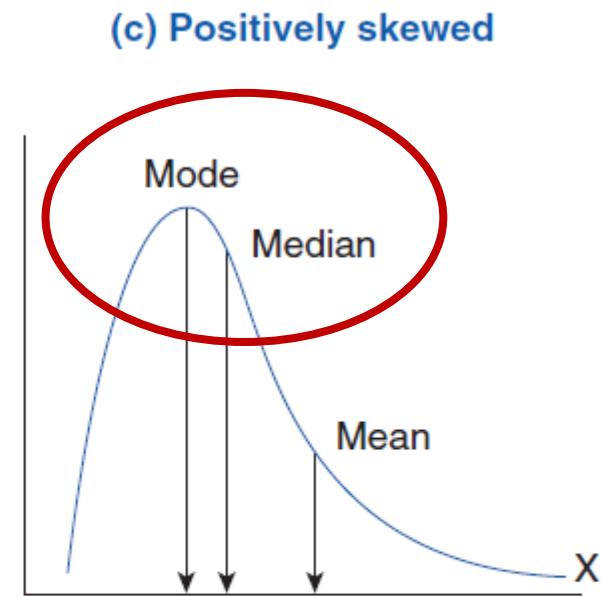
	Exercise	No -Exercise
	20	24
	23	35
	25	41
	30	21
	35	38
	29	23
	37	37
	24	44
	29	32
	31	33
	26	34
	28	42
Mean	28.08333	33.66667
SD	4.680782	7.261007



Normality?



The normal curve
represents a perfectly
symmetrical distribution



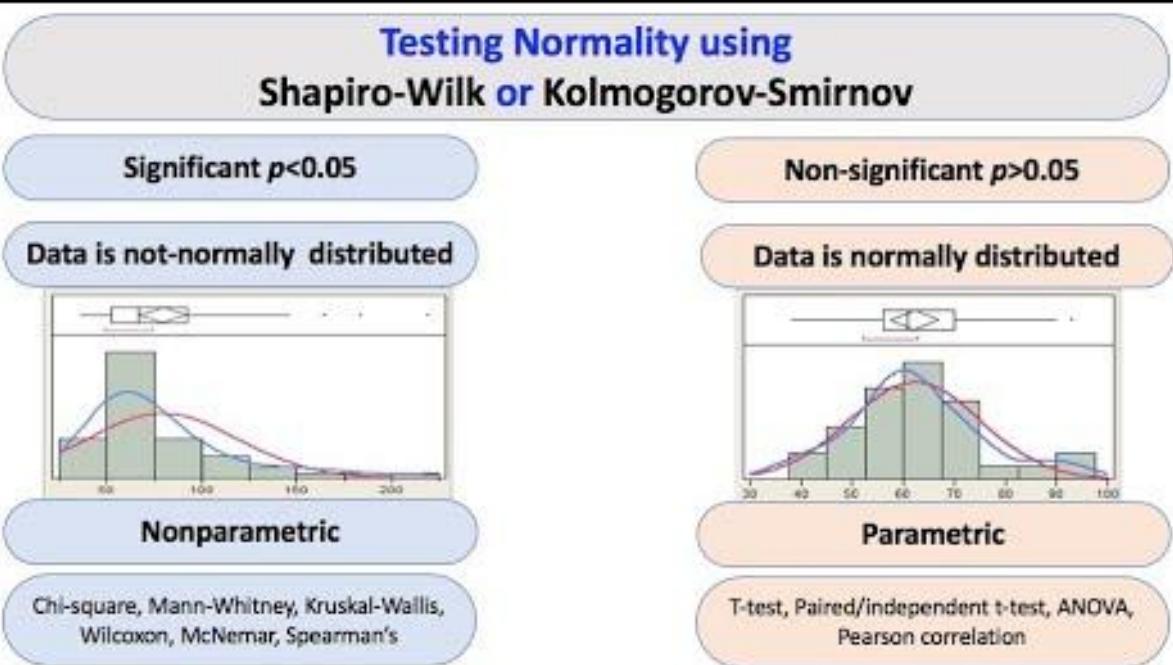
Kolmogorov–Smirnov test ($n \geq 50$)

OR

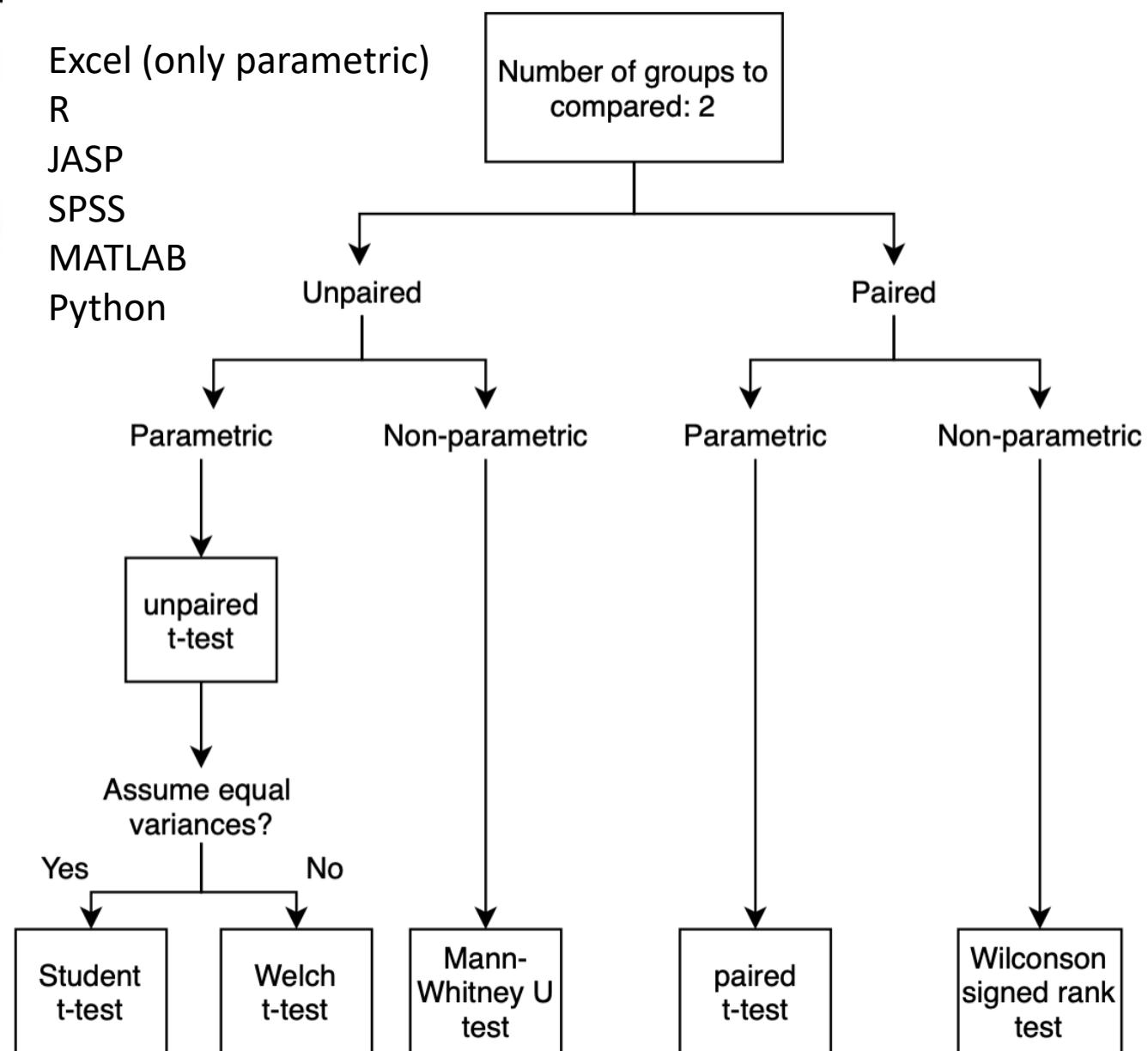
Shapiro–Wilk test ($n < 50$)

The null hypothesis for normality → data is normally distributed

Parametric vs non-parametric



Independent Sample t test



T-Test Example

People who exercise have lower levels of anxiety



Anxiety levels

	Exercise	No -Exercise
	20	24
	23	35
	25	41
	30	21
	35	38
	29	23
	37	37
	24	44
	29	32
	31	33
	26	34
	28	42
Mean	28.08333	33.66667
SD	4.680782	7.261007

	Exercise	No -Exercise
Mean	28.08333333	33.66666667
Variance	23.90151515	57.51515152
Observations	12	12
Pooled Variance	40.70833333	
Hypothesized Mean Diff	0	
df	22	
t Stat	-2.143519905	
P(T<=t) one-tail	0.021690748	
t Critical one-tail	1.717144374	
P(T<=t) two-tail	0.043381495	
t Critical two-tail	2.073873068	

t-value

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

\bar{x}_1 : Mean value of the first group

\bar{x}_2 : Mean value of the second group

n_1 : Size of the first group

n_2 : Size of the second group

s_1 : Standard deviation of the first group

s_2 : Standard deviation of the second group

$$\text{Cohen's Effect size} = \frac{(\text{Mean}_{\text{treatment}} - \text{Mean}_{\text{control}})}{\text{Standard deviation pooled}}$$

Cohen's d effect size interpretation

< 0.1 = trivial effect

0.1 - 0.3 = small effect

0.3 - 0.5 = moderate effect

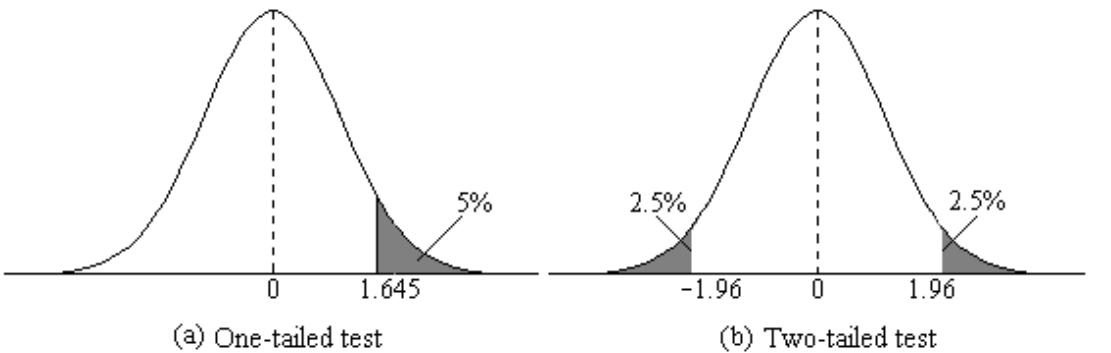
> 0.5 = large difference effect

$$\text{Cohen's d} = (33.66 - 28.083) / 6.107782 = 0.913097$$

$$t(df=22) = -2.14, p=0.04, d = 0.9$$

Critical value $\alpha = 0.05$

df = 22



(a) One-tailed test

(b) Two-tailed test

Statistic	df	Explanation
ANOVA: Mean Sum of Squares Within (MSW)	N - k	N: total # of all data points k: # of groups
ANOVA: Mean Sum of Squares Between (MSB)	k - 1	
χ^2	n - 1	n: Sample Size
χ^2 test for Goodness of Fit	n - 1	k: # of categories
χ^2 test for Independence	(r - 1)(c - 1)	r: # of rows, c: #columns
χ^2 test for Variance	n - 1	n: Sample Size
F	$n_1 - 1$ and $n_2 - 1$	n_1 and n_2 : Sizes of the 2 Samples
t	n - 1	n: Sample Size
1-Sample t-test, and Paired t-test	n - 1	
2 (Independent)-Sample t-test	$n_1 + n_2 - 2$	n_1 and n_2 : Sizes of the 2 Samples

Table T Critical Values of the t Distribution

df	One-Tail = .4 Two-Tail = .8	.25 .5	.1 .2	.05 .1	.025 .05	.01 .02	.005 .01	.0025 .005	.001 .002	.0005 .001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Source: From Biometrika Tables for Statisticians, Vol. 1, Third Edition, edited by E. S. Pearson and H. O. Hartley, 1966, p. 146.
Reprinted by permission of the Biometrika Trustees.

Independent Samples T-Test

t-value

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

\bar{x}_1 : Mean value of the first group

\bar{x}_2 : Mean value of the second group

n_1 : Size of the first group

n_2 : Size of the second group

s_1 : Standard deviation of the first group

s_2 : Standard deviation of the second group

Cohen's Effect size = $(\text{Mean}_{\text{treatment}} - \text{Mean}_{\text{control}})$

Standard deviation pooled

For equal sample size

$$df = (n_1 + n_2 - 2)$$

For unequal sample size

$$\text{degrees of freedom, } df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2} \right)^2}$$

Paired Samples T-Test

Paired Samples t-tests	
$t = \frac{\Sigma(X_{\text{pre}} - X_{\text{post}})}{\text{SE}_{\text{diff}}}$	
$t = \frac{\bar{d}}{\sqrt{s^2/n}}$	

Independent Samples T-Test

t-value

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

\bar{x}_1 : Mean value of the first group

\bar{x}_2 : Mean value of the second group

n_1 : Size of the first group

n_2 : Size of the second group

s_1 : Standard deviation of the first group

s_2 : Standard deviation of the second group

t-Test: Paired Two Sample for Means

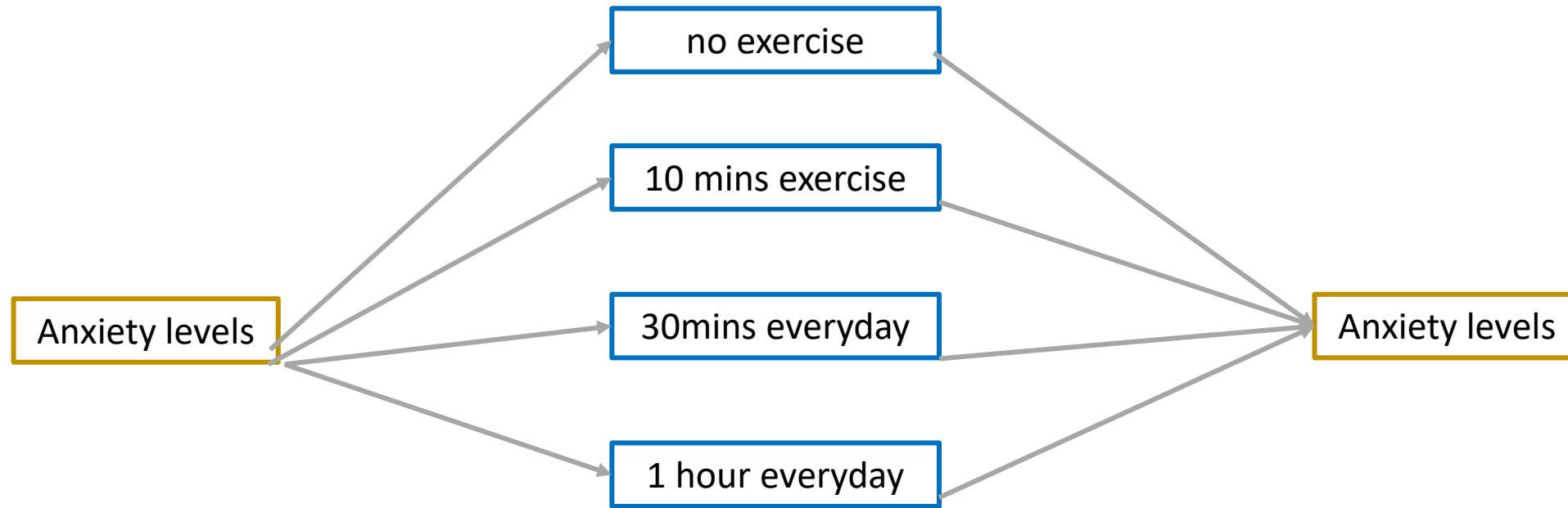
	Variable 1	Variable 2
Mean	28.0833333	33.6666667
Variance	23.9015152	57.5151515
Observations	12	12
Pearson Corr	0.06701871	
Hypothesized	0	
df	11	
t Stat	-2.2120964	
P(T<=t) one-t	0.02451926	
t Critical one	1.79588482	
P(T<=t) two-t	0.04903853	
t Critical two	2.20098516	

Cohen's Effect size = $\frac{\text{Mean}_{\text{difference}}}{\text{SD}_{\text{difference}}}$

DV: Anxiety level

IV: Exercise

IV – 4 levels



Factor = Independent variable

2 Independent Variables - 2 levels each

Exercise – exercise vs control

Time of Day – morning vs evening

Two factorial design

Exercise-morning	Control-morning
Exercise-evening	Control-evening

2x2 factorial design

2 Independent Variables – different levels

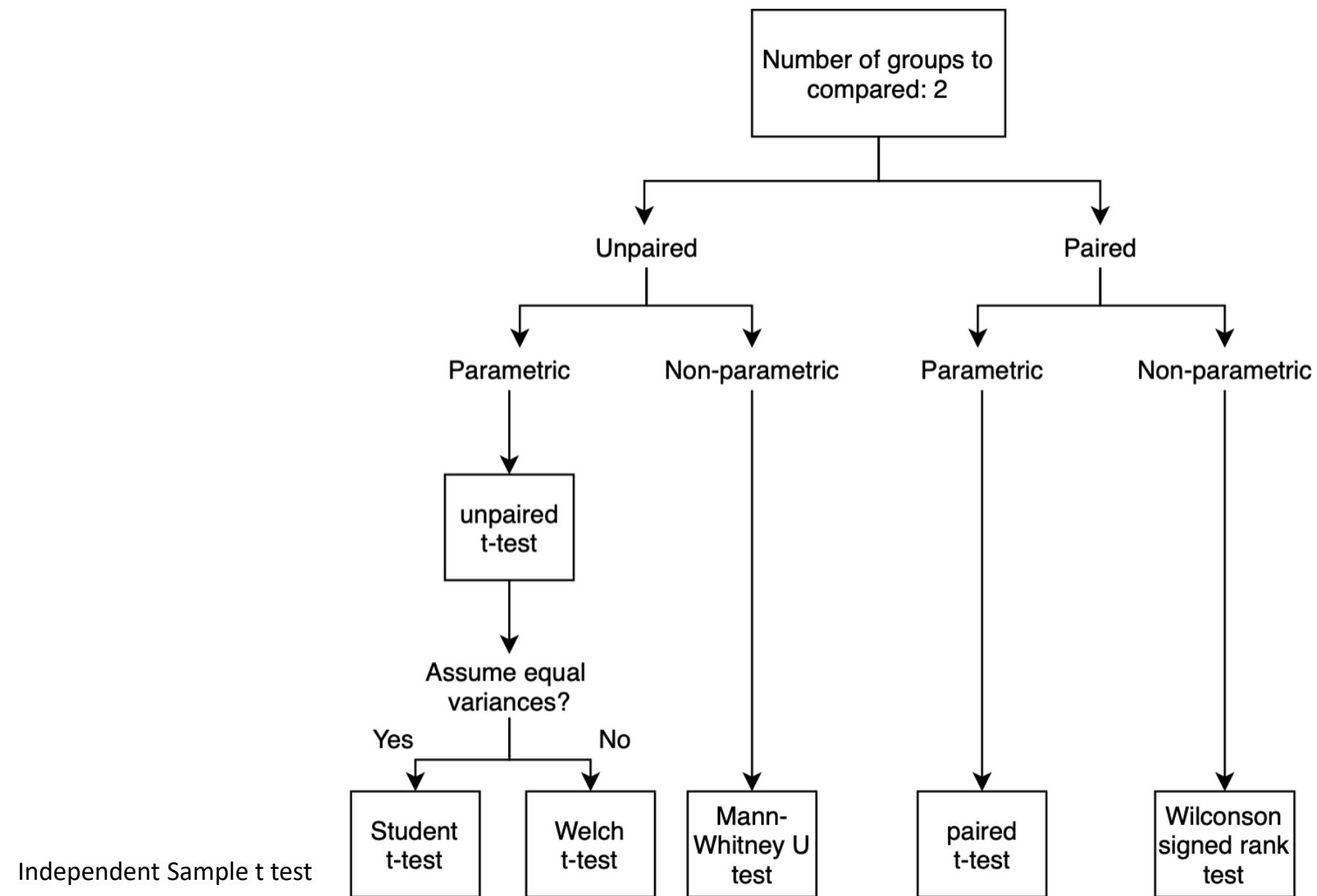
Exercise – 30mins, 1 hour, 2 hours

Time of Day – morning vs evening

Two factorial design

30 mins-morning	1 hr-morning	2 hrs - morning
30 mins-evening	1 hr-evening	2 hrs - evening

3x2 factorial design



The problem of multiple comparisons

BRSM

A drug for memory improvement

- CEO: I think this drug can improve memory
- Scientists: oops, $p>0.05$
- CEO: Hmm... Not to despair. Reanalyze the data and see if it improves concentration
- Scientists: no luck, $p > 0.05$
- CEO: Ok, here's a list of different things to try, 20 of them to be exact.
- Scientists: yay, one of them is $p < 0.05!!$ It seems to improve executive control.
- CEO: See? I told you, now we go raise 100 crores. We call it the "miracle executive control drug"!

In the earlier
example,
 $\alpha = 0.05$

- 20 different comparisons made with the same data
- 1 of them yielded $p < 0.05$
- What is the problem with this?



Another example

- We have a coin, need to test if it is fair
- We toss it 10 times. If we get 9 heads and 1 tail, we might think that it is an unfair coin as the chances of getting it are very low for a fair coin.
- Now, clone that coin 19 times.
- Toss each one of those 20 coins 10 times each. If you get 9 heads and 1 tail for one of them, will you be confident that the coin is unfair?
- No, because with 20 coins, there is a much higher chance of obtaining 9H + 1T for a fair coin by random chance.
- This is the same issue that was present in the drug company example earlier.

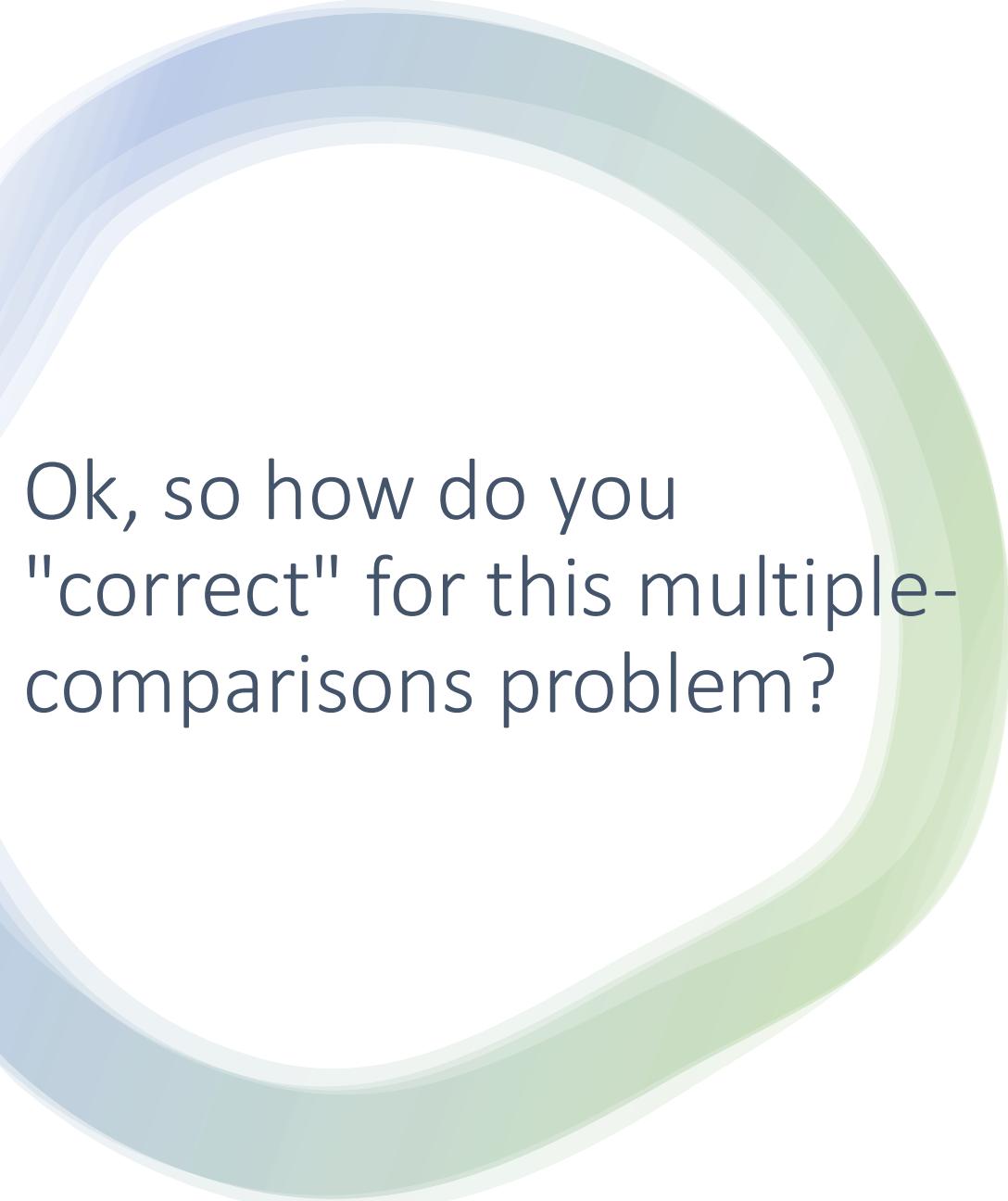
Type I and type II errors

- Type I: When H_0 is rejected even though it is true (a false positive from the perspective of H_1)
- Type II: When H_0 is accepted even though it is false (a false negative from the perspective of H_1)

Error types	Actual fact		
	H_0 true	H_0 false	
Statistical inference	H_0 true	Correct	Type II error (β)
	H_0 false	Type I error (α)	Correct

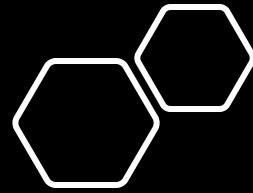
Alpha is a criterion set by us to say that this is the $P(\text{false positive})$ that we can live with under the null hypothesis – for one test.

The more tests you do on the data, the more likely you are to mistakenly claim an effect when there is none --> the multiple comparisons problem.

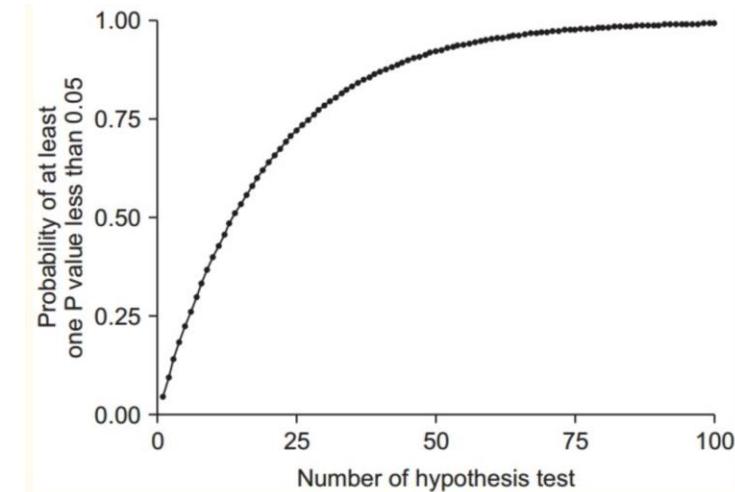


Ok, so how do you
"correct" for this multiple-
comparisons problem?

- Intuition: if you continue to reject the null hypothesis at alpha = 0.05, then when you have multiple tests, you are more likely to endorse effects (i.e., reject the null) when there are none (i.e., when the null is true).
- So what if you adjust the p values you obtain from all your tests such that you reduce the chance that you falsely claim an effect?
- Should you increase or decrease the p values of your multiple tests to reduce the chance of a false positive?
- Ans: increase! So if you obtain a p value of 0.04 for one of many tests you do, and if you somehow adjust that p value to 0.06, you don't falsely claim an effect.



Type 1 error rate
goes up with the
number of tests



$$\text{Inflated } \alpha = 1 - (1 - \alpha)^N, \quad N = \text{number of hypotheses tested}$$

p-value adjustments

- These corrections either control the Family-Wise Error Rate (FWER)
OR the False Discovery Rate (FDR)

Assuming m
different tests

R of them are statistically significant

$m-R$ are not statistically significant

Problem: how many of R are false positives (FP, type 1 errors) and how many are true positives (TP)?

	Fail to reject null hypothesis ($p > 0.05$)	Reject null hypothesis ($p \leq 0.05$)	Total Hypotheses
Null hypothesis is true	TN (True Negative)	FP (Type I error, False Positive)	m_0
Null hypothesis is false	FN (Type II error, False Negative)	TP (True Positive)	$m - m_0$
Total Hypotheses	$m - R$	R	m

Error table for a family of m tests

Family-Wise Error Rate (FWER) vs False Discovery Rate (FDR)

- FWER = Probability of falsely rejecting even one null hypothesis = $P(FP \geq 1)$ across all m tests.
- FDR = Expected proportion of false discoveries among all discoveries = $E[FP/R]$. That is, you take all the rejected hypotheses and find how many of those "discoveries" were false. The expected proportion of this is the FDR.
- FWER = FDR if all null hypotheses are true and so are connected to each other but they are different and the difference is subtle.

Family-Wise Error Rate (FWER)

Assume we have 3 null hypotheses, all of which are true (drug does not improve memory, does not improve concentration, does not improve executive control)

Alpha = 0.05 (our criterion for type 1 errors that we can live with)

Now, for the family of 3 tests, what is the type 1 error probability?

The prob of rejecting any given hypothesis erroneously = 0.05, prob of accepting the null = 0.95

The prob of rejecting any one of three tests erroneously = $1 - 0.95^3 = 0.142$

Controlling for this FWER involves reducing this 0.142 to 0.05 as you would expect for a single test

Bonferroni correction: a simple but very conservative way to control for FWER

- If you have 20 tests, adjust the alpha by dividing by the total number of tests (I.e., each individual test now has to pass a more stringent criterion such that the family-wise type 1 error remains at 0.05)
 - e.g. 50 t-tests. Adjusted alpha = $0.05/50 = 0.001$
 - So any given test of the 50 different tests should be considered significant only if $p < 0.001$ and not $p < 0.05$ so that the overall family-wise type 1 error is at 0.05

$$\text{Adjusted alpha } (\alpha) = \alpha / k \text{ (number of hypothesis tested)}$$



The problems with Bonferroni correction

- Too stringent: The side-effect is that it also increases type 2 error (i.e., we miss out on true effects due to the stringent criterion)
- Assumes independent tests but in many cases, our tests are not independent (e.g. in neuroscience, when testing different brain regions, the regions are not independent, they have correlated activity, etc)
- The correction does not depend on the structure of the data and instead only on the number of tests. For any given true effect, you will change the ability to find that effect drastically by simply changing the number of tests you do.
- **Holm correction:** a sequential procedure that still controls for FWER but does not increase type 2 error as much as Bonferroni does.
- Do you care about not making ANY false positives? Then controlling for FWER is appropriate.
- In many domains, you can live with some false positives (such as in genomics), there the more appropriate quantity to control for is FDR.

Controlling for False Discovery Rate (FDR) - Benjamini-Hochberg procedure

$$P_{(1)} \dots P_{(m)}$$

Benjamini-Hochberg critical value = $(i / m) \cdot Q$

- Step 1: Rank the p values from smallest to largest
- Step 2: Compare against the B-H critical value.
- i = rank, m = number of tests, Q = chosen FDR
- Step 3: The largest p value that is < the critical value is significant and so are all the p values in your list that are less than this p value

Controlling for False Discovery Rate (FDR) - Benjamini-Hochberg procedure

B&H FDR Example

Controlling the FDR at $\delta = 0.05$

Rank (j)	P-value	$(j/m) \times \delta$	Reject H_0 ?
1	0.0008	0.005	1
2	0.009	0.010	1
3	0.165	0.015	0
4	0.205	0.020	0
5	0.396	0.025	0
6	0.450	0.030	0
7	0.641	0.035	0
8	0.781	0.040	0
9	0.900	0.045	0
10	0.993	0.050	0

The independence assumption

Both Bonferroni (for controlling FWER) and Benjamini-Hochberg (for controlling FDR) assume independent tests.

This assumption however is often not true.

Therefore, you might want to control for FWER/FDR using a "non-parametric" technique where you use the structure of your data.

This is done through permutation tests (you use randomization procedures on your data to create null distributions).

The general idea behind using permutation-based simulations for controlling FWER

- m tests --> m "uncorrected" p-values
- Randomize the experiment 1000+ times
- Conduct all m tests each time – we know that all m null hypotheses are true in the randomizations (by construction)
- Determine from the simulations, the target p-value threshold for which you get 5% significant results across the 1000+ simulations -- i.e., calculate the proportion of simulations in which you get at least one significant test.
- Use this threshold now as the criterion for the uncorrected p-values in the true data, all those below the criterion are significant.

A final note
before we
look at
permutation-
based
procedures in
detail

The answer will be similar to parametric methods if the m tests are independent.

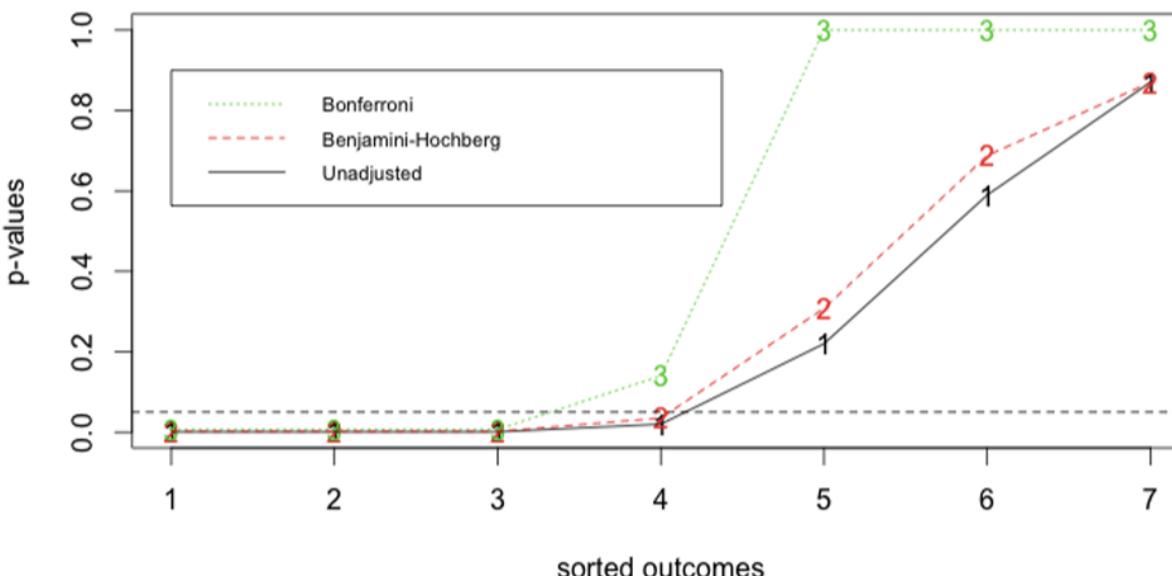
The simulation-based threshold will be less stringent than the parametric versions as more of the m tests become less independent from each other.

You should try this out with your data and get an intuitive grasp.

1. The following array p consists of the observed significance values for multiple correlation tests.

```
p = [0.0050 0.0010 0.0100 0.0005 0.0009 0.0400 0.0560 0.0500  
0.0480 0.0130 0.0370 0.0430 0.0020 0.0250 0.1100 0.0700 0.0800]
```

Apply both Bonferroni and Benjamini-Hochberg correction and create a graph as shown below with the observed/unadjusted p-values, Bonferroni corrected and BH-corrected ones for all the tests sorted and comment on the results (ensure you also plot the black dashed line that represents an alpha level of .05). The relevant function is **p.adjust** in R. The Bonferroni method is known to be a more conservative approach. Do the results of the correction support that?



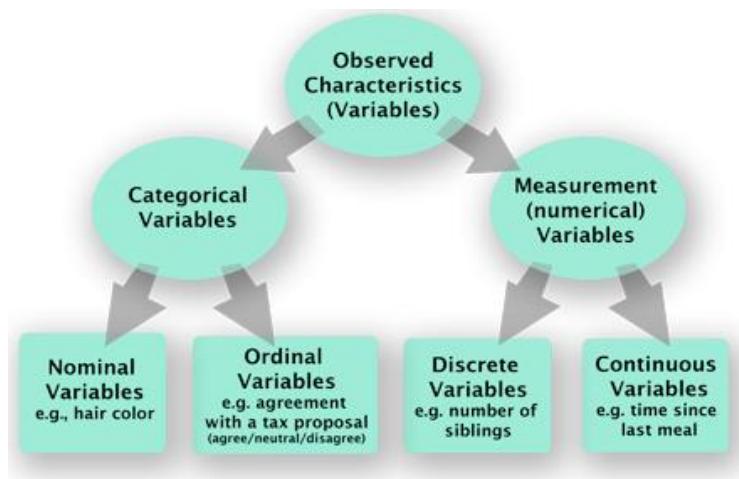
Non-parametric tests

How to deal with Categorical data?

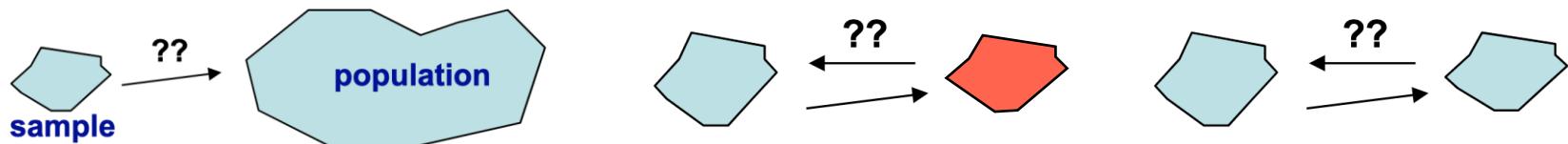
How to deal with cases where parametric assumptions are violated?

Selecting a statistical test

- Different tests are used according to the level of measurement:
 - Interval
 - Ordinal
 - Categorical
- Parametric vs non-parametric (makes no assumption on the population distribution or sample size) assumptions



Selecting a statistical test

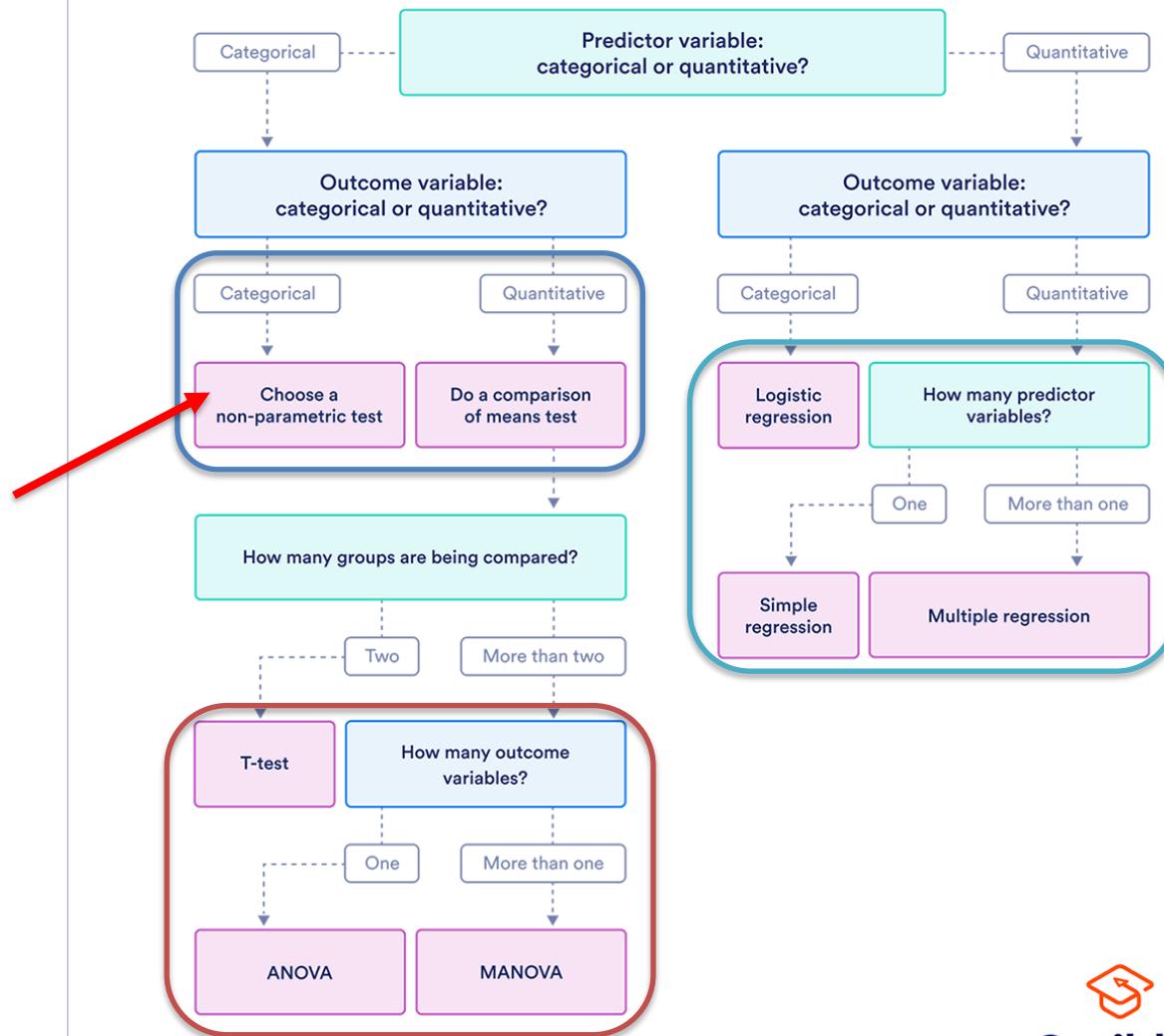


- Different tests are used for varying amount of groups/conditions:
 - two samples
 - > two samples

- Different tests are used for related versus unrelated designs:
 - unrelated samples = between subjects designs
 - related samples = within subjects designs & matched pairs

Choosing a statistical test

This flowchart helps you choose among parametric tests



	Predictor variable	Outcome variable	Research question example
Simple linear regression	<ul style="list-style-type: none"> Continuous 1 predictor 	<ul style="list-style-type: none"> Continuous 1 outcome 	What is the effect of income on longevity?
Multiple linear regression	<ul style="list-style-type: none"> Continuous 2 or more predictors 	<ul style="list-style-type: none"> Continuous 1 outcome 	What is the effect of income and minutes of exercise per day on longevity?
Logistic regression	<ul style="list-style-type: none"> Continuous 	<ul style="list-style-type: none"> Binary 	What is the effect of drug dosage on the survival of a test subject?

Regression tests

Regression tests look for **cause-and-effect** relationships. They can be used to estimate the effect of one or more continuous variables on another variable.

Predictor variable	Outcome variable	Research question example
Paired t-test	<ul style="list-style-type: none"> Categorical 1 predictor 	<ul style="list-style-type: none"> Quantitative groups come from the same population <p>What is the effect of two different test prep programs on the average exam scores for students from the same class?</p>
Independent t-test	<ul style="list-style-type: none"> Categorical 1 predictor 	<ul style="list-style-type: none"> Quantitative groups come from different populations <p>What is the difference in average exam scores for students from two different schools?</p>
ANOVA	<ul style="list-style-type: none"> Categorical 1 or more predictor 	<ul style="list-style-type: none"> Quantitative 1 outcome <p>What is the difference in average pain levels among post-surgical patients given three different painkillers?</p>
MANOVA	<ul style="list-style-type: none"> Categorical 1 or more predictor 	<ul style="list-style-type: none"> Quantitative 2 or more outcome <p>What is the effect of flower species on petal length, petal width, and stem length?</p>

Comparison tests

Comparison tests look for **differences among group means**. They can be used to test the effect of a categorical variable on the mean value of some other characteristic.

T-tests are used when comparing the means of precisely two groups (e.g., the average heights of men and women). ANOVA and MANOVA tests are used when comparing the means of more than two groups (e.g., the average heights of children, teenagers, and adults).

Correlation tests

Correlation tests **check whether variables are related** without hypothesizing a cause-and-effect relationship.

These can be used to test whether two variables you want to use in (for example) a multiple regression test are autocorrelated.

	Variables	Research question example
Pearson's r	• 2 continuous variables	How are latitude and temperature related?

Selecting a statistical test

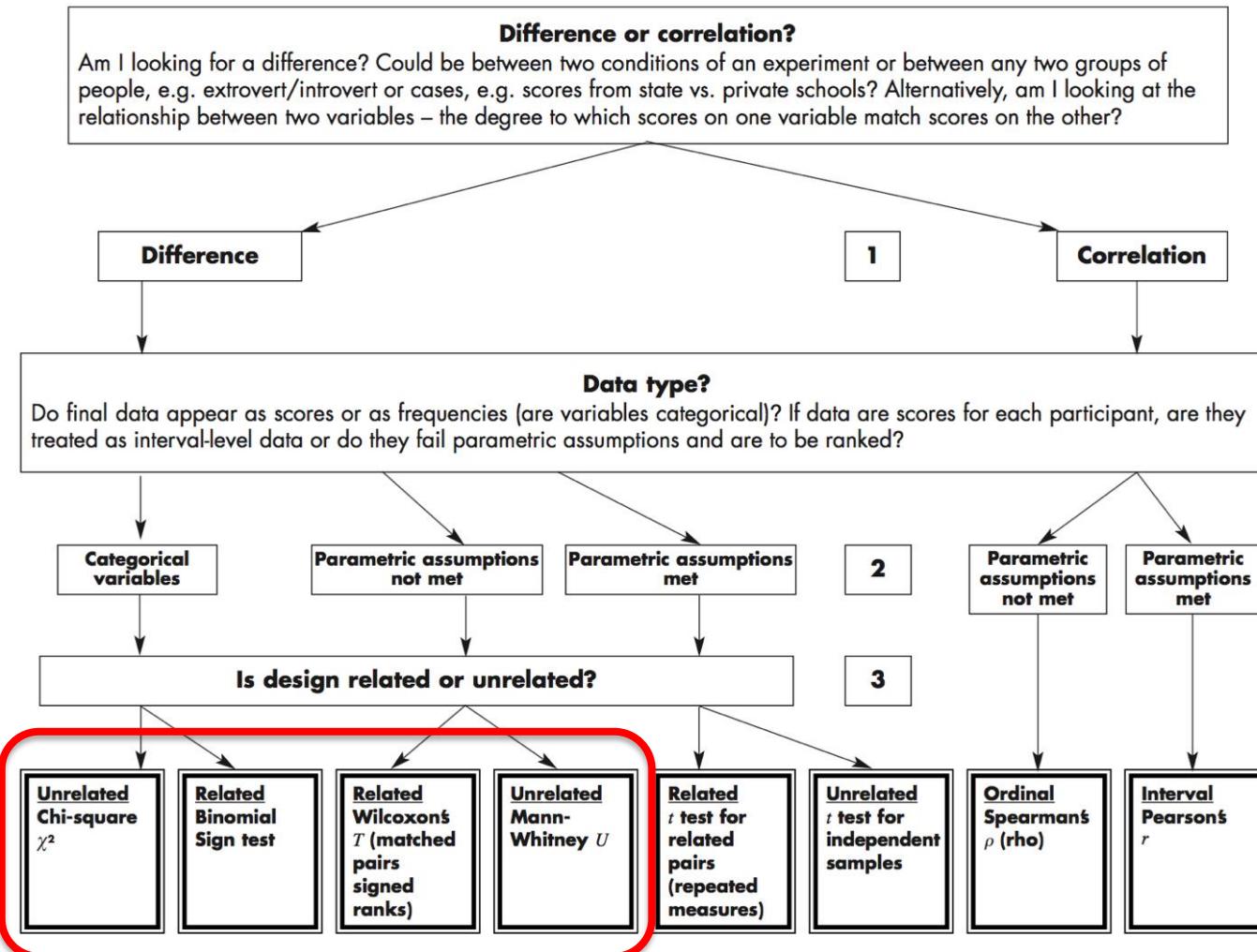


Figure 23.1 Choosing an appropriate two-sample test.

	Predictor variable	Outcome variable	Use in place of...
Spearman's r	<ul style="list-style-type: none"> Quantitative 	<ul style="list-style-type: none"> Quantitative 	Pearson's r
Chi square test of independence	<ul style="list-style-type: none"> Categorical 	<ul style="list-style-type: none"> Categorical 	Pearson's r
Sign test	<ul style="list-style-type: none"> Categorical 	<ul style="list-style-type: none"> Quantitative 	One-sample t -test
Kruskal-Wallis H	<ul style="list-style-type: none"> Categorical 3 or more groups 	<ul style="list-style-type: none"> Quantitative 	ANOVA
ANOSIM	<ul style="list-style-type: none"> Categorical 3 or more groups 	<ul style="list-style-type: none"> Quantitative 2 or more outcome variables 	MANOVA
Wilcoxon Rank-Sum test	<ul style="list-style-type: none"> Categorical 2 groups 	<ul style="list-style-type: none"> Quantitative groups come from 	Independent t-test
Wilcoxon Signed-rank test	<ul style="list-style-type: none"> Categorical 2 groups 	<ul style="list-style-type: none"> Quantitative groups come from the same population 	Paired t-test

Choosing a nonparametric test

Non-parametric tests don't make as many assumptions about the data, and are useful when one or more of the common statistical assumptions are violated. However, the inferences they make aren't as strong as with parametric tests.

