ECE1724H S2: Empirical Software Engineering

Introduction of Quantitative Study

The Edward S. Rogers Sr. Department
of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

Course Project

- **Interim Report: Annotated outline (due 3/11)**
- **Final Report: (due 4/11)**


Just a reminder...


SURVEY

When to use Survey Research?

- To evaluate the frequency of some characteristic across a population
  - E.g. how many companies use agile methods?
- To evaluate the severity of some condition that occurs in a population
  - E.g. what's the average cost overrun of software projects?
- To identify factors that influence a characteristic or condition
  - E.g. What factors cause companies to adopt new ASE tools?

What type of question are you asking?

Exploratory

➜Existence:
- Does X exist?

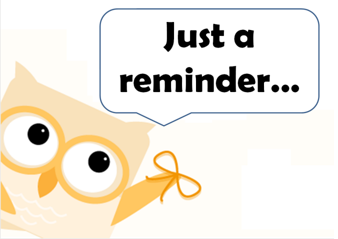
➜Description & Classification
- What is X like?
- What are its properties?
- How can it be categorized?
- How can we measure it?
- What are its components?

➜Descriptive-Comparative
- How does X differ from Y?

Base-rate

➜Frequency and Distribution
- How often does X occur?
- What is an average amount of X?

➜Descriptive-Process
- How does X normally work?
- By what process does X happen?
- What are the steps as X evolves?



Correlation

➜Relationship
- Are X and Y related?
- Do occurrences of X correlate with occurrences of Y?

Causal Relationship

➜Causality
- Does X cause Y?
- Does X prevent Y?
- What causes X?
- What effect does X have on Y?

➜Causality-Comparative
- Does X cause more Y than does Z?
- Is X better at preventing Y than is Z?
- Does X cause more Y than does Z under one condition but not others?

Design

➜Design
- What is an effective way to achieve X?
- How can we improve X?

Agenda for today

- Paper reading presentation for Surveys
- Descriptive statistics
- Hypothesis testing


45%
8%
17%
$
2%
28%

- Henne, B., Harbach, M., & Smith, M. (2013). Location privacy revisited: factors of privacy decisionsPreview the document. Extended Abstracts on Human Factors in Computing Systems.
- Shklovski, I., Mainwaring, S. D., Skúladóttir, H. H., & Borgthorsson, H. (2014). **Leakiness and creepiness in app space: perceptions of privacy and mobile app usePreview the documen**t, Proceedings of the ACM conference on Human factors in computing systems (pp. 2347-2356): ACM.

Research Designs

- Step 1 to **design** empirical research:

  adopt a general (and guiding) approach

- Three main approaches:
  - **Qualitative**
  - **Quantitative**
  - **Mixed Methods**

- Qual. vs Quant.:
  - Not dichotomies
  - Rather, different ends on a continuum

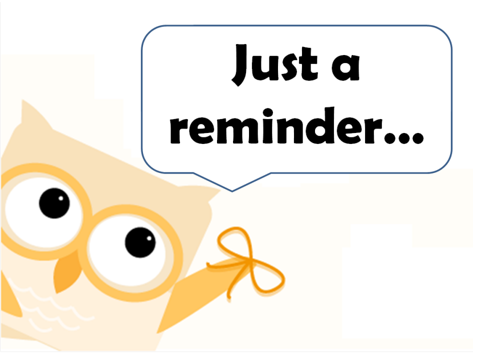
Chapter 1 - Creswell, J. W., & Creswell, J. D. (2017
Research design: Qualitative, quantitative, and mixed methods approaches. Sage publications.

Qualitative vs Quantitative

- Often:
  - Words (qual) vs Numbers (quant)
  - Open-ended questions (qual interview questions)

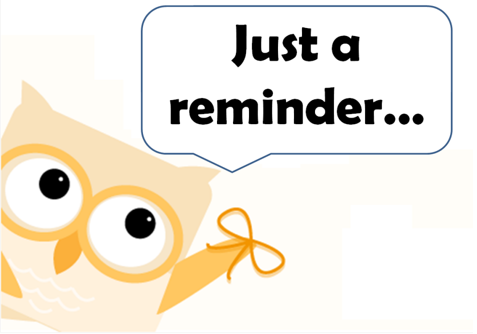
    vs

    Closed-ended questions (quant hypotheses)
  - Inductive (qual) vs Deductive (quant)
- More completely, differences in:
  - **philosophical** assumptions
  - strategies of **inquiry**
  - research **methods**