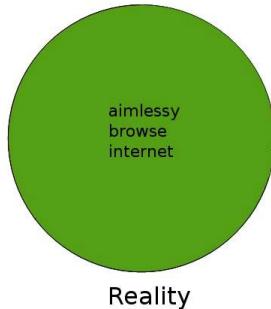
Different types of tests

Test type	Between subjects designs (Independent samples)	Within subject designs (repeated measures/matched pairs)
Non-parametric (for categorical data)	<i>Chi-square test</i>	<i>The binomial sign test</i>
Non-parametric (for ordinal data)	<i>Mann-Whitney U</i>	<i>Wilcoxon Signed-Rank Test</i> <i>The binomial sign test</i>
Parametric	<i>Unrelated t-test (level of data: interval)</i>	<i>Related t-test (level of data: interval)</i>

Chi-Square Test

Theoretical
categorical distribution
vs
I
Observed
categorical distribution

Weekend in college



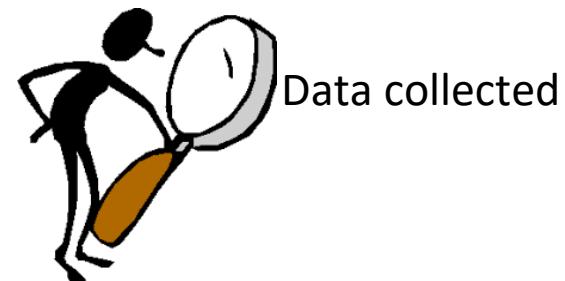
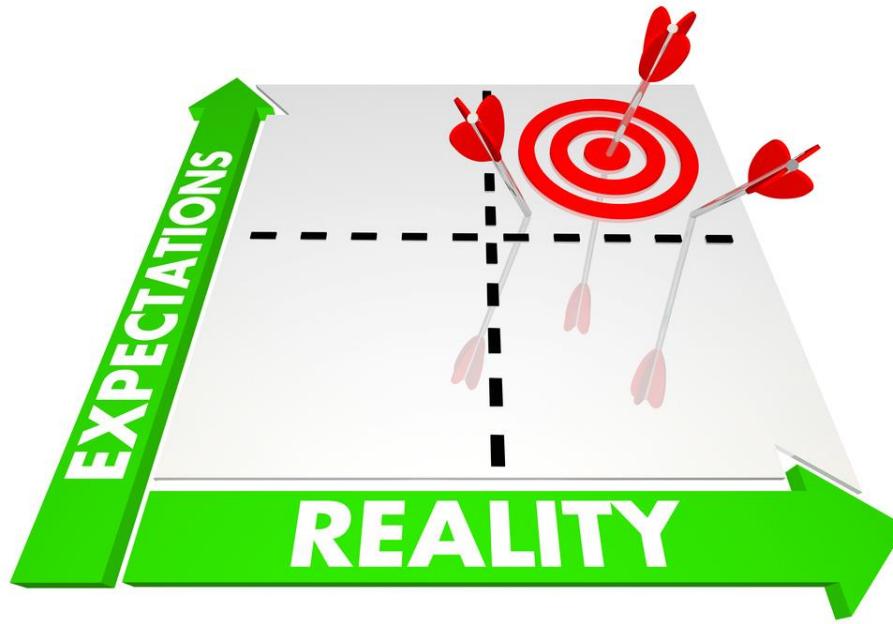
preference for one brand



Chi-Square test

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

H_0
 H_A



Chi-Square Test - Applications

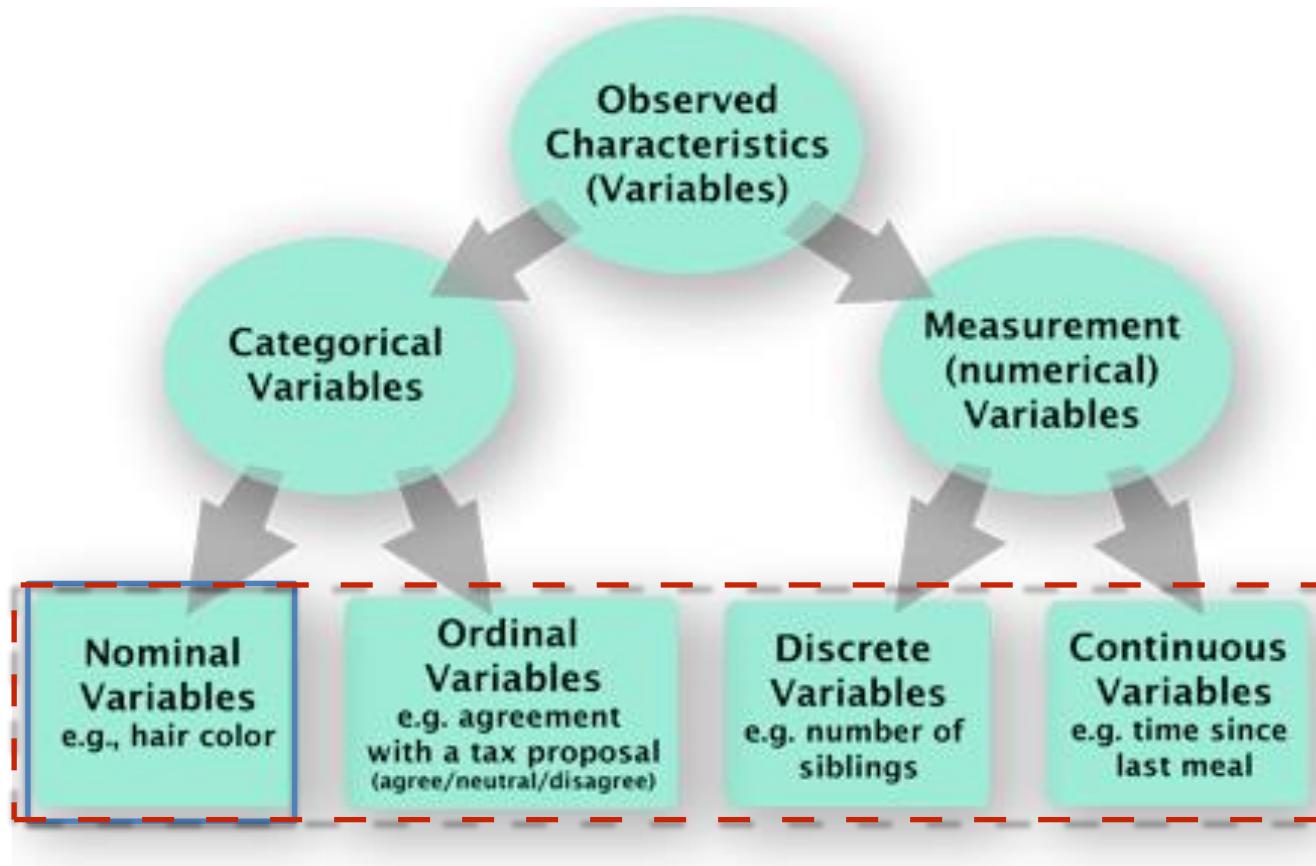
- Goodness-of-fit:
 - compare the observed sample distribution with the expected probability distribution
 - H_0 = no difference from a known population
- Chi-Square fit test:
 - determines how well theoretical distribution fits the empirical distribution
 - H_0 = no difference, equal proportions

Chi-Square Test - Applications

- Test for Independence (for two variables):
 - test *relationship* between two separate variables
 - H_0 = there is no relationship between the variables
 - eg: females prefer pepsi more than males



Chi-Square Test



Applications of Chi-square test

- Chi-square test is used to compare categorical variables. There are two type of chi-square test
 - 1. **Goodness of fit test** which determines if a sample matches the population.
 - 2. A chi-square fit **test for two independent variables** is used to compare two variables in a contingency table to check if the data fits.
 - The hypothesis being tested for chi-square is
 - Null: Variable A and Variable B are independent
 - Alternate: Variable A and Variable B are not independent.
- What to expect?
 - A small chi-square value means that data fits / variables independent
 - A high chi-square value means that data doesn't fit / variables not independent



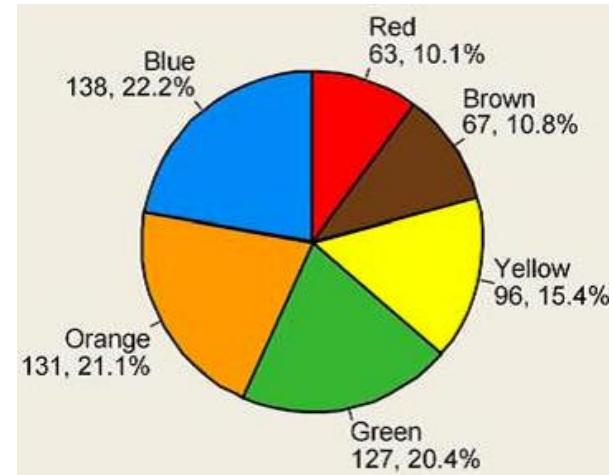
Chi-Square test/goodness-of-fit

EXAMPLE



H₀

M&Ms Color Distribution %
according to their website



H₀: The color distribution is equal

revised H₀:

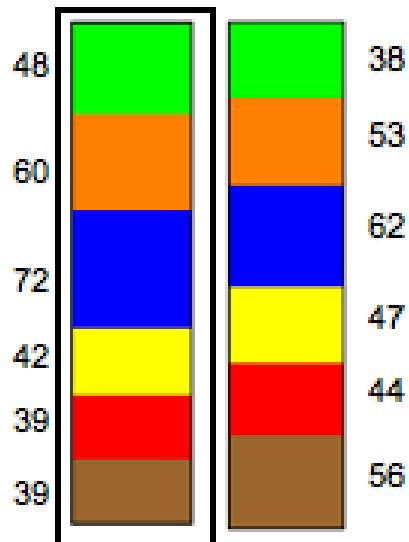
The color distribution is 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, 16% green

H_A: The color distribution is different from 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, 16% green

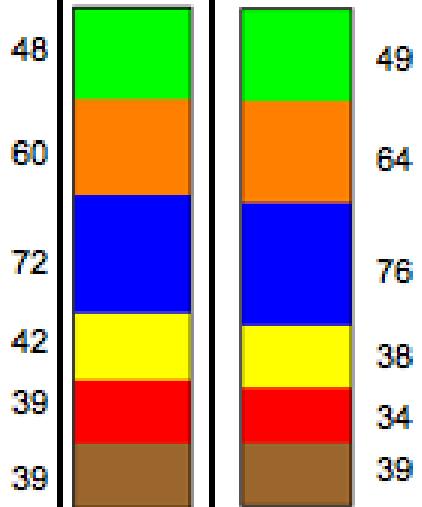


$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$

Null
Distribution Sample
#1



Null
Distribution Sample
#2



$df = ?$

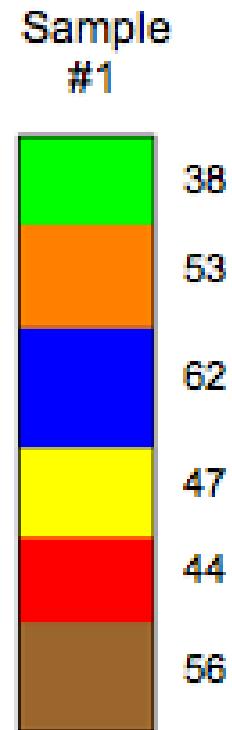
$$\chi^2 = 12.94$$

$$\chi^2 = 1.53$$

H_0 ?

Chi-Square distribution & df

- Degrees of freedom for goodness-of-fit
 - number of cells you would need to calculate all other cell values, assuming we know marginal values
- $df = C-1$, $C = \text{no. of categories}$



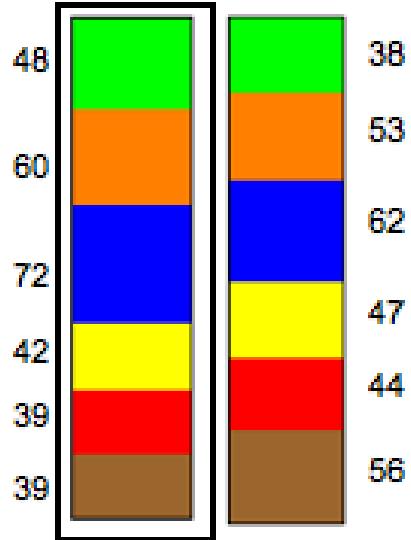
$df = ?$

<https://www.khanacademy.org/math/statistics-probability/inference-categorical-data-chi-square-tests/chi-square-goodness-of-fit-tests/v/chi-square-distribution-introduction>

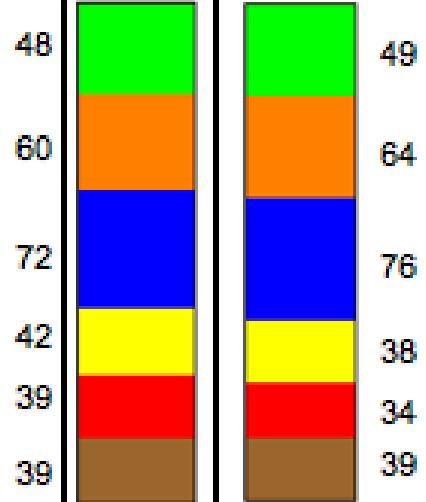


$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$

Null
Distribution Sample
#1



Null
Distribution Sample
#2



$$\chi^2 = 12.94$$

REJECTED H_0

$$\chi^2 = 1.53$$

? ACCEPTED

$$df = 5$$

Critical values of the Chi-square distribution with d degrees of freedom

d	Probability of exceeding the critical value			d		
	0.05	0.01	0.001			
1	3.841	6.635	10.828	11	19.675	24.725
2	5.991	9.210	13.816	12	21.026	26.217
3	7.815	11.345	16.266	13	22.362	27.688
4	9.488	13.277	18.467	14	23.685	29.141
5	11.070	15.086	20.515	15	24.996	30.578
6	12.592	16.812	22.458	16	26.296	32.000
7	14.067	18.475	24.322	17	27.587	33.409
8	15.507	20.090	26.125	18	28.869	34.805
9	16.919	21.666	27.877	19	30.144	36.191
10	18.307	23.209	29.588	20	31.410	37.566

Degrees of Freedom (df)

- number of independent pieces of information that go into the estimate of a parameter
- df depends on
 - particular calculation you will be performing
 - what you already know before making calculation

<https://www.youtube.com/watch?v=rATNoxKg1yA>

https://www.youtube.com/watch?v=rATNoxKg1yA&ab_channel=JamesGilbert

H_A: artists typically tend to be Aries or Cancer

EXAMPLE

H₀:

Category	Observed	Expected
Aries	29	21.333
Taurus	24	21.333
Gemini	22	21.333
Cancer	19	21.333
Leo	21	21.333
Virgo	18	21.333
Libra	19	21.333
Scorpio	20	21.333
Sagittarius	23	21.333
Capricorn	18	21.333
Aquarius	20	21.333
Pisces	23	21.333

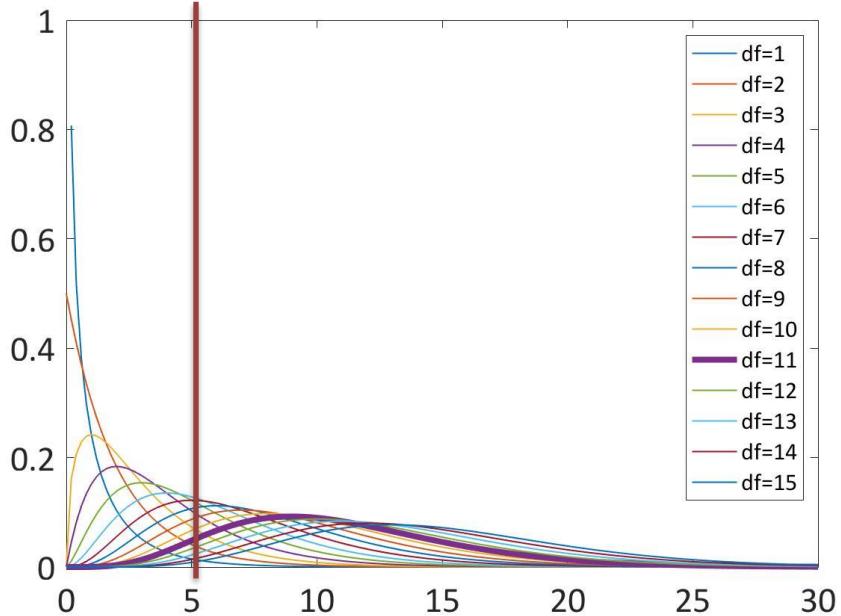
df = ?



Please submit on
Moodle [07-Mar-
2025]:
Is the Null Hypothesis
Accepted or Rejected
based on Chi-Square
Test



256 artists

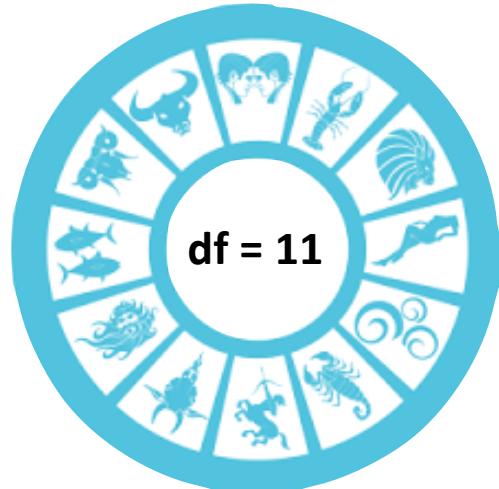


Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12

zodiac signs are evenly distributed across artists

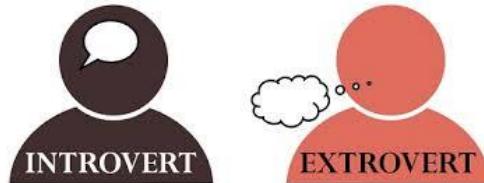
H_0



Chi-Square test for independence

test *relationship* between two separate variables

H_{01} = there is no relationship between extroversion and comfort level of dancing in public



test *difference* between two conditions

H_{02} = there is no difference in comfort level of dancing in public between introverts and extraverts

Chi-Square (example)

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

	Extroverts	Introverts	TOTAL
Not comfortable	10	40	50
comfortable	40	10	50
TOTAL	50	50	100

observed frequencies of Introverts and Extroverts who say they would or would not feel comfortable dancing in public

Chi-Square (example)

	Extroverts	Introverts	TOTAL
Not comfortable	10	40	50
comfortable	40	10	50
TOTAL	50	50	100

expected frequencies if the null hypothesis were true?

Chi-Square (example)

$\frac{\text{row total} \times \text{column total}}{\text{total } n \text{ for table}}$

	Extroverts	Introverts	TOTAL
Not comfortable	25 10	25 40	50
comfortable	25 40	25 10	50
TOTAL	50	50	100

observed and **expected** frequencies Introverts and Extroverts who say they would or would not feel comfortable dancing in public

Chi-Square (example)

	Extroverts	Introverts	TOTAL
Comfortable	10	40	50
Not Comfortable	40	10	50
TOTAL	50	50	100

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\frac{(40 - 25)^2}{25} + \frac{(10 - 25)^2}{25} + \frac{(10 - 25)^2}{25} + \frac{(40 - 25)^2}{25}$$
$$9 + 9 + 9 + 9$$

$$\chi^2 = 36$$

df=?

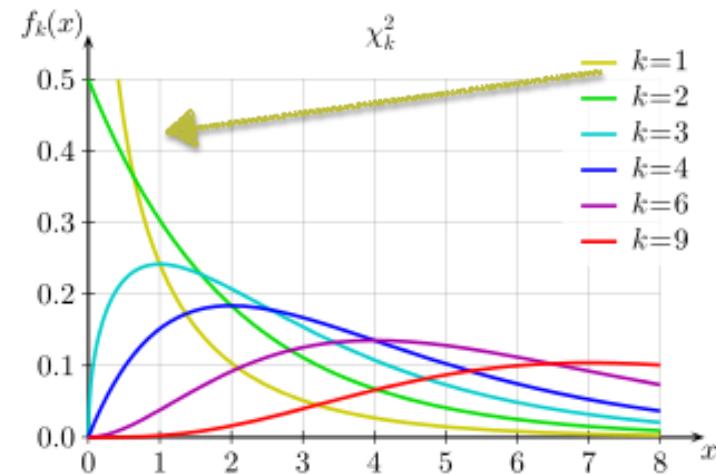
Chi-Square (example)

- Degrees of freedom for independence
 - number of cells you would need to calculate all other cell values, assuming we know marginal values

$$df = (R-1)(C-1)$$

$$df = (2-1)(2-1) = 1$$

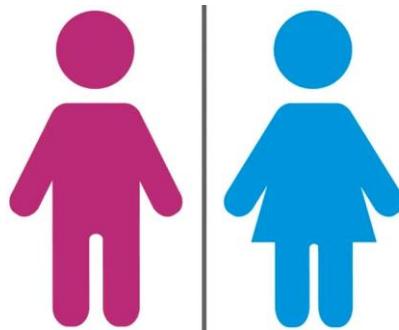
- Our chi-square is significant
 - Introverts tend to feel more comfortable dancing in public compared to Extroverts (surprise!)
 - Cooked-up data!!



$\chi^2 = 36$
 H_0 REJECTED

EXAMPLE

Gender Study



H₀: There is no relationship between gender and willingness to use mental health services

H_{A1}: The distribution of reported willingness to use mental health services for males has proportions that are different from those in females

H₀: The distribution of reported willingness to use mental health services has the same proportions for males and females

H_A: The distribution of reported willingness to use mental health services for males has proportions that are different from those in females

Contingency Table

row total × column total
total *n* for table

Willingness to Use Mental Health Services (n=150)

	No	Maybe	Yes	Total
Males	17 12	32 30	11 18	60
Female	13 18	43 45	34 27	90
Total	30	75	45	150

$$df = (R-1)(C-1)$$

$$df = 2$$

Note: How did we estimate “expected” values?

30 said NO - so we would expect equal distribution of 15 and 15.
BUT only 40% are males so 40% of 30 makes it 12.

EXAMPLE

Gender Study

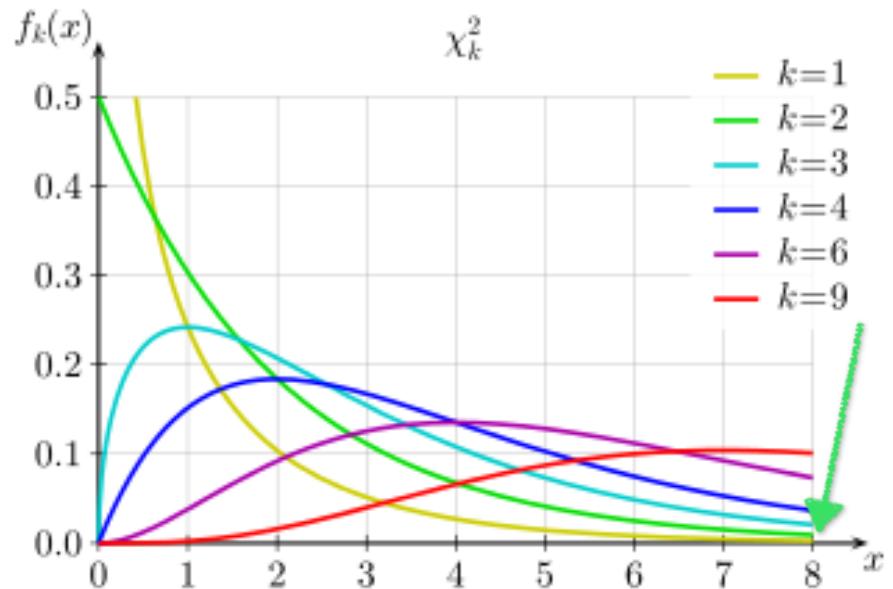
Willingness to Use Mental Health Services

$$\chi^2 = 8.23$$

$$df = 2$$

$$H_0$$

REJECTED



Males are less willing to use Mental Health Services

Effect Size

Effect size in Chi square

- For a 2×2 table \rightarrow Phi Coefficient

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Correlation between two categorical variables

Phi of 0.1 small, 0.3 medium, 0.5 large

- For larger tables \rightarrow Cramer's V coefficient ($> 2 \times 2$)

$$V = \sqrt{\frac{\chi^2}{n \times df^{*}}}$$

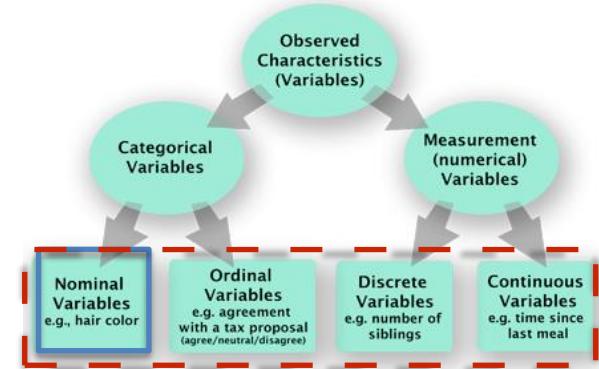
Df* is the smallest of C-1, R-1

$$\begin{aligned} \text{Phi} &= \sqrt{8.23/150} \\ &= 0.23 \end{aligned}$$

Results showed a significant difference between males' and females' attitude toward using mental health services,
 $\chi^2 (2, n = 150) = 8.23, p < .05, V = 0.23$

Chi-Square Test and Correlation

Participant	Self-Esteem X	Academic Performance Y
A	13	73
B	19	88
C	10	71
D	22	96
E	20	90
F	15	82
.	.	.
.	.	.
.	.	.



		Level of Self-Esteem				
		High	Medium	Low		
Academic Performance	High	17	32	11	60	n = 150
	Low	13	43	34	90	
		30 75 45				

Chi-square and independent measures *t* and ANOVA

Participant	Self-Esteem <i>X</i>	Academic Performance <i>Y</i>
A	13	73
B	19	88
C	10	71
D	22	96
E	20	90
F	15	82
.	.	.
.	.	.
.	.	.

		Level of Self-Esteem			$n = 150$
		High	Medium	Low	
Academic Performance	High	17	32	11	60
	Low	13	43	34	90
		30	75	45	

Median Test for Independent Samples

- non-parametric alternative to independent measures *t*-test (or ANOVA) to determine significant group differences
- H_0 = different samples come from population that share a common median

Self-Esteem Scores for Children at Three Levels of Academic Performance							
High		Medium				Low	
22	14	22	13	24	20	11	19
19	18	18	22	10	16	13	15
12	21	19	15	14	19	20	16
20	18	11	18	11	10	10	18
23	20	12	19	15	12	15	11

Median Test for Independent Samples

- calculate median for combined group ($n = 40$)
- within each group, perform median (17) split and fill contingency table

Self-Esteem Scores for Children at Three Levels of Academic Performance							
High		Medium			Low		
22	14	22	13	24	20	11	19
19	18	18	22	10	16	13	15
12	21	19	15	14	19	20	16
20	18	11	18	11	10	10	18
23	20	12	19	15	12	15	11

	Academic Performance		
	High	Medium	Low
Above Median	8	9	3
Below Median	2	11	7

Median Test for Independent Samples

		Academic Performance		
		High	Medium	Low
Above Median	High	8 5	9 10	3 5
	Below Median	2 5	11 10	7 5

$$\chi^2 = 5.4 \quad df = 2 \quad \chi^2 = 5.99 (p < .05)$$

→ not sufficient evidence to conclude that there are significant differences among the self-esteem for these three groups of students

Chi-Square test

Limitations

- Observations must be unique to one cell (Between subjects)
 - each person must fall into only one cell
 - not valid for within subject designs (repeated measures/matched pairs)
- Only frequencies can be studied, not means, percentages, ratios, etc.
- Low **expected** frequencies cause problems (should be ≥ 5)
 - loss of statistical power
- No group should contain less than 10 (or 5) (try to regroup instead)
- Not apt for low sample size.
- Informs of presence or absence (probability of occurrence) of association but doesn't measure strength of association

Life after chi-squared: an introduction to log-linear analysis.

Streiner DL¹, Lin E.

Author information

1 Kunin-Lunenfeld Applied Research Unit, Baycrest Centre for Geriatric Care, North York, Ontario. dstreiner@rotman-baycrest.on.ca

Abstract

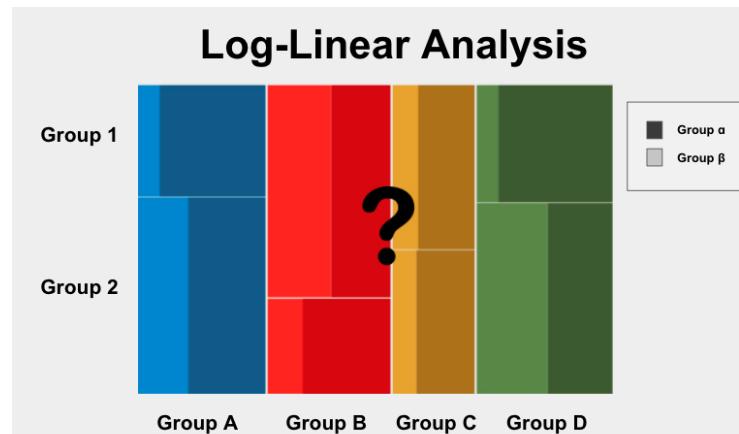
Chi-squared tests are used to examine the relationships among categorical variables. However, they are difficult to use and interpret when more than 2 variables are involved. In such cases, it is better to use a related statistic, called log-linear analysis. This article is an introduction to log-linear models, illustrating how they can be used to tease apart relationships among several variables in looking at the factors associated with photophobia.

Age category of car			
	New	Old	Total
Male drivers			
Behaviour at amber light			
Stopped	79	63	142
Did not stop	87	95	182
Total	166	158	324
Female drivers			
Behaviour at amber light			
Stopped	95	83	178
Did not stop	51	94	145
Total	146	177	323
Total old/new cars:	312	335	647

Table 18.15 Stopping behaviour of male and female drivers in old and new cars.

Log-Linear Analysis

- variable of interest is proportional or categorical
- have two or more options
- no assumptions of IV or DV
- used for both hypothesis testing and model building



Different types of tests (Summary)

Test type	Between subjects designs (Independent samples)	Within subject designs (repeated measures/matched pairs)
Non-parametric (for categorical data)	Chi-square	<i>The binomial sign test</i>
Non-parametric (for ordinal data)	<i>Mann-Whitney U</i>	<i>Wilcoxon Signed-Rank test</i> <i>The binomial sign test</i>
Parametric	<i>Unrelated t-test (level of data: interval)</i>	<i>Related t-test (level of data: interval)</i>

Non-parametric tests

How to deal with Categorical data?

How to deal with cases where parametric assumptions are violated?

Recap

- Chi-square test
 - Goodness of fit
 - Independence of variables (2 variables) (for unrelated case)
 - Effect size
 - Median test for independence of samples
 - Log-Linear analysis (>2 categorical variables)
 - Binomial sign test (for related categorical variables)

Selecting a statistical test

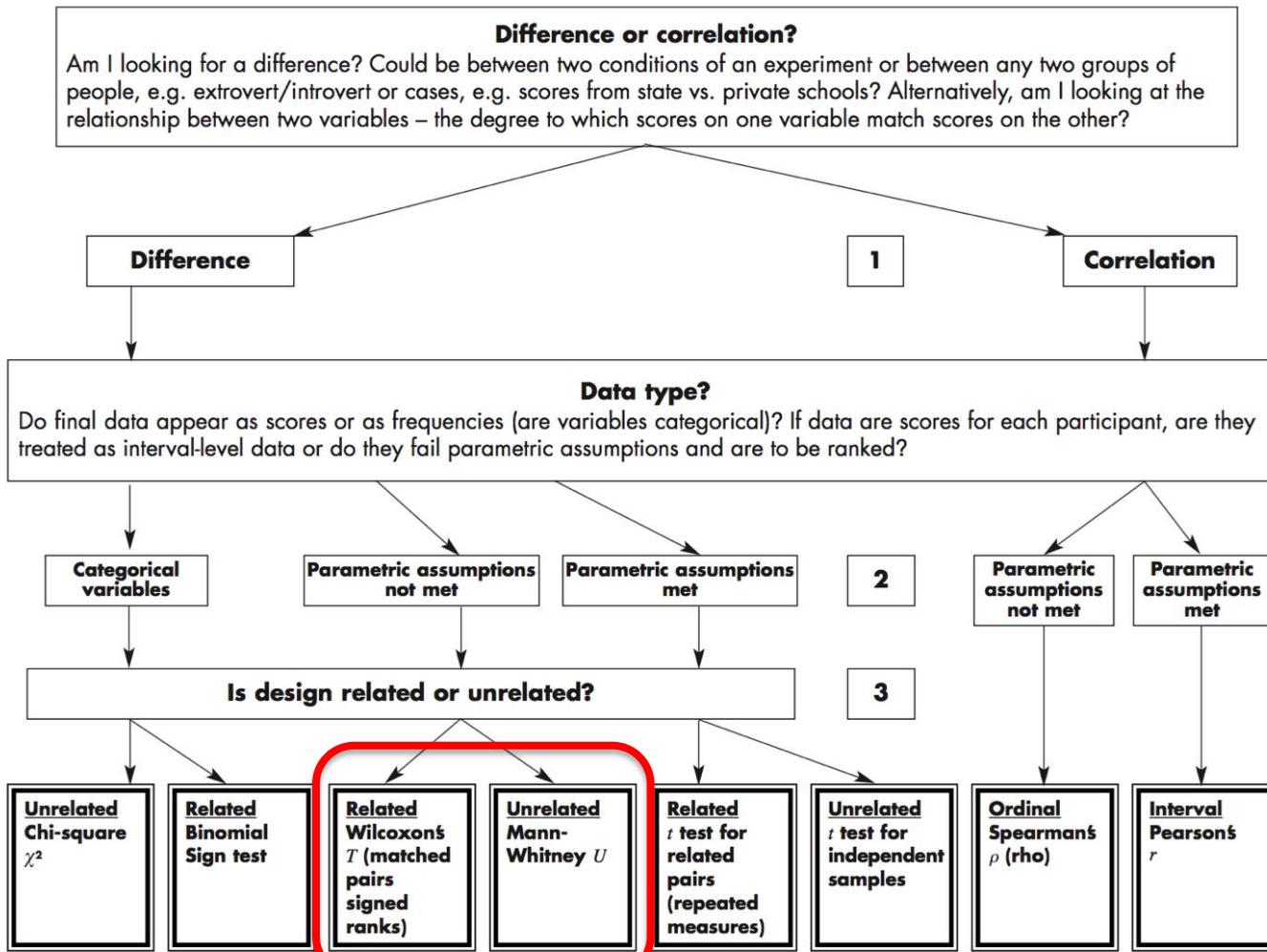


Figure 23.1 Choosing an appropriate two-sample test.

Predictor variable	Outcome variable	Use in place of...
Spearman's r	• Quantitative	• Quantitative Pearson's r
Chi square test of independence	• Categorical	• Categorical Pearson's r
Sign test	• Categorical	• Quantitative One-sample t -test
Kruskal-Wallis H	• Categorical • 3 or more groups	• Quantitative ANOVA
ANOSIM	• Categorical • 3 or more groups	• Quantitative • 2 or more outcome variables MANOVA
Wilcoxon Rank-Sum test	• Categorical • 2 groups	• Quantitative • groups come from Independent t-test
Wilcoxon Signed-rank test	• Categorical • 2 groups	• Quantitative • groups come from the same population Paired t-test

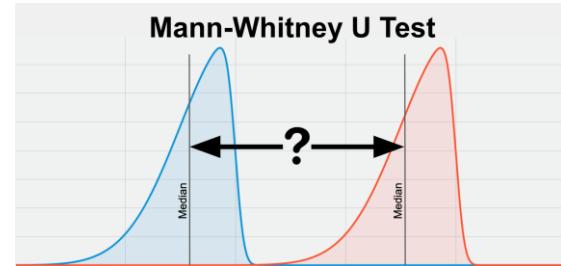
Choosing a nonparametric test

Non-parametric tests don't make as many assumptions about the data and are useful when one or more of the common statistical assumptions are violated. However, the inferences they make aren't as strong as with parametric tests.

Different types of tests

Test type	Between subjects designs (Independent samples)	Within subject designs (repeated measures/matched pairs)
Non-parametric (for categorical data)	Chi-square	<i>The binomial sign test</i>
Non-parametric (for ordinal data)	<i>Mann-Whitney U</i>	<i>Wilcoxon Signed-Rank test</i> <i>The binomial sign test</i>
Parametric	<i>Unrelated t-test (level of data: interval)</i>	<i>Related t-test (level of data: interval)</i>

Mann-Whitney U Test



- between subjects design
- skewed distribution
- used on ordinal non-normal data
- ***assumption:***
 - a real difference between two populations should cause the scores in one sample to be generally larger than the other;
 - if two samples are combined and all scores are ranked, then the larger ranks should be concentrated in one sample and smaller ranks in the other
 - eg: Likert items (e.g., a 7-point scale from "strongly agree" through to "strongly disagree")

Mann-Whitney U test

- ex: children's tendency to stereotype according to traditional gender roles if they have working mothers vs not

Full-time jobs		No job outside home	
Score	Points	Score	Points
17	9	19	6
32	7	63	0
39	6.5	78	0
27	8	29	4
58	6	39	1.5
25	8	59	0
31	7	77	0
		81	0
		68	0

Totals: $51.5 = U_1$ $11.5 = U_2$

U is the lower of 51.5 and 11.5, so *U* is 11.5

- the observed *U* value should be less than or equal to critical *U* value in order to reject H_0

Mann-Whitney U Table

Alpha = .05 (two-tailed)

> two groups - Kruskal-Wallis test

Mann-Whitney U test

- ex: children's tendency to stereotype according to traditional gender roles if they have working mothers vs not

Full-time jobs		No job outside home	
Score	Points	Score	Points
17	9	19	6
32	7	63	0
39	6.5	78	0
27	8	29	4
58	6	39	1.5
25	8	59	0
31	7	77	0
		81	0
		68	0
Totals:	$51.5 = U_1$		$11.5 = U_2$
	U is the lower of 51.5 and 11.5, so U is 11.5		

critical U value = 12

$\alpha < .05$

children of working mothers are less likely to use gender-role stereotypes

REJECTED
H₀

Kruskal-Wallis Test

Aim



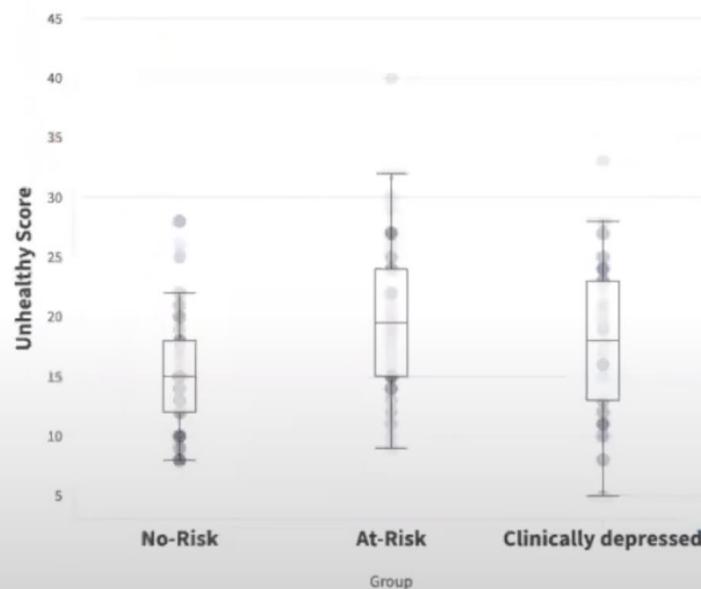
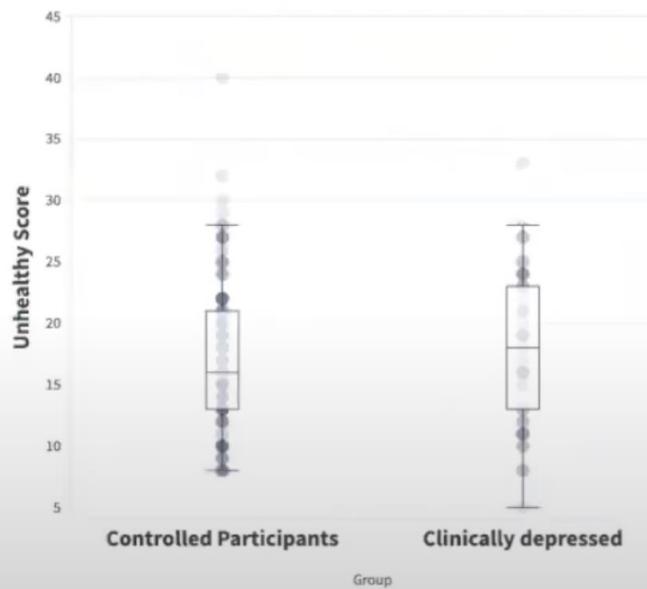
Clinically Depressed Cohort (DC)



Controlled Participants (CP)

To investigate musical engagement strategies via HUMS in DC compared to control participants (CP) from the community

Results: Group Differences for *Unhealthy Scores*



Different types of tests

Test type	Between subjects designs (Independent samples)	Within subject designs (repeated measures/matched pairs)
Non-parametric (for categorical data)	Chi-square	<i>The binomial sign test</i>
Non-parametric (for ordinal data)	<i>Mann-Whitney U</i>	<i>Wilcoxon Signed-Rank Test</i>
Parametric	<i>Unrelated t-test (level of data: interval)</i>	<i>Related t-test (level of data: interval)</i>

Wilcoxon Signed-Rank Test

- ordinal level (tests based on rank order)
- within subjects design (related, repeated-measures/matched pairs)
- null hypothesis as the claim that the two populations from which scores are sampled are identical
- most of the time this is more specifically that the two medians are equal (not means because we are working at the ordinal level)
- the observed W value should be less than or equal to critical W value in order to reject H_0

Note: Can also be used for discrete related samples.

If the paired data violate parametric assumptions,
Convert the paired scores into difference and rank them.

Wilcoxon Signed-Rank Test

- example:
 - assess if students performed better in the mock exam than the final GRE exam

H_0 : Population median difference = 0

H_1 : Population median difference > 0 (1-tail)

Wilcoxon Signed-Rank Test

Student	Mock	Real	Diff(d)	Rank
1	316	320	-4	-4.5
2	324	319	5	6
3	317	318	-1	-1.5
4	323	314	9	10
5	333	333	0	n/a
6	329	321	8	9
7	328	311	17	12
8	319	309	10	11
9	320	318	2	3
10	314	321	-7	-8
11	309	315	-6	-7
12	323	319	4	4.5
13	335	334	1	1.5

$$T_+ = 57 \quad T_- = 21$$

$$W_{\text{stat}} = \min(T_+, T_-) = 21$$

(> critical W value 17
 $\alpha < .05$)

n	Two-Tailed Test		One-Tailed Test	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15

H_0

ACCEPTED

Different types of tests

Test type	Between subjects designs (Independent samples)	Within subject designs (repeated measures/matched pairs)
Non-parametric (for categorical data)	Chi-square	<i>The binomial sign test</i>
Non-parametric (for ordinal data)	<i>Mann-Whitney U</i>	<i>Wilcoxon Signed-Rank Test</i> <i>The binomial sign test</i>
Parametric	<i>Unrelated t-test (level of data: interval)</i>	<i>Related t-test (level of data: interval)</i>

The Binomial Sign Test

Categorical data

- Within subjects design
- Items are dichotomous and nominal
- may be reduced from interval or ordinal level
- two dependent samples should be paired or matched

The Binomial Sign Test

A	B	C	D	E	
Client	Self-image rating before therapy	Self-image rating after 3 months' therapy	Difference (C – B)	Sign of difference	
a	3	7	4	+	
b	12	18	6	+	
c	9	5	-4	-	
d	7	7	0		
e	8	12	4	+	$S = 1$
f	1	5	4	+	
g	15	16	1	+	
h	10	12	2	+	
i	11	15	4	+	
j	10	17	7	+	

Table 17.6 Self-image scores before and after three months' therapy.

- the observed S value should be less than or equal to critical S value in order to reject H_0

Numbers in the table represent $p(X=x)$ for a binomial distribution with n trials and probability of success p .

Binomial probabilities:		p										
n	x	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9
8	0	0.430	0.168	0.100	0.058	0.017	0.004	0.001	0.000	0.000	0.000	0.000
	1	0.383	0.336	0.267	0.198	0.090	0.031	0.008	0.001	0.000	0.000	0.000
	2	0.149	0.294	0.311	0.296	0.209	0.109	0.041	0.010	0.004	0.001	0.000
	3	0.033	0.147	0.208	0.254	0.279	0.219	0.124	0.047	0.023	0.009	0.000
	4	0.005	0.046	0.087	0.136	0.232	0.273	0.232	0.136	0.087	0.046	0.005
	5	0.000	0.009	0.023	0.047	0.124	0.219	0.279	0.254	0.208	0.147	0.033
	6	0.000	0.001	0.004	0.010	0.041	0.109	0.209	0.296	0.311	0.294	0.149
	7	0.000	0.000	0.000	0.001	0.008	0.031	0.090	0.198	0.267	0.336	0.383
	8	0.000	0.000	0.000	0.000	0.001	0.004	0.017	0.058	0.100	0.168	0.430
9	0	0.387	0.134	0.075	0.040	0.010	0.002	0.000	0.000	0.000	0.000	0.000
	1	0.387	0.302	0.225	0.156	0.060	0.018	0.004	0.000	0.000	0.000	0.000
	2	0.172	0.302	0.300	0.267	0.161	0.070	0.021	0.004	0.001	0.000	0.000
	3	0.045	0.176	0.234	0.267	0.251	0.164	0.074	0.021	0.009	0.003	0.000
	4	0.007	0.066	0.117	0.172	0.251	0.246	0.167	0.074	0.039	0.017	0.001
	5	0.001	0.017	0.039	0.074	0.167	0.246	0.251	0.172	0.117	0.066	0.007
	6	0.000	0.003	0.009	0.021	0.074	0.164	0.251	0.267	0.234	0.176	0.045
	7	0.000	0.000	0.001	0.004	0.021	0.070	0.161	0.267	0.300	0.302	0.172
	8	0.000	0.000	0.000	0.000	0.004	0.018	0.060	0.156	0.225	0.302	0.387
	9	0.000	0.000	0.000	0.000	0.000	0.002	0.010	0.040	0.075	0.134	0.387
10	0	0.349	0.107	0.056	0.028	0.006	0.001	0.000	0.000	0.000	0.000	0.000
	1	0.387	0.268	0.188	0.121	0.040	0.010	0.002	0.000	0.000	0.000	0.000
	2	0.194	0.302	0.282	0.233	0.121	0.044	0.011	0.001	0.000	0.000	0.000
	3	0.057	0.201	0.250	0.267	0.215	0.117	0.042	0.009	0.003	0.001	0.000
	4	0.011	0.088	0.146	0.200	0.251	0.205	0.111	0.037	0.016	0.006	0.000
	5	0.001	0.026	0.058	0.103	0.201	0.246	0.201	0.103	0.058	0.026	0.001
	6	0.000	0.006	0.016	0.037	0.111	0.205	0.251	0.200	0.146	0.088	0.011
	7	0.000	0.001	0.003	0.009	0.042	0.117	0.215	0.267	0.250	0.201	0.057
	8	0.000	0.000	0.000	0.001	0.011	0.044	0.121	0.233	0.282	0.302	0.194
	9	0.000	0.000	0.000	0.000	0.002	0.010	0.040	0.121	0.188	0.268	0.387
	10	0.000	0.000	0.000	0.000	0.000	0.001	0.006	0.028	0.056	0.107	0.349
11	0	0.314	0.086	0.042	0.020	0.004	0.000	0.000	0.000	0.000	0.000	0.000
	1	0.384	0.236	0.155	0.093	0.027	0.005	0.001	0.000	0.000	0.000	0.000
	2	0.213	0.295	0.258	0.200	0.089	0.027	0.005	0.001	0.000	0.000	0.000
	3	0.071	0.221	0.258	0.257	0.177	0.081	0.023	0.004	0.001	0.000	0.000
	4	0.016	0.111	0.172	0.220	0.236	0.161	0.070	0.017	0.006	0.002	0.000
	5	0.002	0.039	0.080	0.132	0.221	0.226	0.147	0.057	0.027	0.010	0.000
	6	0.000	0.010	0.027	0.057	0.147	0.226	0.221	0.132	0.080	0.039	0.002
	7	0.000	0.002	0.006	0.017	0.070	0.161	0.236	0.220	0.172	0.111	0.016
	8	0.000	0.000	0.001	0.004	0.023	0.081	0.177	0.257	0.258	0.221	0.071
	9	0.000	0.000	0.000	0.001	0.005	0.027	0.089	0.200	0.258	0.295	0.213
	10	0.000	0.000	0.000	0.000	0.001	0.005	0.027	0.093	0.155	0.236	0.384
	11	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.020	0.042	0.086	0.314

[From the Binomial Distribution Table]
The p-value is 0.01, which is smaller than the **alpha-level** of 0.05. We can reject the null hypothesis and say there is a significant difference.

S-Table Lookup



n	One tailed, $\alpha = 0.005$	$\alpha = 0.01$	$\alpha = 0.025$	$\alpha = 0.05$
	Two tailed, $\alpha = 0.01$	$\alpha = 0.02$	$\alpha = 0.05$	$\alpha = 0.10$
8	0	0	0	1
9	0	0	1	1
10	0	0	1	1
11	0	1	1	2
12	1	1	2	2
13	1	1	2	3
14	1	2	3	3
15	2	2	3	3
16	2	2	3	4
17	2	3	4	4
18	3	3	4	5
19	3	4	4	5
20	3	4	5	5
21	4	4	5	6
22	4	5	5	6
23	4	5	6	7
24	5	5	6	7
25	5	6	6	7

The Binomial Sign Test

A	B	C	D	E	
Client	Self-image rating before therapy	Self-image rating after 3 months' therapy	Difference (C – B)	Sign of difference	
a	3	7	4	+	
b	12	18	6	+	
c	9	5	-4	-	
d	7	7	0		
e	8	12	4	+	$S = 1$
f	1	5	4	+	
g	15	16	1	+	
h	10	12	2	+	
i	11	15	4	+	
j	10	17	7	+	

Table 17.6 Self-image scores before and after three months' therapy.

critical S value = 1

$\alpha \leq .05$

REJECTED
 H_0

	Predictor variable	Outcome variable	Use in place of...
Spearman's r	<ul style="list-style-type: none"> Quantitative 	<ul style="list-style-type: none"> Quantitative 	Pearson's r
Chi square test of independence	<ul style="list-style-type: none"> Categorical 	<ul style="list-style-type: none"> Categorical 	Pearson's r
Sign test	<ul style="list-style-type: none"> Categorical 	<ul style="list-style-type: none"> Quantitative 	One-sample t -test
Kruskal-Wallis H	<ul style="list-style-type: none"> Categorical 3 or more groups 	<ul style="list-style-type: none"> Quantitative 	ANOVA
ANOSIM	<ul style="list-style-type: none"> Categorical 3 or more groups 	<ul style="list-style-type: none"> Quantitative 2 or more outcome variables 	MANOVA
Wilcoxon Rank-Sum test	<ul style="list-style-type: none"> Categorical 2 groups 	<ul style="list-style-type: none"> Quantitative groups come from 	Independent t-test
Wilcoxon Signed-rank test	<ul style="list-style-type: none"> Categorical 2 groups 	<ul style="list-style-type: none"> Quantitative groups come from the same population 	Paired t-test

Choosing a nonparametric test

Non-parametric tests don't make as many assumptions about the data, and are useful when one or more of the common statistical assumptions are violated. However, the inferences they make aren't as strong as with parametric tests.

Permutation Tests

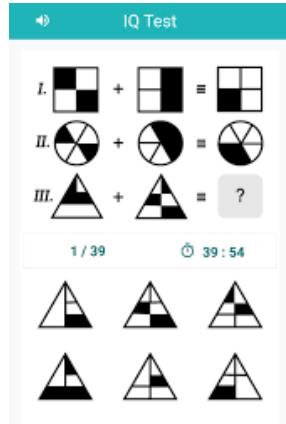
- rely on randomizations of the observed data and typically seek to quantify the null distribution in order to perform hypothesis testing
- permute the data in a way that removes some aspects of the statistical structure and evaluate how likely is the observed statistic to occur if the null hypothesis was true
- the test statistic is compared against a theoretical distribution of test statistics expected under the H_0 .
- determine the statistical significance of a model by computing a test statistic on the dataset and then for many random permutations of that data
 - > If the model is significant, the original test statistic value should lie at one of the tails of the null hypothesis distribution.

Why choose Permutation Tests?

- small sample size
- assumptions (for parametric approach) not met
- test statistic other than comparing means/medians
- difficult to estimate SE for test statistic

EXAMPLE

H_A = Engineering students have higher IQ than Art students



Group 1: CS

Group 2: Art

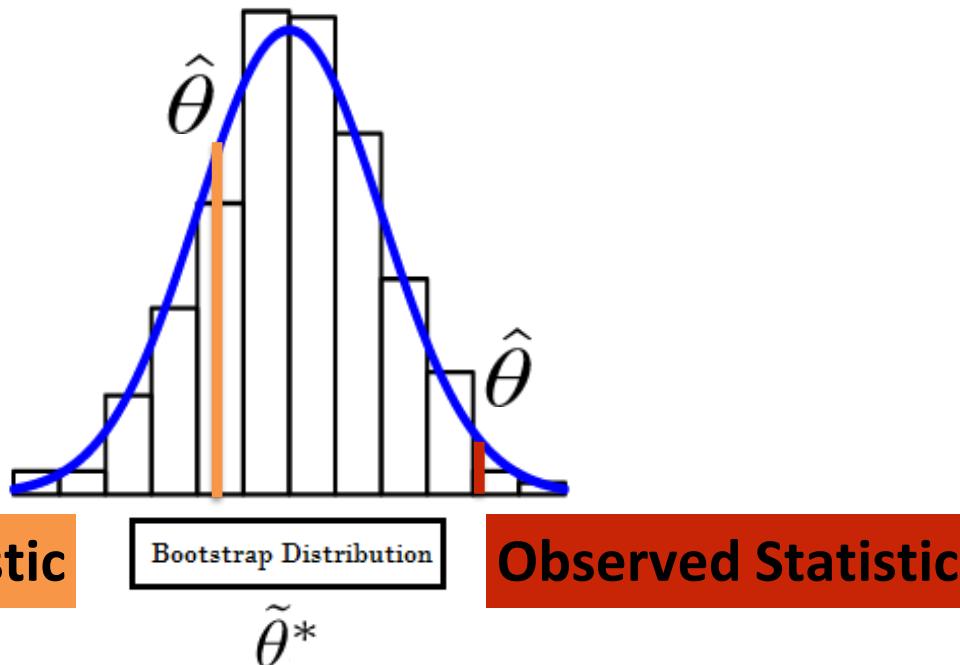
Result: $\text{mean(IQ}_{\text{CS}}\text{)} > \text{mean(IQ}_{\text{Art}}\text{)}$

how certain am I that i can reject the null-hypothesis?
what is the probability that this result can appear due to chance alone?

Permutation Tests (Group Differences)

- eg: context of *difference in mean* of both groups
 1. randomly permute (or “shuffle”) the data into both groups
 2. recalculate the difference in mean
 3. repeat steps 1-2 several times to obtain the resulting samples to characterize the null distribution (i.e. the distribution we would expect if there were no statistical relationship between x and y)
 4. significance estimation: evaluate the proportion of times you get at least a value equal to or more extreme than the observed/actual statistic (*difference in mean*)

likely to accept null hypothesis or not?



Estimate

$\hat{\theta}$ = difference in medians, t-statistic, mode, std, etc..

COMMENT · 20 MARCH 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

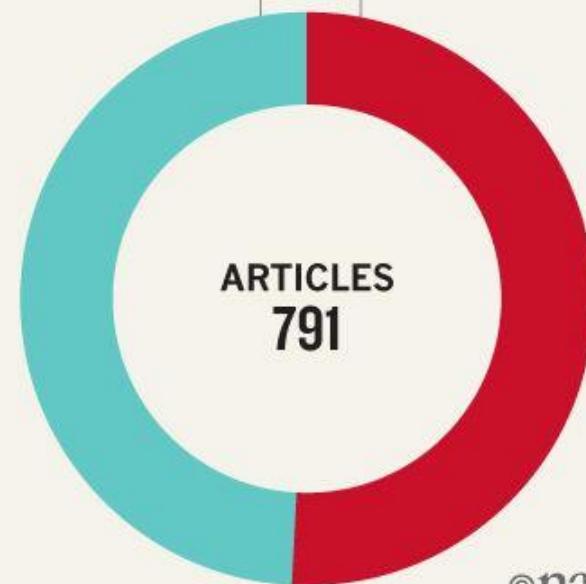
Valentin Amrhein , Sander Greenland & Blake McShane

WRONG INTERPRETATIONS

An analysis of 791 articles across 5 journals* found that around half mistakenly assume non-significance means no effect.

Appropriately interpreted
49%

Wrongly interpreted
51%



*Data taken from: P. Schatz et al. *Arch. Clin. Neuropsychol.* **20**, 1053–1059 (2005); F. Fidler et al. *Conserv. Biol.* **20**, 1539–1544 (2006); R. Hoekstra et al. *Psychon. Bull. Rev.* **13**, 1033–1037 (2006); F. Bernardi et al. *Eur. Sociol. Rev.* **33**, 1–15 (2017).



COMMENT · 20 MARCH 2019

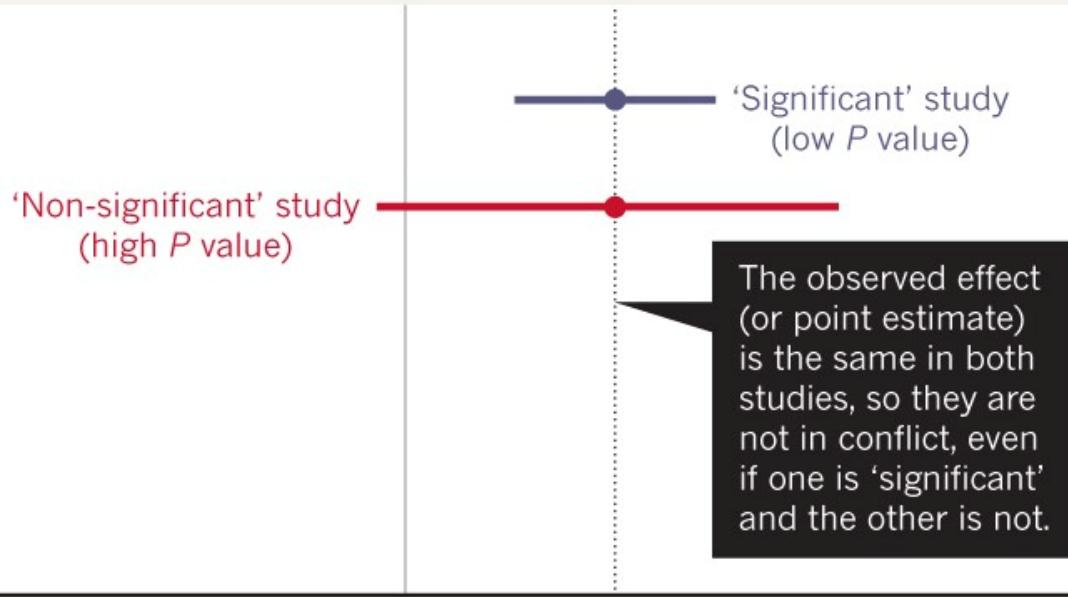
Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Valentin Amrhein , Sander Greenland & Blake McShane

BEWARE FALSE CONCLUSIONS

Studies currently dubbed ‘statistically significant’ and ‘statistically non-significant’ need not be contradictory, and such designations might cause genuine effects to be dismissed.



p-hacking!!

1. Stop collecting data once $p < .05$
2. Analyze many measures, but report only those with $p < .05$.
3. Collect and analyze many conditions, but only report those with $p < .05$.
4. Use covariates to get $p < .05$.
5. Exclude participants to get $p < .05$.
6. Transform the data to get $p < .05$.

Different types of tests

Test type	Between subjects designs (Independent samples)	Within subject designs (repeated measures/matched pairs)
Non-parametric (for categorical data)	Chi-square	<i>The binomial sign test</i>
Non-parametric (for ordinal data)	<i>Mann-Whitney U</i>	<i>Wilcoxon Signed-Rank Test</i>
Parametric	<i>Unrelated t-test (level of data: interval)</i>	<i>Related t-test (level of data: interval)</i>

Next Class

ANOVA

BRSM

DV: Anxiety level

IV: Exercise

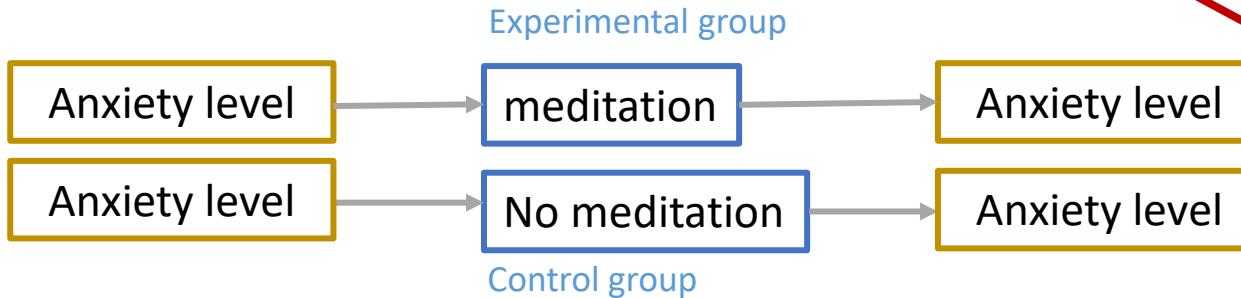
People who exercise have lower levels of anxiety



Within group/Repeated measures
(crossover design)

- Participant fatigue
- Longer experimental duration
- Carry over effects

Exercise lowers anxiety



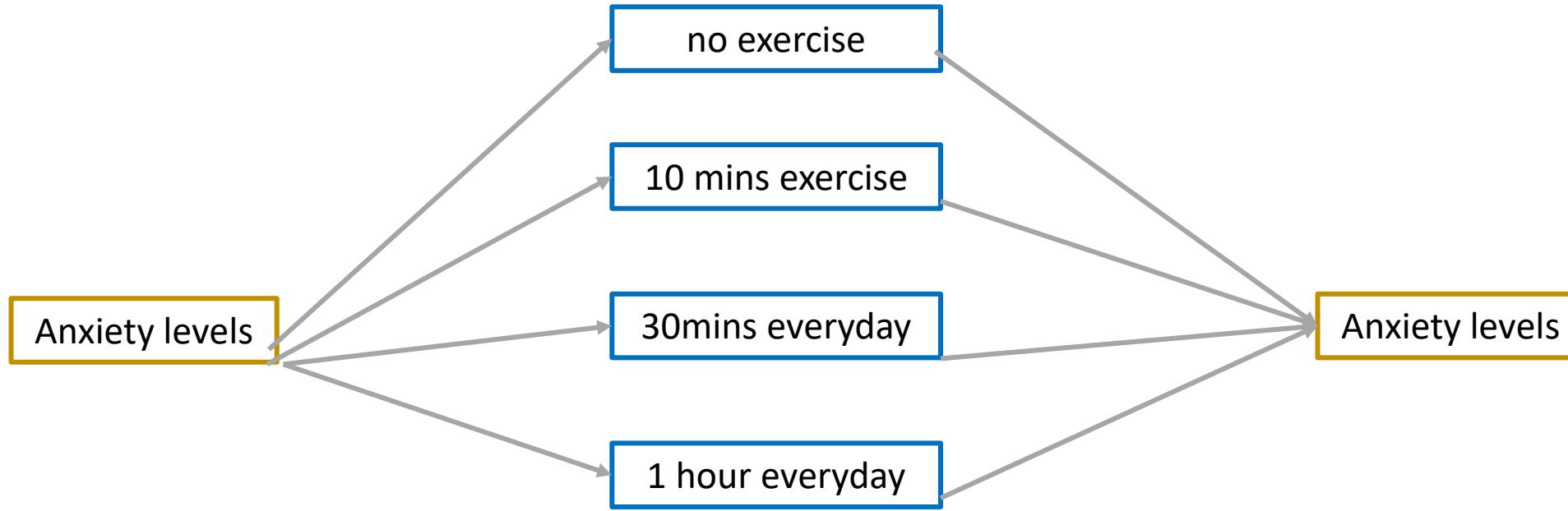
Mixed design
Between groups & Within groups

DV: Anxiety level

IV: Exercise

1 factor, 4 levels

IV – 4 levels



Can we perform multiple t-tests?

Can you do a t-test on this data?

Factor = Independent variable

2 Independent Variables - 2 levels each

Exercise – exercise vs control

Time of Day – morning vs evening

Two factorial design

Exercise-morning	Control-morning
Exercise-evening	Control-evening

2x2 factorial design

2 Independent Variables – different levels

Exercise – 30mins, 1 hour, 2 hours

Time of Day – morning vs evening

Two factorial design

30 mins-morning	1 hr-morning	2 hrs - morning
30 mins-evening	1 hr-evening	2 hrs - evening

3x2 factorial design

Can we perform multiple t-tests?

Why not just perform multiple t-tests?

To avoid Type I error – false positive

- For 'k' independent groups there are $k(k-1)/2$ possible t-tests
 - For 5 groups = $5(5-1)/2= 10$ t-tests
 - For 4 groups = $4(4-1)/2= 6$ t-tests
 - For 3 groups = $3(3-1)/2= 3$ t-tests
- Using too many t-test comparisons increases the chances of finding random significant effects which may be due to chance. In reality there may be no difference between the groups/conditions

Risk of family-wise error – Increase Type I error

Bonferroni Correction

$$\frac{0.05 (\alpha)}{\text{No. of comparisons}} = 0.05/3 = 0.0167 \text{ (new } \alpha \text{ value)}$$

Sample Variance

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

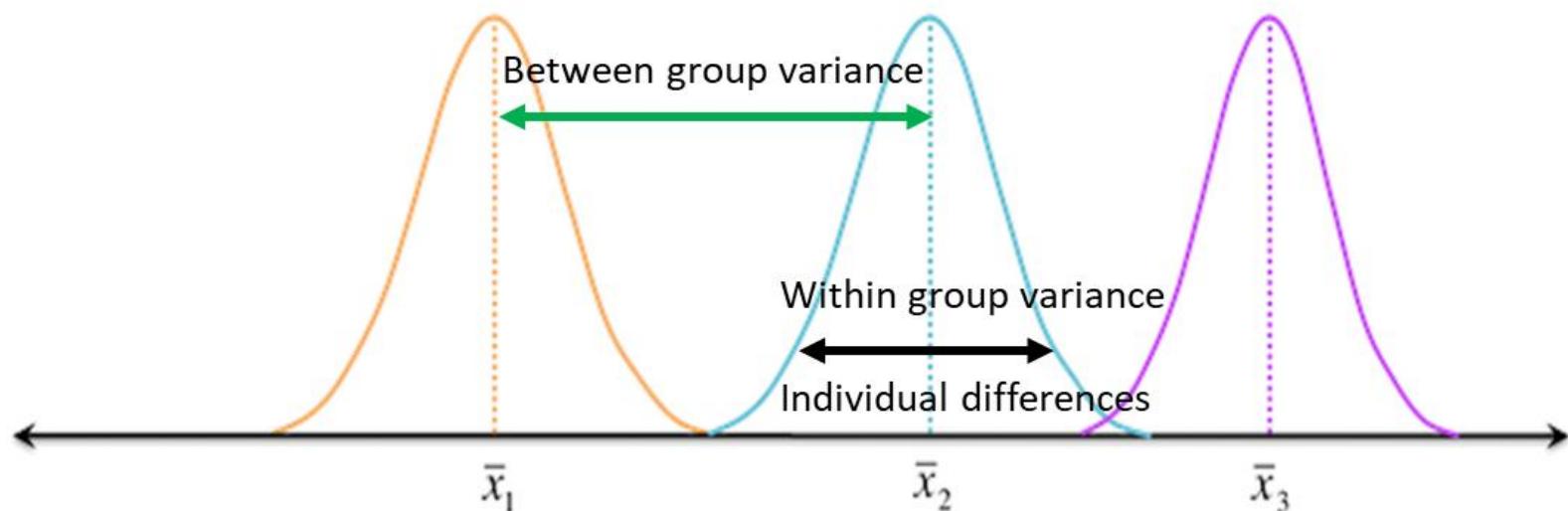
s^2 = sample variance

x_i = value of i^{th} element

\bar{x} = sample mean

n = sample size

ANOVA ANalysis Of VAriance

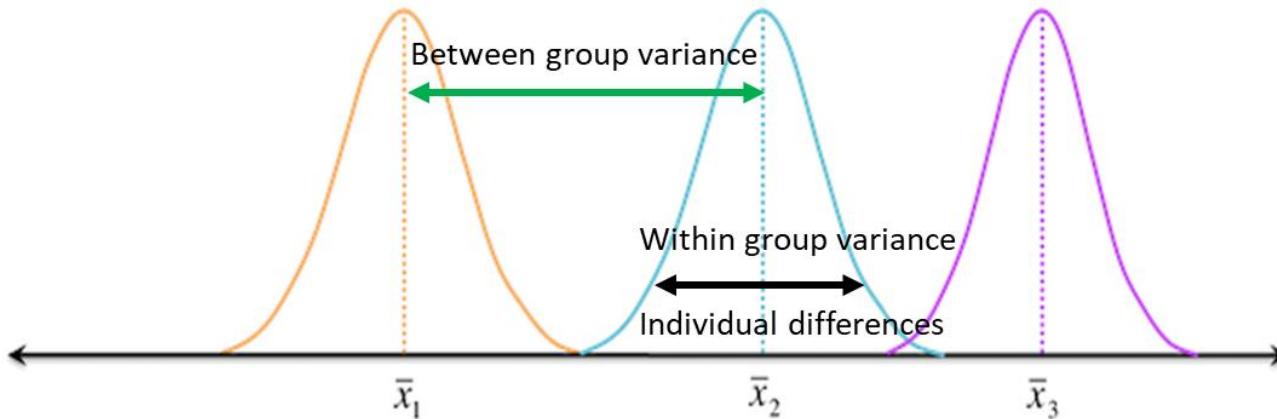


ANOVA performs all three comparisons simultaneously in one test.

No matter how many different means are being compared, ANOVA uses one test with one alpha level to evaluate the difference in variance

$$F = \frac{\text{variance (differences) between sample means}}{\text{variance (differences) within sample}}$$

= difference due to treatment/experimental condition
individual differences in each condition



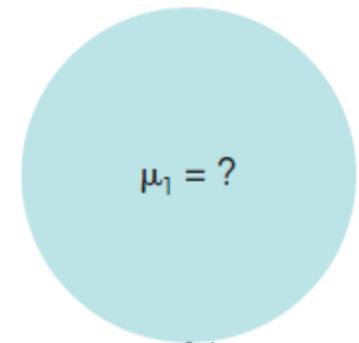
Sample Variance
$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
s^2 = sample variance
x_i = value of i^{th} element
\bar{x} = sample mean
n = sample size

One Way ANOVA

IV (factor) – Type of treatment

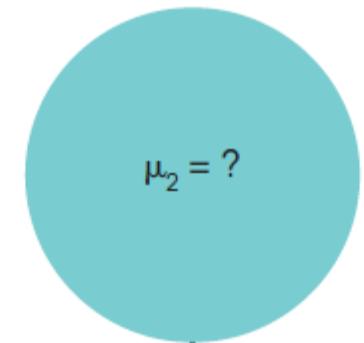
Counseling

Population 1
(Treatment 1)



Anti-anxiety meds

Population 2
(Treatment 2)

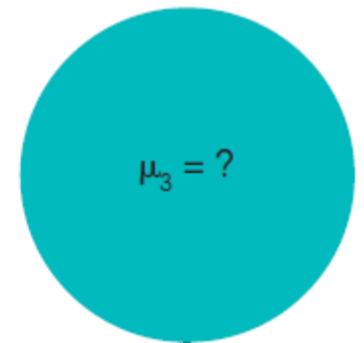


Counseling

+

Anti-anxiety meds

Population 3
(Treatment 3)



DV – Level of anxiety

Ho - Null hypothesis – anxiety levels are equal across all groups after treatment (no difference between treatments)

H₁ - Alternate hypothesis ?

$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}} = \frac{\text{systematic treatment effects + random, unsystematic differences}}{\text{random, unsystematic differences}} \geq 1 \text{ (treatment had an effect)}$$

$$F = \frac{0 + \text{random, unsystematic differences}}{\text{random, unsystematic differences}} \leq 1 \text{ (Treatment had no effect)}$$

	Source	SS (ss)	df	s² (MS)	F
Effect of treatment	Between treatments	$\sum n_i(\bar{X}_i - \bar{\bar{X}})^2$	$k-1$	$\frac{SS_b}{df_b} = \text{MSB}$	$F = \frac{\text{MSB}}{\text{MSW}}$
Random differences/error	Within treatments	$\sum (X_{ij} - \bar{X}_i)^2$	$N-k$	$\frac{SS_w}{df_w} = \text{MSW}$	
	Total	$\sum (X_{ij} - \bar{\bar{X}})^2$	$N-1$		

X_{ij} = an individual observation

k = the number of groups

□

\bar{X}_i = the mean of the i^{th} group

n_i = the number of subjects in the i^{th} group

$\bar{\bar{X}}$ = the grand mean

N = the number of subjects total

Source	SS	df	s^2 (MS)	F
Between treatments	$\sum n_i(\bar{X}_i - \bar{\bar{X}})^2$	$k-1$	$\frac{SS_b}{df_b}$	s_b^2 / s_w^2
Within treatments	$\sum (X_{ij} - \bar{X}_i)^2$	$N-k$	$\frac{SS_w}{df_w}$	
Total	$\sum (X_{ij} - \bar{\bar{X}})^2$	$N-1$		

X_{ij} = an individual observation

k = the number of groups

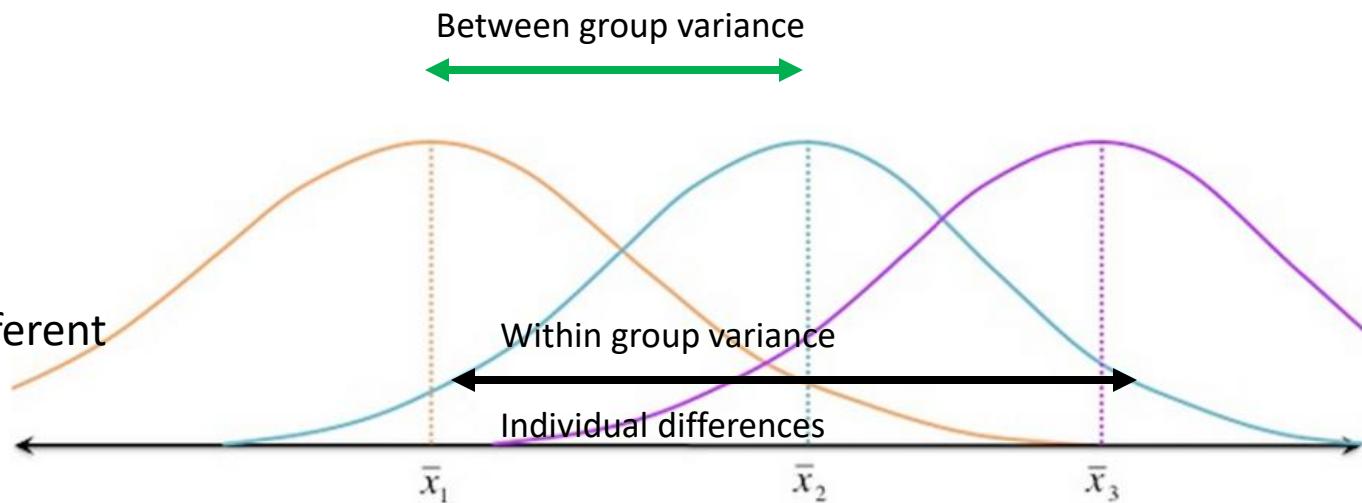
\bar{X}_i = the mean of the i^{th} group

n_i = the number of subjects in the i^{th} group

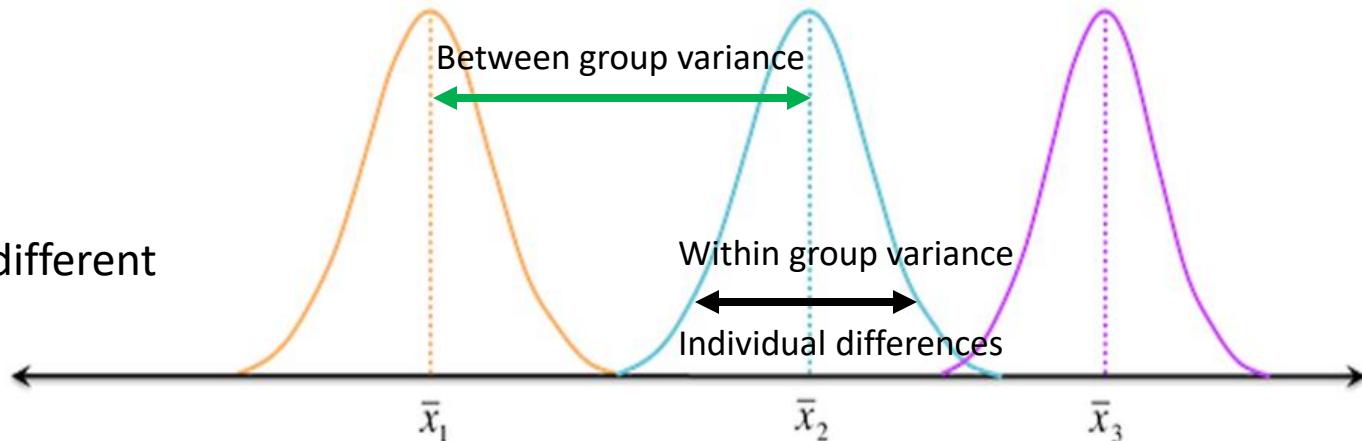
$\bar{\bar{X}}$ = the grand mean

N = the number of subjects total

Group/Treatments are less likely to be significantly different



Groups/Treatments are more likely to be significantly different



Calculating variances

Source	SS	df	s^2 (MS)	F
Between treatments	$\sum n_i (\bar{X}_i - \bar{\bar{X}})^2$	$k-1$	$\frac{SS_b}{df_b}$	$F = \frac{MSB}{MSW}$
Within treatments	$\sum (X_{ij} - \bar{X}_i)^2$	$N-k$	$\frac{SS_w}{df_w}$	
Total	$\sum (X_{ij} - \bar{\bar{X}})^2$	$N-1$		

groups	scores	means	diff	diff_squared
T1	20	11	4	16
T1	11	11	4	16
T1	2	11	4	16
T2	6	5	-2	4
T2	2	5	-2	4
T2	7	5	-2	4
T3	2	5	-2	4
T3	11	5	-2	4
T3	2	5	-2	4
Sums	63	63	0	72
Means	7	7	0	8

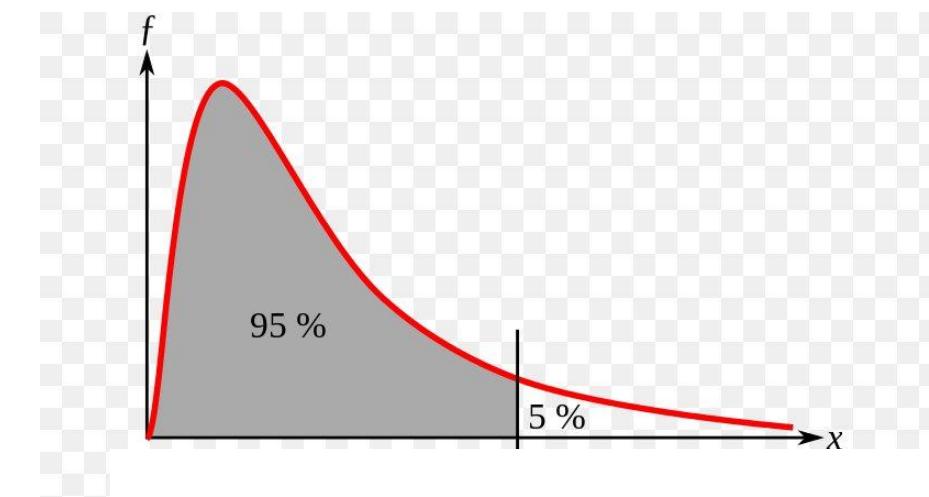
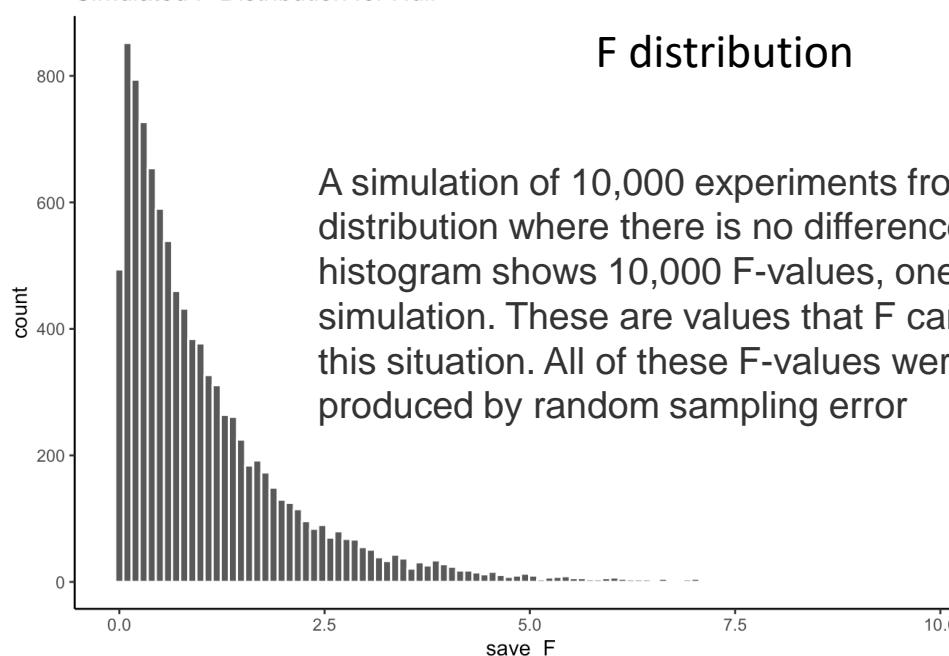
SSbetween

groups	scores	diff	diff_squared
T1	20	13	169
T1	11	4	16
T1	2	-5	25
T2	6	-1	1
T2	2	-5	25
T2	7	0	0
T3	2	-5	25
T3	11	4	16
T3	2	-5	25
Sums	63	0	302
Means	7	0	33.556

SStotal = SSbetween + SSwithin

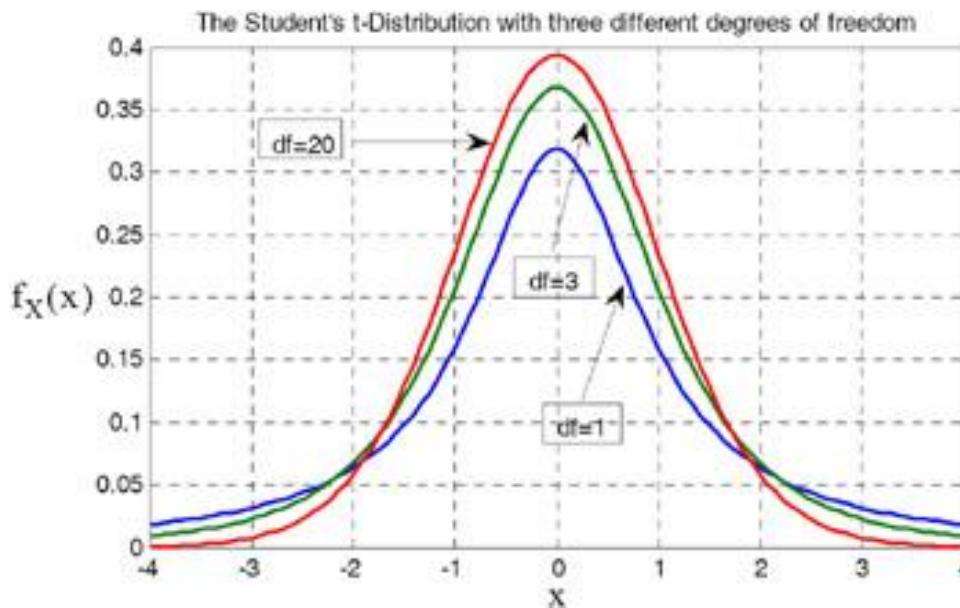
groups	scores	means	diff	diff_squared
T1	20	11	-9	81
T1	11	11	0	0
T1	2	11	9	81
T2	6	5	-1	1
T2	2	5	3	9
T2	7	5	-2	4
T3	2	5	3	9
T3	11	5	-6	36
T3	2	5	3	9
Sums	63	63	0	230
Means	7	7	0	25.556

SSwithin



Why is F distribution positively skewed?

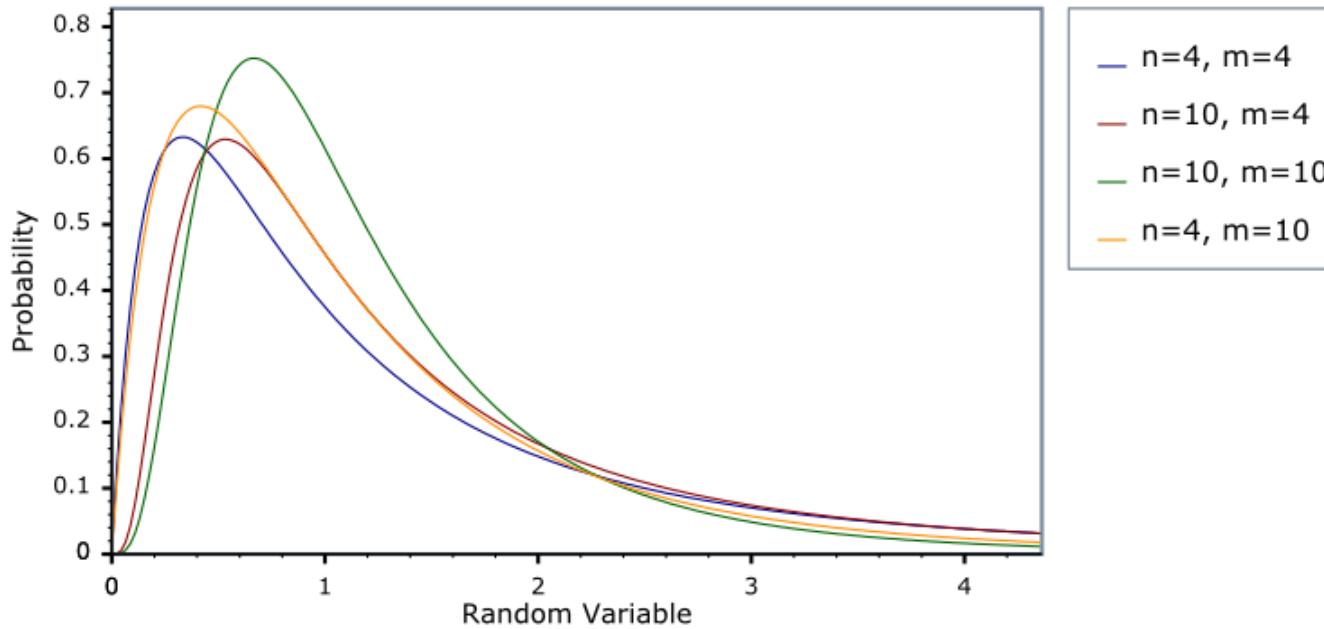
F distribution has only one tail – can only tell whether there is difference between the groups or not. Does not tell which group is better or worse.



t distribution
(can be one tailed or two-tailed)

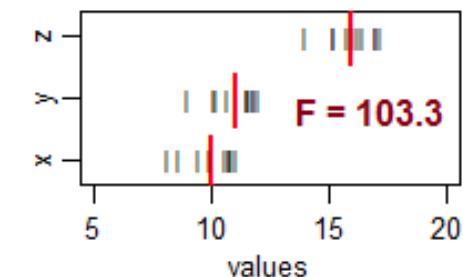
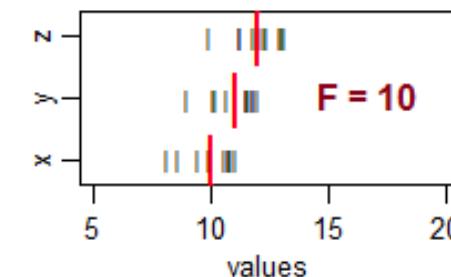
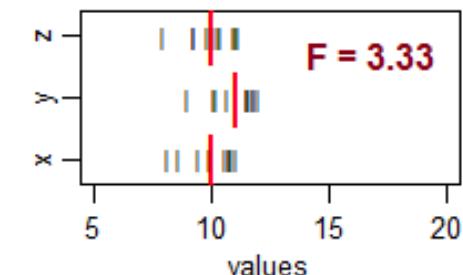
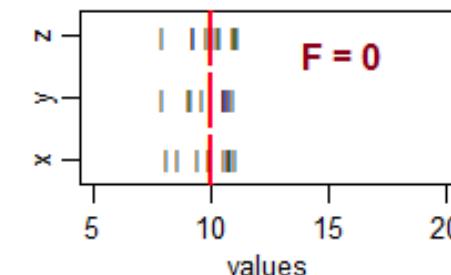
$$F = t^2$$

F Distribution PDF



$F < 1$ – no effect

$F > 1$ – there might be an effect

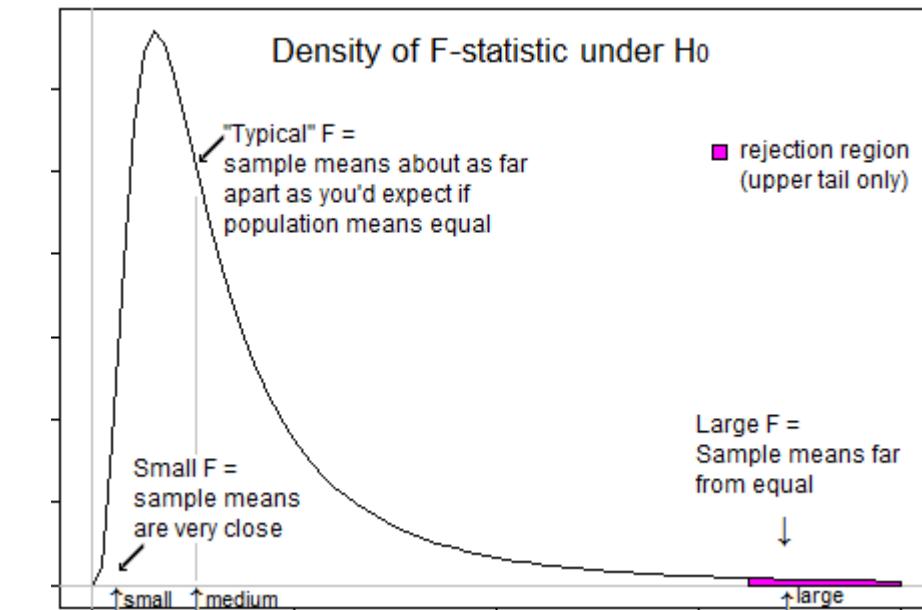


Groups	Count	Sum	Average	Variance
T1	3	33	11	81
T2	3	15	5	7
T3	3	15	5	27

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	72	2	36	0.93913	0.441736	5.143253
Within Groups	230	6	38.33333			
Total	302	8				

$$F = \frac{SS_B}{SS_W}$$

$$SS_W$$



ANOVA					
Source of Variation	SS	df	MS	F	P-value
Between Groups	72	2	36	0.93913	0.441736
Within Groups	230	6	38.33333		
Total	302	8			

$$F = \frac{SSB}{SSW}$$

of the F Distribution

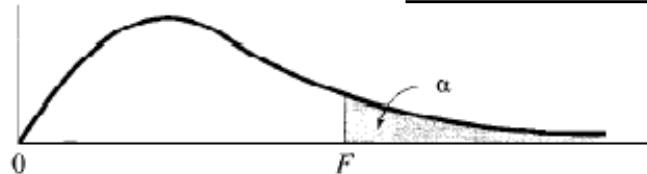


Table 1 $\alpha = 0.05$

Degrees of Freedom for Denominator	Degrees of Freedom for Numerator															
	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50
1	161.4	199.5	215.8	224.8	230.0	233.8	236.5	238.6	240.1	242.1	245.2	248.4	248.9	250.5	250.8	252.6
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.44	19.46	19.47	19.48	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.58
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.70
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.44
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.75
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.32
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.02
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.80
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.64
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.51
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.40
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.31
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.24
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.18
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.23	2.19	2.15	2.12
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.18	2.15	2.10	2.08
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.14	2.11	2.06	2.04
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.11	2.07	2.03	2.00
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.97
21	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	2.02	1.98	1.94	1.91
22	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.97	1.94	1.89	1.86

One way ANOVA showed that type of treatment had no effect on the level of anxiety $F_{(2,6)} = 0.93$, $p=0.44$

$F < 1$ (no effect)



Just another Example!

One way (One factor, One IV) ANOVA

H₀ – exam performance not affected by type of schooling

FAKE DATA

Exam performance		
Home school	Boarding school	Regular Day school
89	85	91
75	78	88
49	59	84
87	77	81
84	63	91
68	88	75
88	71	69
78	73	93
77	69	95
93	80	85
67	72	87
79	68	84
69	66	83
88	59	80
91	70	77

H₁ – Type of schooling affects exam performance

Groups	Count	Sum	Average	Variance
Home school	15	1182	78.8	141.1714
Boarding school	15	1078	71.86667	73.98095
Regular Day school	15	1263	84.2	50.45714

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1146.711	2	573.3556	6.475922	0.003537	3.219942
Within Groups	3718.533	42	88.53651			
Total	4865.244	44				

$$F_{(2,42)} = 6.47, p < 0.01$$

or

$$F_{(2,42)} = 6.47, p = .003$$

Effect size for ANOVA

$$\eta^2 = \frac{SS_{Between}}{SS_{Total}}$$

Eta-squared

$$= \frac{1146.711}{4865.244} = 0.236$$

$$F_{(2,42)} = 6.47, p=.003, \eta^2 = .24$$

Type of schooling explains 24% of variance in exam performance

Table I Values of Effect Sizes and Their Interpretation

Kind of Effect Size	Small	Medium	Large
r	.10	.30	.50
d	0.20	0.50	0.80
η^2_p	.01	.06	.14
f^2	.02	.15	.35

Source: Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159. doi:10.1037/0033-2909.112.1.155

We know there is difference between the groups,
but which groups perform better or worse?

Why not just perform multiple t-tests?

To avoid Type I error – false positive

- For 'k' independent groups there are $k(k-1)/2$ possible t-tests
 - For 5 groups = $5(5-1)/2= 10$ t-tests
 - For 4 groups = $4(4-1)/2= 6$ t-tests
 - For 3 groups = $3(3-1)/2= 3$ t-tests
- Using too many t-test comparisons increases the chances of finding random significant effects which may be due to chance. In reality there may be no difference between the groups/conditions

SOLUTION?

Hypothesis Driven
(like a one-tailed test)

Option 1 (Planned Contrasts): Pre-planned, therefore limited no. of comparisons.
**You are not comparing all groups to one another, very specific comparisons
(so the risk of Type 1 error is controlled)**

Exploratory
(like a two-tailed test)

Option 2 (Post Hoc tests) – All possible comparisons can be using special tests (to avoid Type I error).

Planned Comparison

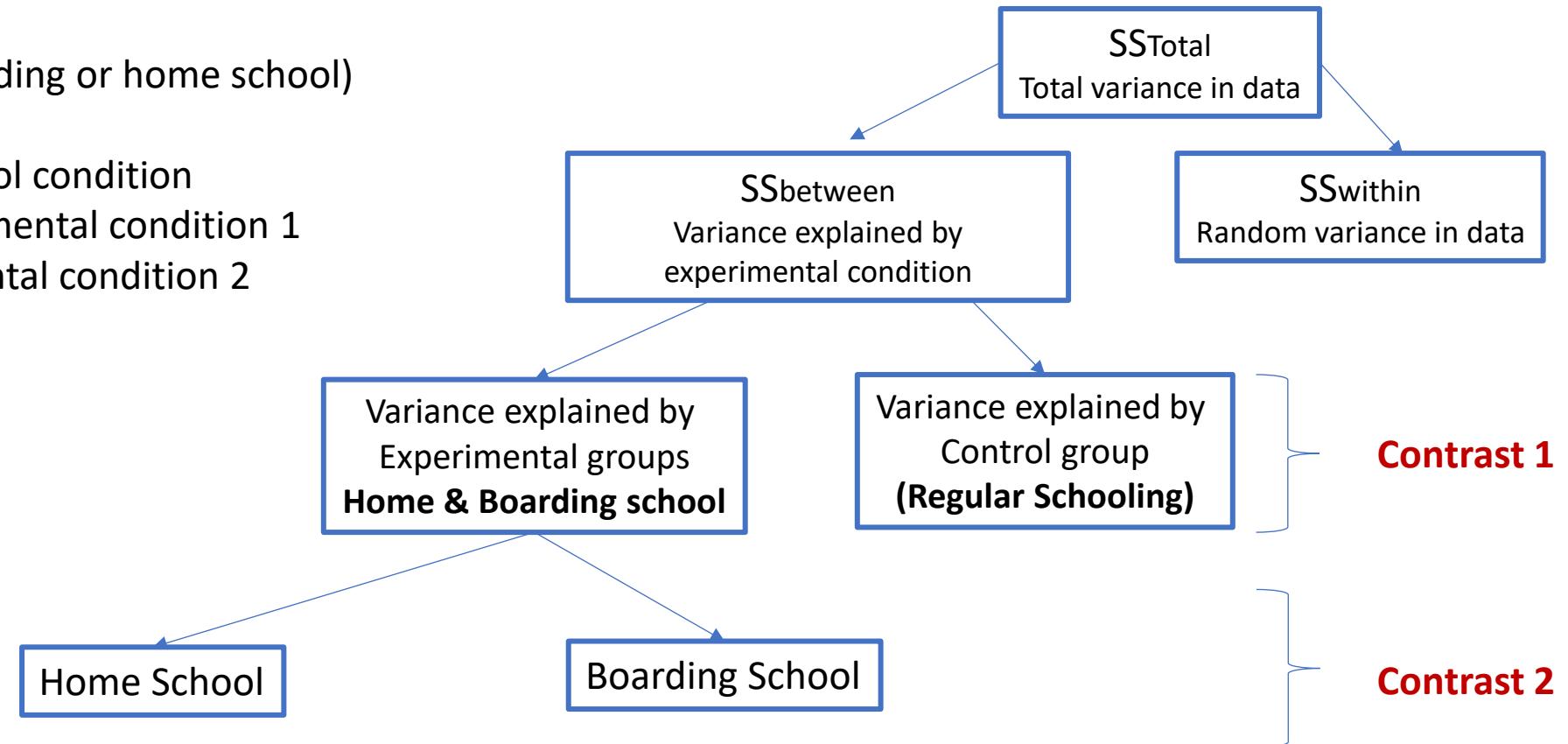
Planned comparison (contrast) – prior to experiment (based on the literature)

Regular schooling > (boarding or home school)

Regular schooling – control condition

Boarding school – Experimental condition 1

Home school – Experimental condition 2



But as the no. of planned comparisons increase (>2 comparisons), the alpha level has to adjusted/Bonferroni correction, again to avoid Type I error.