Planning Checklist

- ✓ Pick a topic
- ✓ Identify the research question(s)
- ✓ Check the literature
- ✓ Identify your philosophical stance
- ✓ Identify appropriate theories
- ✓ Choose the method(s)
- ❍ Design the study
  - ✓ Unit of analysis?
  - ✓ Target population?
  - ✓ Sampling technique?
  - ✓ Data collection techniques?
  - Metrics for key variables?
  - Handle confounding factors
- ❍ Critically appraise the design for threats to validity
- ❍ Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- ❍ Recruit subjects / field sites
- ❍ Conduct the study
- ❍ Analyze the data
- ❍ Write up the results and publish them
- ❍ Iterate

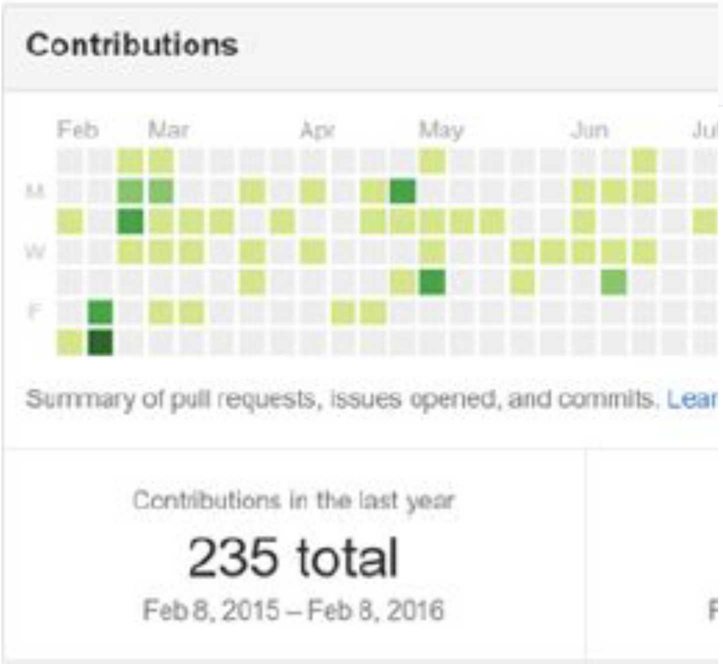


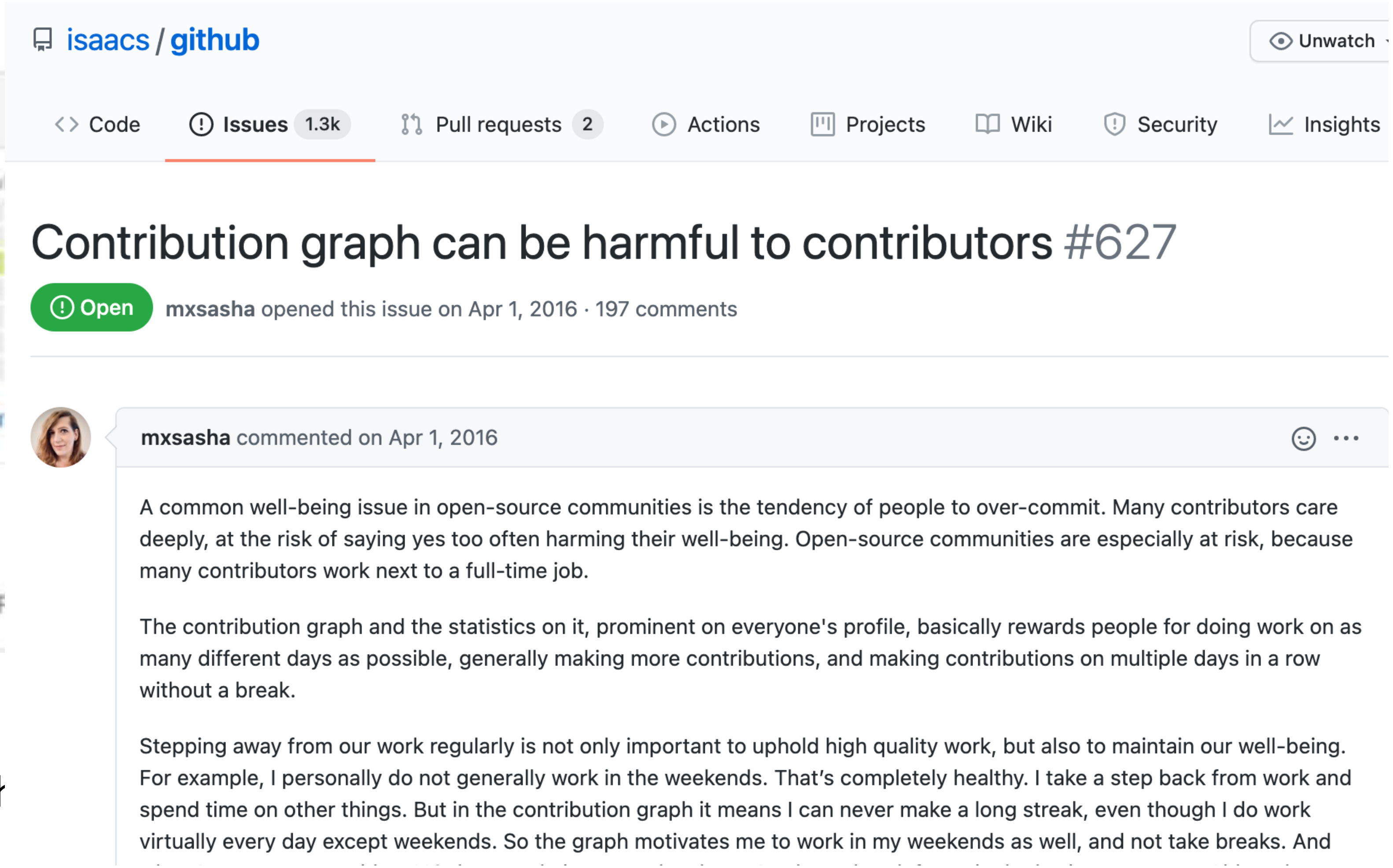
Agenda for today

- Paper reading presentation for Surveys
- Descriptive statistics
- Hypothesis testing

Measurement

"Measuring programming progress by lines of code is like measuring aircraft building progress by weight."





isaacs / github

Unwatch

Code | Issues 1.3k | Pull requests 2 | Actions | Projects | Wiki | Security | Insights

Contribution graph can be harmful to contributors #627

Open **mxsasha** opened this issue on Apr 1, 2016 · 197 comments

mxsasha commented on Apr 1, 2016

A common well-being issue in open-source communities is the tendency of people to over-commit. Many contributors care deeply, at the risk of saying yes too often harming their well-being. Open-source communities are especially at risk, because many contributors work next to a full-time job.

The contribution graph and the statistics on it, prominent on everyone's profile, basically rewards people for doing work on as many different days as possible, generally making more contributions, and making contributions on multiple days in a row without a break.

Stepping away from our work regularly is not only important to uphold high quality work, but also to maintain our well-being. For example, I personally do not generally work in the weekends. That's completely healthy. I take a step back from work and spend time on other things. But in the contribution graph it means I can never make a long streak, even though I do work virtually every day except weekends. So the graph motivates me to work in my weekends as well, and not take breaks. And

Contributing graphs considered h

https://www.hanselman.com/

Context of Measurement

- The meaning of measurements will vary depending on whether they derive from observation or experiment.
  - Observation: sampling difficulties → bias
  - Experiment: a lack of generalizability
- The available measurements are not immediately connected with the phenomena of interest
  - Leading/Lagging -- difficult to use in managing day-to-day operations
- Side-effects -- the numbers are "massaged" or the work process altered in order to produce the "right" results
- Good measurements are **actionable**

(Ch 6) Statistical Methods and Measurement . from F. Shull et al. (eds.), Guide to Advanced Empirical Software Engineering. Springer 2008

Creating Effective Metrics

- **Precise, Reliable, Valid**
- **Well-motivated** -- a metric must provide at least a partial answer to a specific question, a question which itself is aimed at some particular research or management goal.