Bonferroni correction

Adjust the α level by the no. of comparisons

For 3 comparisons, $\alpha/3 = 0.05/3 = 0.0167 \sim 0.01$

Conduct 3 t-tests with $\alpha = 0.01$

Groupwise comparisons	T-test	Bonferroni correction corrected p value
Home vs Boarding	0.07780999	0.016666667
Boarding vs Regular	0.00019644	
Regular vs Home	0.14204177	

Good for small no. of comparisons, else risk of Type II error

Post-hoc test (Tukey's)

Tukey's test requires that the sample size, n , be the same for all treatments.

Groups	Count	Sum	Average	Variance
Home school	15	1182	78.8	141.1714
Boarding school	15	1078	71.86667	73.98095
Regular Day school	15	1263	84.2	50.45714

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1146.711	2	573.3556	6.475922	0.003537	3.219942
Within Groups	3718.533	42	88.53651			
Total	4865.244	44				

$$\text{Tukey's } HSD = q \sqrt{\frac{MS_{\text{within}}}{n}} = 3.44 \sqrt{(88.53/15)} = 6.95$$

q – studentized range statistic

The mean difference between any two samples must be more than 6.95 (at $\alpha=0.05$) to be significant.

$$M_{\text{Home}} - M_{\text{Regular}} = |78.8 - 84.2| = 5.4 \text{ (not significantly different)}$$

$$M_{\text{Boarding}} - M_{\text{Regular}} = |71.866 - 84.2| = 12.33 \text{ (significantly different)}$$

$$M_{\text{Home}} - M_{\text{Boarding}} = |78.8 - 71.87| = 6.93 \text{ (not significantly different)}$$

TABLE B.5 The Studentized Range Statistic (q)*

*The critical values for q corresponding to $\alpha = .05$ (lightface type) and $\alpha = .01$ (boldface type).

df for Error Term	k = Number of Treatments										
	2	3	4	5	6	7	8	9	10	11	12
5	3.64 5.70	4.60 6.98	5.22 7.80	5.67 8.42	6.03 8.91	6.33 9.32	6.58 9.67	6.80 9.97	6.99 10.24	7.17 10.48	7.32 10.70
6	3.46 5.24	4.34 6.33	4.90 7.03	5.30 7.56	5.63 7.97	5.90 8.32	6.12 8.61	6.32 8.87	6.49 9.10	6.65 9.30	6.79 9.48
7	3.34 4.95	4.16 5.92	4.68 6.54	5.06 7.01	5.36 7.37	5.61 7.68	5.82 7.94	6.00 8.17	6.16 8.37	6.30 8.55	6.43 8.71
8	3.26 4.75	4.04 5.64	4.53 6.20	4.89 6.62	5.17 6.96	5.40 7.24	5.60 7.47	5.77 7.68	5.92 7.86	6.05 8.03	6.18 8.18
9	3.20 4.60	3.95 5.43	4.41 5.96	4.76 6.35	5.02 6.66	5.24 6.91	5.43 7.13	5.59 7.33	5.74 7.49	5.87 7.65	5.98 7.78
10	3.15 4.48	3.88 5.27	4.33 5.77	4.65 6.14	4.91 6.43	5.12 6.67	5.30 6.87	5.46 7.05	5.60 7.21	5.72 7.36	5.83 7.49
11	3.11 4.39	3.82 5.15	4.26 5.62	4.57 5.97	4.82 6.25	5.03 6.48	5.20 6.67	5.35 6.84	5.49 6.99	5.61 7.13	5.71 7.25
12	3.08 4.32	3.77 5.05	4.20 5.50	4.51 5.84	4.75 6.10	4.95 6.32	5.12 6.51	5.27 6.67	5.39 6.81	5.51 6.94	5.61 7.06
13	3.06 4.26	3.73 4.96	4.15 5.40	4.45 5.73	4.69 5.98	4.88 6.19	5.05 6.37	5.19 6.53	5.32 6.67	5.43 6.79	5.53 6.90
14	3.03 4.21	3.70 4.89	4.11 5.32	4.41 5.63	4.64 5.88	4.83 6.08	4.99 6.26	5.13 6.41	5.25 6.54	5.36 6.66	5.46 6.77
15	3.01 4.17	3.67 4.84	4.08 5.25	4.37 5.56	4.59 5.80	4.78 5.99	4.94 6.16	5.08 6.31	5.20 6.44	5.31 6.55	5.40 6.66
16	3.00 4.13	3.65 4.79	4.05 5.19	4.33 5.49	4.56 5.72	4.74 5.92	4.90 6.08	5.03 6.22	5.15 6.35	5.26 6.46	5.35 6.56
17	2.98 4.10	3.63 4.74	4.02 5.14	4.30 5.43	4.52 5.66	4.70 5.85	4.86 6.01	4.99 6.15	5.11 6.27	5.21 6.38	5.31 6.48
18	2.97 4.07	3.61 4.70	4.00 5.09	4.28 5.38	4.49 5.60	4.67 5.79	4.82 5.94	4.96 6.08	5.07 6.20	5.17 6.31	5.27 6.41
19	2.96 4.05	3.59 4.67	3.98 5.05	4.25 5.33	4.47 5.55	4.65 5.73	4.79 5.89	4.92 6.02	5.04 6.14	5.14 6.25	5.23 6.34
20	2.95 4.02	3.58 4.64	3.96 5.02	4.23 5.29	4.45 5.51	4.62 5.69	4.77 5.84	4.90 5.97	5.01 6.09	5.11 6.19	5.20 6.28
24	2.92 3.96	3.53 4.55	3.90 4.91	4.17 5.17	4.37 5.37	4.54 5.54	4.68 5.69	4.81 5.81	4.92 5.92	5.01 6.02	5.10 6.11
30	2.89 3.89	3.49 4.45	3.85 4.80	4.10 5.05	4.30 5.24	4.46 5.40	4.60 5.54	4.72 5.65	4.82 5.76	4.92 5.85	5.00 5.93
40	2.86 3.82	3.44 4.37	3.79 4.70	4.04 4.93	4.23 5.11	4.39 5.26	4.52 5.39	4.63 5.50	4.73 5.60	4.82 5.69	4.90 5.76
60	2.83 3.76	3.40 4.28	3.74 4.59	3.98 4.82	4.16 4.99	4.31 5.13	4.44 5.25	4.55 5.36	4.65 5.45	4.73 5.53	4.81 5.60
120	2.80 3.70	3.36 4.20	3.68 4.50	3.92 4.71	4.10 4.87	4.24 5.01	4.36 5.12	4.47 5.21	4.56 5.30	4.64 5.37	4.71 5.44
∞	2.77 3.64	3.31 4.12	3.63 4.40	3.86 4.60	4.03 4.76	4.17 4.88	4.28 4.99	4.39 5.08	4.47 5.16	4.55 5.23	4.62 5.29

Other post-hoc tests

Games-Howell Test

For unequal variance (unequal sample size)

Calculations are the same as Tukey's but df is calculated using the formula used for unequal sample t-test (Slide 1)

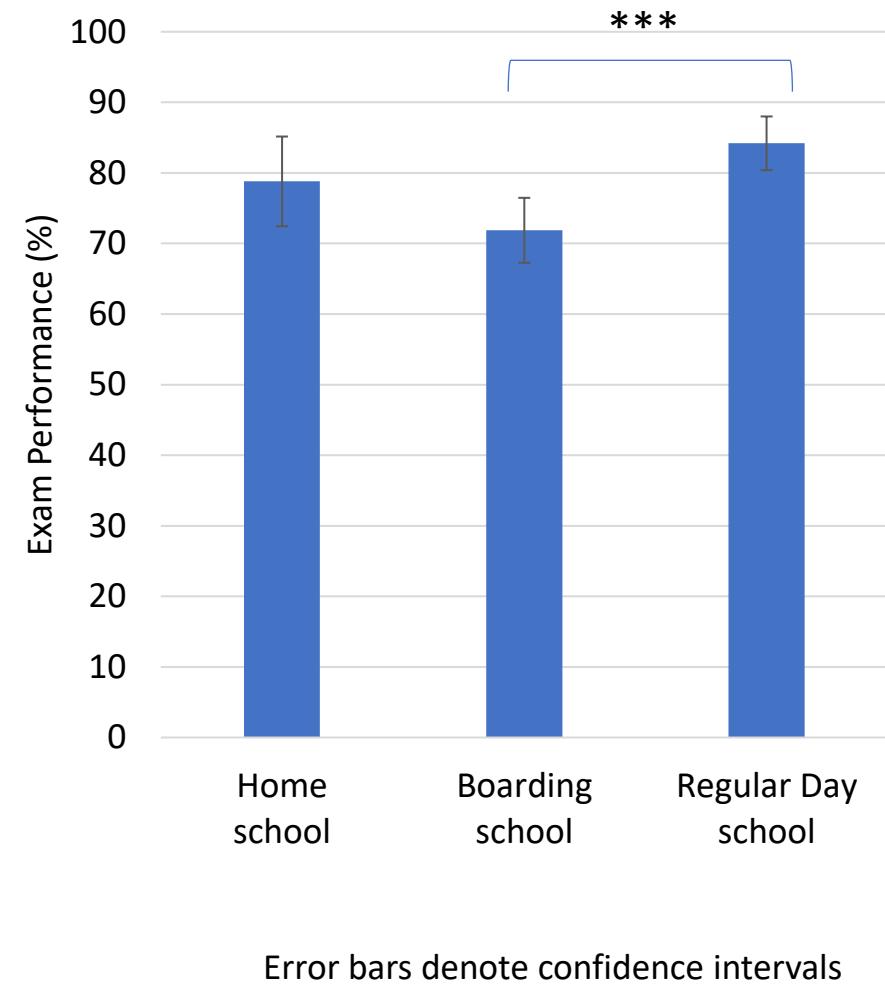
For unequal sample size

$$\text{degrees of freedom, } df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2} \right)^2}$$

Reporting results

- Using a one way ANOVA we observed that the schooling method has a significant effect on exam performance $F_{(2,42)} = 6.47$, $p=.003$, $\eta^2 = .24$
- Using Bonferroni post-hoc test, we found that regular school resulted in better exam performance than boarding school ($p<.001$). There was no significant difference between the other groups.

* $p<0.05$
** $p<0.01$
*** $p<0.001$



ANOVA assumptions

1. The populations from which the samples are selected must be normal
(parametric vs non-parametric)

– Shapiro-Wilk test / Kolmogorov-Smirnov test

If violated – use Kruskal –Wallis Test

Typically for n>25 in each group, normality can be overlooked in ANOVA

2. The populations from which the samples are selected must have equal variances* (**homogeneity of variance**).

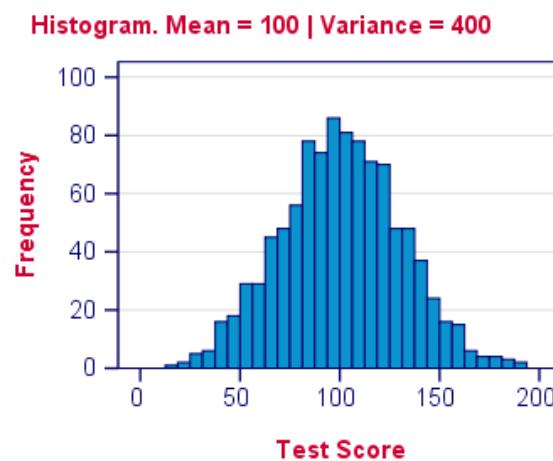
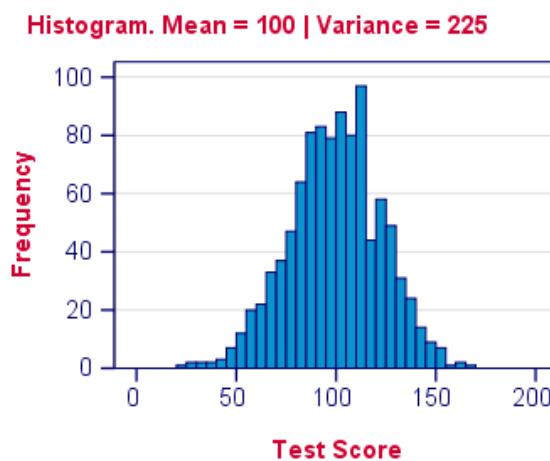
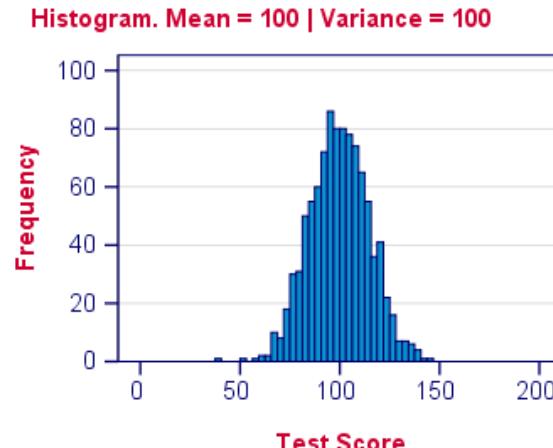
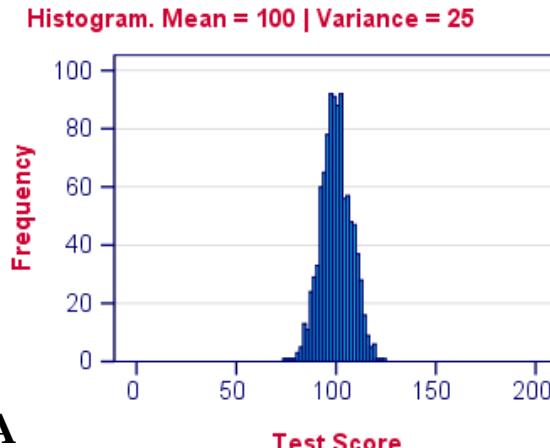
– Levene's or Hartley's F-max test for homogeneity of variance

If violated, solution -

1. Collect more samples and equate samples in all groups.
2. Data transformations - natural log or square root transformations

Consequences -- once you transform a variable and conduct your analysis, you can only interpret the transformed variable. You cannot provide an interpretation of the results based on the untransformed variable values.

NORMAL DISTRIBUTIONS WITH SIMILAR MEANS, DIFFERENT VARIANCES.



ANOVA is a robust statistical test, slight violations of assumptions has minor effects on the test outcomes. As long as the largest variance < 4-5 times smallest variance, ANOVA results are valid

Homogeneity of variances (homoscedasticity)

$$F\text{-max} = \frac{s^2(\text{largest})}{s^2(\text{smallest})}$$

$F\text{-max} \sim 1.00 \rightarrow$ sample variances are similar and homogenous

Look up $df=(n-1)$, k in the Fmax table

If your calculated value < table value, the variance is homogeneous.

If your calculated value > table value, the variance is not homogeneous.

TABLE B.3 Critical Values for the F-Max Statistic*

*The critical values for $\alpha = .05$ are in lightface type, and for $\alpha = .01$, they are in boldface type.

$n - 1$	$k = \text{Number of Samples}$										
	2	3	4	5	6	7	8	9	10	11	12
4	9.60	15.5	20.6	25.2	29.5	33.6	37.5	41.4	44.6	48.0	51.4
	23.2	37.	49.	59.	69.	79.	89.	97.	106.	113.	120.
5	7.15	10.8	13.7	16.3	18.7	20.8	22.9	24.7	26.5	28.2	29.9
	14.9	22.	28.	33.	38.	42.	46.	50.	54.	57.	60.
6	5.82	8.38	10.4	12.1	13.7	15.0	16.3	17.5	18.6	19.7	20.7
	11.1	15.5	19.1	22.	25.	27.	30.	32.	34.	36.	37.
7	4.99	6.94	8.44	9.70	10.8	11.8	12.7	13.5	14.3	15.1	15.8
	8.89	12.1	14.5	16.5	18.4	20.	22.	23.	24.	26.	27.
8	4.43	6.00	7.18	8.12	9.03	9.78	10.5	11.1	11.7	12.2	12.7
	7.50	9.9	11.7	13.2	14.5	15.8	16.9	17.9	18.9	19.8	21.
9	4.03	5.34	6.31	7.11	7.80	8.41	8.95	9.45	9.91	10.3	10.7
	6.54	8.5	9.9	11.1	12.1	13.1	13.9	14.7	15.3	16.0	16.6
10	3.72	4.85	5.67	6.34	6.92	7.42	7.87	8.28	8.66	9.01	9.34
	5.85	7.4	8.6	9.6	10.4	11.1	11.8	12.4	12.9	13.4	13.9
12	3.28	4.16	4.79	5.30	5.72	6.09	6.42	6.72	7.00	7.25	7.48
	4.91	6.1	6.9	7.6	8.2	8.7	9.1	9.5	9.9	10.2	10.6
15	2.86	3.54	4.01	4.37	4.68	4.95	5.19	5.40	5.59	5.77	5.93
	4.07	4.9	5.5	6.0	6.4	6.7	7.1	7.3	7.5	7.8	8.0
20	2.46	2.95	3.29	3.54	3.76	3.94	4.10	4.24	4.37	4.49	4.59
	3.32	3.8	4.3	4.6	4.9	5.1	5.3	5.5	5.6	5.8	5.9
30	2.07	2.40	2.61	2.78	2.91	3.02	3.12	3.21	3.29	3.36	3.39
	2.63	3.0	3.3	3.5	3.6	3.7	3.8	3.9	4.0	4.1	4.2
60	1.67	1.85	1.96	2.04	2.11	2.17	2.22	2.26	2.30	2.33	2.36
	1.96	2.2	2.3	2.4	2.4	2.5	2.5	2.6	2.6	2.7	2.7

Levene's Test (more robust)

$$W = \frac{(N - k)}{(k - 1)} \cdot \frac{\sum_{i=1}^k N_i (Z_{i\cdot} - Z_{..})^2}{\sum_{i=1}^k \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2}$$

where

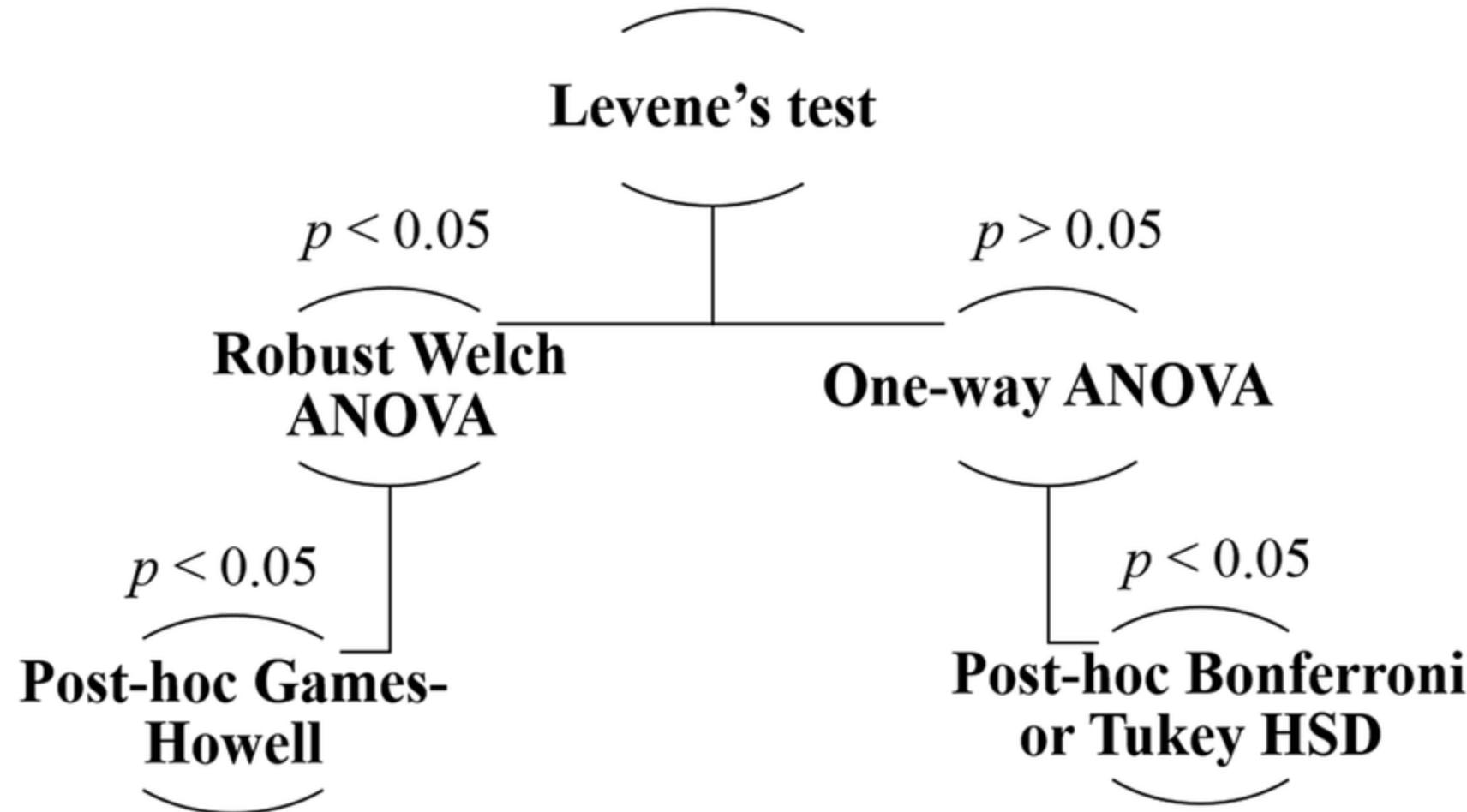
- k is the number of different groups to which the sampled cases belong,
- N_i is the number of cases in the i th group,
- N is the total number of cases in all groups,
- Z_{ij} is the value of the measured variable for the j th case from the i th group,

• $Z_{ij} = |Y_{ij} - \bar{Y}_{i\cdot}|$, $\bar{Y}_{i\cdot}$ is a mean of the i -th group,

• $Z_{i\cdot} = \frac{1}{N_i} \sum_{j=1}^{N_i} Z_{ij}$ is the mean of the Z_{ij} for group i ,

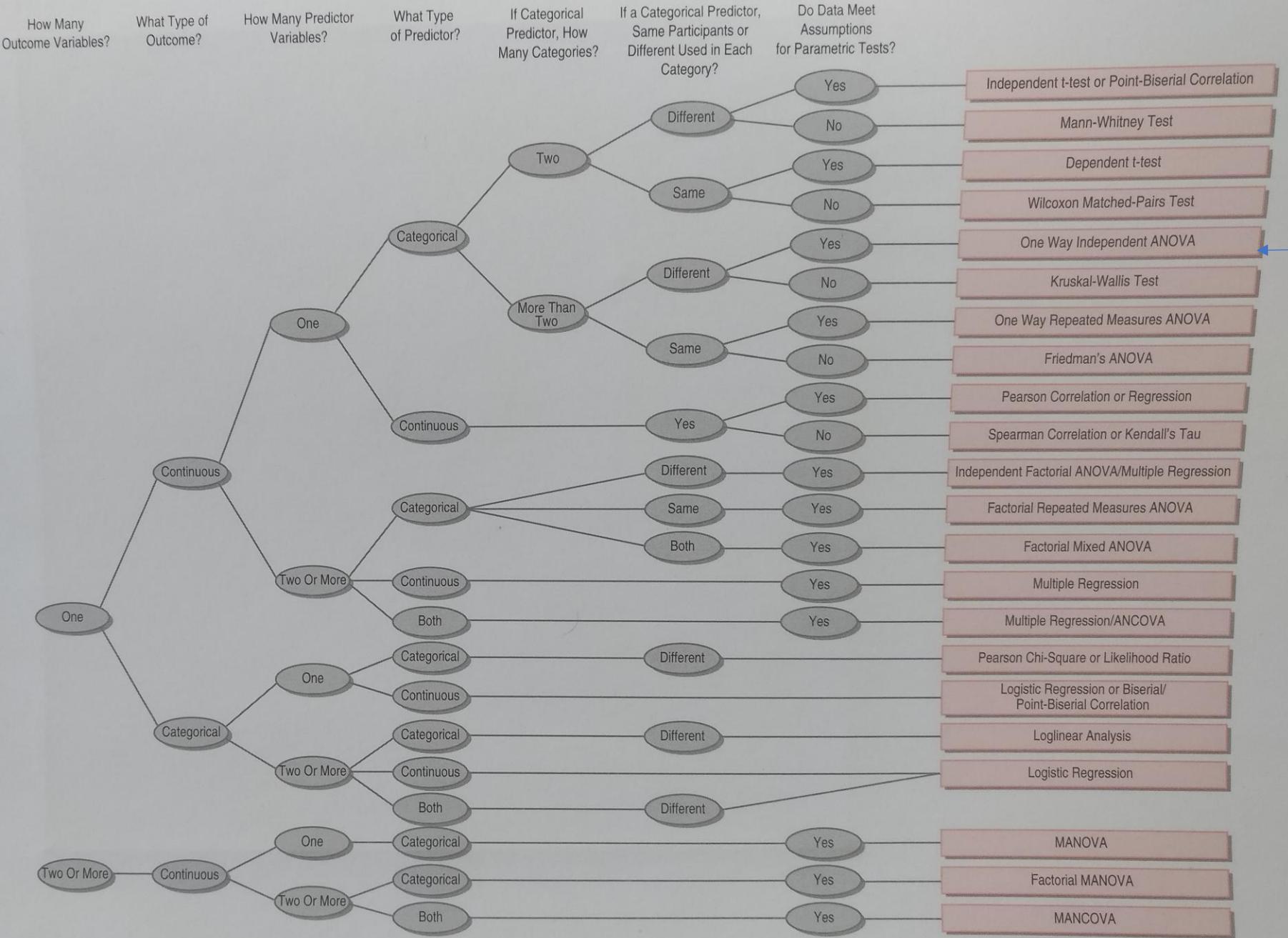
• $Z_{..} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{N_i} Z_{ij}$ is the mean of all Z_{ij} .

<https://www.youtube.com/watch?v=gL7K4vZq0Z4>



Steps in ANOVA

- Check for normality (parametric vs non-parametric tests)
- Check homogeneity of variances (spread of the data is equal across groups or not)
- Choose the appropriate test (ANOVA, Kruskal-Wallis, Welch)
- Only if main effect (F) significant, use a post-hoc test
- Report effect size for significant effects
- Plot analyzed data





Just another Example!

One way (One factor, One IV) ANOVA

Testing homogeneity

H_0 – variance across groups is equal

H_1 – variance across groups is unequal

Exam performance		
Home school	Boarding school	Regular Day school
89	85	91
75	78	88
49	59	84
87	77	81
84	63	91
68	88	75
88	71	69
78	73	93
77	69	95
93	80	85
67	72	87
79	68	84
69	66	83
88	59	80
91	70	77

Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
exam_performance	Based on Mean	1.675	2	42	.200
	Based on Median	1.648	2	42	.205
	Based on Median and with adjusted df	1.648	2	35.806	.207
	Based on trimmed mean	1.690	2	42	.197

Tests of Normality

	school_code	Kolmogorov-Smirnov ^a			<50 samples Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
exam_performance	home school	.155	15	.200*	.905	15	.115
	boarding school	.114	15	.200*	.969	15	.847
	regular scchool	.100	15	.200*	.975	15	.923

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

(When normality is violated)

Kruskal-Wallis Test

$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} \right) - 3(N+1)$$

Where, N = Total observation in all groups (total sample size); k = Number of groups; n_j = sample size for j th group, and R_j is the sum of ranks of j th group

(a) Original Numerical Scores

I	II	III	$N = 15$
14	2	26	
3	14	8	
21	9	14	
5	12	19	
16	5	20	
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	

(b) Ordinal Data (Ranks)

I	II	III	$N = 15$
9	1	15	
2	9	5	
14	6	9	
3.5	7	12	
11	3.5	13	

$$\begin{aligned} T_1 &= 39.5 & T_2 &= 26.5 & T_3 &= 54 \\ n_1 &= 5 & n_2 &= 5 & n_3 &= 5 \end{aligned}$$

$$H = \frac{12}{N(N+1)} \left(\sum \frac{T^2}{n} \right) - 3(N+1)$$

$$\begin{aligned} H &= \frac{12}{15(16)} \left(\frac{39.5^2}{5} + \frac{26.5^2}{5} + \frac{54^2}{5} \right) - 3(16) \\ &= 0.05(312.05 + 140.45 + 583.2) - 48 \\ &= 0.05(1035.7) - 48 \\ &= 51.785 - 48 \\ &= 3.785 \end{aligned}$$

Use chi-square distribution

With $df = (k-1) = 2$

$H_{critical} = 5.99$ for $\alpha = .05$

$H (3.785) < 5.99$
Accept H_0 .

Since the data were not normally distributed, Kruskal-Wallis test for non-parametric data was used to evaluate differences among the three treatments. The outcome of the test indicated no significant differences among the treatment conditions, $H = 3.785 (2, N = 15), p > .05$.

(When homogeneity is violated)

Welch ANOVA Test

$$F = \frac{\frac{1}{k-1} \sum_{j=1}^k w_j (\bar{x}_j - \bar{x}')^2}{\frac{1}{k^2-1} \sum_{j=1}^k \left(\frac{1}{n_j-1}\right) \left(1 - \frac{w_j}{w}\right)^2}$$

$$w_j = \frac{n_j}{s_j^2} \quad w = \sum_{j=1}^k w_j \quad \bar{x}' = \frac{\sum_{j=1}^k w_j \bar{x}_j}{w}$$

$$F \sim F(k-1, df)$$

$$df = \frac{k^2 - 1}{3 \sum_{j=1}^k \left(\frac{1}{n_j-1}\right) \left(1 - \frac{w_j}{w}\right)^2}$$

Welch ANOVA est. $\omega^2 = \frac{df_{bet}(F-1)}{df_{bet}(F-1) + N_T}$

Robust Tests of Equality of Means

exam_performance

	Statistic ^a	df1	df2	Sig.
Welch	8.954	2	26.995	.001

a. Asymptotically F distributed.

The degrees of freedom for Welch's t-test takes into account the difference between the two standard deviations.

df with decimal places → round off to look up in the F table

One way Repeated Measures ANOVA

Advantage?

Disadvantage?

Experimental
Design is equally
important

subjects	T1 counseling	T2 Anti-anxiety meds	T3 both
1	20	11	2
2	6	2	7
3	2	11	2

subjects	T1 counseling	T2 counselling	T3 counselling
1	20	11	2
2	6	2	7
3	2	11	2

One way Repeated Measures ANOVA

$$SS_{total} = SS_{between} + SS_{within}$$



$$SS_{Total} = SS_{between} + SS_{subjects} + SS_{error}$$

	Dependent variable (DV)		
Participants	Timepoint 1	Timepoint 2	Timepoint 3
1	20	11	2
2	6	2	7
3	2	11	2

Source of variance	SS	df	MS	F
Between	$SS_{between}$	k-1	$MS_{between} = \frac{SS_{between}}{k-1}$	$F = \frac{MS_{between}}{MS_{error}}$
Within Subjects	SS_{within}			
	$SS_{subjects}$	n-1		
Error (left-over error)	$SS_{error} = SS_{within} - SS_{subjects}$	(k-1)(n-1)	$MS_{error} = \frac{SS_{error}}{(k-1)(n-1)}$	
Total	SS_{total}	N-1		

subjects	conditions	scores	means	diff	diff_squared
1	A	20	28/3	9.33	2.33
2	A	11	8	1	1
3	A	2	3.66	-3.34	11.1556
1	B	6	9.33	2.33	5.4289
2	B	2	8	1	1
3	B	7	3.66	-3.34	11.1556
1	C	2	9.33	2.33	5.4289
2	C	11	8	1	1
3	C	2	3.66	-3.34	11.1556
Sums		63	62.97	-0.0299	52.75
Means		7	6.997	-0.0033	5.8615

SSsubjects

$$SSwithin = SSsubjects + Serror$$

$$Serror = 230 - 52.75 = 177.25$$

subjects	conditions	scores	means	diff	diff_squared
1	A	20	11	4	16
2	A	11	11	4	16
3	A	2	11	4	16
1	B	6	5	-2	4
2	B	2	5	-2	4
3	B	7	5	-2	4
1	C	2	5	-2	4
2	C	11	5	-2	4
3	C	2	5	-2	4
Sums		63	63	0	72
Means		7	7	0	8

SSbetween

subjects	conditions	scores	diff	diff_squared
1	A	20	13	169
2	A	11	4	16
3	A	2	-5	25
1	B	6	-1	1
2	B	2	-5	25
3	B	7	0	0
1	C	2	-5	25
2	C	11	4	16
3	C	2	-5	25
Sums		63	0	302
Means		7	0	33.556

SStotal

subjects	conditions	scores	means	diff	diff_squared
1	A	20	11	-9	81
2	A	11	11	0	0
3	A	2	11	9	81
1	B	6	5	-1	1
2	B	2	5	3	9
3	B	7	5	-2	4
1	C	2	5	3	9
2	C	11	5	-6	36
3	C	2	5	3	9
Sums		63	63	0	230
Means		7	7	0	25.556

SSwithin

Tests of Within-Subjects Effects

Source of variance	SS	df	MS	F	p
Between	52.67	2	26.33	0.594	.59
Error (left-over error)	177.33	4	44.33		

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
factor1	Sphericity Assumed	52.667	2	26.333	.594	.594
	Greenhouse-Geisser	52.667	1.814	29.032	.594	.584
	Huynh-Feldt	52.667	2.000	26.333	.594	.594
	Lower-bound	52.667	1.000	52.667	.594	.521
Error(factor1)	Sphericity Assumed	177.333	4	44.333		
	Greenhouse-Geisser	177.333	3.628	48.877		
	Huynh-Feldt	177.333	4.000	44.333		
	Lower-bound	177.333	2.000	88.667		

Using a one way repeated measures ANOVA we observed that there was no difference in scores in the 3 timepoints $F_{(2,4)} = 0.594$, $p=.59$.

Mauchly's sphericity test

Sphericity → condition where the variances of related groups (levels T1, T2 , T3) are equal.

- Analogous to homogeneity of variances
- Used specifically in repeated measures testing

$$\text{Df}_{\text{between OR}} \quad df_{\text{time/condition}} = (k - 1)$$
$$df_{\text{error}} = (k - 1)(n - 1)$$

$$df_{\text{time/condition}} = \hat{\varepsilon}(k - 1)$$
$$df_{\text{error}} = \hat{\varepsilon}(k - 1)(n - 1)$$

homogenous

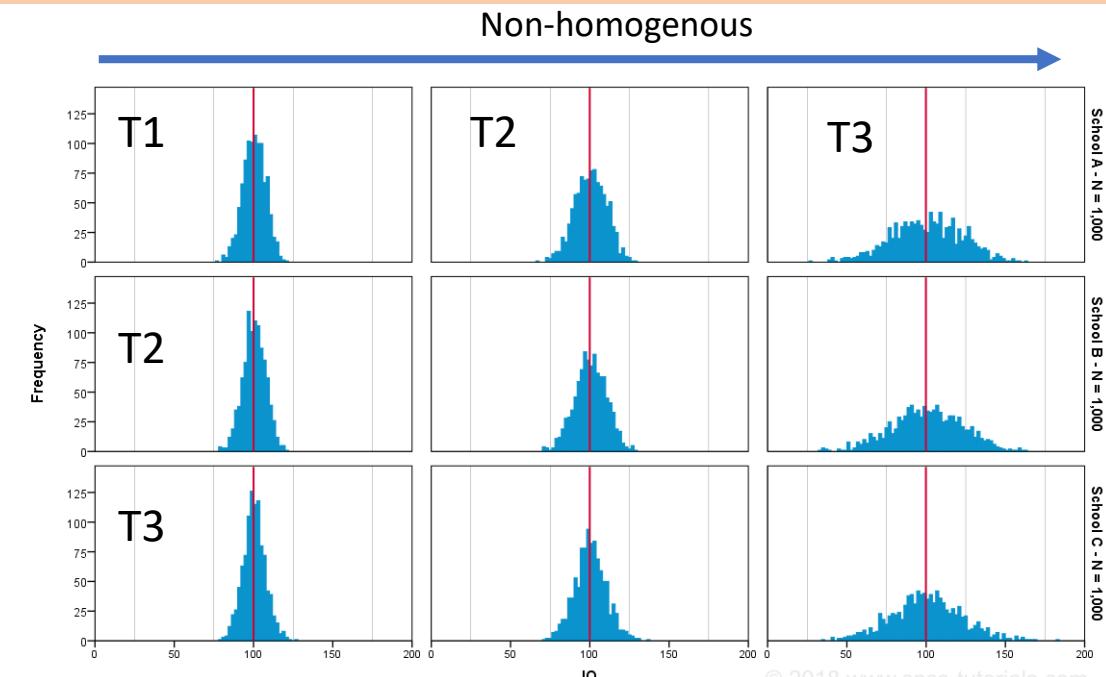
Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Huynh-Feldt	Lower-bound
factor1	.898	.108	2	.947	.907	1.000	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: factor1

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.



Mauchly's Test of Sphericity^a

Steps in ANOVA

- Check for normality
- Check homogeneity of variances (different participants in different groups, Factorial ANOVA)
- Check for sphericity of variances (same participants across groups, repeated measures ANOVA)
- Choose the appropriate test
- Only if main effect (F) significant, use a post-hoc test
- Report effect size for significant effects
- Plot analyzed data



Just another Example!

Test/Exam Scores

Student	Reread	Answer Prepared Questions	Create and Answer Questions
A	2	5	8
B	3	9	6
C	8	10	12
D	6	13	11
E	5	8	11
F	6	9	12

One way (One DV) repeated measures ANOVA

Mauchly's Test of Sphericity^a

Measure: test_score

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Epsilon ^b
factor1	.372	3.957	2	.138	.614	.712

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: factor1

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Tests of Within-Subjects Effects

Source	Measure: test_score					
	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
factor1	Sphericity Assumed	84.000	2	42.000	19.091	<.001
	Greenhouse-Geisser	84.000	1.228	68.380	19.091	.004
	Huynh-Feldt	84.000	1.424	58.991	19.091	.002
	Lower-bound	84.000	1.000	84.000	19.091	.007
Error(factor1)	Sphericity Assumed	22.000	10	2.200		
	Greenhouse-Geisser	22.000	6.142	3.582		
	Huynh-Feldt	22.000	7.120	3.090		
	Lower-bound	22.000	5.000	4.400		

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
reread	.176	6	.200*	.955	6	.783
answer_questions	.184	6	.200*	.957	6	.799
create_answer_question s	.325	6	.047	.827	6	.101

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

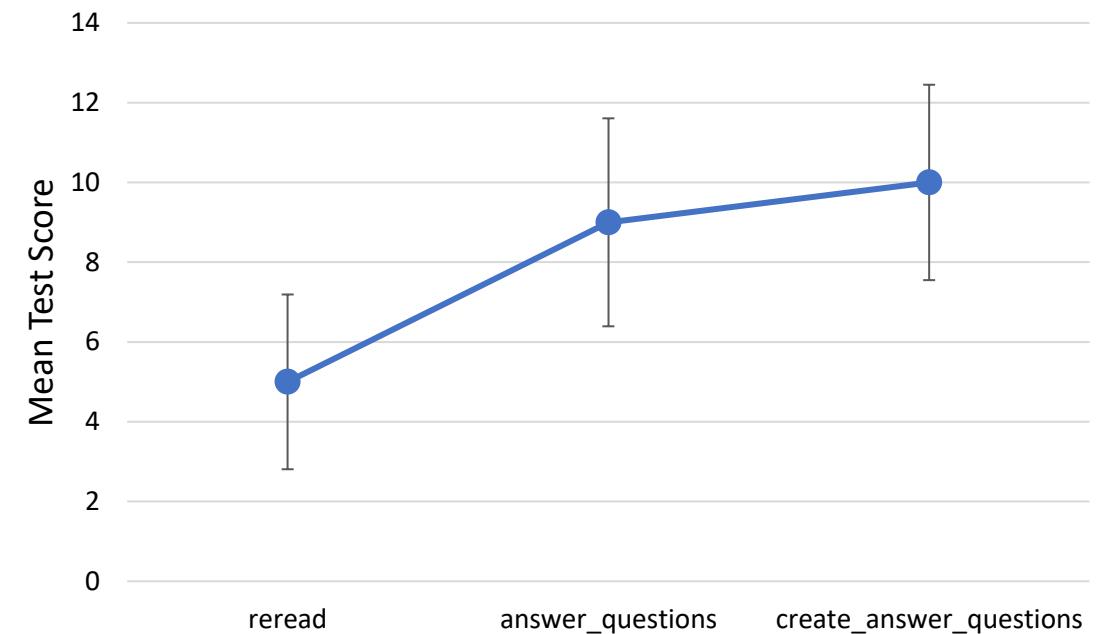
Using a one way repeated measures ANOVA we observed that strategy for studying in preparation for a test had an effect on exam score $F_{(2,10)} = 19.09$, $p < .001$, $\eta^2 = .79$

Tests of Within-Subjects Contrasts

Measure: test_score		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Source	factor1						
factor1	Linear	75.000	1	75.000	93.750	<.001	.949
	Quadratic	9.000	1	9.000	2.500	.175	.333
Error(factor1)	Linear	4.000	5	.800			
	Quadratic	18.000	5	3.600			

Using a one way repeated measures ANOVA we observed that strategy for studying in preparation for a test had an effect on exam score $F_{(2,10)} = 19.09$, $p < .001$, $\eta^2 = .79$

Within-subjects contrasts revealed that there was a linear trend $F_{(1,5)} = 93.75$, $p < .001$, $\eta^2 = .95$



Error bars denote standard deviations

Friedman's Test (non-normal repeated measures)

$$M = \frac{12}{Nk(k+1)} \sum R_i^2 - 3N(k+1)$$

Where, k = number of columns (treatments)

n = number of rows (blocks)

R_j = sum of the ranks

Related-Samples Friedman's Two-Way Analysis of Variance by Ranks Summary

Total N	6
Test Statistic	9.333
Degree Of Freedom	2
Asymptotic Sig.(2-sided test)	.009

Student	Reread	Answer Prepared Questions	Create and Answer Questions
A	2	5	8
B	3	9	6
C	8	10	12
D	6	13	11
E	5	8	11
F	6	9	12

Rank		
1	2	3
1	3	2
1	2	3
1	3	2
1	2	3
1	2	3
Sum = 6	14	16

		IV – categorical	DV – continuous (interval, ratio)
		Independent factor 1 IV > 2 groups	Dependent (Related) Samples 1 DV > 2 timepoints
Parametric (normal)	Homogeneous	One way ANOVA	Repeated measures ANOVA
	Non homogenous	Welch ANOVA	Sphericity correction
Non-parametric		<i>Kruskal-Wallis</i> ANOVA	Friedman's ANOVA

In class activity

Exam performance		
Home school	Boarding school	Regular Day school
89	85	91
75	78	88
49	59	84
87	77	81
84	63	91
68	88	75
88	71	69
78	73	93
77	69	95
93	80	85
67	72	87
79	68	84
69	66	83
88	59	80
91	70	77

- Check for normality
- Check homogeneity of variances (different participants in different groups, Factorial ANOVA)
- Check for sphericity of variances (same participants across groups, repeated measures ANOVA)
- Choose the appropriate test
 - Only if main effect (F) significant, use a post-hoc test
 - Report effect size for significant effects
 - Plot analyzed data



Just another Example!

One way (One factor, One IV) ANOVA

H₀ – exam performance not affected by type of schooling

FAKE DATA

Exam performance		
Home school	Boarding school	Regular Day school
89	85	91
75	78	88
49	59	84
87	77	81
84	63	91
68	88	75
88	71	69
78	73	93
77	69	95
93	80	85
67	72	87
79	68	84
69	66	83
88	59	80
91	70	77

H₁ – Type of schooling affects exam performance

Groups	Count	Sum	Average	Variance
Home school	15	1182	78.8	141.1714
Boarding school	15	1078	71.86667	73.98095
Regular Day school	15	1263	84.2	50.45714

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1146.711	2	573.3556	6.475922	0.003537	3.219942
Within Groups	3718.533	42	88.53651			
Total	4865.244	44				

$$F_{(2,42)} = 6.47, p < 0.01$$

or

$$F_{(2,42)} = 6.47, p = .003$$

- Check for normality
- Check homogeneity of variances (different participants in different groups)
- Choose the appropriate test
- Only if main effect (F) significant, use a post-hoc test
- Report effect size for significant effects (eta/partial eta squared)
- Plot analyzed data

Effect size for ANOVA

$$\eta^2 = \frac{SS_{Between}}{SS_{Total}}$$

Eta-squared

$$= \frac{1146.711}{4865.244} = 0.236$$

$$F_{(2,42)} = 6.47, p=.003, \eta^2 = .24$$

Type of schooling explains 24% of variance in exam performance

Table I Values of Effect Sizes and Their Interpretation

Kind of Effect Size	Small	Medium	Large
r	.10	.30	.50
d	0.20	0.50	0.80
η^2_p	.01	.06	.14
f^2	.02	.15	.35

Source: Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159. doi:10.1037/0033-2909.112.1.155

We know there is difference between the groups,
but which groups perform better or worse?

Factorial ANOVA

BRSM

Does time of day affect reaction time?

Independent Variables

(Factor)

Time of day

Morning, Afternoon, Evening and Night Labels

Morning



Afternoon



Evening



Night



Dependent Variables

(Response)

Reaction time

Which test?

One way ANOVA

Explanatory Variable(s)

We're interested in the relationship between one or more explanatory variables and a response variable.



Response Variable

Covariate

But there may exist some other variable (a *covariate*) that also affects the response variable.

Analysis of Covariance (ANCOVA)

Independent Variables

(Factor)

Time of day

(Covariate)

**Sleep
(hours slept the night before)**

(

Dependent Variable

(Response)

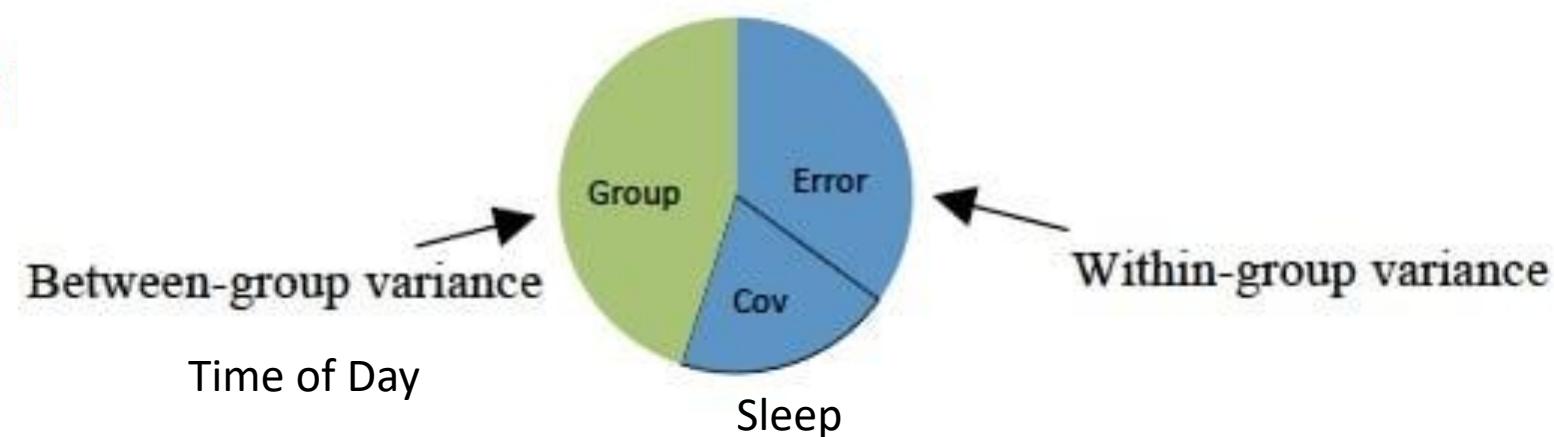
Reaction time



Advantages of ANCOVA

- Reduces Error Variance
 - By explaining some of the unexplained variance (SS_R) the error variance in the model can be reduced.
- Greater Experimental Control:
 - By holding known extraneous variables constant, we gain greater insight into the effect of the predictor variable(s).