- *"What is the expected amount of time for a specific class of defects to go from the initial reported state to the Repaired state?"*

- *"What percent of all customer reported defects are in the Repaired state within two days of being first reported?"*

- **a single metric is usually not sufficient**

(Ch 6) Statistical Methods and Measurement . from F. Shull et al. (eds.), Guide to Advanced Empirical Software Engineering. Springer 2008

Defining a Metric

- Simple
  - **counts** (e.g., number of units shipped this year)
  - **dimensional measures** (e.g., this year's support costs, in dollars)
  - **categories** (e.g., problem types)
  - **rankings** (e.g., problem severity)
- Compound
  - **ratios** (e.g., defects per thousand units)
  - **rates** (time-based ratios such as number of problem reports per month)
  - **proportions or percentages** (e.g., proportion of customers responding "very satisfied" to a survey question)
  - **linear algebraic combinations** (e.g., mean repair cost – the sum of all repair costs divided by the total number of repairs)
  - **indices** (dimensionless measures typically based on a sum and then standardized to some baseline value). Whereas

How will you measure things?

Type	Meaning	Example	Admissible Operations
Nominal Scale	Unordered categories	Gender, Political preferences, Place of residence	=
Ordinal Scale	Ranking of objects into ordered categories (intervals between the values are not necessarily of the same size)	Satisfaction, Happiness, grades	=, <, >
Interval Scale	Differences between points on the scale are meaningful (equal intervals)	Celsius, Fahrenheit Temperature, IQ (intelligence scale), SAT scores	=, <, >, difference, mean
Ratio Scale	Ratios between points on the scale are meaningful (ordered, equal intervals with a zero point)	weight, height, sales figures, ruler measurements, number of children	=, <, >, difference, mean, ratio

Levels of Measurement

Offers:	Nominal	Ordinal	Interval	Ratio
The sequence of variables is established	–	Yes	Yes	Yes
Mode	Yes	Yes	Yes	Yes
Median	–	Yes	Yes	Yes
Mean	–	–	Yes	Yes
Difference between variables can be evaluated	–	–	Yes	Yes
Addition and Subtraction of variables	–	–	Yes	Yes
Multiplication and Division of variables	–	–	–	Yes
Absolute zero	–	–	–	Yes

https://www.questionpro.com/blog/nominal-ordinal-interval-ratio/

Quantitative interpretation

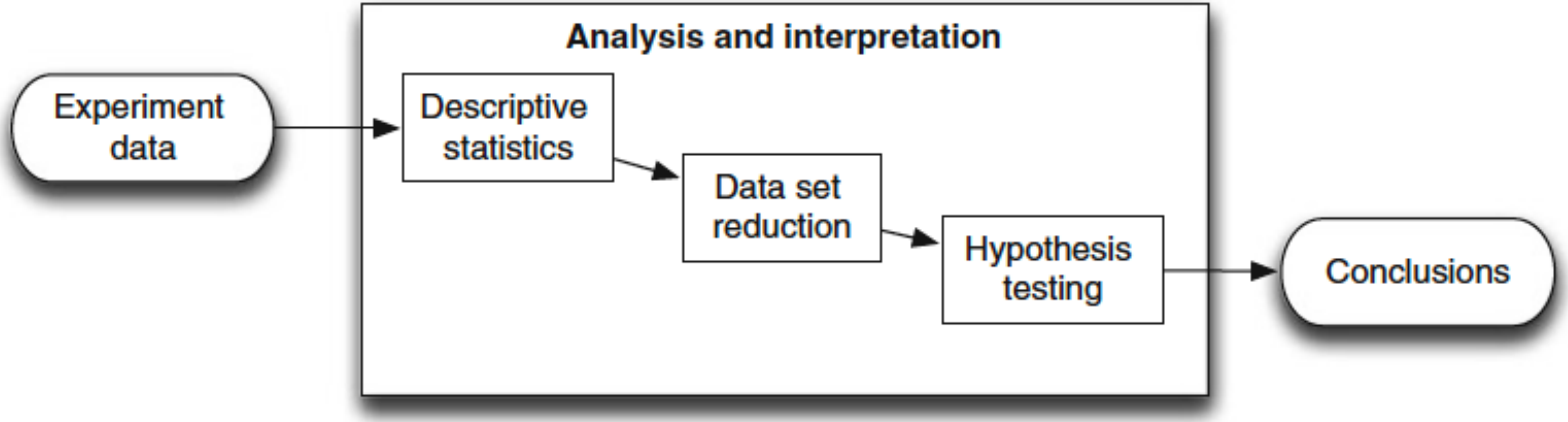
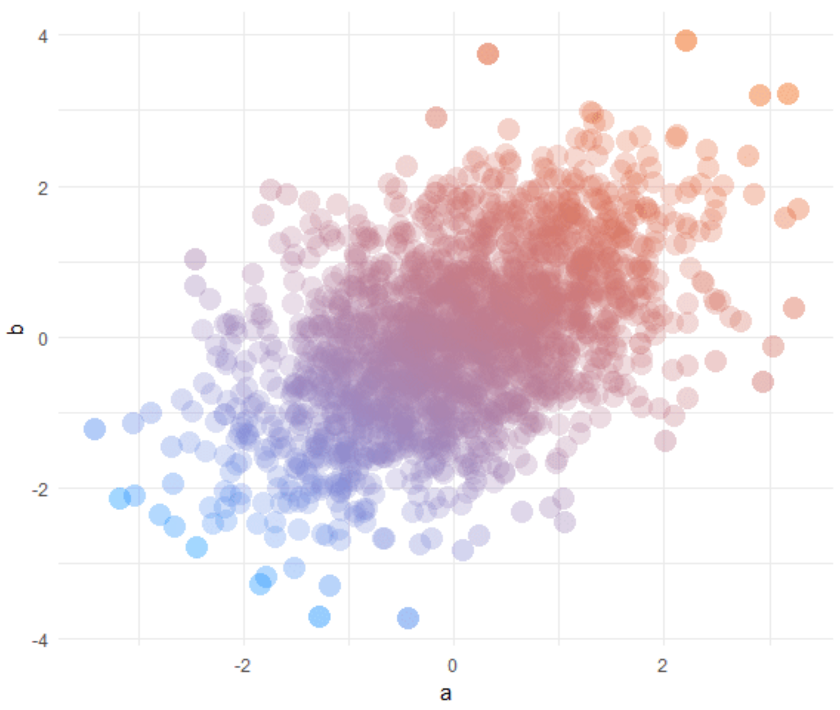


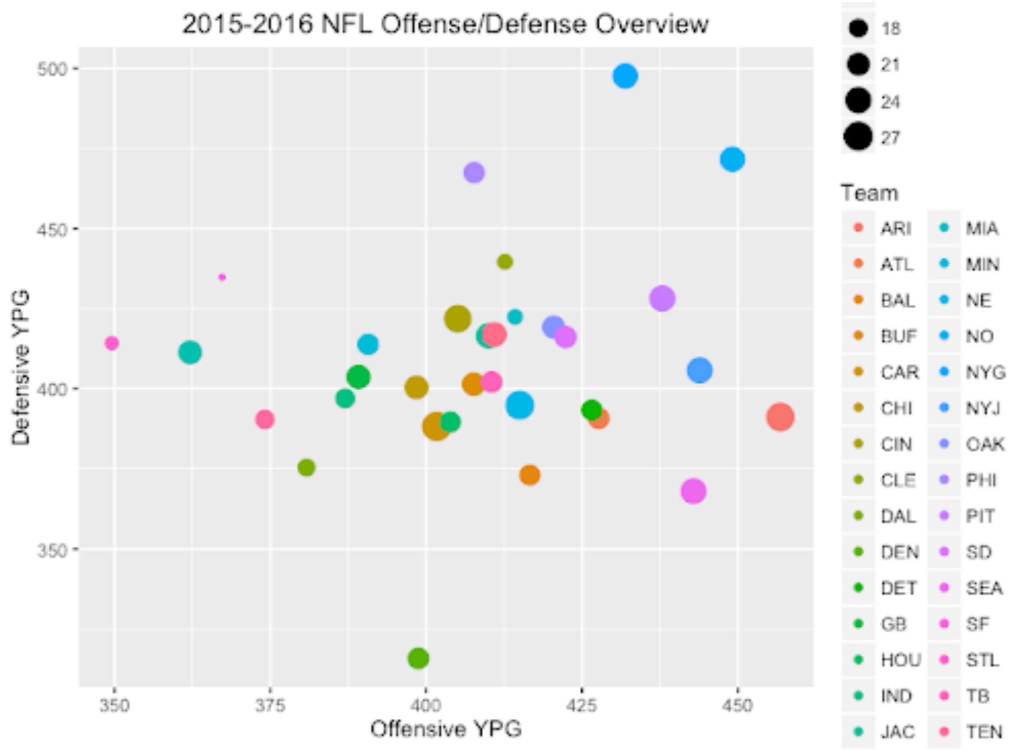
Fig. 10.1 Three steps in quantitative interpretation