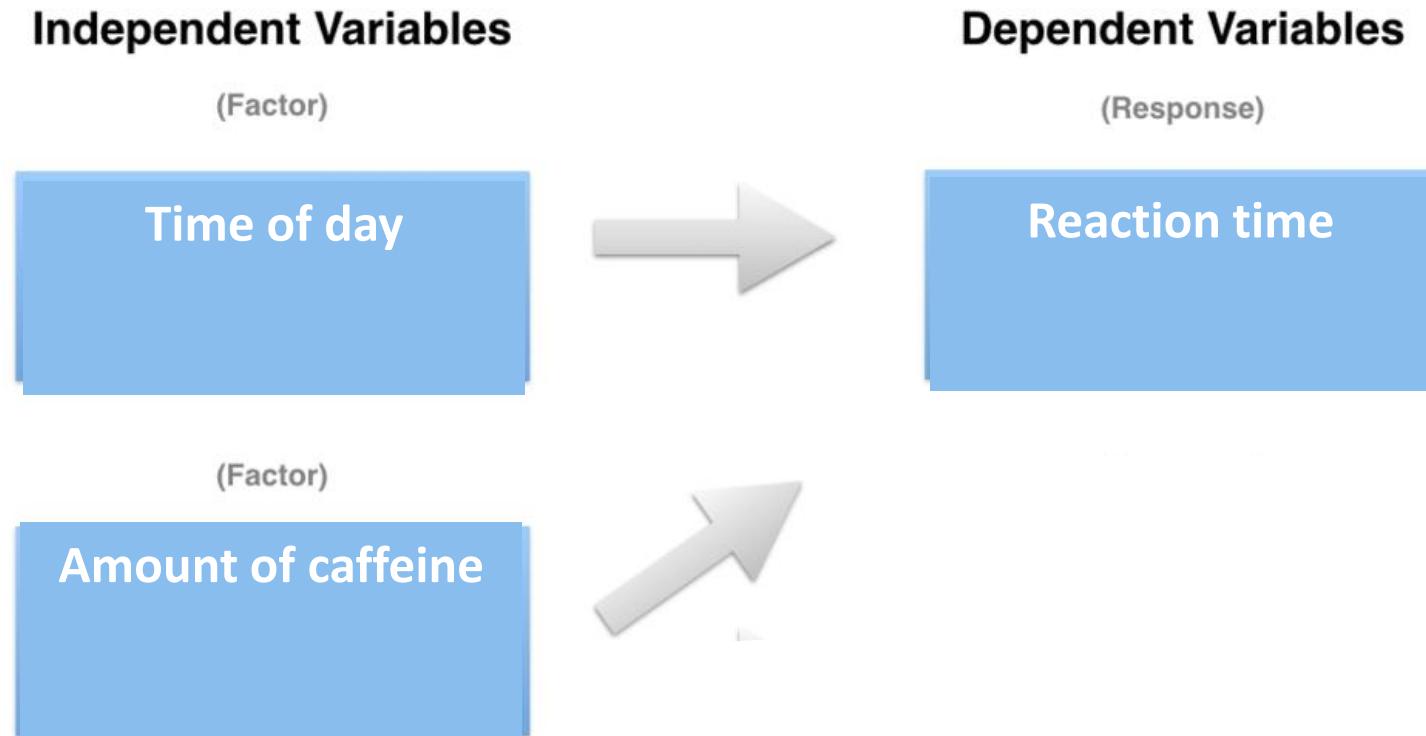
ANDY FIELD



Do people with private health insurance visit their physicians more frequently than people with no insurance or other types of insurance?

Factorial (Univariate) ANOVA

≥ 2 IVs, 1 DV



Factorial ANOVA

≥ 2 IVs, 1 DV

IV – categorical
DV – continuous (interval, ratio)

2x2 Design		Time of Day	
Caffeine	Some Caffeine	Morning	Afternoon
	No Caffeine	Reaction time	Reaction time
		Reaction time	Reaction time

2 Independent variables, 1 Dependent variable

2 x 2 ANOVA

2x3 Design		Time of Day	
Caffeine	1 coffee	Morning	Afternoon
	2 coffees	Reaction time	Reaction time
	3 coffees	Reaction time	Reaction time

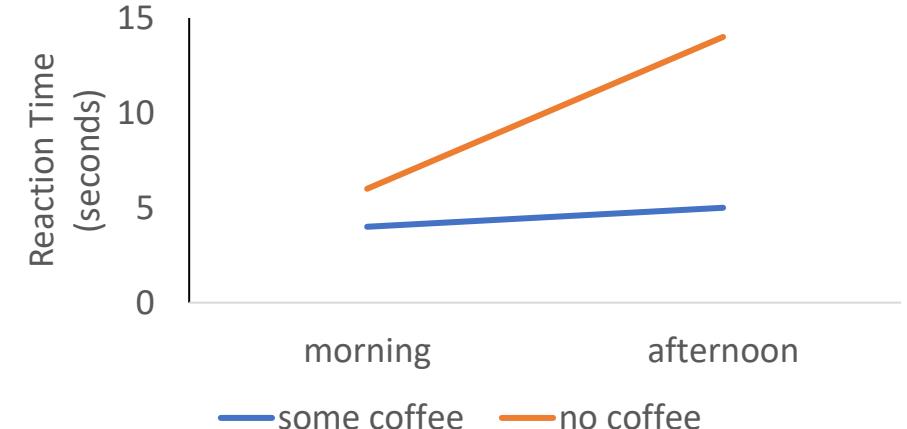
2 Independent variables, 1 Dependent variable

3 x 2 ANOVA

Factorial ANOVA

Interaction between time of day and amount of caffeine intake

2x2 Design		Time of Day	
		Morning	Afternoon
Caffeine	Some Caffeine	Reaction time	Reaction time
	No Caffeine	Reaction time	Reaction time



Source	SS	df	S ²	F
(Between) Row	$\sum [N_{\text{row}}(M_{\text{row}} - M_o)^2]$	rows-1	$\frac{SS_r}{df_r}$	$\frac{S^2_r}{S^2_w}$
(Between) Column	$\sum [N_{\text{col}}(M_{\text{col}} - M_o)^2]$	columns-1	$\frac{SS_c}{df_c}$	$\frac{S^2_c}{S^2_w}$
(Between) Inter-action	$\sum [N_{\text{cell}}(M_{\text{cell}} - M_o)^2]$ - $SS_{\text{row}} - SS_{\text{col}}$	(rows-1)(columns-1)	$\frac{SS_i}{df_i}$	$\frac{S^2_i}{S^2_w}$
Within	$\sum (X - M_{\text{cell}})^2$	N - cells	$\frac{SS_w}{df_w}$	X: individual score N: number of scores M _o : overall mean
Total	$\sum (X - M_o)^2$	N - 1		

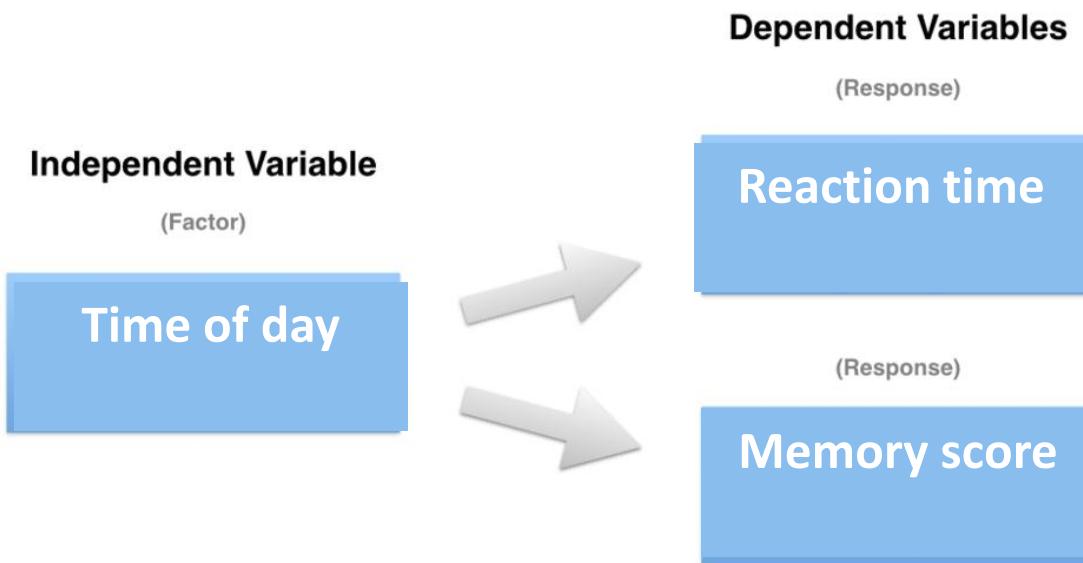
Source	SS	df	S ²	F
(Between) Row	$\sum [N_{\text{row}}(M_{\text{row}} - M_o)^2]$	rows-1	$\frac{SS_r}{df_r}$	$\frac{S^2_r}{S^2_w}$
(Between) Column	$\sum [N_{\text{col}}(M_{\text{col}} - M_o)^2]$	columns-1	$\frac{SS_c}{df_c}$	$\frac{S^2_c}{S^2_w}$
(Between) Inter-action	$\sum [N_{\text{cell}}(M_{\text{cell}} - M_o)^2]$ - $SS_{\text{row}} - SS_{\text{col}}$	(rows-1)(columns-1)	$\frac{SS_i}{df_i}$	$\frac{S^2_i}{S^2_w}$
Within	MEANS:			
Total	Row Means X: individual score N: number of scores M _o : overall mean			

Multivariate ANOVA (MANOVA)

≥ 2 DVs

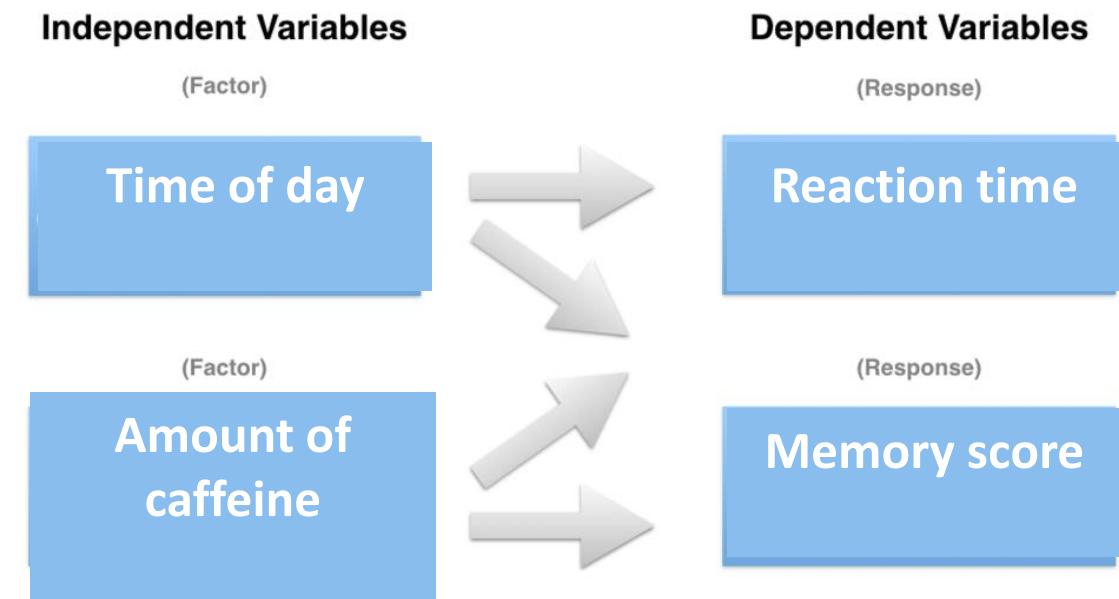
1 IV + 2 DVs

ONE-WAY MANOVA EXAMPLE

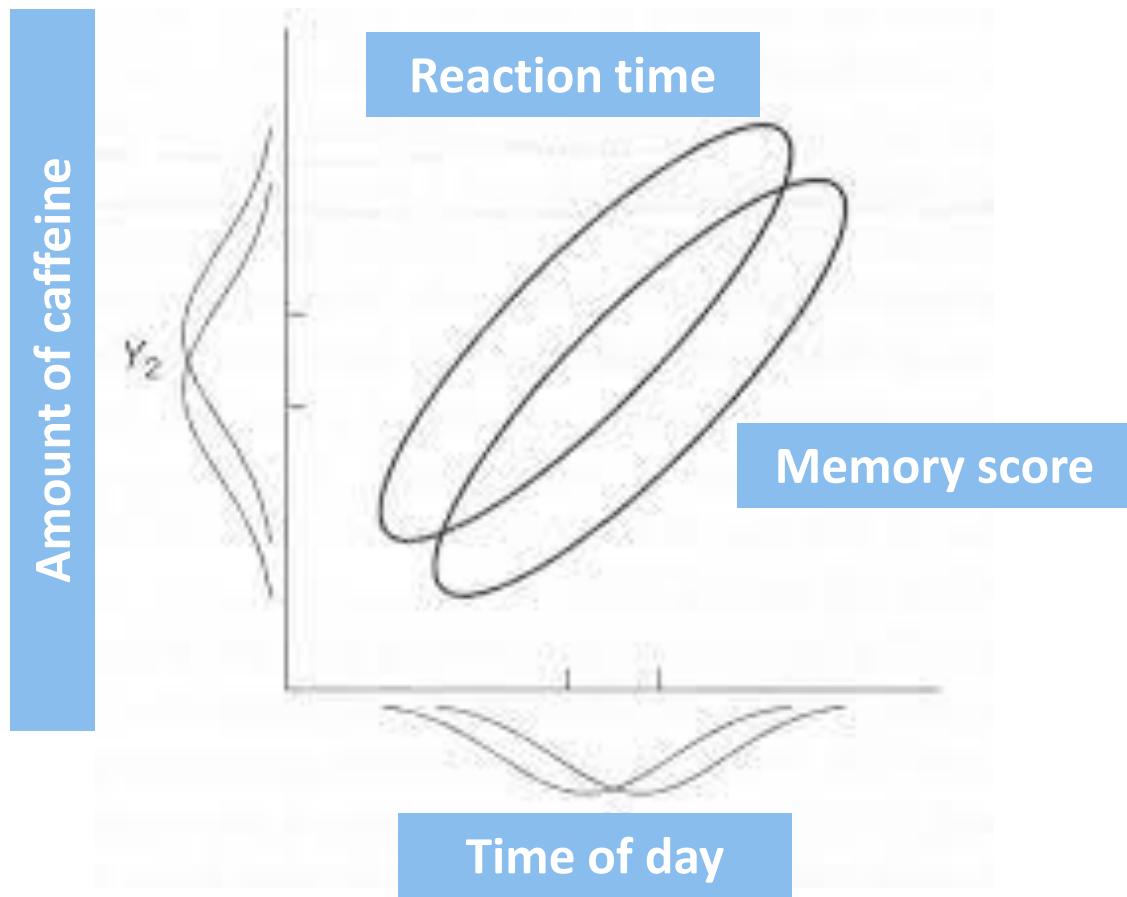


2 IVs + 2 DVs

TWO-WAY MANOVA EXAMPLE



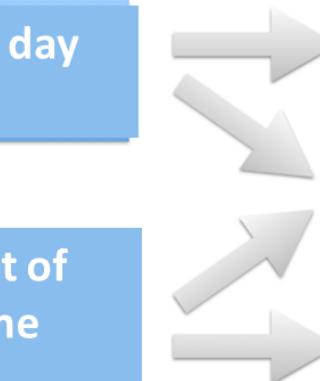
TWO-WAY MANOVA EXAMPLE



Independent Variables (Factor)

Time of day

Amount of caffeine



Dependent Variables (Response)

Reaction time

Memory score

Main effect - Time of Day?

Main effect - Amount of Caffeine?

Interaction? – MANOVA brings out interaction effects

linear combination of DVs to increase effects of IVs

Is MANOVA different from Repeated measures?

- Multiple DVs but not related
- They could be correlated/covariate
(Homogeneity of Covariances)

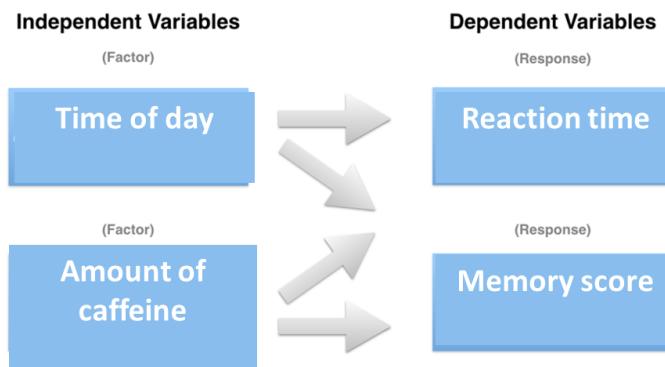
homogeneous variance → Pillai's trace test statistic

Non-homogenous variance → Wilks' lambda test statistic

Assumption of sphericity

- Same participant is repeated tested
- Relationship between multiple timepoints of DV
- Sphericity assumes these timepoints are homogenous
- Individual differences are removed

TWO-WAY MANOVA EXAMPLE



Reaction time → coffee → reaction time

Spot the difference

Younger vs older adults



Spot the difference

Younger vs older adults

Effect of distraction?



Spot the difference

Younger vs older adults

Effect of distraction?

IV 1 – age (young, old)

IV 2 – no distraction, unpleasant sounds

DV – no. of differences spotted (attention)



Spot the difference

Younger vs older adults

Effect of distraction?

Effect of reward?



Spot the difference

Younger vs older adults

Effect of distraction?

IV 1 – age (young, old)

IV 2 – no distraction, unpleasant sounds

DV – no. of differences spotted

Effect of reward?

IV 3– no reward, low reward, high reward



Spot the difference

Younger vs older adults

Effect of distraction?

Can reward change attention levels in distractive environments?



Effect of distraction?

IV 1 – age (young, old)

IV 2 – no distraction, unpleasant sounds

DV – no. of differences spotted

Can reward change attention in
distractive environments?

Spot the difference

Younger vs older adults



Old

Young

Attention → no reward → attention → reward → attention

distraction

no distraction

- Scenario 1: A fitness instructor wants to test the effectiveness of a performance-enhancing herbal supplement on students in his exercise class. The instructor gives Group A The herbal supplement and Group B receives the placebo. The students' fitness level is compared before and after six weeks of consuming the supplement or the sugar pill. His data do not support his hypothesis that the herbal supplement had an effect on fitness levels.

- Scenario 2: A social psychologist is interested discovering if whether women who are taller have a stronger career orientation. She measures height and gives a survey to women that measures many personality variables, including career orientation. Her data support her hypothesis that women who are taller had stronger career orientations.

Several weeks after Allen conducted a classroom experiment on the effectiveness of various metals in releasing hydrogen gas from hydrochloric acid, he read that the gas company was burying sheets of magnesium next to iron pipelines in order to prevent rusting. Allen wondered if other active metals would also be effective in preventing rust.

To investigate, he placed each of the following into a separate test tube containing water: one iron nail; one iron nail wrapped with an aluminum strip; one iron nail wrapped with a magnesium strip; and one iron nail wrapped in a lead strip. He used the same amounts of water from the same source, equal amounts of the metal wraps and the same type of iron nails. At the end of five days, he described the amounts of rusting either as small, moderate or large. He also recorded the color of the water.

Class Data

Gender	Years spent at IIIT (including covid years)	How stressful has the semester been for you?	Do you engage in physical/sports activities everyday?	How many hours do you spend listening to music every day?	In the past week, how many hours (avg) have you slept every night ?
female	4	5	yes	0.5	5
male	4	5	no	4	4.5
male	1.5	3	yes	4	7
male	3	3	yes	2	8
male	1.5	4	not everyday, sometimes	0	6
male	2.5	3	yes	1	7
female	4	5	yes	1	7
female	5	3	not everyday, sometimes	2	7
male	5.5	4	not everyday, sometimes	2	7
male	0.5	5	not everyday, sometimes	0	4
male	1	3	not everyday, sometimes	4	7
male	0.5	4	yes	2	5
male	3	3	yes	1	7
female	4	2	not everyday, sometimes	0.5	6
female	1	4	yes	1	6
male	3.5	1	no	0	6.5
male	3	4	yes	2	8
female	3	3	yes	2	6
male	1.5	2	not everyday, sometimes	0	5
male	0	4	yes	0.5	8
non-binary	4	4	not everyday, sometimes	2	7
male	1.5	4	not everyday, sometimes	1	6

Q1. Do students listen to more music after 2 years in IIIT?

Q2. Does exercise affect night sleep?

Q3. Which factors predict the stress experienced by students at IIIT?

		IV – categorical	DV – continuous (interval, ratio)
		Independent factor	Dependent (Related) Samples
		1 IV	1 DV
		> 2 groups	> 2 timepoints
Parametric (normal)	Homogeneous	One way ANOVA	Repeated measures ANOVA
	Non homogenous	Welch ANOVA	Sphericity correction
Non-parametric		<i>Kruskal-Wallis</i> ANOVA	Friedman's ANOVA

	≥ 2 IVs & 1 DV	≥ 1 IV, repeated DV	≥ 1 IV & ≥ 2 DVs
	Factorial ANOVA	Mixed ANOVA (>1 factors + Repeated measures)	MANOVA

IV= factor (these terms are used interchangeably)

VariableX	VariableY	Type of correlation
Nominal	Nominal	Phi coefficient
Nominal	Ordinal	Rank-biserial coefficient
Nominal	Interval	Point-biserial
Ordinal	Ordinal	Spearman rank correlation coefficient
Interval	Interval	Pearson product-moment correlation coefficient

Test type	Between subjects designs (Independent samples)	Within subject designs (repeated measures/ matched pairs)
Non-parametric (for categorical data)	Chi-square test	The binomial sign test
Non-parametric (for ordinal data)	Mann-Whitney U	Wilcoxon Signed-Rank Test The binomial sign test
Parametric	Unrelated t-test (level of data: interval)	Related t-test (level of data: interval)

IV – categorical

DV – continuous (interval, ratio)

		Independent factor	Dependent (Related) Samples
Parametric (normal)	Homogeneous	1 IV > 2 groups	1 DV > 2 timepoints
	Non homogenous	One way ANOVA	Repeated measures ANOVA
Non-parametric		Welch ANOVA	Sphericity correction
		Kruskal-Wallis ANOVA	Friedman's ANOVA

	>=2 IVs & 1 DV	>=1 IV, repeated DV	>=1 IV & >= 2 DVs
	Factorial ANOVA	Mixed ANOVA (Factors + Repeated measures)	MANOVA

REGRESSION

BRSM

SIMPLE LINEAR REGRESSION

- Interval/ratio scale predictors and outcome variables

SCATTERPLOT

Imagine a line through these points that capture the correlation you're thinking about

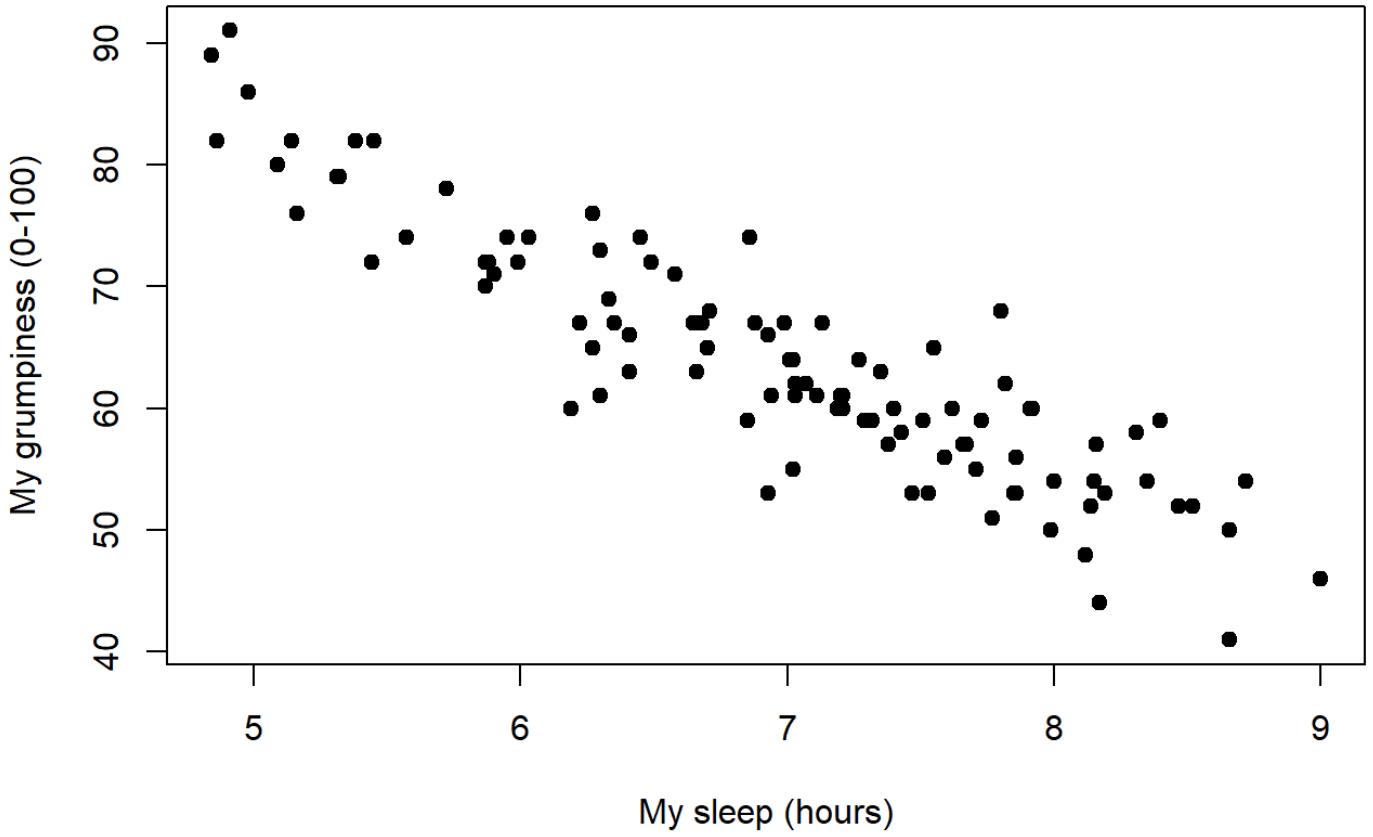


Figure 15.1: Scatterplot showing grumpiness as a function of hours slept.

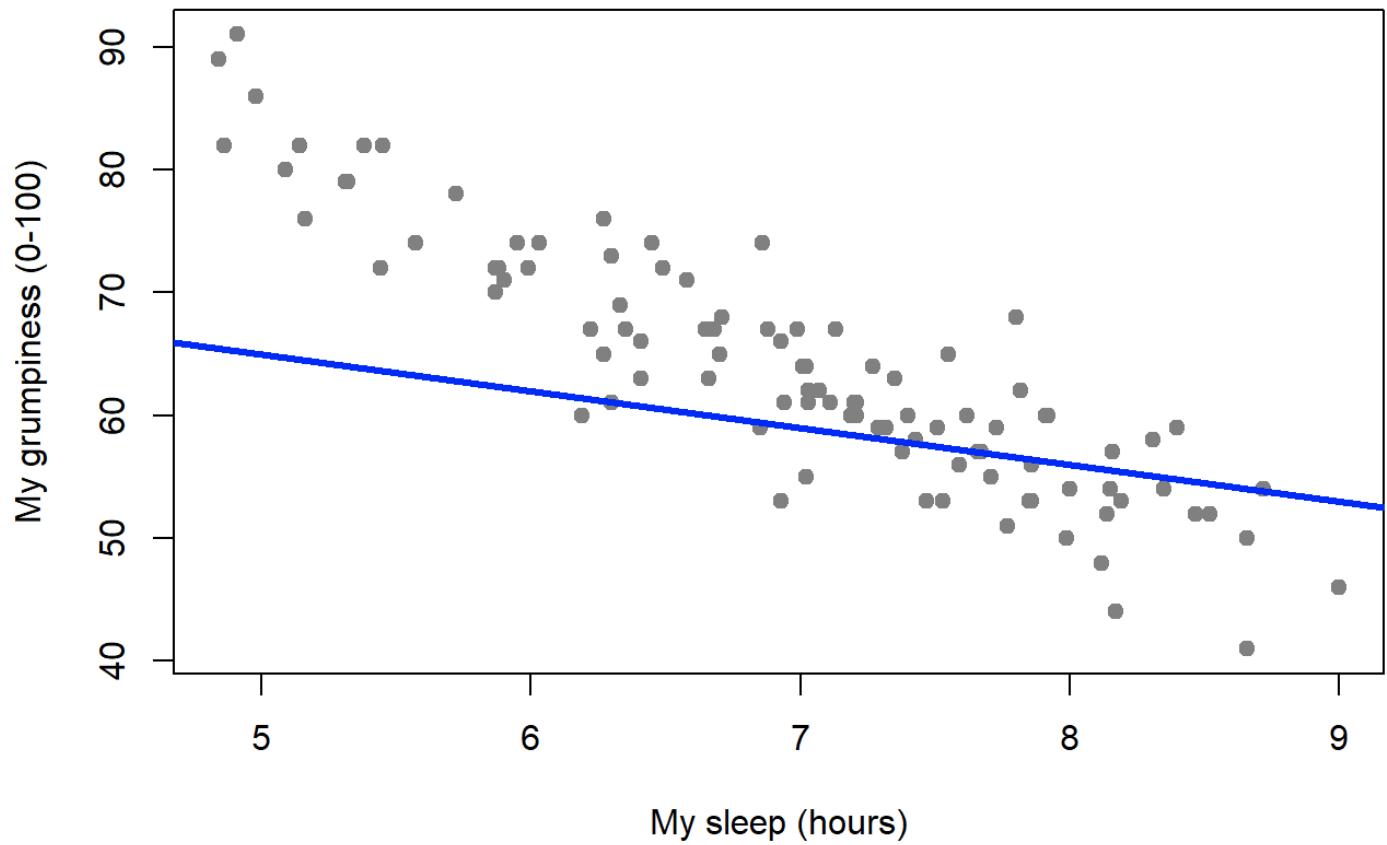
BEST-FITTING REGRESSION LINE



A POOR-FITTING LINE



Not The Best Fitting Regression Line!



SIMPLE LINEAR REGRESSION

- Related to the idea of correlations

$$y = mx + c$$

$$\hat{Y}_i = b_1 X_i + b_0$$

Hat --> predicted

$$\epsilon_i = Y_i - \hat{Y}_i$$

b1 --> regression coefficient

Error --> residual

$$Y_i = b_1 X_i + b_0 + \epsilon_i$$

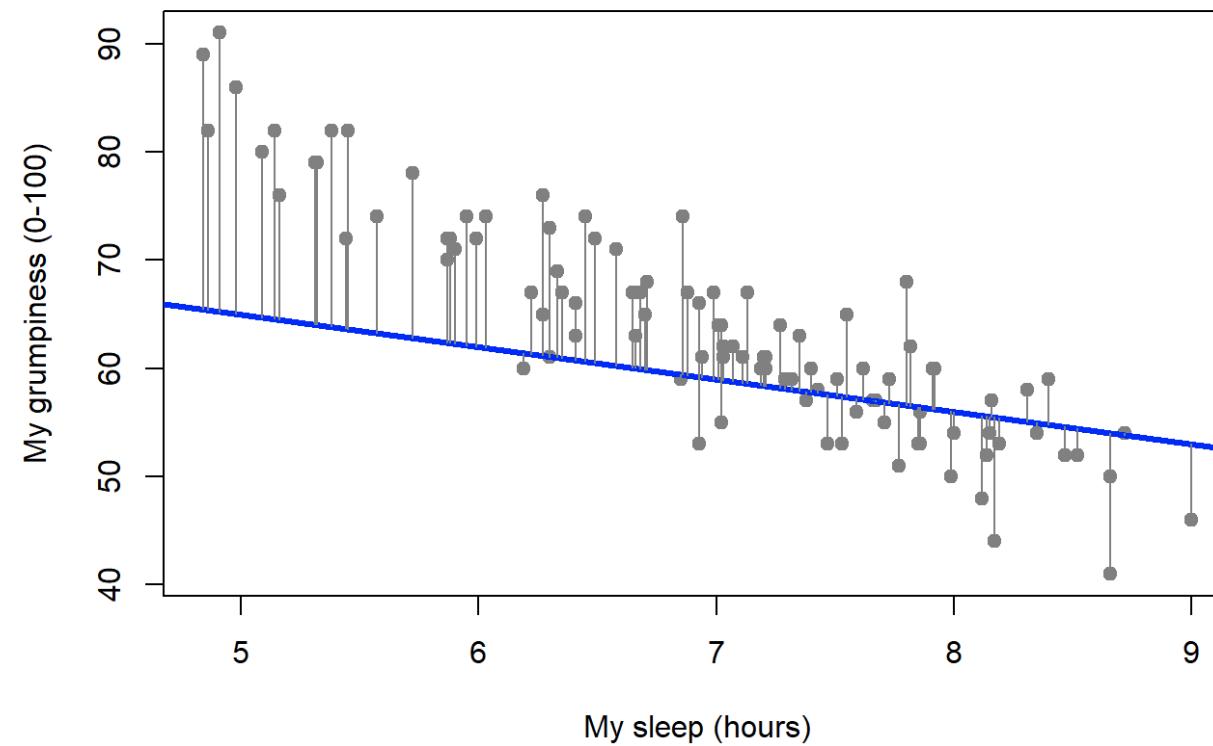
RESIDUALS RELATED TO THE BEST-FITTING REGRESSION LINE



Figure 15.4: A depiction of the residuals associated with the best fitting regression line

RESIDUALS RELATED TO A POOR-FITTING LINE

Regression Line Distant from the Data



HOW DO WE ESTIMATE THE REGRESSION COEFFICIENTS?

- Intuition regarding residuals?
- Small residuals
- Quantity to minimize: sum of squares of errors (residuals)
- This is called Ordinary Least Squares Regression
- Many other ways to estimate regression coefficients

R FORMULA

```
regression.1 <- lm( formula = dan.grump ~ dan.sleep,  
                     data = parenthood )
```

```
print( regression.1 )
```

```
##  
## Call:  
## lm(formula = dan.grump ~ dan.sleep, data = parenthood)  
##  
## Coefficients:  
## (Intercept)  dan.sleep  
##      125.956     -8.937
```

$$\hat{Y}_i = -8.94 X_i + 125.96$$

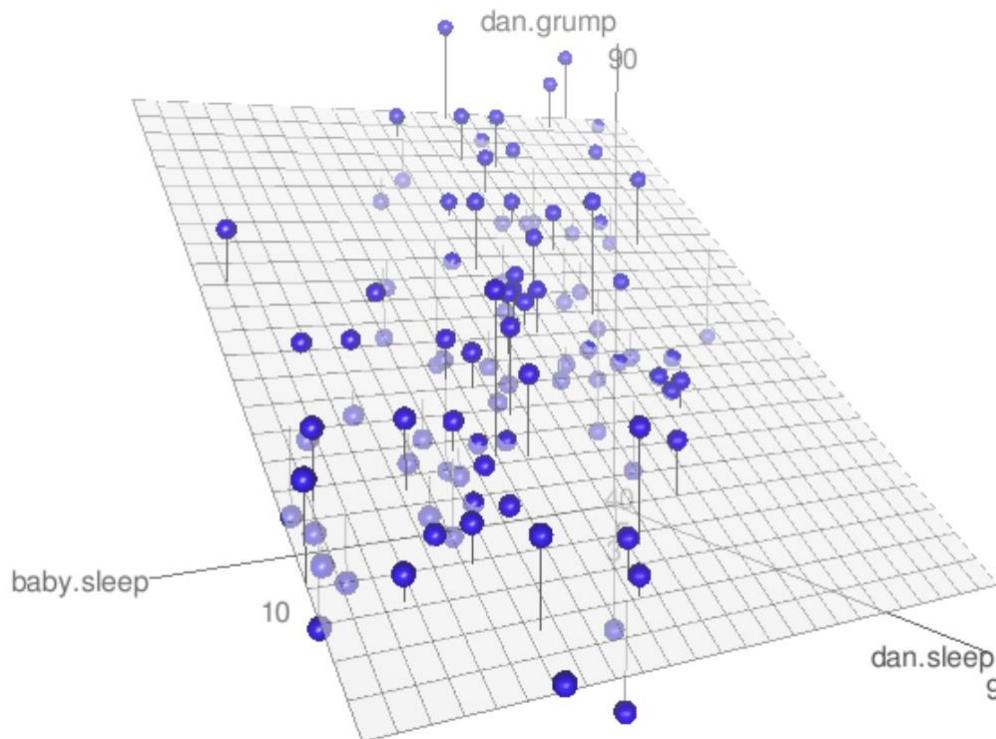
PLAY TIME: GUESS THE REGRESSION

- <https://sophieehill.shinyapps.io/eyeball-regression/>

MULTIPLE LINEAR REGRESSION (MLR)

- When you have more than one predictor variable

$$Y_i = b_2 X_{i2} + b_1 X_{i1} + b_0 + \epsilon_i$$



R FORMULA

```
regression.2 <- lm( formula = dan.grump ~ dan.sleep + baby.sleep,  
                     data = parenthood )
```

```
print( regression.2 )
```

```
##  
## Call:  
## lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)  
##  
## Coefficients:  
## (Intercept)    dan.sleep    baby.sleep  
##   125.96557     -8.95025      0.01052
```

MLR WITH K VARIABLES

$$Y_i = \left(\sum_{k=1}^K b_k X_{ik} \right) + b_0 + \epsilon_i$$

HOW DO YOU KNOW IF THE REGRESSION DOES A GOOD JOB?

```
##  
## Call:  
## lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)  
##  
## Coefficients:  
## (Intercept)    dan.sleep    baby.sleep  
##   125.96557     -8.95025      0.01052
```

- Can you infer how good the regression line fit is based on the coefficients?
- No, these just help you predict \hat{Y} , how good this prediction is needs to be quantified.

R SQUARED

$$SS_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

Sum of squared residuals (SSR), which we hope is small.

How small should this be? What do we compare against?

The outcome variable Y itself is quite variable. If the SSR << the variability in Y , that is a good sign. If the SSR is the same as the variability in Y , that is a bad sign.

$$SS_{tot} = \sum_i (Y_i - \bar{Y})^2$$

R SQUARED

$$SS_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

$$SS_{tot} = \sum_i (Y_i - \bar{Y})^2$$

So construct something that is 0 if the fit is the worst and 1 if the fit is the best.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

The coefficient of determination

The proportion of variance in the outcome variable accounted for by the predictor

WHAT IS THE RELATIONSHIP BETWEEN REGRESSION AND CORRELATION?

- The squared Pearson correlation and the R square value from the regression are the same for the case of one predictor.

WHAT IS ONE EASY WAY TO IMPROVE R SQUARE?

- Add more predictors!
- The R square will never decrease by adding more predictors.
- However, this added complexity of the model should be accounted for in your measure of goodness of fit.
- Adjusted R square: constructed such that additional variables will improve adj R square only if the added variables significantly improve the predictions more than what you'd expect by chance.

$$\text{adj. } R^2 = 1 - \left(\frac{\text{SS}_{res}}{\text{SS}_{tot}} \times \frac{N - 1}{N - K - 1} \right)$$

WHAT SHOULD YOU REPORT: R SQUARE OR ADJ. R SQUARE?

- 
- R square: straightforward to interpret as the proportion of variance in the outcome variable accounted for by the predictors but does not account for complexity and added degrees of freedom due to added predictor variables.
 - Adj . R square: not straightforward to interpret but is a measure of goodness of fit that is not biased by added complexity of the model.

NEXT: HYPOTHESIS TESTS FOR REGRESSION MODELS AND COEFFICIENTS

- So far: interpreting regression coefficients, and evaluating overall goodness of fit, but we do not know if a regression coefficient of 3.4 for instance is statistically significant (i.e., statistically meaningfully greater than 0).
- We also need to do a statistical test for the model as a whole by comparing it against a trivial model, as it is possible that the use of a more trivial model can also lead to comparable R squares in some situations.

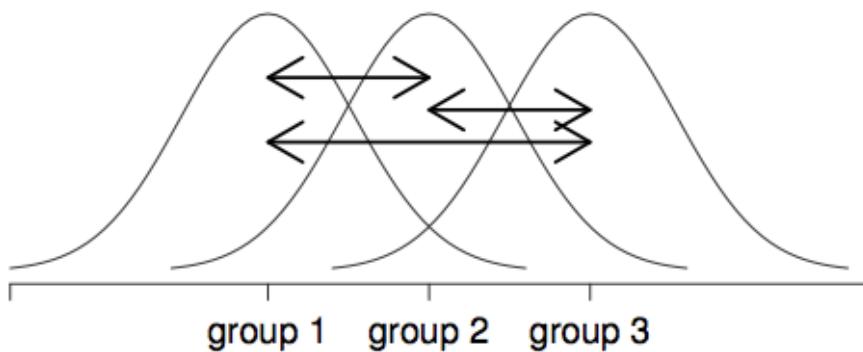
HYPOTHESIS TESTS FOR THE ENTIRE REGRESSION MODEL

- The null model: $H_0 : Y_i = b_0 + \epsilon_i$
- The alternative model: $H_1 : Y_i = \left(\sum_{k=1}^K b_k X_{ik} \right) + b_0 + \epsilon_i$
- To construct the test, we start by dividing the total sum of squares of the outcome variable just as it is done in ANOVA

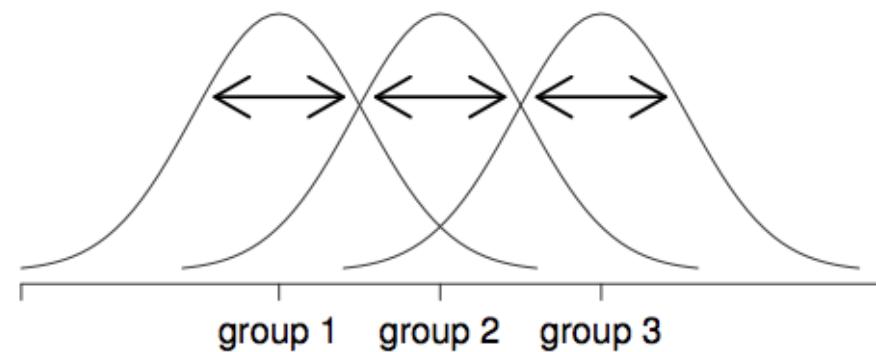
$$SS_{mod} = SS_{tot} - SS_{res}$$

A REMINDER ABOUT ANOVA

Between-group variation
(i.e., differences among group means)



Within-group variation
(i.e., deviations from group means)



A REMINDER ABOUT ANOVA

$$SS_{tot} = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

$$SS_w = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$$

$$\begin{aligned} SS_b &= \sum_{k=1}^G \sum_{i=1}^{N_k} (\bar{Y}_k - \bar{Y})^2 \\ &= \sum_{k=1}^G N_k (\bar{Y}_k - \bar{Y})^2 \end{aligned}$$

$$SS_w + SS_b = SS_{tot}$$

A REMINDER ABOUT ANOVA

	df	sum of squares	mean squares	F-statistic	p-value
between groups	$df_b = G - 1$	$SS_b = \sum_{k=1}^G N_k (\bar{Y}_k - \bar{Y})^2$	$MS_b = \frac{SS_b}{df_b}$	$F = \frac{MS_b}{MS_w}$	[complicated]
within groups	$df_w = N - G$	$SS_w = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$	$MS_w = \frac{SS_w}{df_w}$	-	-

Large F --> ??

BACK TO LINEAR REGRESSION AND SUM OF SQUARES

$$\text{SS}_{mod} = \text{SS}_{tot} - \text{SS}_{res}$$

$$df_{mod} = K. \quad df_{res} = N - K - 1.$$

$$F = \frac{\text{MS}_{mod}}{\text{MS}_{res}}$$

Similar interpretation as in the ANOVA case.
High value of F --> the alternative model
performs better than the null

SO WE HAVE JUST TESTED THE REGRESSION MODEL AS A WHOLE

- If the F test is not significant, then either the model is a poor one or your data has issues.
- If it is significant, it still doesn't mean you know for sure your predictors all explain the outcome. Need to do statistical tests for individual regression coefficients.

```
> print( regression.2 )

Call:
lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)

Coefficients:
(Intercept)    dan.sleep    baby.sleep
  125.96557     -8.95025      0.01052
```

TESTING FOR INDIVIDUAL REGRESSION COEFFICIENTS

- CLT
- Normally distributed sampling distribution of the estimator of b , centered on b .
- If we can then come up with a standard error for this estimator, then we can construct a t-statistic
- Turns out we can do this, a complicated formula, but note that this SE depends on both predictor and outcome variables and is also sensitive to violations of homogeneity of variance assumptions (later).

$$H_0 : b = 0 \quad H_1 : b \neq 0 \quad t = \frac{\hat{b}}{\text{SE}(\hat{b})}$$

HYPOTHESIS TEST FOR COEFFICIENTS IN R

```
> summary( regression.2 )
```

```
Call:  
lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)
```

Residuals:

Min	1Q	Median	3Q	Max
-11.0345	-2.2198	-0.4016	2.6775	11.7496

Regression assumption: residuals are normally distributed around 0.
So check if median around 0, 1Q and 3Q approx equidistant from 0...

HYPOTHESIS TEST FOR COEFFICIENTS IN R

```
> summary( regression.2 )
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	125.96557	3.04095	41.423	<2e-16 ***
dan.sleep	-8.95025	0.55346	-16.172	<2e-16 ***
baby.sleep	0.01052	0.27106	0.039	0.969

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 4.354 on 97 degrees of freedom

HYPOTHESIS TEST FOR COEFFICIENTS IN R

```
> summary( regression.2 )
```

Residual standard error: 4.354 on 97 degrees of freedom
Multiple R-squared: 0.8161, Adjusted R-squared: 0.8123
F-statistic: 215.2 on 2 and 97 DF, p-value: < 2.2e-16

A global assessment of the model

CONFIDENCE INTERVALS FOR REGRESSION COEFFICIENTS

$$CI(\hat{b}) = \hat{b} \pm (t_{crit} \times SE(\hat{b}))$$

N-K-1 degrees of freedom

Critical t value for 97.5th percentile, to construct
a 95% CI.