C. Wohlin et al., Experimentation in Software Engineering, Springer-Verlag Berlin Heidelberg 2012

Descriptive Statistics

Scatter plot





Descriptive Statistics

- Measure of central tendency (mean, median, mode)
- Measure of dispersion (variance, standard deviation, coefficient of variant)
- Measure of dependency (regression, covariance, correlation coefficient,...)

Measures of Central Tendency

most *representative or typical* of all values in a group

"average"

MODE

- **most frequent data point**
- mode exists as a data point
- unaffected by extreme values
- useful for qualitative data
- may have more than 1 value

MEDIAN

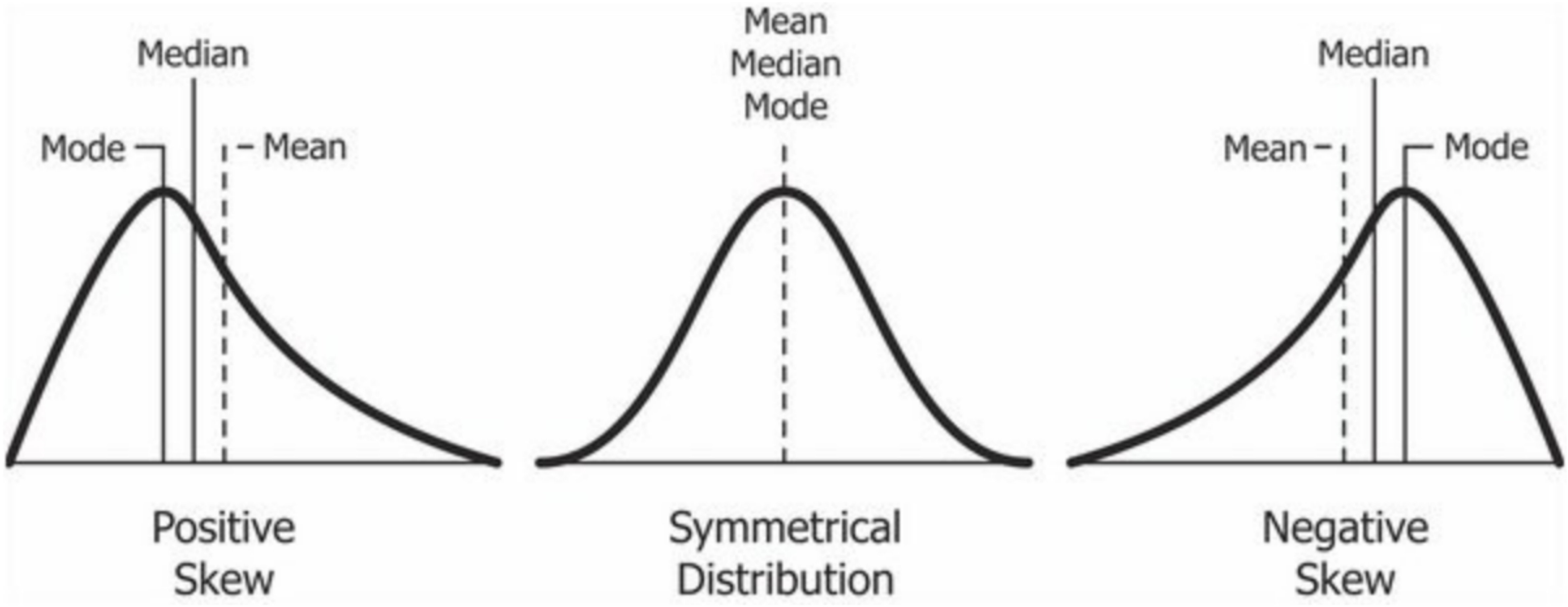
- **value that divides ranked data points into halves: 50% larger than it, 50% smaller**
- may not exist as a data point in the set
- influenced by position of items, but not their values

MEAN

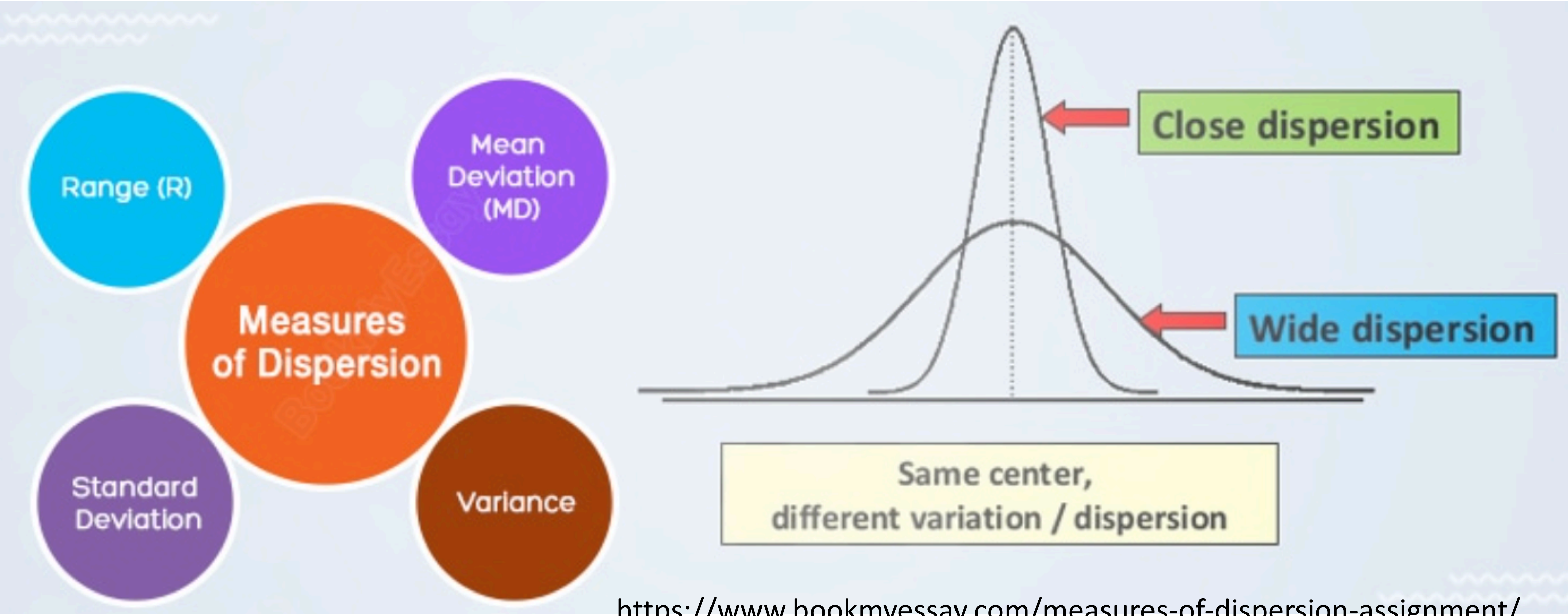
$$\bar{x} = \frac{\Sigma x}{N}$$

- most stable measure
- affected by extreme values
- may not exist as a data point in the set

https://www.slideshare.net/toledo98/measures-of-central-tendency-and-variability

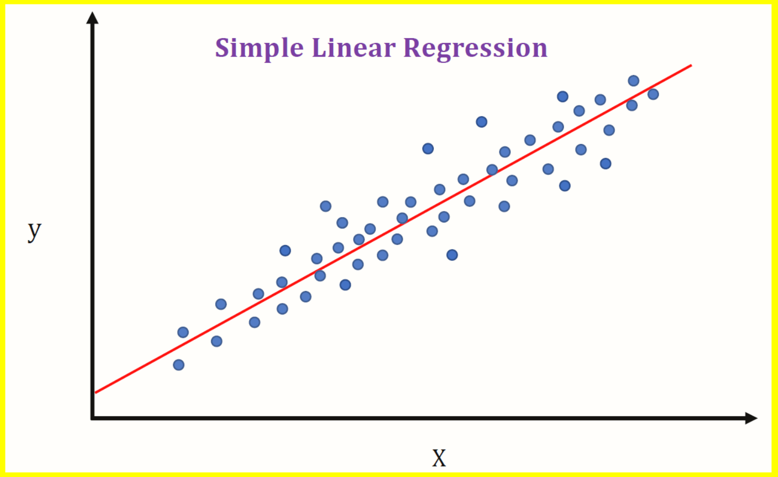

Median
Mode
Mean
Positive
Skew
Mean
Median
Mode
Symmetrical
Distribution
Median
Mean
Mode
Negative
Skew