```
> confint( object = regression.2,
+           level = .99
+ )
      0.5 %    99.5 %
(Intercept) 117.9755724 133.9555593
dan.sleep   -10.4044419  -7.4960575
baby.sleep   -0.7016868   0.7227357
```

HOW DO YOU COMPARE REGRESSION COEFFICIENTS FOR PREDICTORS THAT HAVE DIFFERENT UNITS AND HAVE TOTALLY DIFFERENT SCALES?

- e.g. Predicting intelligence scores using years of education and income.
- Income: may vary from tens of thousands p.a to several lakhs.
- Years of education: 0–15 years
- Comparing regression coefficients from these predictors would be difficult. Say 0.25 for income and 0.89 for years of education.
- Here, we can use standardized regression (basically z-score your variables and do the regression).
- Easier: just run the regular regression and standardize the coefficients -- $\beta_X = b_X \times \frac{\sigma_X}{\sigma_Y}$

INTERPRETING STANDARDIZED REGRESSION COEFFICIENTS

- $\text{IQ} \sim b_1 * \text{income} + b_2 * \text{years of education} + b_0$
- Standardized coeffs are usually denoted by betas.
- A change in income by 1 s.d. (of income) corresponds to beta 1 s.d. change in IQ when years of education is held constant.
- Can directly compare b_1 and b_2 in terms of how much each variable affects IQ (in terms of IQ s.d.)
- However, 1 s.d. change in income and 1 s.d. change in years of education - are they comparable quantities? This is not too straightforward. So while standardized regression is supposed to help you put different predictors on the same scale, you have to be judicious with its use.

FINAL SECTION OF THE BASICS: THE ASSUMPTIONS OF LINEAR REGRESSION

- **Normality:** residuals are normally distributed. The variables can be non-normal!
- **Linearity:** the relationship between X and Y is more or less linear
- **Homogeneity of variance:** We assume that the residuals are i.i.d with mean 0 and the same s.d. Not easy to test this, but we will check whether the s.d. of the residuals are the same at each level of X and Y instead --> homogeneity of variance.

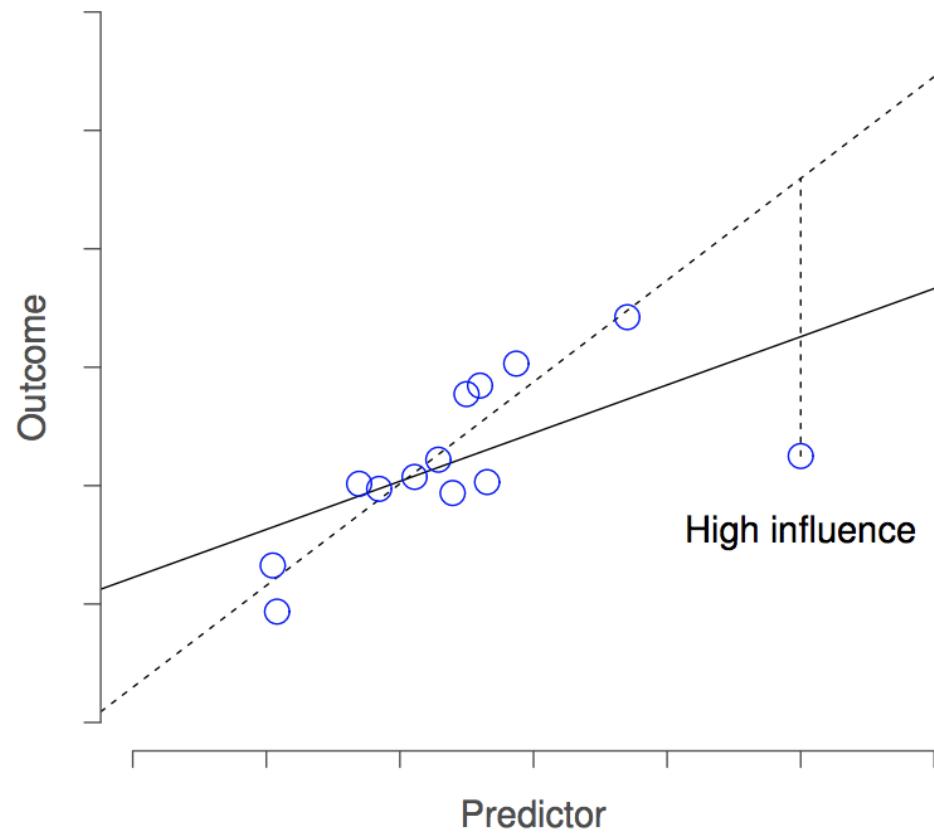
OTHER DESIRABLE FEATURES FOR REGRESSION (BUT NOT STRICT ASSUMPTIONS)

- Uncorrelated predictors - collinear/correlated predictors makes it hard to interpret the regression output in many cases.
- No large outliers - Is the regression being influenced heavily by one or two points?

REGRESSION DIAGNOSTICS

Checking for outlier influence: Cook's distance

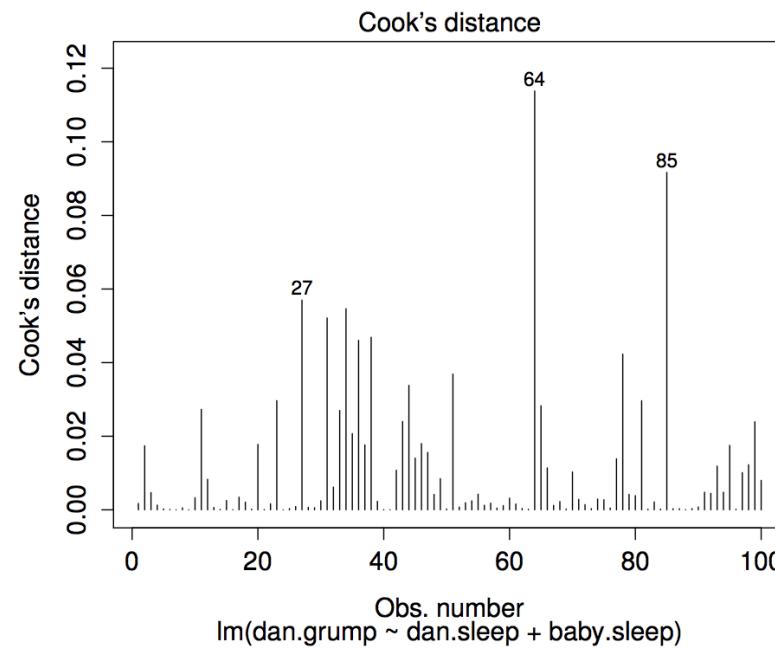
$$D_i = \frac{\epsilon_i^{*2}}{K + 1} \times \frac{h_i}{1 - h_i}$$



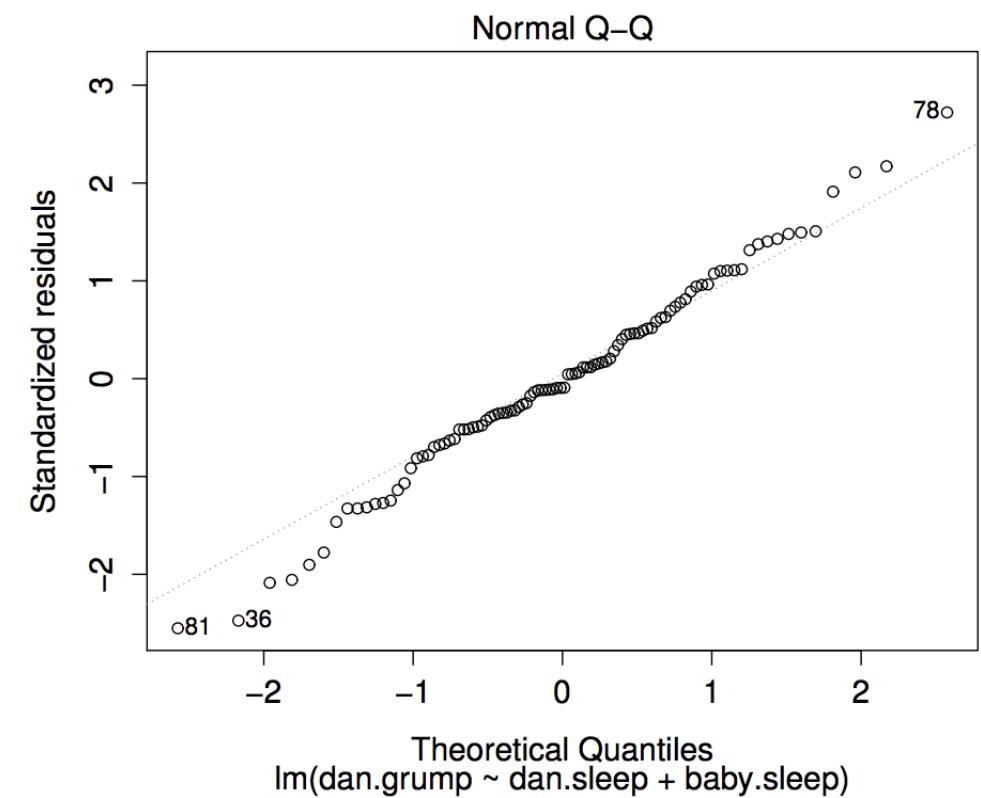
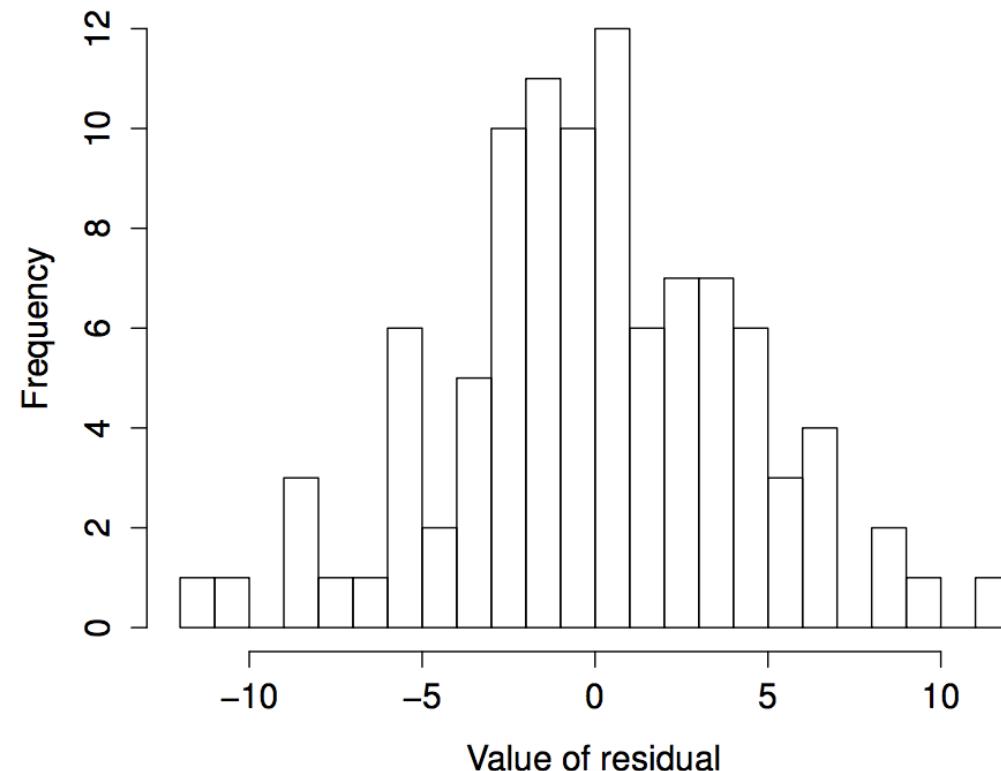
COOK'S DISTANCE PLOTS: CHECKING FOR OUTLIER INFLUENCE

```
> plot(x = regression.2, which = 4)
```

Cook's distance > 1 might indicate problems
If you get a point like that, try removing that
data point and re-running your regression. If
the coefficients and results change by a lot, you
can tell that the outlier had a huge influence.

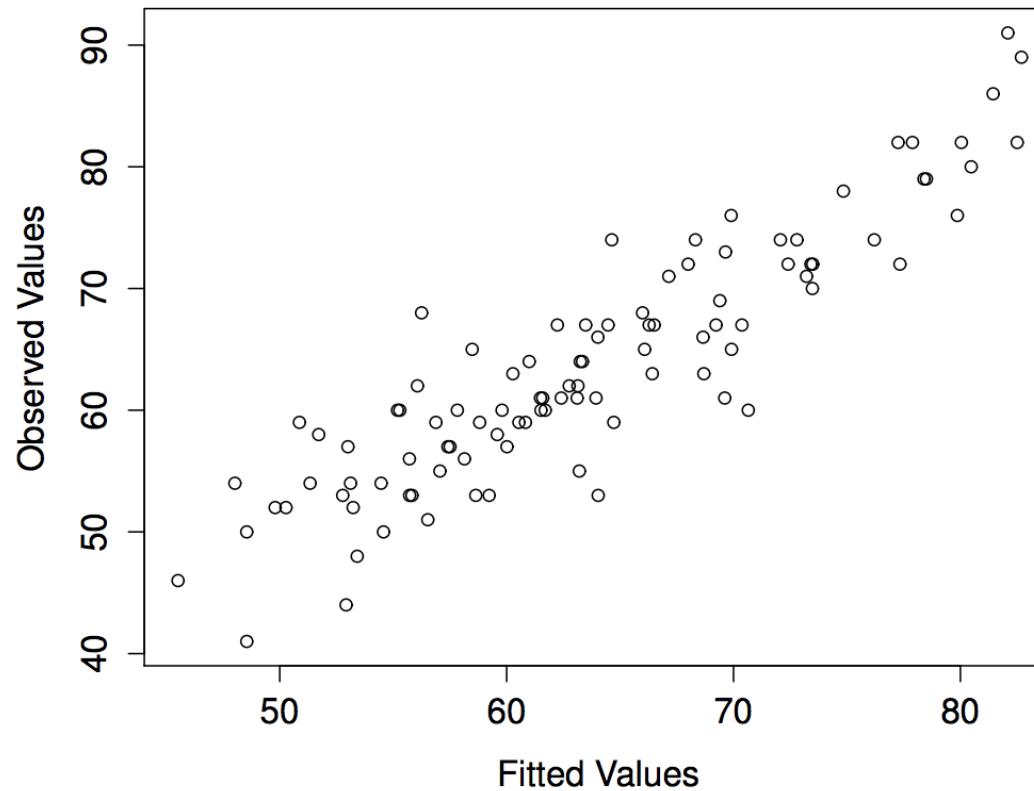


NORMALITY OF RESIDUALS

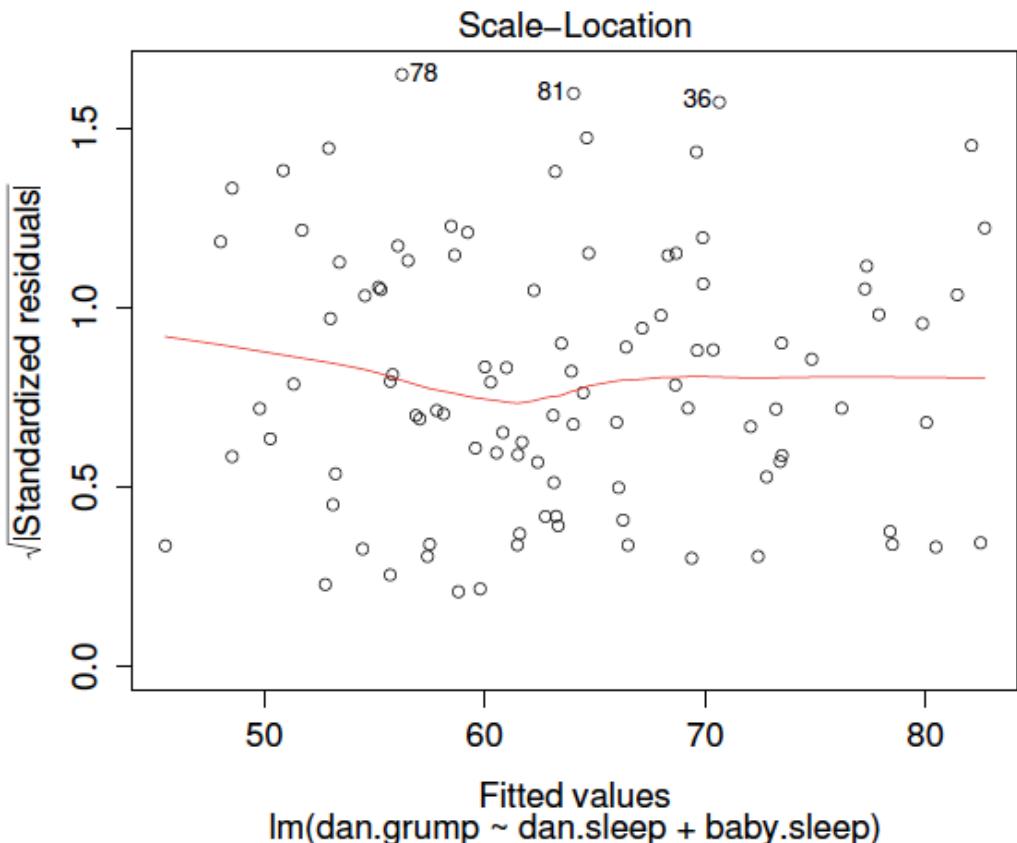


Also do the Shapiro-Wilk test, etc to test for Normality

CHECKING LINEARITY



CHECKING FOR HOMOGENEITY OF VARIANCE



```
> ncvTest( regression.2 )
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.09317511    Df = 1    p = 0.7601788
```

HOW TO DEAL WITH VIOLATIONS OF THE HOMOGENEITY OF VARIANCE ASSUMPTION?

- The problem? Our SE estimates of the estimators of the regression coefficients will no longer be correct as they are based on this homogeneity assumption.
- So will have to use other estimators for computing this SE
- These have been figured out: using heteroscedasticity corrected covariance matrix
- "sandwich estimators"

```
> coeftest( regression.2, vcov= hccm )
```

CHECKING FOR COLLINEARITY

- VIF
- Typical rules of thumb: >5 or 10..

$$\text{VIF}_k = \frac{1}{1 - R_{(-k)}^2}$$

```
> regression.3 <- lm( day ~ baby.sleep + dan.sleep + dan.grump, parenthood )
```

```
> vif( regression.3 )
baby.sleep  dan.sleep  dan.grump
1.651064    6.102337    5.437903
```

COMPARING REGRESSION MODELS: MODEL SELECTION AND OCCAM'S RAZOR

$$AIC = \frac{SS_{res}}{\hat{\sigma}^2} + 2K$$

STEP REGRESSION

- Backward: specify the full model first and then remove predictors one at a time in different ways and pick the model with the lowest AIC

```
> full.model <- lm( formula = dan.grump ~ dan.sleep + baby.sleep + day,
+                     data = parenthood
+ )

> step( object = full.model,      # start at the full model
+        direction = "backward"  # allow it remove predictors but not add them
+ )
```

Start: AIC=299.08
dan.grump ~ dan.sleep + baby.sleep + day

STEP REGRESSION

	Df	Sum of Sq	RSS	AIC
- baby.sleep	1	0.1	1837.2	297.08
- day	1	1.6	1838.7	297.16
<none>			1837.1	299.08
- dan.sleep	1	4909.0	6746.1	427.15

STEP REGRESSION

Step: AIC=297.08
dan.grump ~ dan.sleep + day

	Df	Sum of Sq	RSS	AIC
- day	1	1.6	1838.7	295.17
<none>			1837.2	297.08
- dan.sleep	1	8103.0	9940.1	463.92

STEP REGRESSION

Step: AIC=295.17
dan.grump ~ dan.sleep

	Df	Sum of Sq	RSS	AIC
<none>		1838.7	295.17	
- dan.sleep	1	8159.9	9998.6	462.50

STEP REGRESSION: FINAL CHOSEN MODEL

Call:

```
lm(formula = dan.grump ~ dan.sleep, data = parenthood)
```

Coefficients:

(Intercept)	dan.sleep
125.956	-8.937

FORWARD STEP REGRESSION

- Also possible
- The answers from forward and backward regression need not always be the same! So be careful when using this, always use your intuition about interpretability of the resulting models as well in addition to all these numbers and diagnostics.

COMPARING TWO REGRESSION MODELS IN GENERAL

```
> M0 <- lm( dan.grump ~ dan.sleep + day, parenthood )
> M1 <- lm( dan.grump ~ dan.sleep + day + baby.sleep, parenthood )
```

```
> AIC( M0, M1 )
      df      AIC
M0  4 582.8681
M1  5 584.8646
```

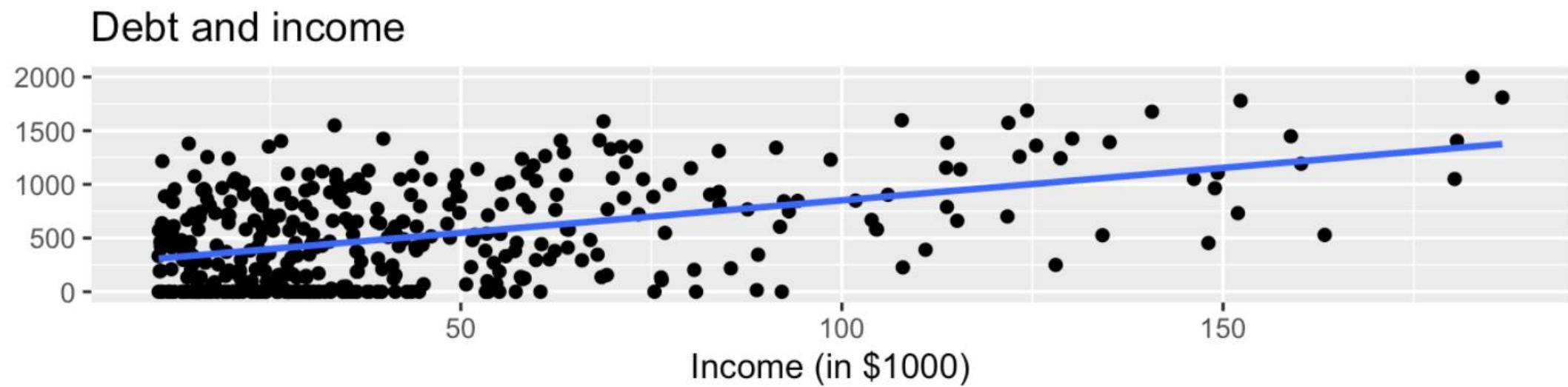
SUMMARY

- Basic ideas in linear regression and how regression models are estimated
- Multiple linear regression
- Measuring the overall performance of a regression model using R²
- Hypothesis tests for regression models
- Calculating confidence intervals for regression coefficients, and standardised coefficients
- The assumptions of regression and how to check them
- Selecting a regression model

OTHER RESOURCES

- For a fast-paced and more technical introduction, check out Chapter 1 in Roback and Legler's *Beyond multiple linear regression: Applied generalized linear models and multilevel models in R* (<https://github.com/proback/BeyondMLR>)
- For an introduction from a Bayesian perspective, Check out Chapters 4 and 5 in McElreath's *Statistical rethinking* (<https://osf.io/2h6ut/>). You can also find him lecturing on the material in these playlists: https://www.youtube.com/channel/UCNJK6_DZvcMqNSzQdEkvzA/playlists.

A FINAL NOTE FOR THE DAY: GUESS THE REGRESSION COEF FOR INCOME



WOAH! -- SIMPSON'S PARADOX

TABLE 6.17: Multiple regression results

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-385.179	19.465	-19.8	0	-423.446	-346.912
credit_limit	0.264	0.006	45.0	0	0.253	0.276
income	-7.663	0.385	-19.9	0	-8.420	-6.906

Credit limit and 4 credit limit brackets.

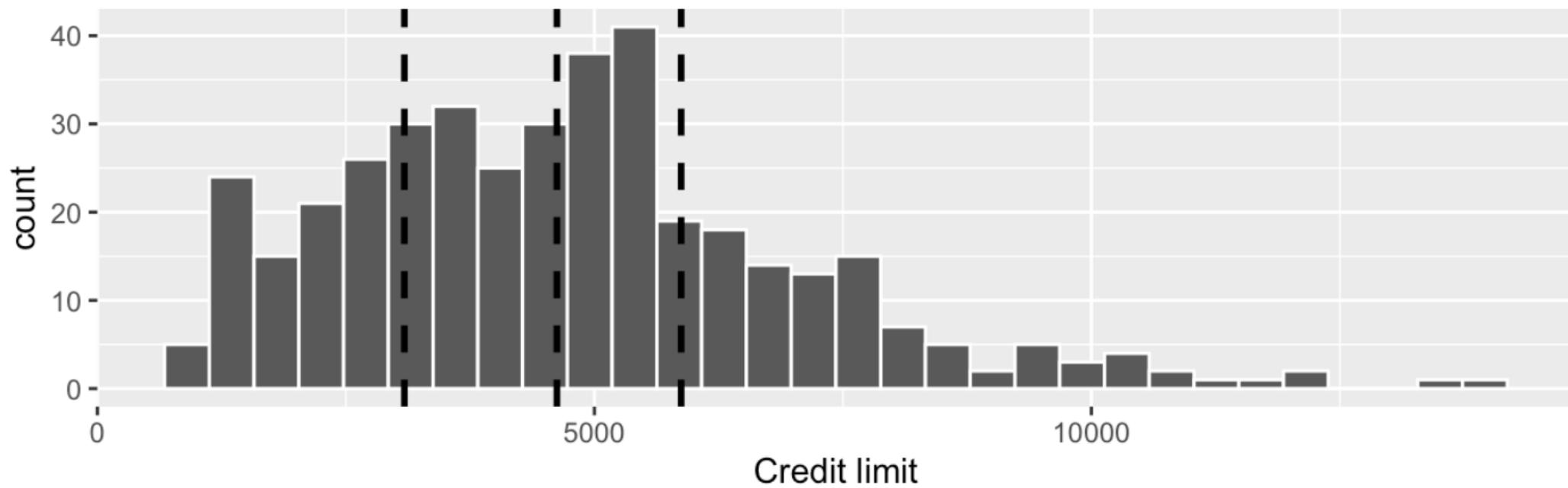
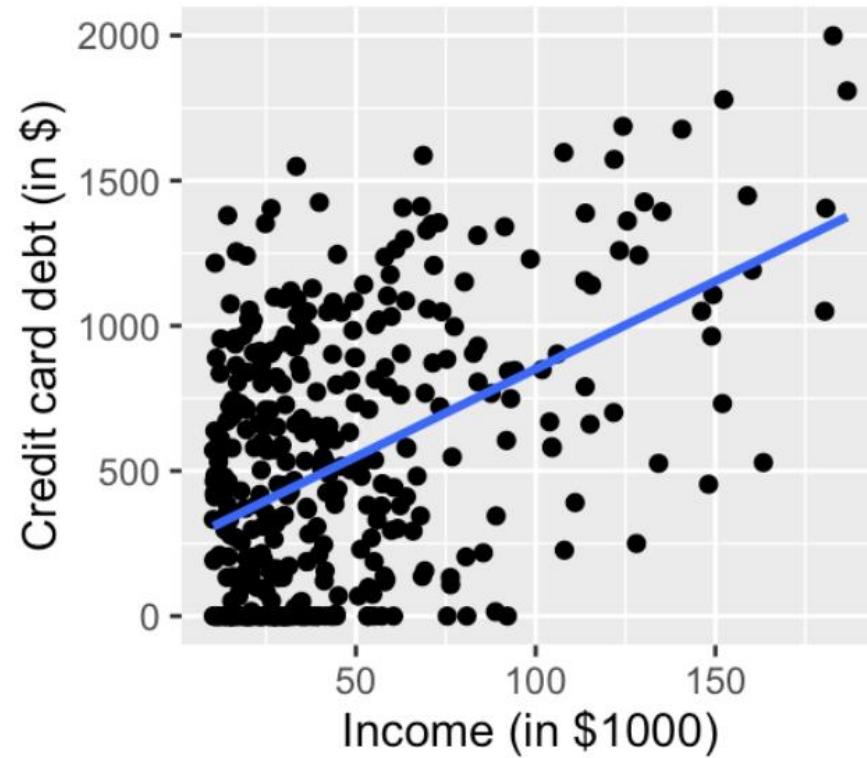


FIGURE 6.10: Histogram of credit limits and brackets.

Two scatterplots of credit car



Credit limit bracket

- low
- med-low
- med-high
- high

NEXT CLASS

- Dealing with other types of variables, interactions, etc
- Practicals
- Now/homework: Simulate some data with $y \leftarrow b_1x_1 + b_2x_2 + b_0$, add some errors drawn from a normal distribution
- Now fit these simulated data using regression
- Make x_1 and x_2 correlated, redo, calculate VIF
- Simulate heteroscedasticity? Redo normal regression and compare with regression using the heteroscedasticity corrected covariance matrix option and compare the results.

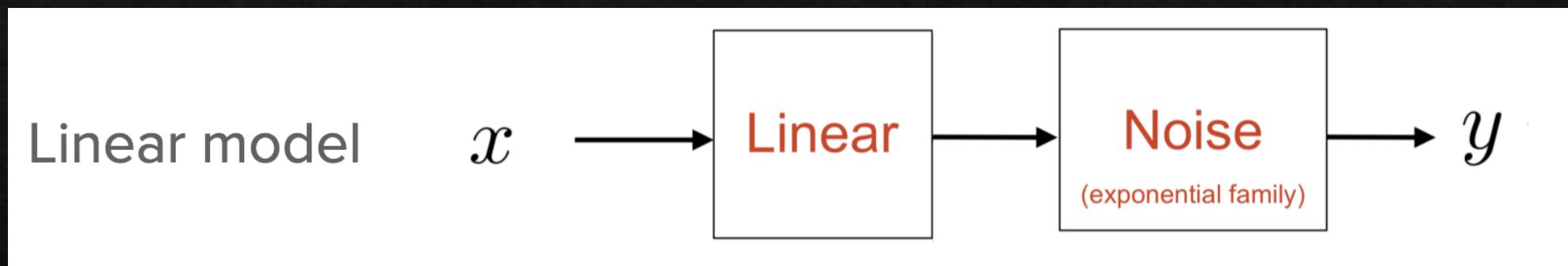
Regression - practicals

BRSM

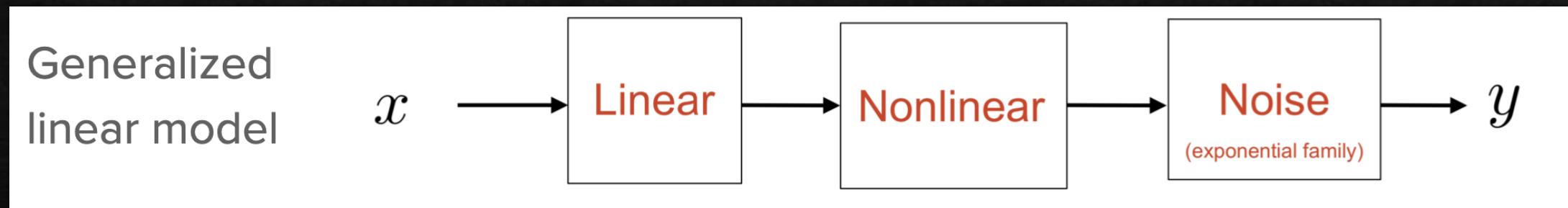
Agenda

- ❖ GLM
- ❖ Assignment

Linear Model



Generalized Linear Model



Generalized Linear Model

Example: nonlinear Gaussian model $y = f(\theta x) + \eta$ where $\eta \sim \mathcal{N}(0, \sigma^2)$


nonlinear
 f^{-1} : link function

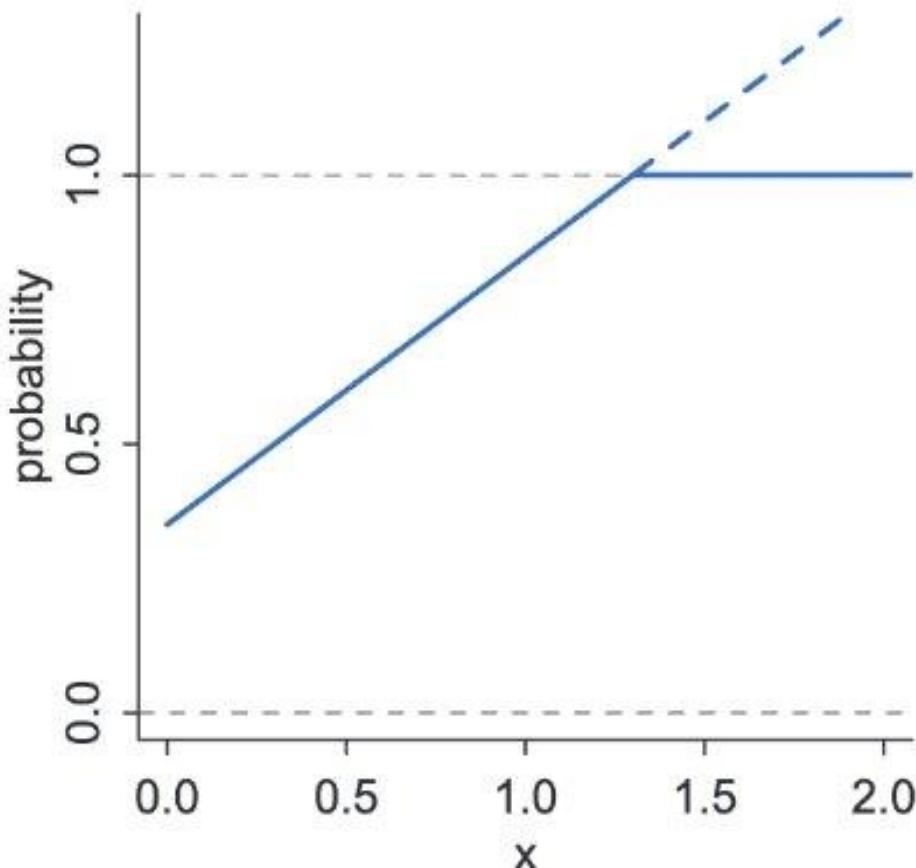


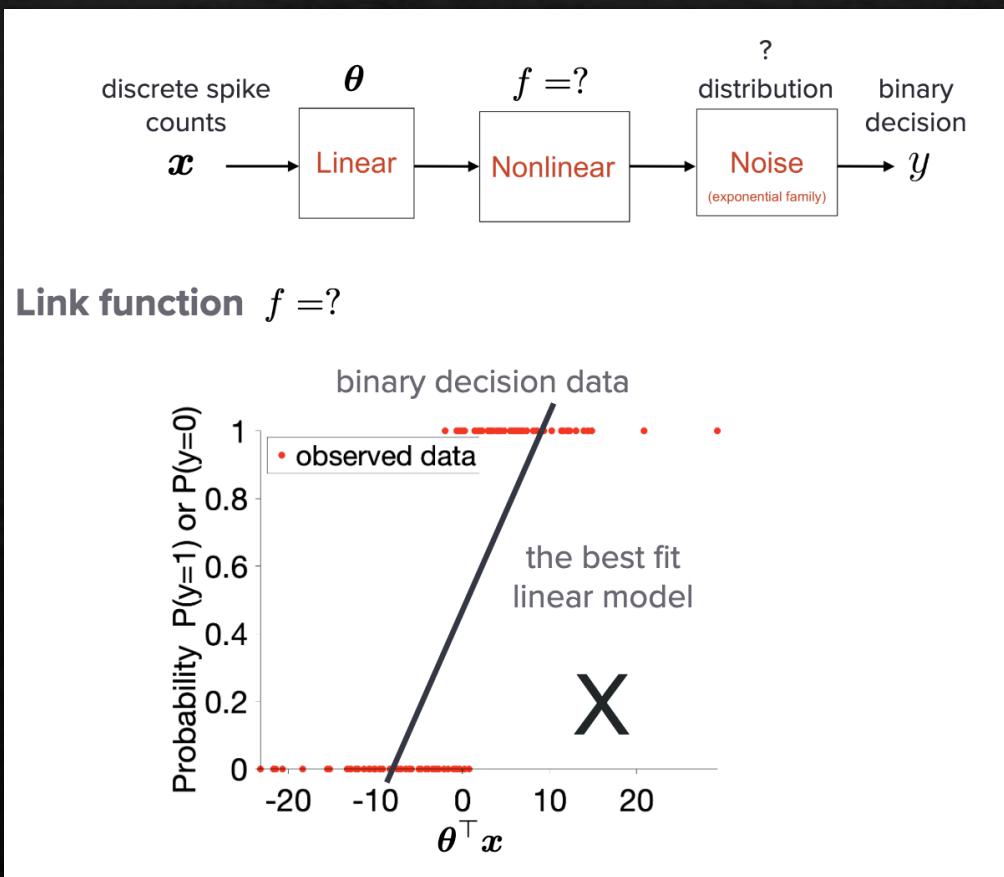
FIGURE 9.5. Why we need link functions. The solid blue line is a linear model of a probability mass. It increases linearly with a predictor, x , on the horizontal axis. But when it reaches the maximum probability mass of 1, at the dashed boundary, it will happily continue upwards, as shown by the dashed blue line. In reality, further increases in x could not further increase probability, as indicated by the horizontal continuation of the solid trend.

$$y_i \sim \text{Binomial}(n, p_i)$$

$$f(p_i) = \alpha + \beta x_i$$

<https://osf.io/2h6ut>

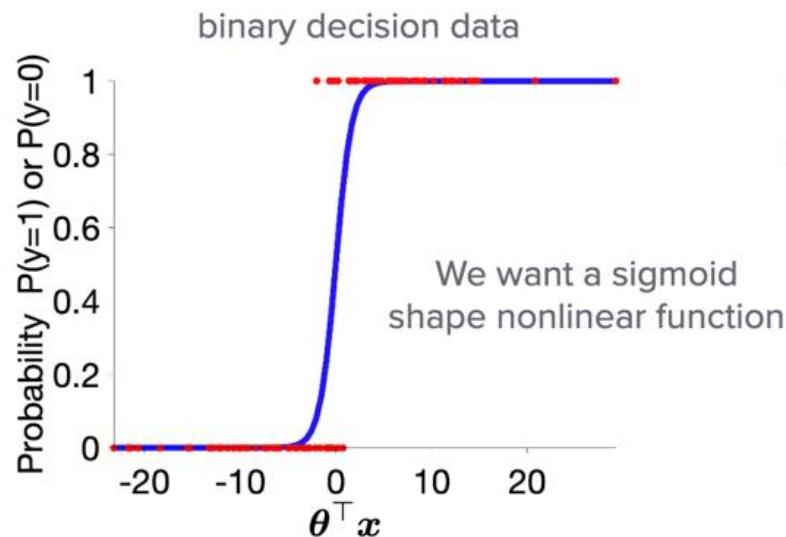
Binary Outcome Variable



Good mathy introduction:
https://www.youtube.com/watch?v=X-ix97pw0xY&ab_channel=MITOpenCourseWare

Binary Outcome Variable

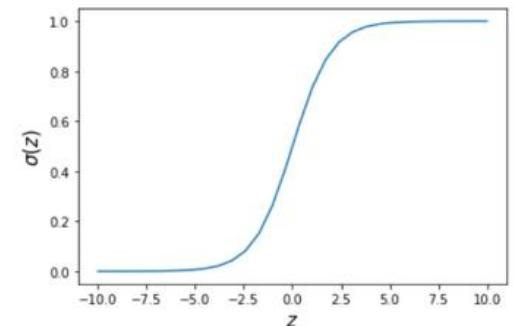
Link function $f = ?$



We define $f = \sigma(\cdot)$ which is a "squashing" function called the sigmoid function.

$$\sigma(z) = \frac{1}{\exp(-z) + 1}$$

Notice $0 \leq f(\boldsymbol{\theta}^\top \mathbf{x}) = \frac{1}{\exp(-\boldsymbol{\theta}^\top \mathbf{x}) + 1} \leq 1$



Binary Outcome Variable

Distribution of the observation noise

$f(\boldsymbol{\theta}^\top \mathbf{x}) = \sigma(\boldsymbol{\theta}^\top \mathbf{x}) = \frac{1}{\exp(-\boldsymbol{\theta}^\top \mathbf{x}) + 1}$ only gives us a probability-like value, not a binary decision.

Bernoulli distribution: generate a binary value with some input probability value.



single coin flip



outcome y:

head

tail

probability:

p

1-p

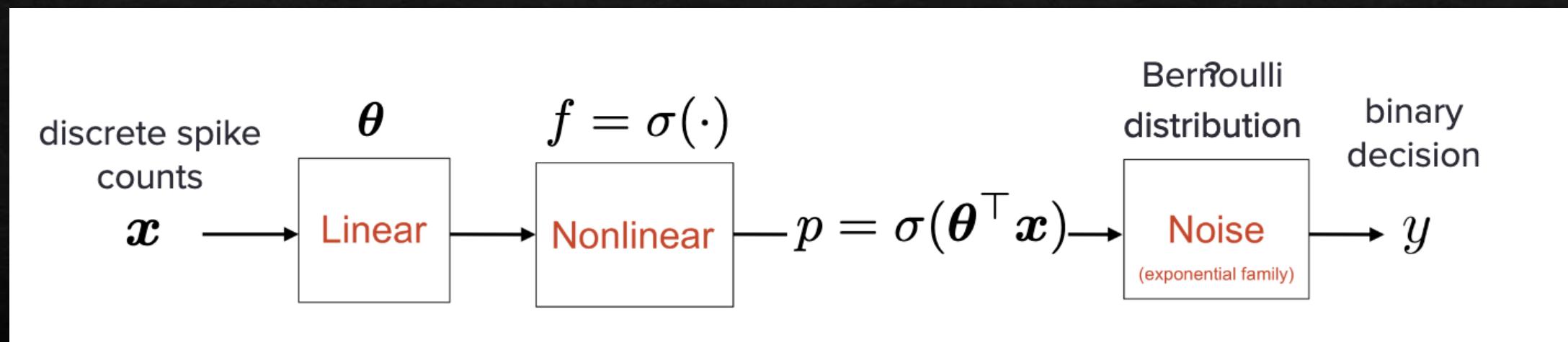
The probability mass function for y is

$$P(y|p) = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{if } y = 0 \end{cases}$$

Alternatively,

$$P(y|p) = p^y (1 - p)^{1-y}$$

Binary Outcome Variable



Binary Outcome Variable

Bernoulli GLM:

(coin flipping model,
 $y = 0 \text{ or } 1$)

$$p_t = f(\vec{x}_t \cdot \vec{k})$$

nonlinearity

$$p(y_t = 1 | \vec{x}_t) = p_t$$

probability of
spike at bin t

Logistic regression:

$$f(x) = \frac{1}{1 + e^{-x}}$$

logistic function

- so logistic regression is a special case of a Bernoulli GLM

Binary Outcome Variable: Logistic Regression

logit(p) = $\beta_0 + \beta_1 * \text{female}$						
Logistic regression						
			Number of obs	=	200	
			LR chi2(1)	=	3.10	
			Prob > chi2	=	0.0781	
Log likelihood =	-109.80312		Pseudo R2	=	0.0139	
<hr/>						
hon		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
<hr/>						
female		.5927822	.3414294	1.74	0.083	-.0764072 1.261972
intercept		-1.470852	.2689555	-5.47	0.000	-1.997995 -.9437087
<hr/>						

$$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 * \text{math} + \beta_2 * \text{female} + \beta_3 * \text{read}$$

$$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 * \text{female} + \beta_2 * \text{math} + \beta_3 * \text{female} * \text{math}$$

Classic Linear Model to Generalized Linear Model

LM:

- 1) *Random Component* : Each component of \underline{Y} is independent and normally distributed.
The mean μ_i allowed to differ, but all Y_i have common variance σ_e^2
- 2) *Systematic Component* : The n covariates combine to give the "linear predictor"

$$\underline{\eta} = \beta \mathbf{X}$$

- 3) *Link Function* : The relationship between the random and systematic components is specified via a link function. In linear model, link function is identity fnc.

$$E[\underline{Y}] = \mu = \underline{\eta}$$

GLM:

- 1) *Random Component* : Each component of \underline{Y} is independent and from one of the exponential family of distributions
- 2) *Systematic Component* : The n covariates are combined to give the "linear predictor"

$$\underline{\eta} = \beta \mathbf{X}$$

- 3) *Link Function* : The relationship between the random and systematic components is specified via a link function g , that is differentiable and monotonic

$$E[\underline{Y}] = \mu = g^{-1}(\underline{\eta})$$

Poisson Regression

models how the mean of a discrete (count) response variable Y depends on a set of explanatory variables

$$\log \lambda_i = \beta_0 + \beta x_i$$

- **Random component** - The distribution of Y is Poisson with mean λ .
- **Systematic component** - x is the explanatory variable (can be continuous or discrete) and is linear in the parameters. As with the above example, this can be extended to multiple variables or non-linear transformations.
- **Link function** - the log link is used.

Binary Logistic Regression

Binary logistic regression models how the odds of "success" for a binary response variable Y depend on a set of explanatory variables:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i$$

- **Random component** - The distribution of the response variable is assumed to be binomial with a single trial and success probability $E(Y) = \pi$.
- **Systematic component** - x is the explanatory variable (can be continuous or discrete) and is linear in the parameters. As with the above example, this can be extended to multiple variables of non-linear transformations.
- **Link function** - the log-odds or logit link, $\eta = g(\pi) = \log\left(\frac{\pi_i}{1 - \pi_i}\right)$, is used.

Lab Practice + Assignment

❖ Housing.csv

longitude
latitude
housing*medianage*
total_rooms
total_bedrooms
population
households
median_income
median*housevalue*
ocean_proximity

Assignment (15 marks)

Part 1 (10 marks)

- ❖ Visualize some correlations between variables in the data set (2 marks)
- ❖ Pick 2 linear regression models (i.e., sets of predictors) to predict median house value
- ❖ Check for collinearity using VIF to remove highly correlated variables from the models (1 mark)
- ❖ Plot the distribution of the residuals against the fitted values to check for heteroscedasticity (1 mark)
- ❖ Use ncvTest or equivalent to test for heteroscedasticity (1 mark) (<https://www.rdocumentation.org/packages/car/versions/3.0-12/topics/ncvTest>)
- ❖ Test for normality of the residuals (use at least one of Wald test, Q-Q plots, etc). 1 mark
- ❖ Compare the 2 models using AIC and pick the best model. 1 mark
- ❖ Report the coefficients of the winning model and their statistics (including confidence intervals) and interpret the resulting model coefficients. 3 marks

Part 2 (5 marks)

- ❖ Binary.csv
- ❖ Predict admission using GRE, GPA, and undergrad institution ranks
- ❖ Admission = 1 or 0. Hence use logistic regression (GLM)
- ❖ Report the statistics, confidence intervals, etc for the logistic regression and interpret the results (what are the most significant variables that predict whether someone will get admitted? Explain in terms odds by exponentiating the coefficients) - **3 marks**
- ❖ Can you test an interaction effect? Let's say GPA matters even more if you are from a lower ranked institution (lower GPAs may be tolerated if you are from a higher ranked institution). So include a GPA*rank term in the model and try to interpret the resulting coefficient. - **2 marks**

Logistic Regression – Qs with binary or binomial responses

- ❖ Are students with poor grades more likely to binge watch Netflix series?
- ❖ Is exposure to a particular chemical associated with a cancer diagnosis?
- ❖ Are the number of votes for a political candidate associated with the amount of money raised by their party?

- ❖ Binomial responses: # of successes in N independent trials, each with probability p of success.

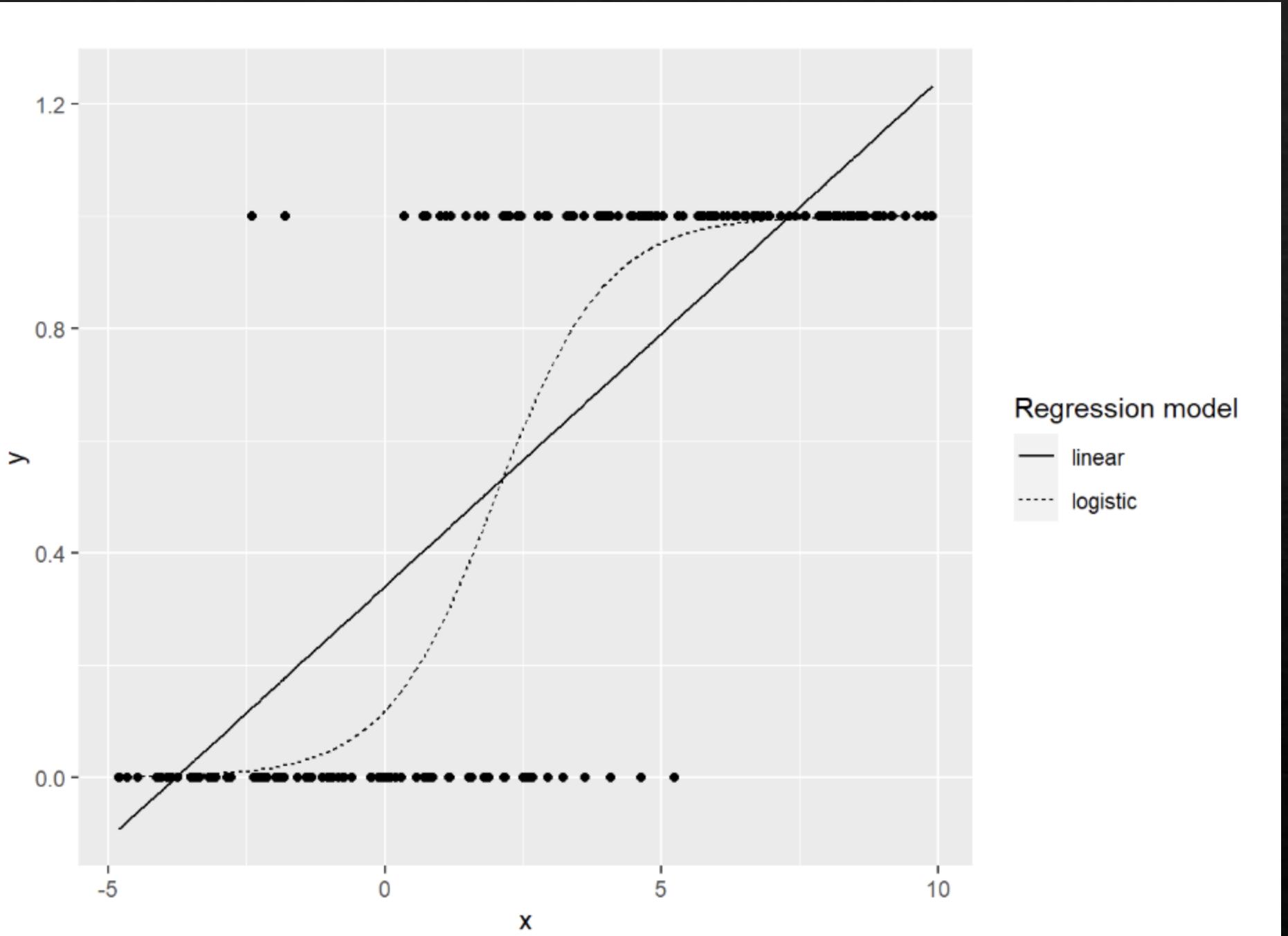
Logistic Regression – Assumptions

Binary Response - The response variable is dichotomous (two possible responses) or the sum of dichotomous responses.

Independence - The observations must be independent of one another.

Variance Structure - By definition, the variance of a binomial random variable is $np(1-p)$, so that variability is highest when $p=.5$.

Linearity - The log of the odds ratio, $\log(p/1-p)$, must be a linear function of x .



$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i$$

where the observed values $Y_i \sim \text{binomial}$ with $p = p_i$ for a given x_i and $n = 1$ for binary responses.

Do goalkeepers save more goals when their team is behind?

	Saves	Scores	Total
Behind	2	22	24
Not Behind	39	141	180
Total	41	163	204

(Source: Roskes et al. 2011.)

Modeling Odds

- ❖ How can we quantify the goalkeeper's performance?
- ❖ Odds that he saves a penalty kick when his team is behind = 2/22

$$\text{Odds} = \frac{\#\text{successes}}{\#\text{failures}} = \frac{\#\text{successes}/n}{\#\text{failures}/n} = \frac{p}{1-p}.$$

Modeling Odds

- ◆ However, Odds are strictly positive. Cannot model it directly as a linear function since we want to model something that can take values from -inf to +inf.
- ◆ So, we will model $\log(\text{odds})$.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Modeling Odds

- ❖ define $X=0$ for not behind and $X=1$ for being behind in the game.

$$\log\left(\frac{p_X}{1 - p_X}\right) = \beta_0 + \beta_1 X$$

$$\log\left(\frac{p_0}{1 - p_0}\right) = \beta_0,$$

Modeling Odds

- ❖ $X=0$ for not behind

$$\log\left(\frac{p_0}{1-p_0}\right) = \beta_0,$$

Modeling Odds

- ❖ $X=1$ for being behind

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1.$$

Modeling Odds

- ❖ What does β_1 stand for?
- ❖ e^{β_1} = odds ratio (ratio of odds of success under one condition and the other condition)

$$\beta_1 = (\beta_0 + \beta_1) - \beta_0 = \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) = \log\left(\frac{p_1/(1-p_1)}{p_0/(1-p_0)}\right).$$

Logistic Regression: Estimating coefficients

$$\log\left(\frac{p_X}{1 - p_X}\right) = \beta_0 + \beta_1 X$$

$$p_X = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Table 6.1: Soccer goalkeepers' penalty kick saves when their team is and is not behind.

	Saves	Scores	Total
Behind	2	22	24
Not Behind	39	141	180
Total	41	163	204

(Source: Roskes et al. 2011.)

$$\text{Lik}(p_1, p_0) = \binom{24}{22} p_1^{22} (1 - p_1)^2 \binom{180}{141} p_0^{141} (1 - p_0)^{39}$$

Logistic Regression: Estimating coefficients - MLE

$$\text{Lik}(p_1, p_0) = \binom{24}{22} p_1^{22} (1 - p_1)^2 \binom{180}{141} p_0^{141} (1 - p_0)^{39}$$

$$\text{Lik}(\beta_0, \beta_1) \propto \\ \left(\frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \right)^{22} \left(1 - \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \right)^2 \left(\frac{e^{\beta_0}}{1 + e^{\beta_0}} \right)^{141} \left(1 - \frac{e^{\beta_0}}{1 + e^{\beta_0}} \right)^{39}$$

$$\hat{\beta}_0 = 1.2852 \text{ and } \hat{\beta}_1 = 1.1127.$$

Logistic Regression: Interpreting the coefs

$$\hat{\beta}_0 = 1.2852 \text{ and } \hat{\beta}_1 = 1.1127.$$

Exponentiate β_1 to get odds ratio

Odds ratio ~ 3 .

Three times likely to score when the goalkeeper's team is behind compared to when they're ahead.

CI, p values, model comparisons using AIC/BIC etc as discussed before.

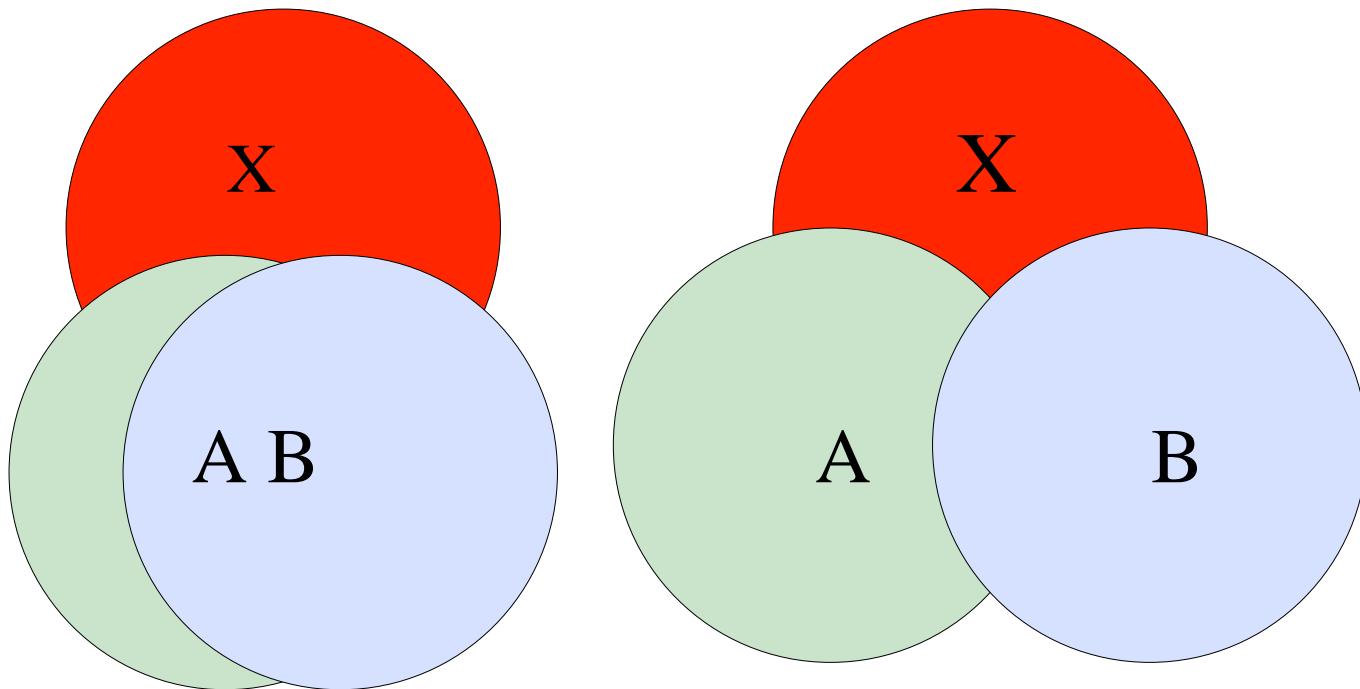
READ: <https://bookdown.org/roback/bookdown-BeyondMLR/ch-logreg.html#introduction-to-logistic-regression>

Part 2 (5 marks)

- ❖ Binary.csv
- ❖ Predict admission using GRE, GPA, and undergrad institution ranks
- ❖ Admission = 1 or 0. Hence use logistic regression (GLM)
- ❖ Report the statistics, confidence intervals, etc for the logistic regression and interpret the results (what are the most significant variables that predict whether someone will get admitted? Explain in terms odds by exponentiating the coefficients) - **3 marks**
- ❖ Can you test an interaction effect? Let's say GPA matters even more if you are from a lower ranked institution (lower GPAs may be tolerated if you are from a higher ranked institution). So include a GPA*rank term in the model and try to interpret the resulting coefficient. - **2 marks**

Multicollinearity Data Reduction (FA & PCA)

Multicollinearity



- high degree of correlation amongst IVs
 - ex: height and weight, household income and water consumption, mileage and price of a car

Multicollinearity in IVs

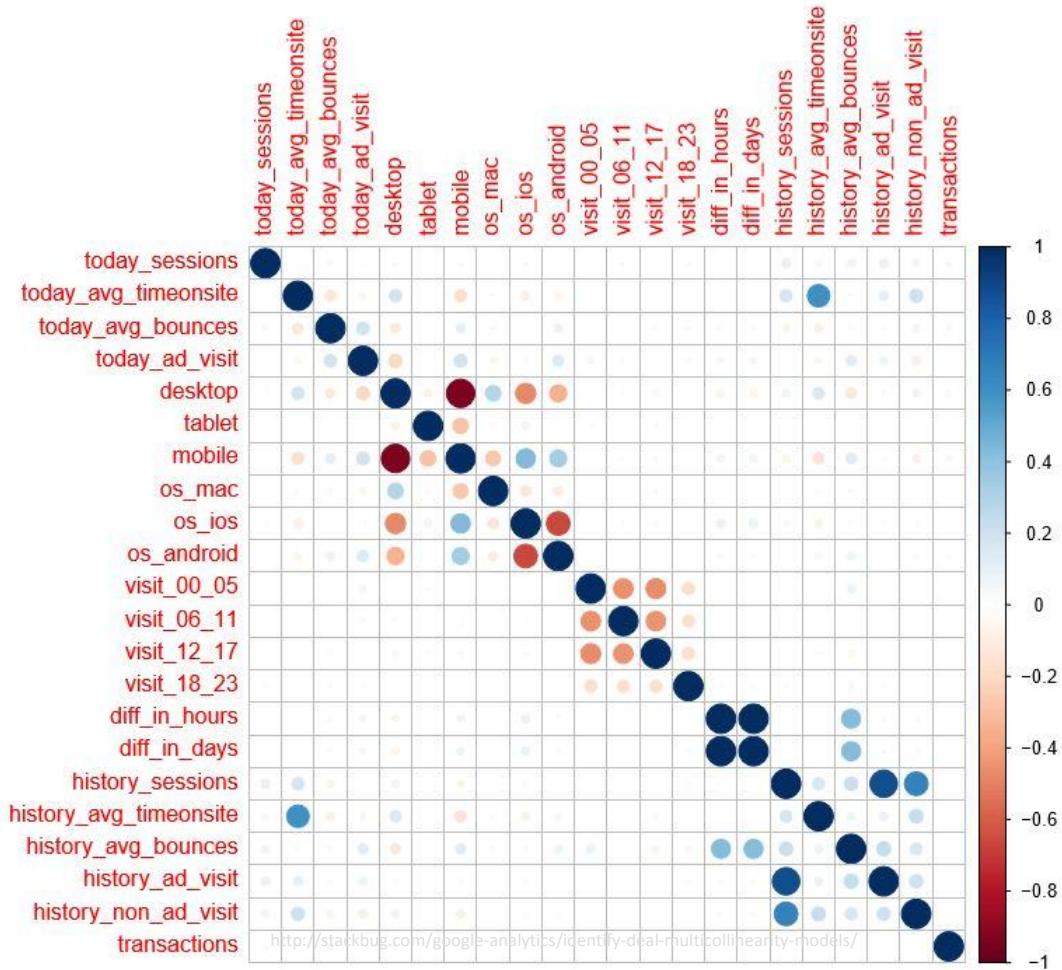
- causes unwanted effects
 - saps statistical power of the analysis
 - can cause switch in signs of the coefficients (in regression), overestimate standard errors, reduced precision in estimating the coefficients' effects, etc...
 - will result in less reliable statistical inferences
- higher number of IV —> increase in sample size required
- what can you do?
 - removing highly correlated IVs / features / items / predictors
 - combine them/uncover latent dimensions **[PCA, FA]**

Multicollinearity

- Some Solutions:
 - **Feature or Variable Selection**
 - **Reduce by Combining Variables**
- choice depends upon
 - research inquiry
 - interpretability

Feature or Variable Selection

- **Correlation:** helps identify collinear variables



where R_j^2 is the R²-value obtained by regressing the j^{th} predictor on the remaining predictors.

Feature or Variable Selection

- **Variance Inflation Factor (VIF)**
 - The R-square term tells us
 - how predictable one IV is from the set of other IVs
 - 1 = not correlated.
 - Between 1 and 5 = moderately correlated.
 - Greater than 5 = highly correlated.

$$VIF_j = \frac{1}{1 - R_j^2}$$

Where, R_j^2 is the R²-value obtained by regressing the j^{th} predictor on the remaining predictors.

EXAMPLE

Feature or Variable Selection

	Gender	Age	Years of service	Education level	Salary
0	0.0	27.0	1.7	0.0	39343.0
1	1.0	26.0	1.1	1.0	43205.0
2	1.0	26.0	1.2	0.0	47731.0
3	0.0	27.0	1.6	1.0	46525.0
4	0.0	26.0	1.5	1.0	40891.0

	variables	VIF
0	Gender	2.207155
1	Age	13.706320
2	Years of service	10.299486
3	Education level	2.409263

	variables	VIF
0	Gender	1.863482
1	Years of service	2.478640
2	Education level	2.196539

Dropping Age

	variables	VIF
0	Gender	2.168068
1	Education level	2.407695
2	Age_at_joining	3.326991

(Age - Years of service)
Combining Age & Service

Multicollinearity

- **Solutions:**
 - *Feature or Variable Selection*
 - *Reduce by Combining Variables*
- choice depends upon
 - research inquiry
 - interpretability vs model performance

Feature Set Reduction

- Why?
 - increase in dimensions -> complex data -> harder to interpret
 - additional variables = additional processing time and space
 - avoid curse of dimensionality -> amount of data needed to support the result often grows exponentially with the dimensionality
 - reduce overfitting
 - help eliminate irrelevant features
 - easier visualisation

Research Question?

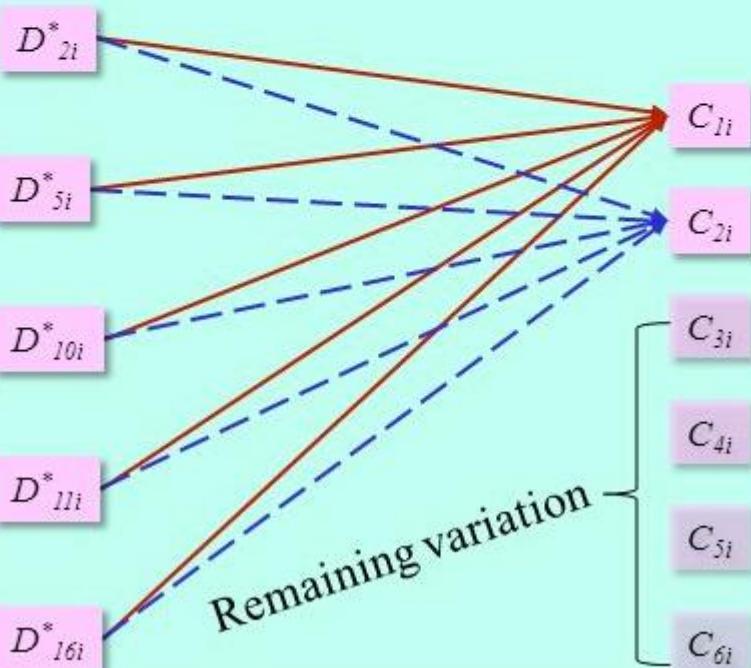
Rather than asking ... “Can We Forge These Several Indicators Together Into A Smaller Number Of Composites With Defined Statistical Properties?”

Then, we would need ...
Principal Components Analysis (PCA)

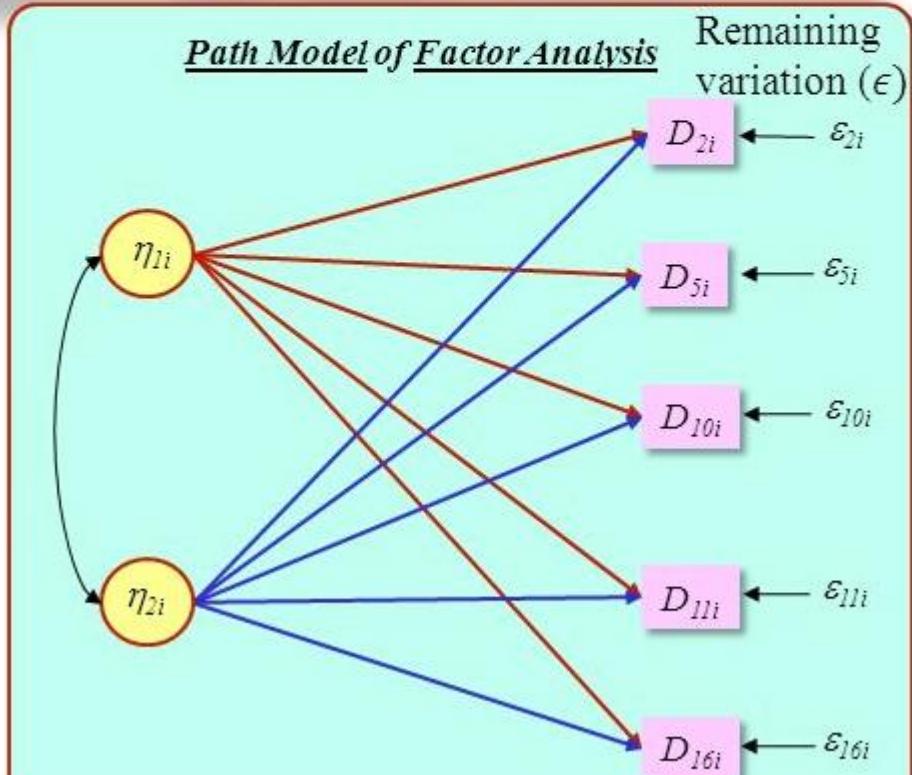
We could ask ... “Are There A Number Of Unseen (Latent) Factors (Constructs) Acting “Beneath” These Indicators To Forge Their Observed Values?”

Instead, we would need ...
Factor Analysis (CFA or EFA?)

Path Model of Principal Components Analysis

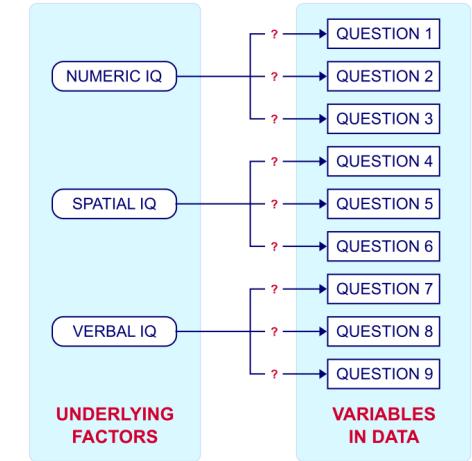


Path Model of Factor Analysis



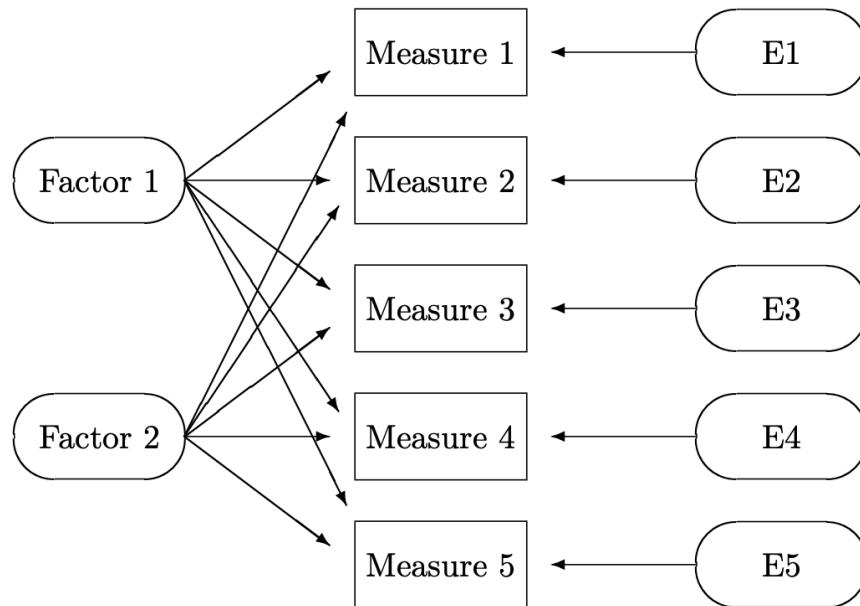
Factor Analysis

- idea—> there are underlying “latent” variables or “factors”, and several variables might be measures of the same factor
- underlying/latent dimensions are not directly observable
- hidden constructs/factors give rise to observed variables



Factor Analysis

- condense information into **factors** with minimum information loss
- predetermined no. of factors (intrinsic dimensionality estimation)



Factor Analysis

$$\mathbf{X} = \mu + \mathbf{Lf} + \epsilon$$

$$\begin{aligned} X_1 &= \mu_1 + l_{11}f_1 + l_{12}f_2 + \cdots + l_{1m}f_m + \epsilon_1 \\ X_2 &= \mu_2 + l_{21}f_1 + l_{22}f_2 + \cdots + l_{2m}f_m + \epsilon_2 \\ &\vdots \\ X_p &= \mu_p + l_{p1}f_1 + l_{p2}f_2 + \cdots + l_{pm}f_m + \epsilon_p \end{aligned}$$

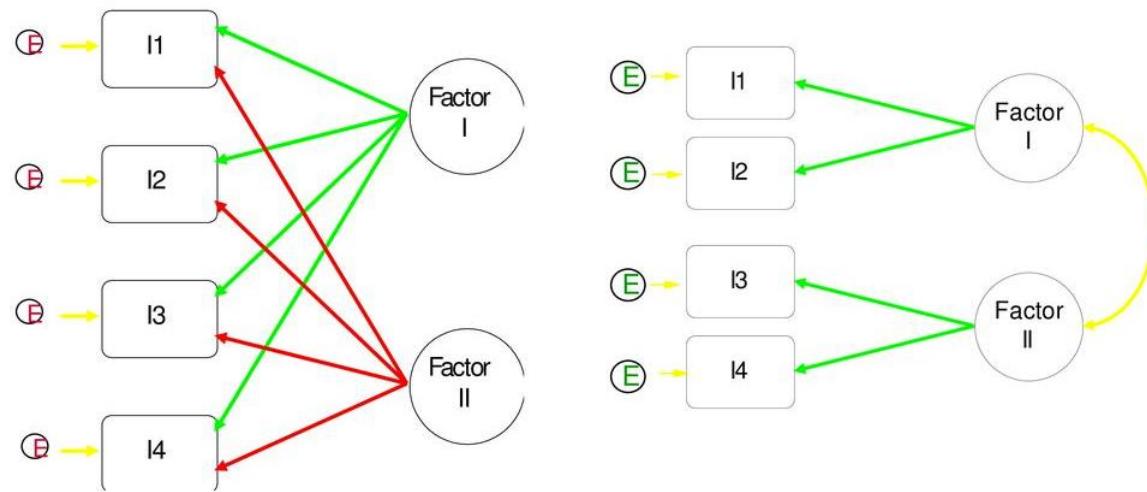
$$\mathbf{L} = \begin{pmatrix} l_{11} & l_{12} & \dots & l_{1m} \\ l_{21} & l_{22} & \dots & l_{2m} \\ \vdots & \vdots & & \vdots \\ l_{p1} & l_{p2} & \dots & l_{pm} \end{pmatrix} = \text{matrix of factor loadings}$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix} = \text{vector of specific factors}$$

*error terms,
what the Factors
cannot explain
in each variable*

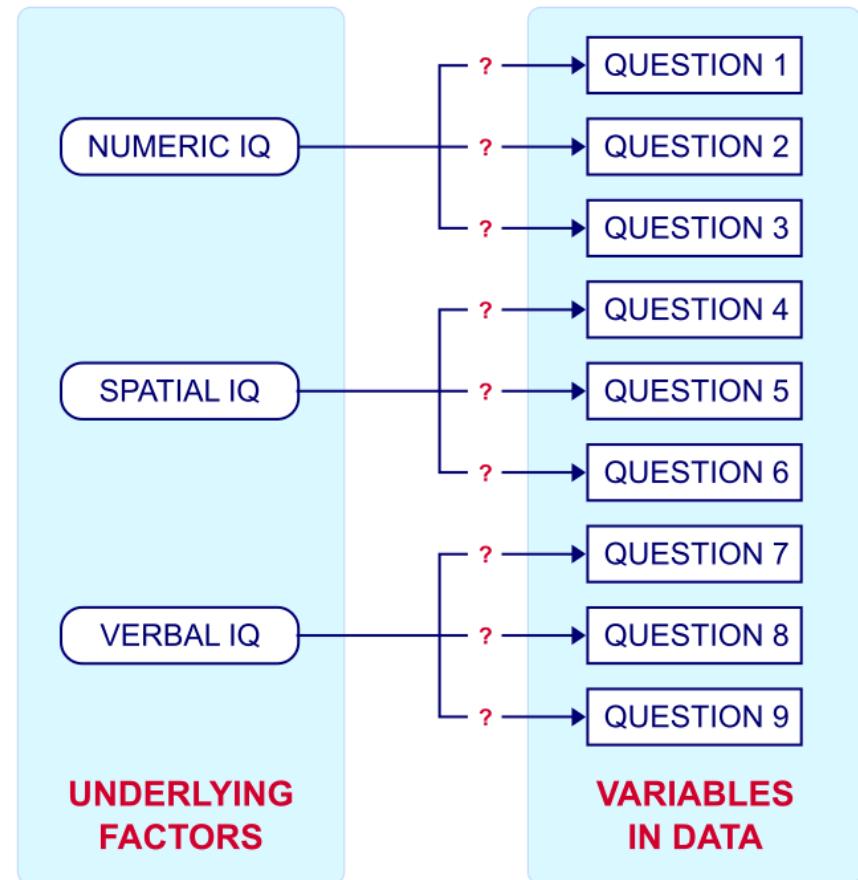
Factor Analysis

- **Exploratory Factor Analysis:** *data-driven*
 - explore underlying structure
- **Confirmatory Factor Analysis:** *theory-driven*
 - confirm or reject pre-established theory

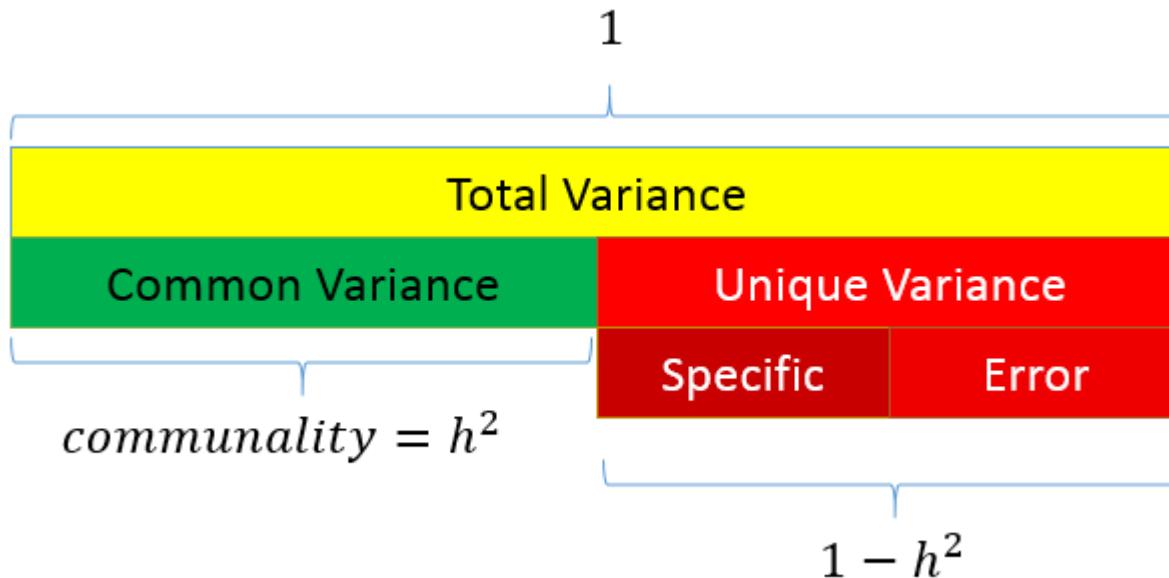


EFA: Factor Analysis Types

- **R-Type** (commonly used)
 - covariation or correlation between variables



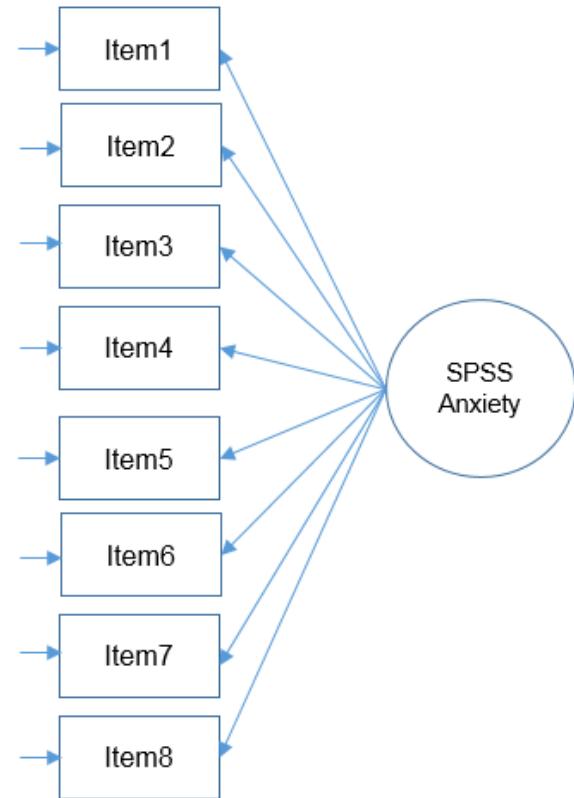
Factor Analysis



The total variance is made up to common variance and unique variance, and unique variance is composed of specific and error variance. If the total variance is 1, then the communality is h^2 and the unique variance is $1-h^2$.

EXAMPLE

1. Statistics makes me cry
2. My friends will think I'm stupid for not being able to cope with SPSS
3. Standard deviations excite me
4. I dream that Pearson is attacking me with correlation coefficients
5. I don't understand statistics
6. I have little experience with computers
7. All computers hate me
8. I have never been good at mathematics



Do all these items actually measure what we call “SPSS Anxiety”?

EXAMPLE

	My friends will think I'm stupid for not being able to cope with SPSS	Standard deviations excite me	I dream that Pearson is attacking me with correlation coefficients	I don't understand statistics	I have little experience with computers	All computers hate me	I have never been good at mathematics	
Statistics makes me cry	1							
My friends will think I'm stupid for not being able to cope with SPSS	-.099	1						
Standard deviations excite me	-.337	.318	1					
I dream that Pearson is attacking me with correlation coefficients	.436	-.112	-.380	1				
I don't understand statistics	.402	-.119	-.310	.401	1			
I have little experience with computers	.217	-.074	-.227	.278	.257	1		
All computers hate me	.305	-.159	-.382	.409	.339	.514	1	
I have never been good at mathematics	.331	-.050	-.259	.349	.269	.223	.297	1

Inter-scale/item correlation

Factor Matrix^a

EXAMPLE

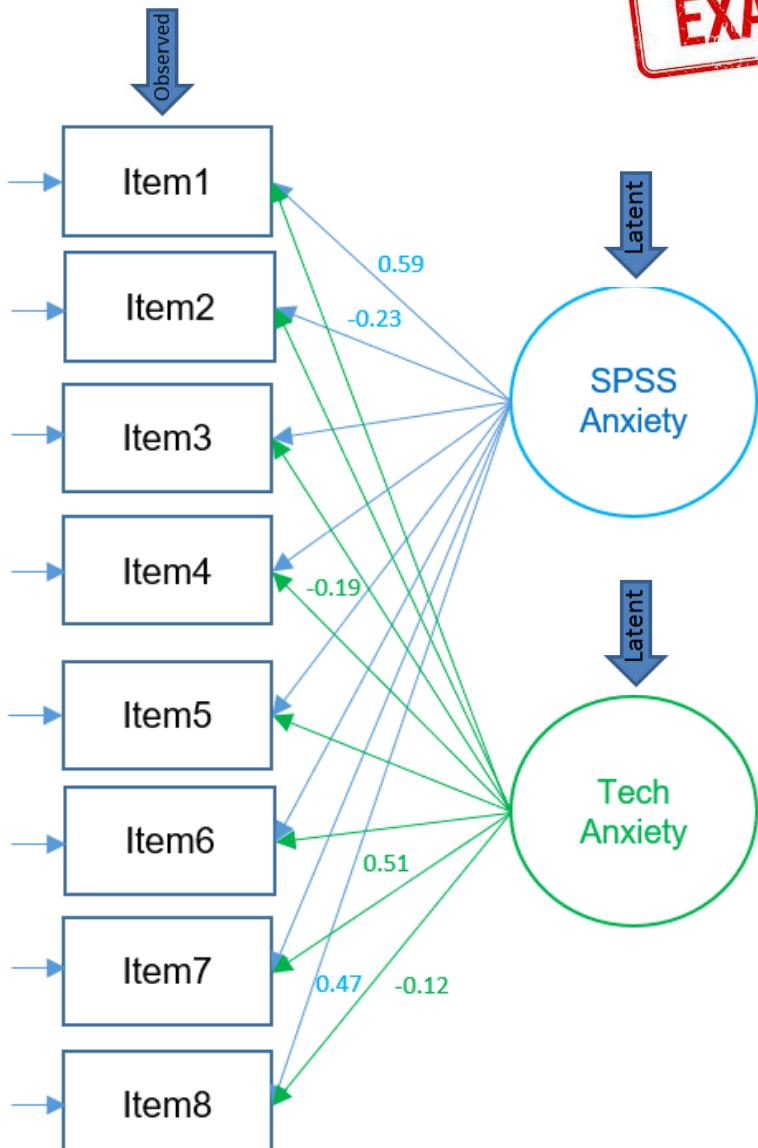
1. Statistics makes me cry
2. My friends will think I'm stupid for not being able to cope with SPSS
3. Standard deviations excite me
4. I dream that Pearson is attacking me with correlation coefficients
5. I don't understand statistics
6. I have little experience with computers
7. All computers hate me
8. I have never been good at mathematics

	Factor	
	1	2
Statistics makes me cry	.588	-.303
My friends will think I'm stupid for not being able to cope with SPSS	-.227	.020
Standard deviations excite me	-.557	.094
I dream that Pearson is attacking me with correlation coefficients	.652	-.189
I don't understand statistics	.560	-.174
I have little experience of computers	.498	.247
All computers hate me	.771	.506
I have never been good at mathematics	.470	-.124

Factor Loadings: the weight of the factor in predicting the variable/correlations between variables and factors

Factor Matrix^a

EXAMPLE



Note: only selected loadings shown

	Factor	1	2
Statistics makes me cry		.588	-.303
My friends will think I'm stupid for not being able to cope with SPSS		-.227	.020
Standard deviations excite me		-.557	.094
I dream that Pearson is attacking me with correlation coefficients		.652	-.189
I don't understand statistics		.560	-.174
I have little experience of computers		.498	.247
All computers hate me		.771	.506
I have never been good at mathematics		.470	-.124

EXAMPLE

Factor Interpretation

F1: customer experience post boarding

F2: airline booking experience and related perks

F3: flight competitive advantage of the airline compared to its competition

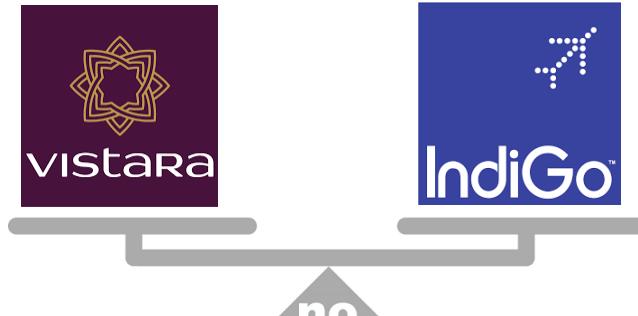
	Factor 1	Factor 2	Factor 3
Great hospitality	0.98	-0.04	0.02
Flight is on time	0.95	-0.01	0.18
Great Food	0.92	0.04	-0.05
Friendly atmosphere	0.62	0.17	-0.33
Frequent flyer program	-0.03	0.97	-0.01
Flights are economic	-0.02	0.96	0.09
No hassles in boarding	-0.07	0.95	0.09
Good flight times	-0.09	0.19	0.96
Seats are comfortable	0.03	0.09	0.95
Loyalty or attachment	-0.19	-0.42	-0.09

ex: factor loadings for an airlines survey

Factor Scores

- composite scores represented by the latent variable which can be used in subsequent statistical analyses (ex: multiple regression, t-tests, etc.)

F1: customer experience post boarding



no significant difference

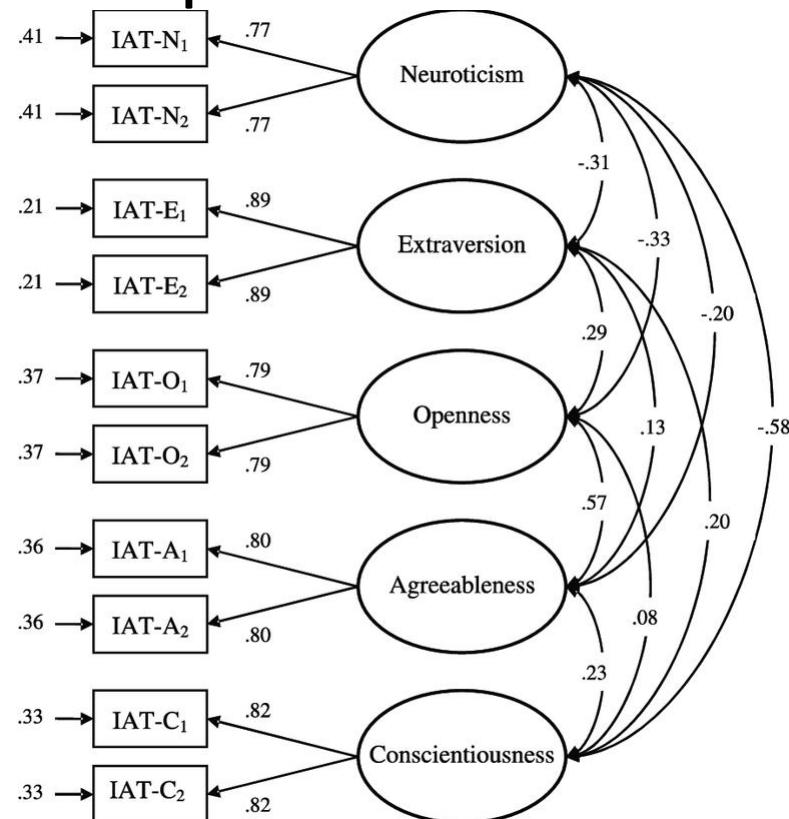
F2: airline booking experience and related perks

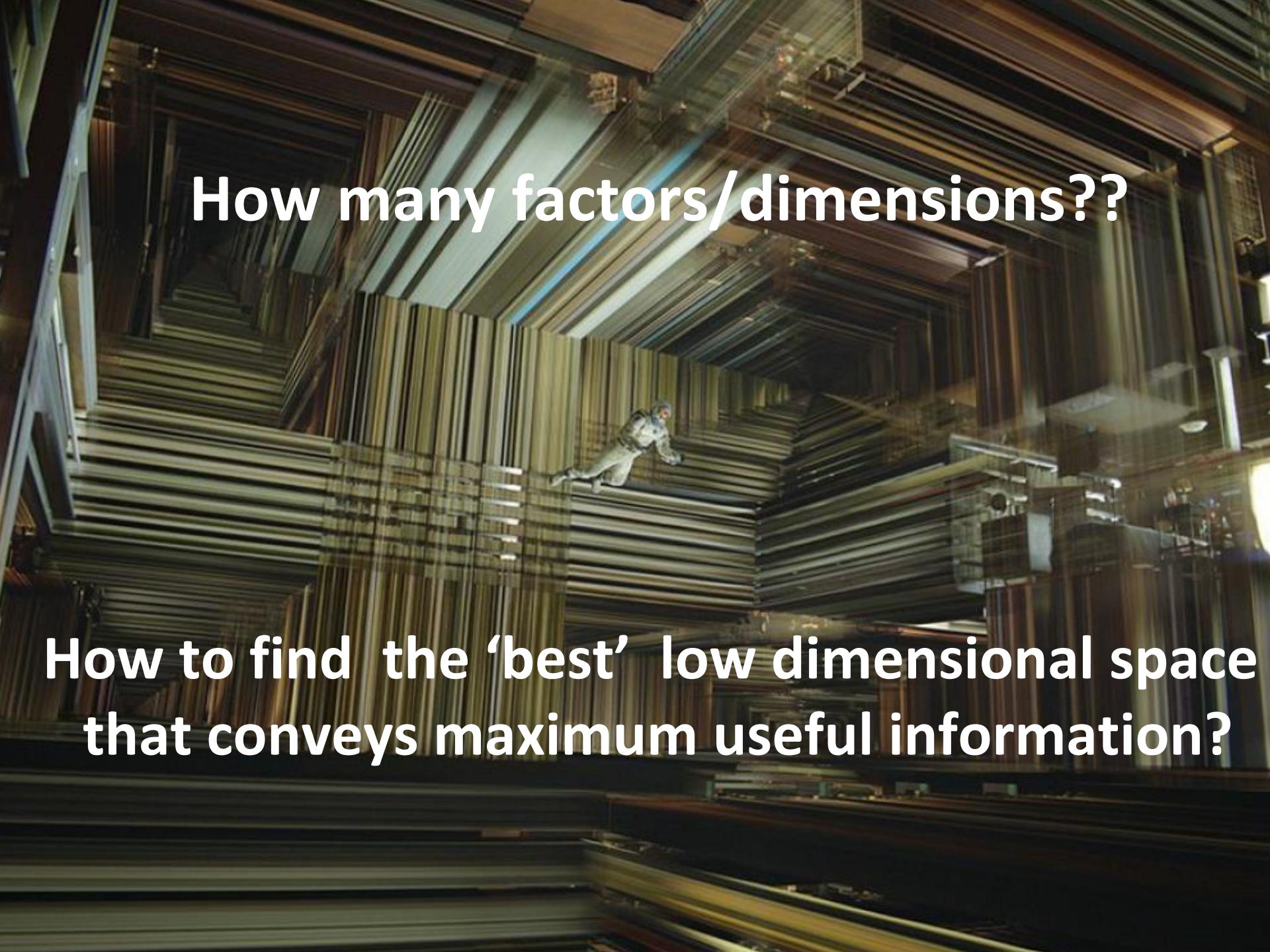


Factor Analysis Types

- **Q-Type**

- similar to clustering of people
- allows identification of groups
- ex: participant X's responses are similar to Y's



A photograph of a person standing on a massive, multi-layered stack of vinyl records. The records are stacked in a grid pattern, creating a deep, perspective-like tunnel effect that extends into the distance. The lighting is dramatic, with strong highlights on the edges of the records and deep shadows in the center of the stack. The person is positioned in the middle ground, looking towards the camera. The background is dark, making the metallic surfaces of the records stand out.

How many factors/dimensions??

**How to find the ‘best’ low dimensional space
that conveys maximum useful information?**

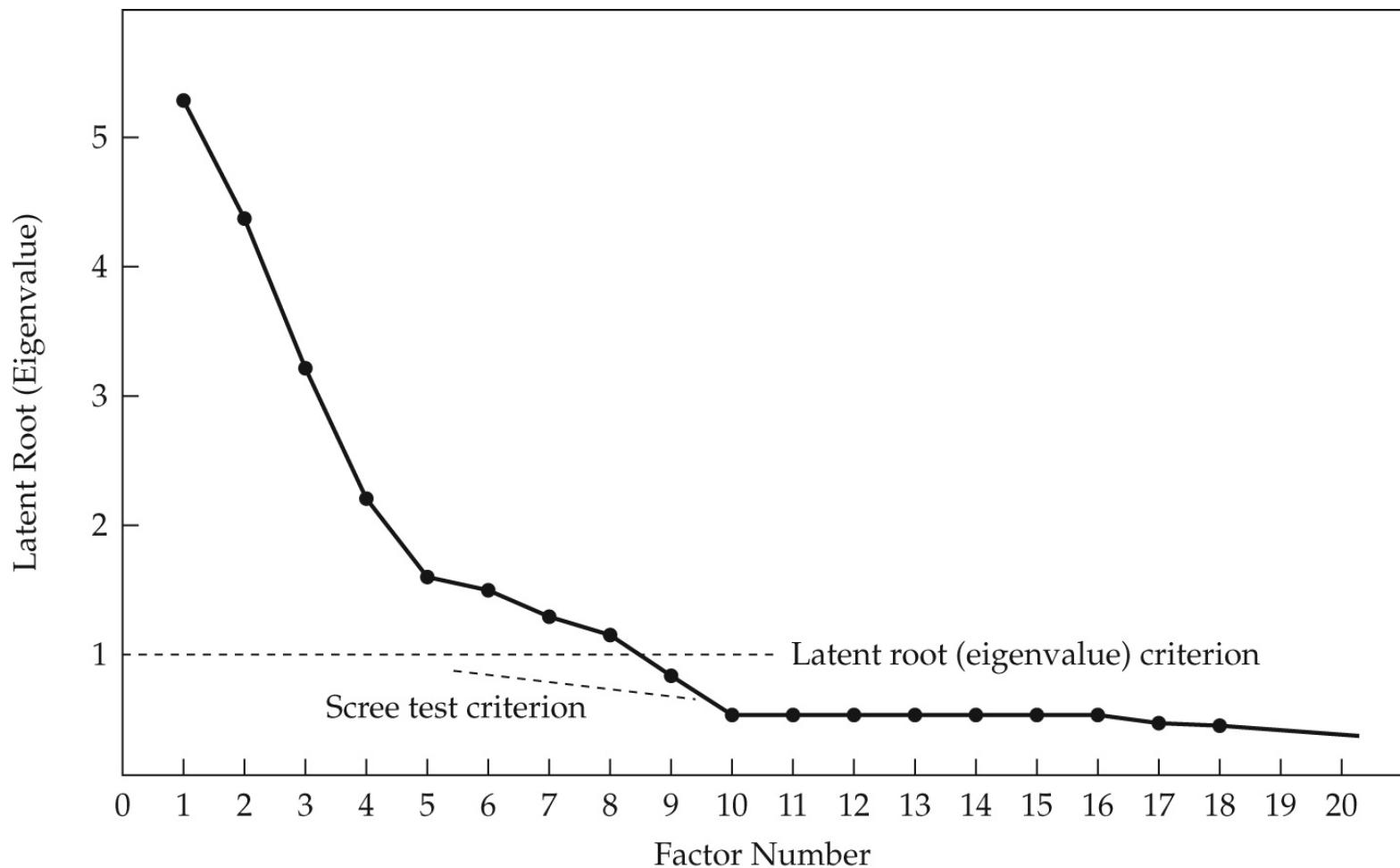
Dimensionality Estimation

- **a priori criterion**
 - define a priori the number of factors to be extracted (testing a hypothesis about the number of factors)
 - trade off - representativeness vs parsimony
- **latent Root criterion**
 - any individual factor should account for the variance of at least one single variable – latent root or eigenvalue >1
- **scree plot/test**
 - point of inflexion in latent root plot

Terminology

- **Scree Plot**
 - plots eigenvalue against component number
 - components with eigenvalues greater than 1 are retained (they are the 'principal' components)
 - components with eigenvalues less than 1 are of little use because they account for less of the variance than the original variable

Scree Plot



Dimensionality Estimation

- **parallel Analysis** (widely used)
 - based on the Monte Carlo simulation
 - creating a random dataset with the same numbers of observations and variables as the original data
 - compare eigenvalues from the random data with original data

Dimensionality Estimation Example

Healthy-Unhealthy Music Scale (HUMS)

Most people believe that music is a helpful part of their lives, but sometimes it's not. When you answer the questions below, please try to recall actual moments when music has been helpful and when it has not.

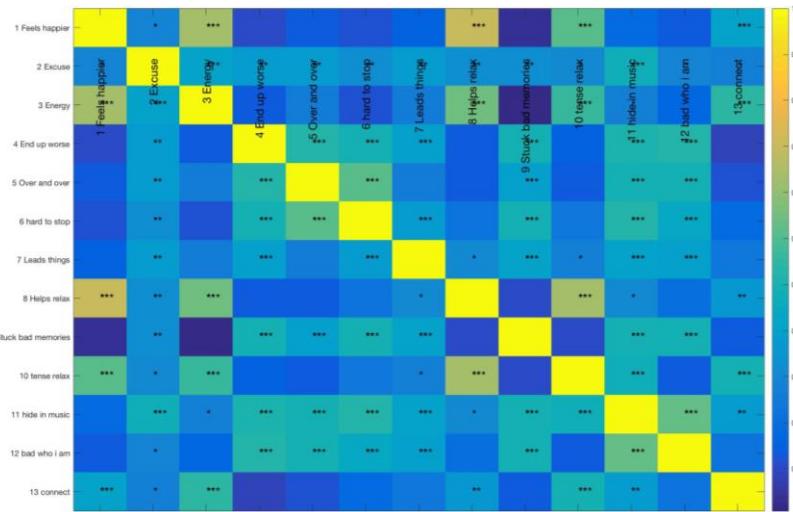
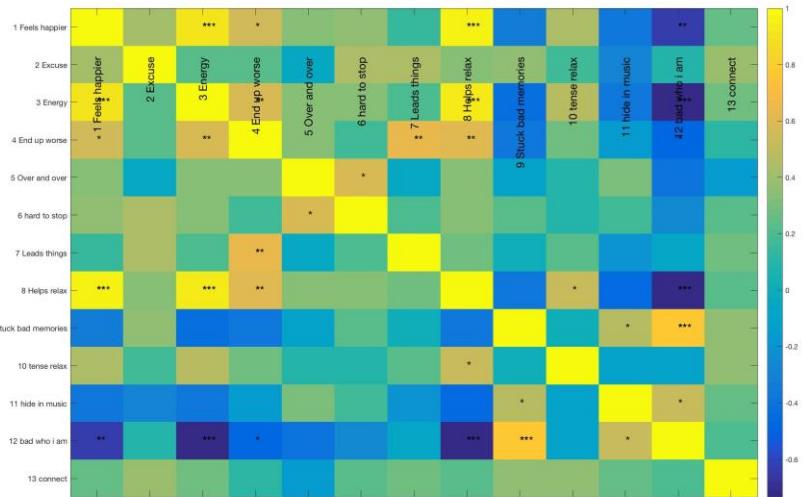
Please read each statement and mark how much it applies to you. Mark only one answer for each question.

	Never	Rarely	Sometimes	Often	Always
1. When I listen to music I get stuck in bad memories	<input type="checkbox"/>				
2. I hide in my music because nobody understands me, and it blocks people out	<input type="checkbox"/>				
3. Music helps me to relax	<input type="checkbox"/>				
4. When I try to use music to feel better I actually end up feeling worse	<input type="checkbox"/>				
5. I feel happier after playing or listening to music	<input type="checkbox"/>				
6. Music gives me the energy to get going	<input type="checkbox"/>				
7. I like to listen to songs over and over even though it makes me feel worse	<input type="checkbox"/>				
8. Music makes me feel bad about who I am	<input type="checkbox"/>				
9. Music helps me to connect with other people who are like me	<input type="checkbox"/>				
10. Music gives me an excuse not to face up to the real world	<input type="checkbox"/>				
11. It can be hard to stop listening to music that connects me to bad memories	<input type="checkbox"/>				
12. Music leads me to do things I shouldn't do	<input type="checkbox"/>				
13. When I'm feeling tense or tired in my body music helps me to relax	<input type="checkbox"/>				

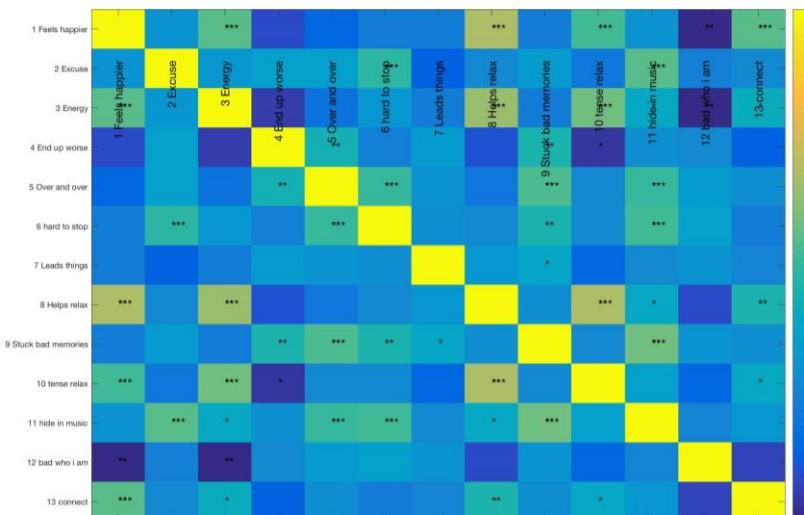
RM class 2018
25 students

Inter-Scale/Item Correlation

141 Indians

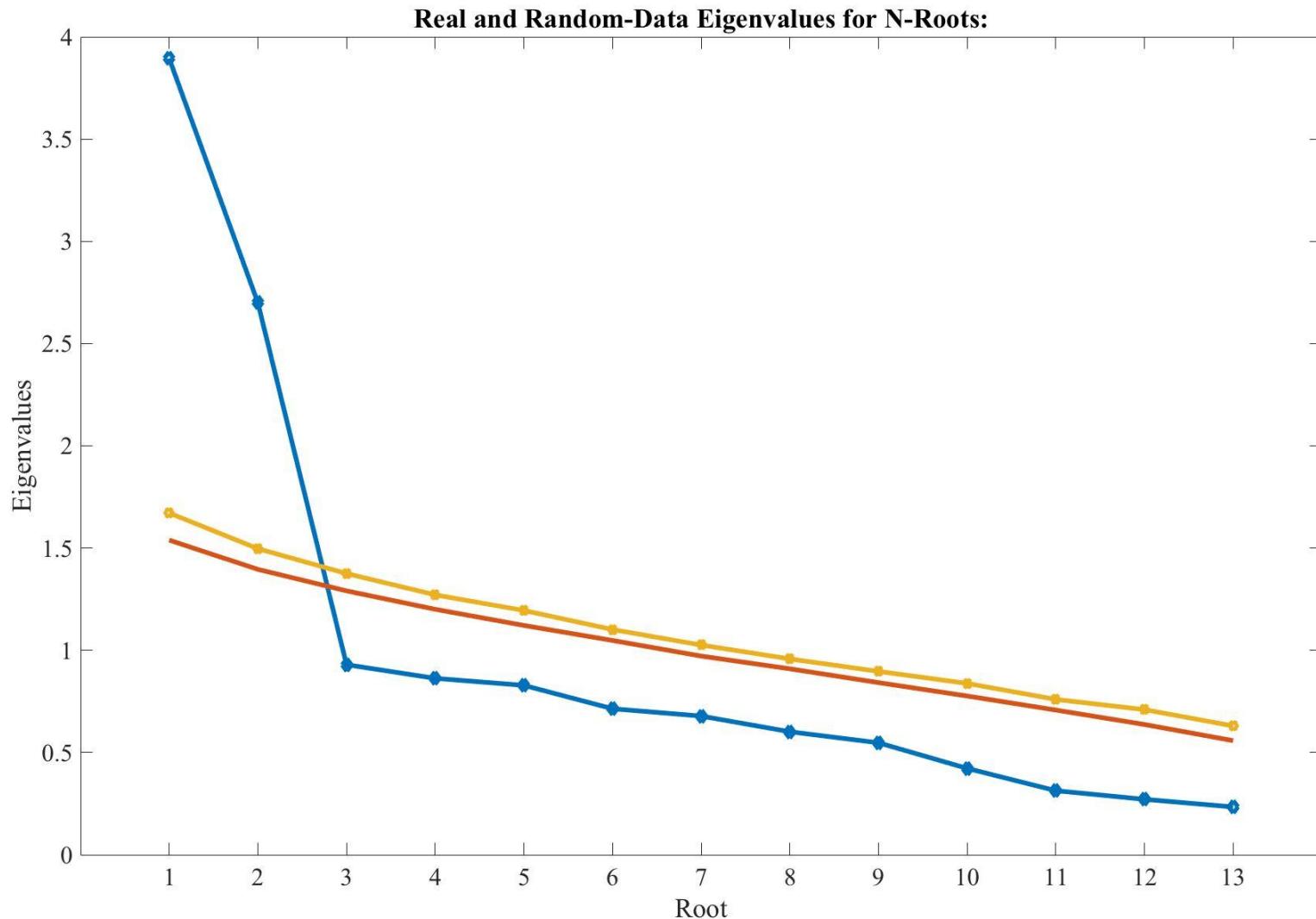


102 British



Parallel Analysis

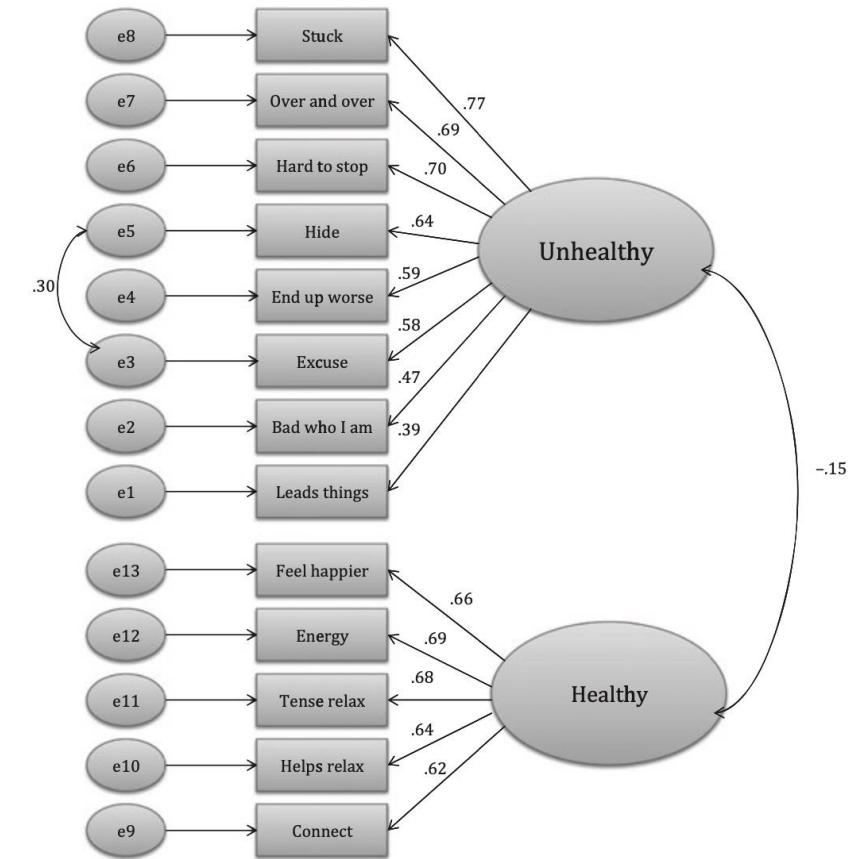
141 Indians



Factor Interpretation

Table 2. The factor loadings (pattern matrix) of the final version of Healthy-Unhealthy Music Scale

Items	F1	F2
When I listen to music I get stuck in bad memories	.760	-.033
I like to listen to songs over and over even though it makes me feel worse	.714	-.092
It can be hard to stop listening to music that connects me to bad memories	.658	.187
I hide in my music because nobody understands me, and it blocks people out	.639	.156
When I try to use music to feel better I actually end up feeling worse	.627	-.163
Music gives me an excuse not to face up to the real world	.571	.249
Music makes me feel bad about who I am	.521	-.186
Music leads me to do things I shouldn't do	.428	-.103
I feel happier after playing or listening to music	-.157	.708
Music gives me the energy to get going	-.005	.692
When I'm feeling tense or tired in my body music helps me to relax	-.028	.667
Music helps me to relax	.040	.621
Music helps me to connect with other people who are like me	-.061	.608



Principal Component Analysis

Research Question?

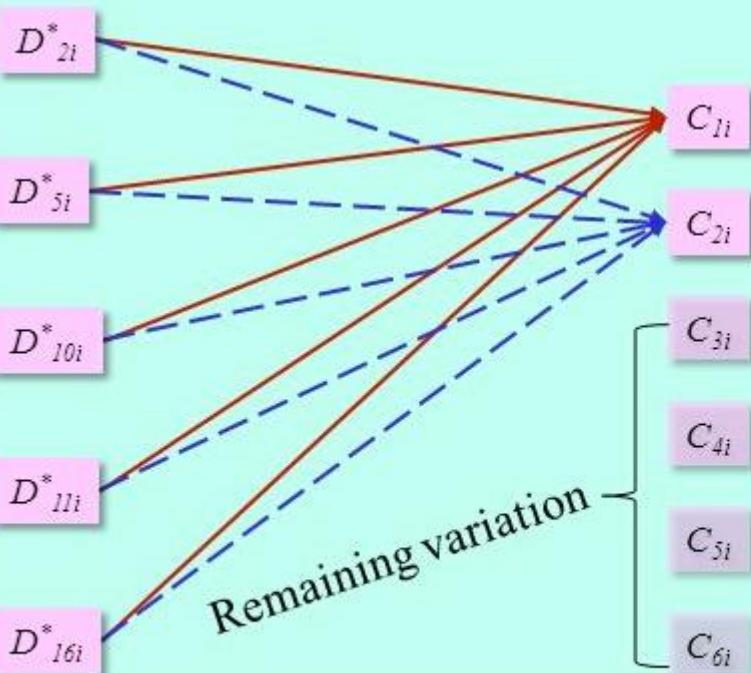
Rather than asking ... “Can We Forge These Several Indicators Together Into A Smaller Number Of Composites With Defined Statistical Properties?”

We could ask ... “Are There A Number Of Unseen (Latent) Factors (Constructs) Acting “Beneath” These Indicators To Forge Their Observed Values?”

Then, we would need ...

Principal Components Analysis (PCA)

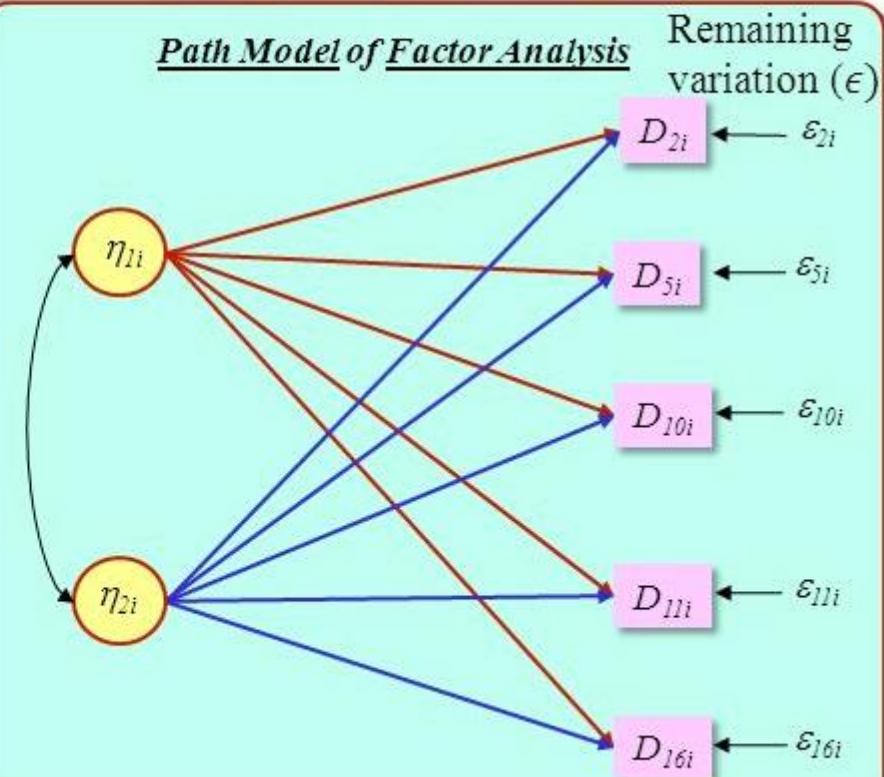
Path Model of Principal Components Analysis



Instead, we would need ...

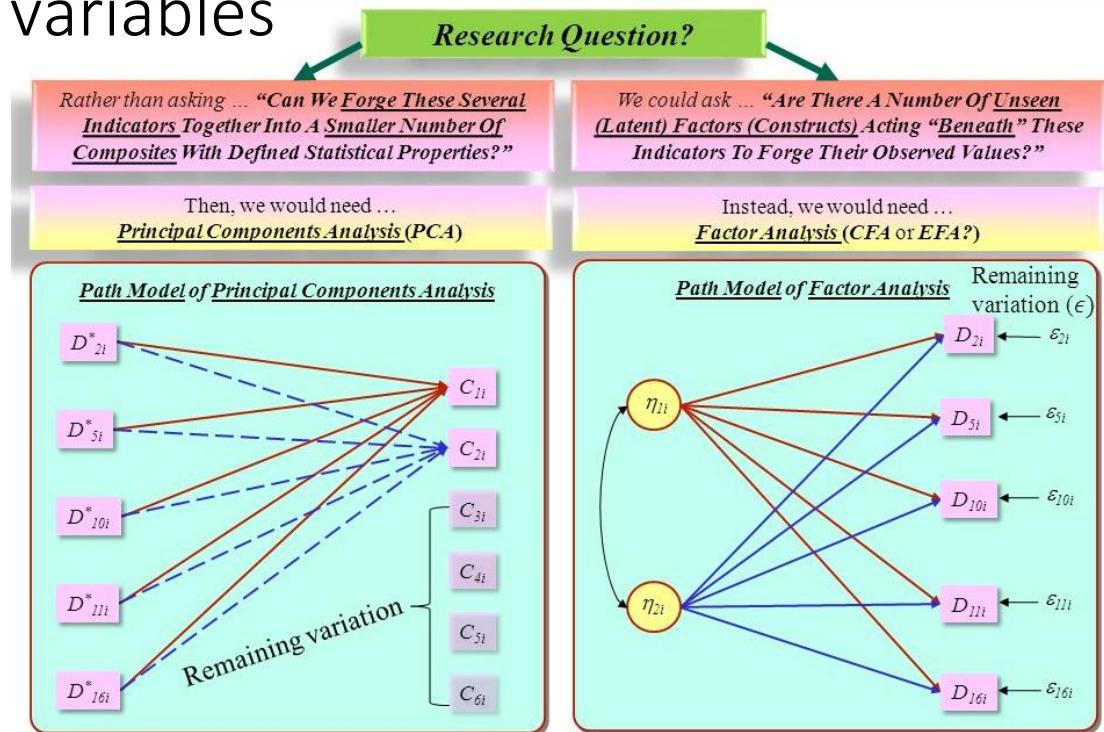
Factor Analysis (CFA or EFA?)

Path Model of Factor Analysis

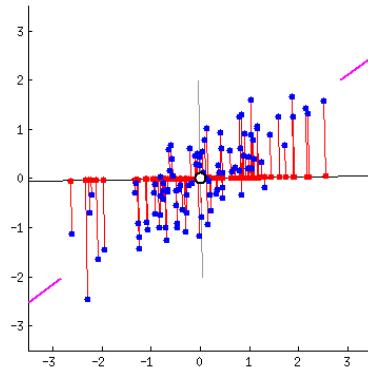


PCA

- idea —> reduce the number of variables of a data set while preserving as much information as possible.
- dimensionality reduction by creating linear combinations of variables



PCA



- Example: Combining two variables into a single component
 - Fit a regression line that represents the 'best' summary of the linear relationship between the variables
 - This line, representing a new component, would capture most of the 'essence' of the two variables

PCA

- If there are more than two variables...
 - this process is repeated until all variables have been assigned to a component
 - gives as many components as variables in decreasing order of variance explained
 - however, only the first few components are likely to be useful..

PCA

- Assumptions:
 - at least interval level data
 - a linear relationship between all variables
 - sampling adequacy (KMO, ~ 15 cases/variable), Bartlett's test of sphericity
 - normally distributed (no outliers)

- Subtract mean from data (center \mathbf{X})
- (Typically) scale each dimension by its variance
 - Helps to pay less attention to magnitude of dimensions
- Compute covariance matrix S
$$S = \frac{1}{N} \mathbf{X}^T \mathbf{X}$$
- Compute k largest eigenvectors of S
- These eigenvectors are the k principal components

<https://www.youtube.com/watch?v=g-Hb26agBFg>

<https://www.youtube.com/watch?v=PFDu9oVAE-g>

Principal Components

- principal components: linear combinations of original variables that result in an axis or a set of axes that explain most of the variability in the dataset
- variables that correlate highly with each other are grouped together into underlying variables, or components
- In mathematical terms, we can say that the first Principal Component is the eigenvector of the covariance matrix corresponding to the maximum eigenvalue

Component Scores & Loadings

- each original variable is assigned a component score and a component loading
- **Component scores** = score/projection on a given component (can be used in subsequent statistical analyses, e.g., regression)
- **Component loadings** = correlation of the original variable with a given component - can be used to determine the importance of a particular variable to a component (Higher loadings = more important)

Dimensionality Estimation

- **Percentage of Variance** criterion
 - achieving a specified cumulative percentage of total variance.
 - typical values – natural sciences ~95%;
 - typical values – social sciences > ~60%
- **Parallel Analysis** (widely used)
 - based on the Monte Carlo simulation
 - creating a random dataset with the same numbers of observations and variables as the original data
 - compare eigenvalues from the random data with original data

Dimensionality Estimation

- **latent Root criterion**
 - any individual factor should account for the variance of at least one single variable – latent root or eigenvalue >1
- **scree plot/test**
 - point of inflexion in latent root plot

Rotation (similar to FA)

- the reference axes of the factors are turned about the origin until some other position has been reached
- the ultimate effect of rotating the factor matrix is to redistribute the variance from earlier factors to later ones to achieve a simpler, theoretically more meaningful factor pattern

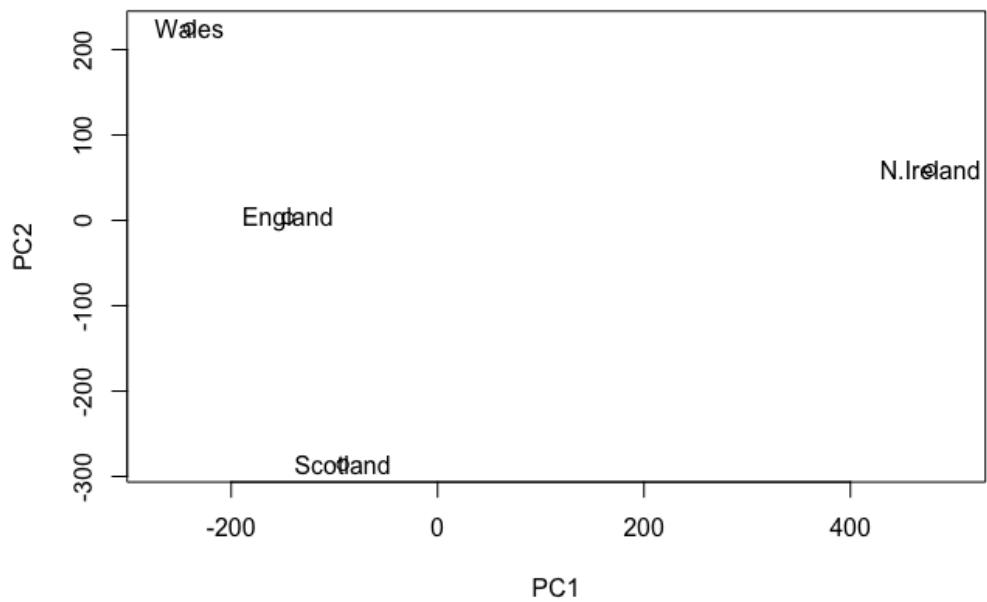
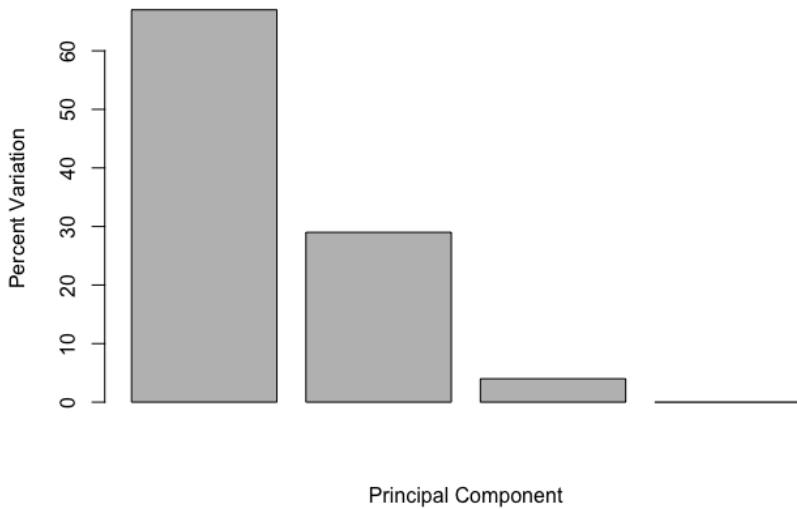
EXAMPLE

Select icon		England	Wales	Scotland	N.Ireland
	Cheese	105	103	103	66
	Carcass_meat	245	227	242	267
	Other_meat	685	803	750	586
	Fish	147	160	122	93
	Fats_and_oils	193	235	184	209
	Sugars	156	175	147	139
	Fresh_potatoes	720	874	566	1033
	Fresh_Veg	253	265	171	143
	Other_Veg	488	570	418	355
	Processed_potatoes	198	203	220	187
	Processed_Veg	360	365	337	334
	Fresh_fruit	1102	1137	957	674
	Cereals	1472	1582	1462	1494
	Beverages	57	73	53	47
	Soft_drinks	1374	1256	1572	1506
	Alcoholic_drinks	375	475	458	135
	Confectionery	54	64	62	41

data set of foods commonly consumed (in grams per person, per week) in different parts of UK

EXAMPLE

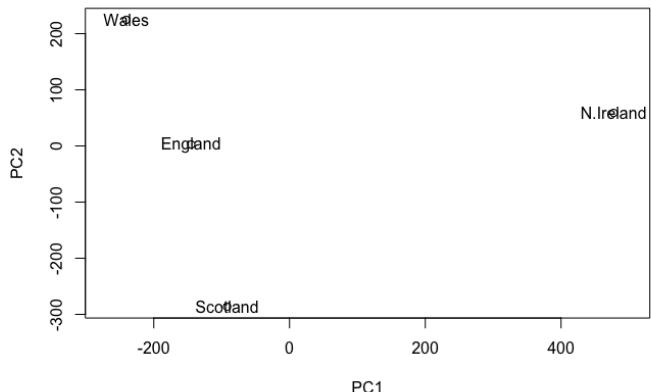
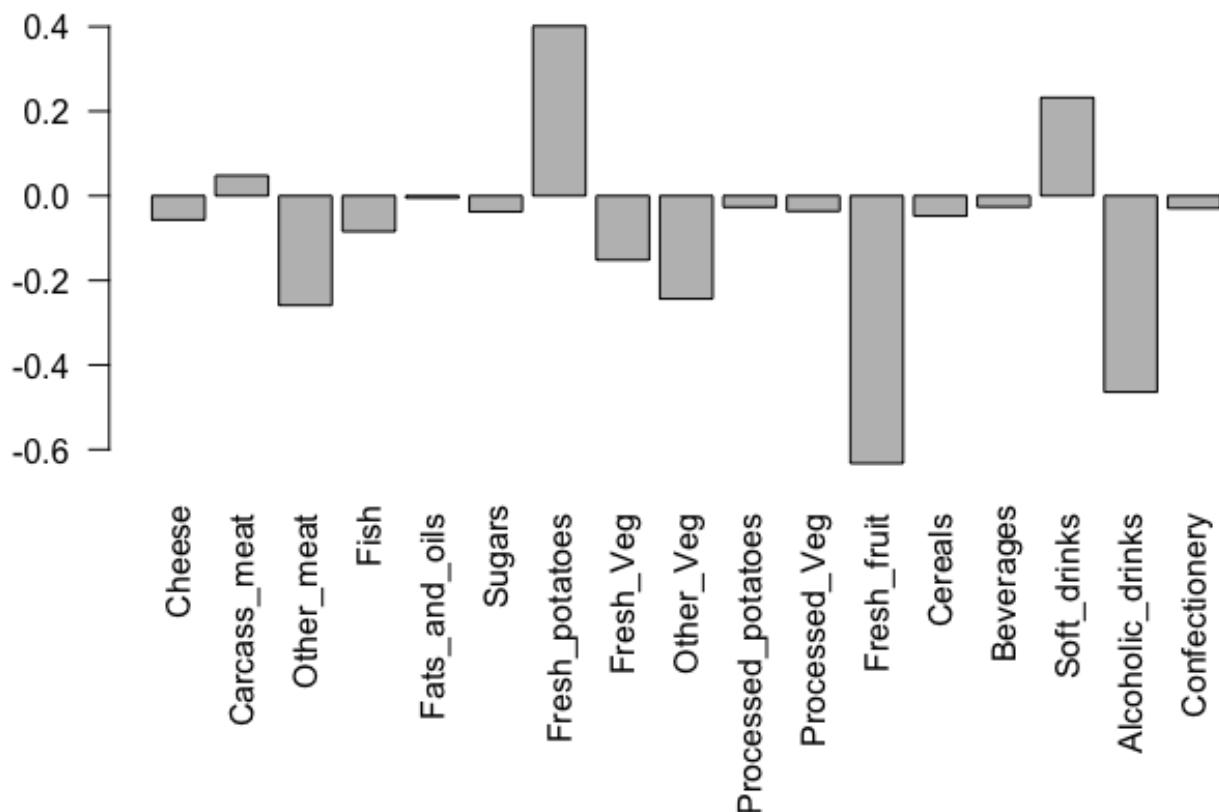
PCA



data set of foods commonly consumed (in grams per person, per week) in different parts of UK

EXAMPLE

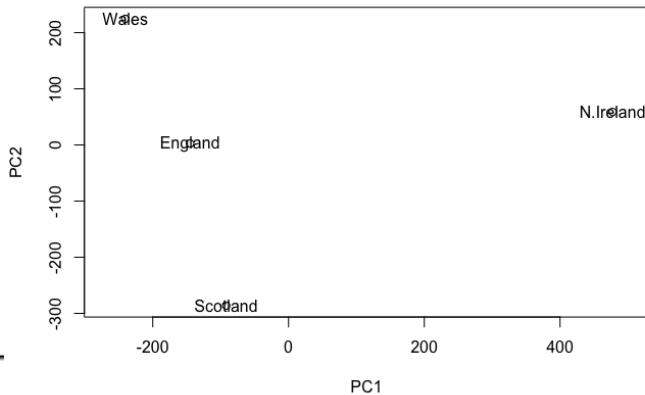
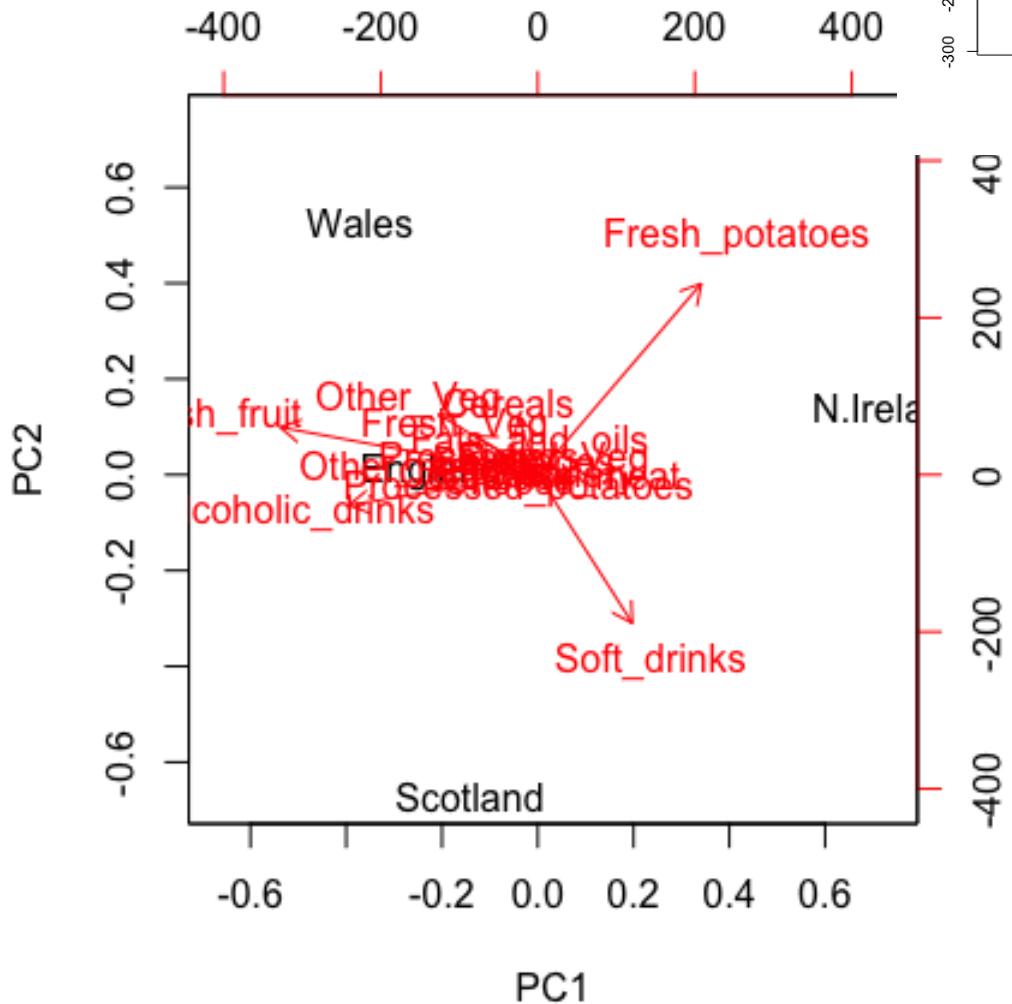
PCA



data set of foods commonly consumed (in grams per person, per week) in different parts of UK

EXAMPLE

Biplot



Factor analysis

Number of factors pre-determined
Many potential solutions
Factor matrix is estimated
Factor scores are estimated
More appropriate when searching for an underlying structure
Factors are not necessarily sorted

Only common variability is taken into account
Estimated factor scores may be correlated

A distinction is made between common and specific variance
Preferred when there is substantial measurement error in variables

Rotation is often desirable as there are many equivalent solutions

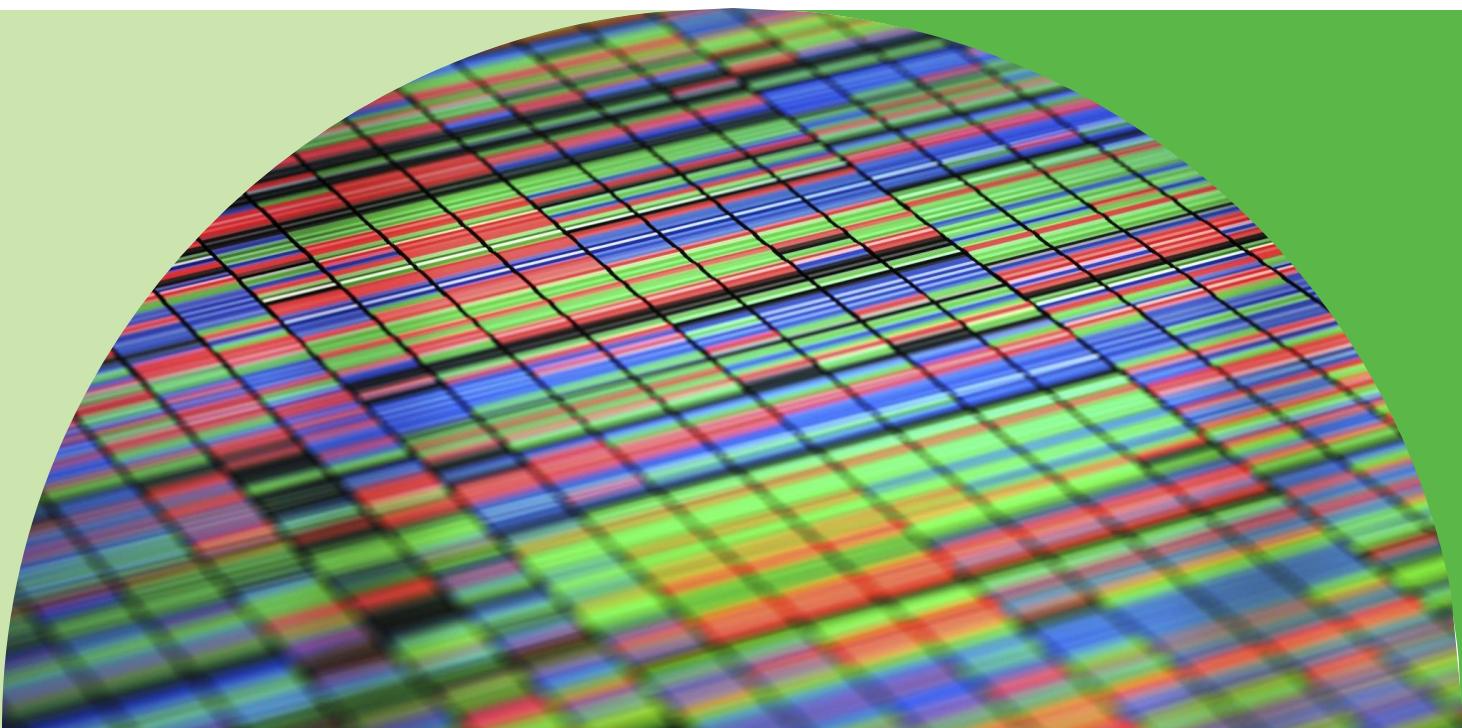
Principal component analysis

Number of components evaluated ex post
Unique mathematical solution
Component matrix is computed
Component scores are computed
More appropriate for data reduction (no prior underlying structure assumed)
Factors are sorted according to the amount of explained variability
Total variability is taken into account

Component scores are always uncorrelated
No distinction between specific and common variability
Preferred as a preliminary method to cluster analysis or to avoid multicollinearity in regression
Rotation is less desirable, unless components are difficult to be interpreted and explained variance is spread evenly across components

Bayesian Statistics

BRSM



Intuitive probability statements

- I'm carrying an umbrella, do you think it will rain?
- Data (d) = I'm carrying an umbrella.
- Hypotheses (h)
- 1) It rains today 2) It does not rain today

What are your **prior** beliefs about rain (hypotheses)?

- Say from historical records, we know that the chance of rain in April in Hyderabad is low (15%).

Hypothesis	Degree of Belief
Rainy day	0.15
Dry day	0.85

Likelihood: Theories about the data (me carrying an umbrella)

- Assumption: I'm not an idiot who randomly carries umbrellas but I can be very forgetful.. So I forget to carry an umbrella on about 70% of rainy days, and might carry an umbrella on 5% of dry days:

$$\text{Likelihood} = P(d|h)$$

Hypothesis	Data	
	Umbrella	No umbrella
Rainy day	0.30	0.70
Dry day	0.05	0.95

A reminder about basic probability rules

English	Notation	Formula
not A	$P(\neg A)$	$= 1 - P(A)$
A or B	$P(A \cup B)$	$= P(A) + P(B) - P(A \cap B)$
A and B	$P(A \cap B)$	$= P(A B)P(B)$

Joint probability of the hypothesis h and the data d
Probability that it is a rainy day AND I'm carrying an umbrella

$$P(d, h) = P(d|h)P(h)$$

$$\begin{aligned} P(\text{rainy, umbrella}) &= P(\text{umbrella|rainy}) \times P(\text{rainy}) \\ &= 0.30 \times 0.15 \\ &= 0.045 \end{aligned}$$

Repeat the exercise for all possibilities

	Umbrella	No-umbrella
Rainy	0.045	0.105
Dry	0.0425	0.8075

Repeat the exercise for all possibilities

	Umbrella	No-umbrella	Total
Rainy	0.0450	0.1050	0.15
Dry	0.0425	0.8075	0.85
Total	0.0875	0.9125	1

	Umbrella	No-umbrella	
Rainy	$P(\text{Umbrella, Rainy})$	$P(\text{No-umbrella, Rainy})$	$P(\text{Rainy})$
Dry	$P(\text{Umbrella, Dry})$	$P(\text{No-umbrella, Dry})$	$P(\text{Dry})$
$P(\text{Umbrella})$		$P(\text{No-umbrella})$	

Repeat the exercise for all possibilities

	Umbrella	No-umbrella	Total
Rainy	0.0450	0.1050	0.15
Dry	0.0425	0.8075	0.85
Total	0.0875	0.9125	1

	Umbrella	No-umbrella	
Rainy	$P(\text{Umbrella, Rainy})$	$P(\text{No-umbrella, Rainy})$	$P(\text{Rainy})$
Dry	$P(\text{Umbrella, Dry})$	$P(\text{No-umbrella, Dry})$	$P(\text{Dry})$
	$P(\text{Umbrella})$	$P(\text{No-umbrella})$	

	d_1	d_2	
h_1	$P(h_1, d_1)$	$P(h_1, d_2)$	$P(h_1)$
h_2	$P(h_2, d_1)$	$P(h_2, d_2)$	$P(h_2)$
	$P(d_1)$	$P(d_2)$	

So far: we have calculated our **prior beliefs before any data was given, in terms of many joint probabilities!**

- We know how confident we are about each of the different possibilities before we observed any data..

Now we are given data about the umbrella

	Umbrella	No-umbrella
Rainy		0
Dry		0
Total	1	0

Posterior Probability

Prior Joint Probabilities before observing the data

	Umbrella	No-umbrella	Total
Rainy	0.0450	0.1050	0.15
Dry	0.0425	0.8075	0.85
Total	0.0875	0.9125	1

$$P(\text{rain and umbrella}) / P(\text{umbrella}) \\ = 0.045/0.0875 = 0.514$$

Posterior Probability after observing the data (that I'm carrying an umbrella)

	Umbrella	No-umbrella
Rainy	0.514	0
Dry	0.486	0
Total	1	0

Posterior Joint probability

$$P(h|d) = P(d,h)/P(d)$$

Marginal probability

Arriving at Bayes' Rule

$P(d,h) = P(d|h) \times P(h)$ from our earlier basic probability rules

$P(h|d) = P(d,h)/P(d) = P(d|h) \times P(h)/P(d)$

So, to update your beliefs given data, you have to go from a prior probability via a likelihood function to a posterior probability

Posterior \sim Likelihood \times Prior

$$P(h|d) = P(d|h) \times P(h)/P(d)$$

**Now that you know Bayes' rule, you can do
Bayesian Hypothesis Tests!**

$$P(h_0|d) = \frac{P(d|h_0)P(h_0)}{P(d)}$$

$$P(h_1|d) = \frac{P(d|h_1)P(h_1)}{P(d)}$$

The Bayes Factor

- Posterior Odds = ratio of posterior probabilities for the hypotheses

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{0.75}{0.25} = 3$$

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{P(d|h_1)}{P(d|h_0)} \times \frac{P(h_1)}{P(h_0)}$$

Bayes Factor

The Bayes Factor or BF is used like the p value, to quantify the strength of evidence provided by the data.

Why not directly report the posterior odds?

- Because the prior beliefs may vary from researcher to researcher!!
- So the polite thing to do is to report BF and anyone can use the priors to calculate the posterior odds from the BF.
- Convention: to assume prior odds = 1, i.e., the null and the alternative hypotheses are equally likely.

Interpreting Bayes Factors

Bayes factor	Interpretation
1 - 3	Negligible evidence
3 - 20	Positive evidence
20 - 150	Strong evidence
>150	Very strong evidence

Why Bayesian Stats?

- Define p value?
- Define confidence interval?
- Wouldn't it be nice if you could say that the parameter lies in this range with 95% probability?
- All you need to do here is to be honest about what prior beliefs you brought to the table before you ran your study and then how your beliefs were updated..
- The dispute seems to be in the flexibility of prior beliefs, but there may be ways to justify these prior beliefs, specify them before you collect data, etc so that people are convinced you did you not use that flexibility to confirm your own biases.

P-values

<https://www.nature.com/articles/s41562-018-0311-x> - our paper arguing why we need to do more work to justify the alpha level we choose for our domain

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	HIGHLY SIGNIFICANT
0.02	HIGHLY SIGNIFICANT
0.03	HIGHLY SIGNIFICANT
0.04	SIGNIFICANT
0.049	SIGNIFICANT
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	ON THE EDGE OF SIGNIFICANCE
0.07	HIGHLY SUGGESTIVE
0.08	HIGHLY SUGGESTIVE
0.09	SIGNIFICANT AT THE $p < 0.10$ LEVEL
0.099	SIGNIFICANT AT THE $p < 0.10$ LEVEL
≥ 0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

P-values

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	
0.06	ON THE EDGE OF SIGNIFICANCE
0.07	
0.08	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.09	
0.099	HEY, LOOK AT THIS INTERESTING
≥0.1	SUBGROUP ANALYSIS

What happens if you peek at the data and say: "oh crap, p = 0.06, better collect some more data"??

p = 0.07, what do you do?

- You conclude that there is no effect, and try to publish it as a null result
- You guess that there might be an effect, and try to publish it as a "borderline significant" result
- You give up and try a new study
- You collect some more data to see if the p value goes up or (preferably!) drops below the magic criterion of p of 0.05

The fourth option?

- Messes up all p-values! They will all be incorrect if you want to interpret them as p-values (controlling for type-1 errors..). Why?

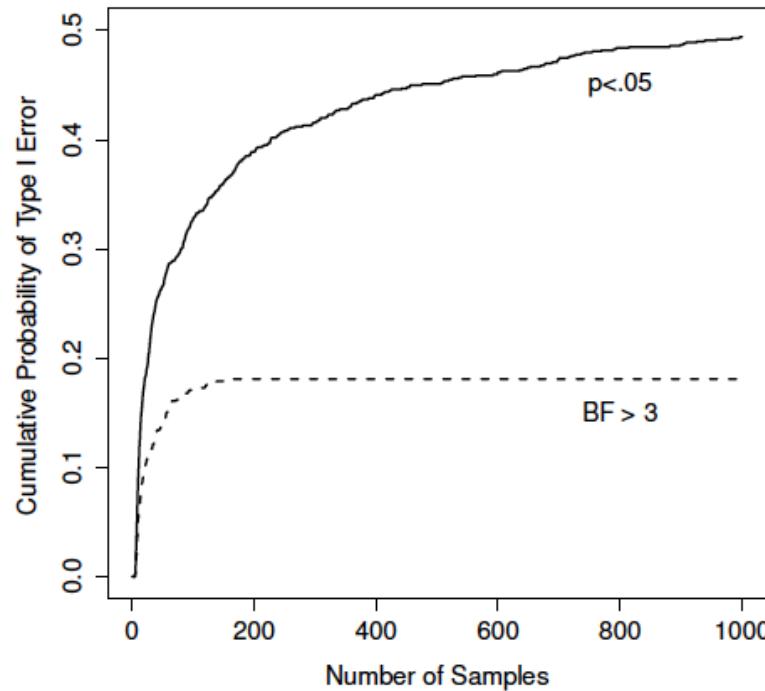
Outcome	Action
p less than .05	Reject the null
p greater than .05	Retain the null

Outcome	Action
p less than .05	Stop the experiment and reject the null
p between .05 and .1	Continue the experiment
p greater than .1	Stop the experiment and retain the null

Let's say you have a tight budget. Max 1000 subjects

- You peek at the data every time you collect an additional subject's worth of data
- Stop if $p < 0.05$
- Let's say you keep doing it and never find an effect and you hit 1000 subjects.
- Assuming the null hypothesis is true, you should be able to get to this conclusion 95% of the times you run such an experiment. This is why you defined $\alpha = 0.05$ at the outset.
- If we simulate this "optional stopping behavior", what is the true type 1 error?

Let's say you have a tight budget. Max 1000 subjects



Ok, but realistically, we don't peek 1000 times. How bad can it get?

- Assume your target is $N = 80$
- You collect $N = 50$. Your will power is exhausted. You decide to peek
- $p < 0.05$
- You decide to stop.
- The actual type 1 error instead of 5% now is 8%. So if you're honest, you have to say $p < 0.08$ in your reporting.
- So that is what happens with just ONE peek.
- So the rules are fairly strict with frequentist null hypothesis testing. No peeking.

Bayesian independent samples t-test

```
> load( "harpo.Rdata" )
> head(harpo)
  grade      tutor
1   65 Anastasia
2   72 Bernadette
3   66 Bernadette
4   74 Anastasia
5   73 Anastasia
6   71 Bernadette
```

Bayesian independent samples t-test

```
> load( "harpo.Rdata" )
> head(harpo)
  grade      tutor
 1    65 Anastasia
 2    72 Bernadette
 3    66 Bernadette
 4    74 Anastasia
 5    73 Anastasia
 6    71 Bernadette
```

Test results:

t-statistic: 2.115
degrees of freedom: 31
p-value: 0.043

Bayesian independent samples t-test

```
> load( "harpo.Rdata" )          > ttestBF( formula = grade ~ tutor, data = harpo )
> head(harpo)
  grade   tutor
1  65 Anastasia
2  72 Bernadette
3  66 Bernadette
4  74 Anastasia
5  73 Anastasia
6  71 Bernadette

                                         Bayes factor analysis
-----
[1] Alt., r=0.707 : 1.754927 ±0%
Against denominator:
  Null, mu1-mu2 = 0
---
Bayes factor type: BFindepSample, JZS
```

Bayesian paired t-test

```
> load(chico)
> head(chico)
  id grade_test1 grade_test2
1 student1     42.9      44.6
2 student2     51.8      54.0
3 student3     71.7      72.3
4 student4     51.6      53.4
5 student5     63.5      63.8
6 student6     58.0      59.3
```

```
> ttestBF(
+   x = chico$grade_test1,
+   y = chico$grade_test2,
+   paired = TRUE
+ )
```

and here's the output:

```
Bayes factor analysis
-----
[1] Alt., r=0.707 : 5992.05 ±0%
Against denominator:
  Null, mu = 0
---
Bayes factor type: BFoneSample, JZS
```

Bayesian Regression

```
> load("parenthood.Rdata")
> head(parenthood)
  dan.sleep baby.sleep dan.grump day
  1      7.59       10.18      56   1
  2      7.91       11.66      60   2
  3      5.14       7.92       82   3
  4      7.71       9.61       55   4
  5      6.68       9.75       67   5
  6      5.99       5.04       72   6
```

Bayesian Regression

```
> model <- lm(  
+   formula = dan.grump ~ dan.sleep + day + baby.sleep,  
+   data = parenthood  
+ )  
  
> summary(model)  
  
BLAH BLAH BLAH  
  
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 126.278707  3.242492 38.945 <2e-16 ***  
dan.sleep    -8.969319  0.560007 -16.016 <2e-16 ***  
day          -0.004403  0.015262 -0.288  0.774  
baby.sleep    0.015747  0.272955  0.058  0.954
```

BLAH BLAH BLAH

Bayesian Regression

```
> regressionBF(  
+   formula = dan.grump ~ dan.sleep + day + baby.sleep,  
+   data = parenthood  
+ )
```

```
Bayes factor analysis  
-----  
[1] dan.sleep : 1.622545e+34 ±0%  
[2] day : 0.2724027 ±0%  
[3] baby.sleep : 10018411 ±0%  
[4] dan.sleep + day : 1.016578e+33 ±0.01%  
[5] dan.sleep + baby.sleep : 9.770233e+32 ±0.01%  
[6] day + baby.sleep : 2340755 ±0%  
[7] dan.sleep + day + baby.sleep : 7.835625e+31 ±0%
```

Against denominator:

 Intercept only

Bayes factor type: BFlinearModel, JZS

Bayesian Regression

Not super helpful comparing everything to the intercept-only model which you typically don't care about much

Highest BF is easy enough to see here but in more complex situations can be harder

So you can use the head function to pick a few best models

However, this is still comparing models to the intercept-only model

```
> models <- regressionBF(  
+   formula = dan.grump ~ dan.sleep + day + baby.sleep,  
+   data = parenthood  
+ )  
  
> head( models, n = 3)  
  
Bayes factor analysis  
-----  
[1] dan.sleep           : 1.622545e+34 ±0%  
[2] dan.sleep + day     : 1.016578e+33 ±0.01%  
[3] dan.sleep + baby.sleep : 9.770233e+32 ±0.01%  
  
Against denominator:  
  Intercept only  
---  
Bayes factor type: BFLinearModel, JZS
```

Bayesian Regression

```
> head( models/max(models), n = 3)
```

```
Bayes factor analysis
```

```
-----  
[1] dan.sleep : 1      ±0%  
[2] dan.sleep + day : 0.06265328 ±0.01%  
[3] dan.sleep + baby.sleep : 0.06021549 ±0.01%
```

```
Against denominator:
```

```
  dan.grump ~ dan.sleep
```

```
---
```

```
Bayes factor type: BFlinearModel, JZS
```

```
> models[1] / models[4]
```

```
Bayes factor analysis
```

```
-----  
[1] dan.sleep : 15.96086 ±0.01%
```

```
Against denominator:
```

```
  dan.grump ~ dan.sleep + day
```

```
---
```

```
Bayes factor type: BFlinearModel, JZS
```

Bayesian Regression: individual coefficients

```
> regressionBF(  
+   formula = dan.grump ~ dan.sleep + baby.sleep,  
+   data = parenthood,  
+   whichModels = "top"  
+ )
```

```
Bayes factor top-down analysis  
-----  
When effect is omitted from dan.sleep + baby.sleep , BF is...  
[1] Omit baby.sleep : 16.60702 ±0.01%  
  
[2] Omit dan.sleep : 1.025401e-26 ±0.01%  
  
Against denominator:  
  dan.grump ~ dan.sleep + baby.sleep  
---  
Bayes factor type: BFlinearModel, JZS
```

Bayesian ANOVA

- Very similar to regression

```
> load("clinicaltrial.Rdata")      > models <- anovaBF(  
> head(clin.trial)                +   formula = mood.gain ~ drug * therapy,  
    drug   therapy mood.gain        +   data = clin.trial  
1 placebo no.therapy      0.5     + )  
2 placebo no.therapy      0.3  
3 placebo no.therapy      0.1  
4 anxifree no.therapy     0.6  
5 anxifree no.therapy     0.4  
6 anxifree no.therapy     0.2
```

```
> models/max(models)  
  
Bayes factor analysis  
-----  
[1] drug                      : 0.3521273 ±0.96%  
[2] therapy                    : 0.001047637 ±0.96%  
[3] drug + therapy             : 1 ±0%  
[4] drug + therapy + drug:therapy : 0.9856421 ±1.62%  
  
Against denominator:  
  mood.gain ~ drug + therapy  
---  
Bayes factor type: BFlinearModel, JZS
```

Summary

- BayesFactor package
- ttestBF()
- regressionBF()
- anovaBF()
- More interpretable Bayes Factors which can be converted to posterior odds ratios via the prior odds ratios.
- Combining p values and Bayes Factors can help you evaluate evidence better.

Project final presentations

- Use the weekend to do some serious analysis of your data for your projects.
- Next week - in-class help with projects
- Endsem - all inclusive, more emphasis on descriptive Qs, slightly more open-ended, will require you to write down your assumptions and justifications.