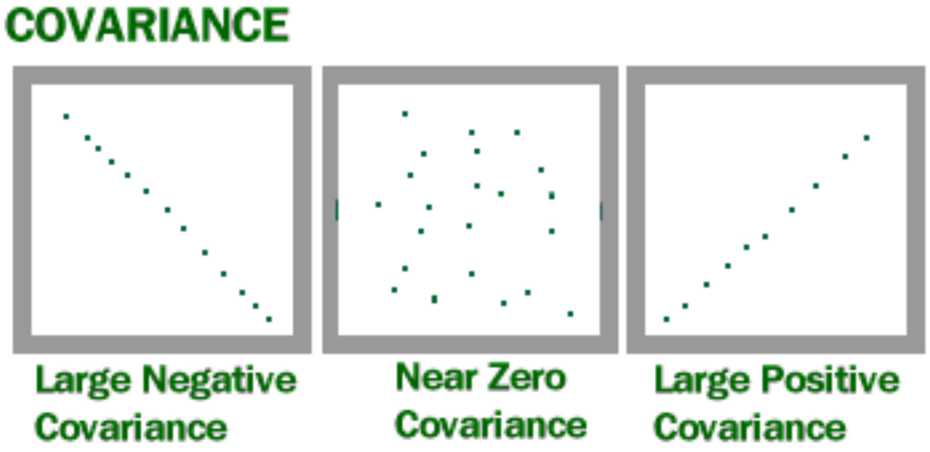
Measures of Dispersion

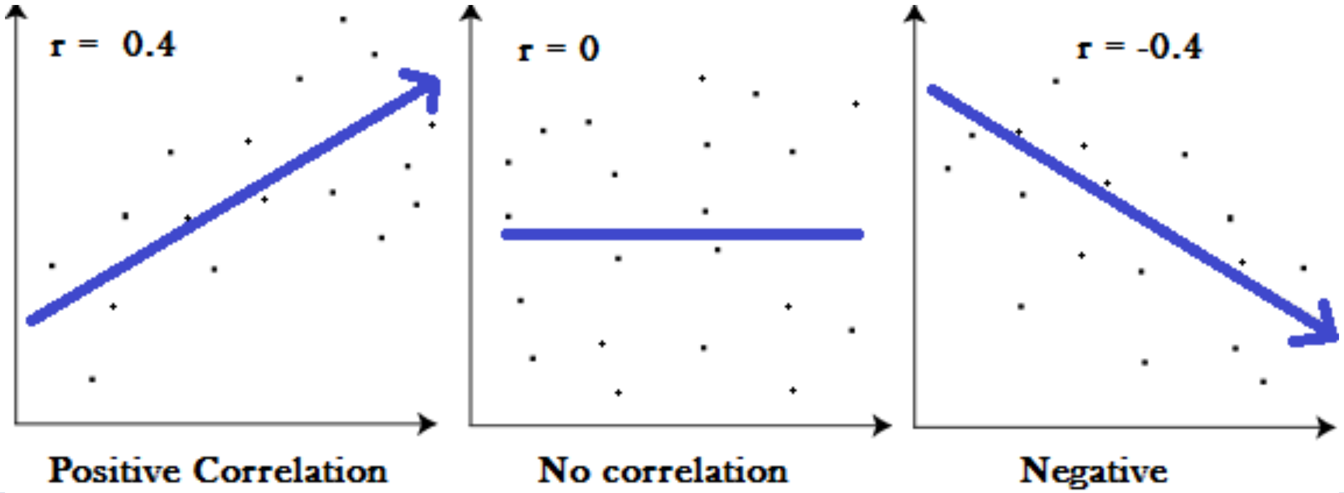


https://www.bookmyessay.com/measures-of-dispersion-assignment/

Measures of Dependency/association

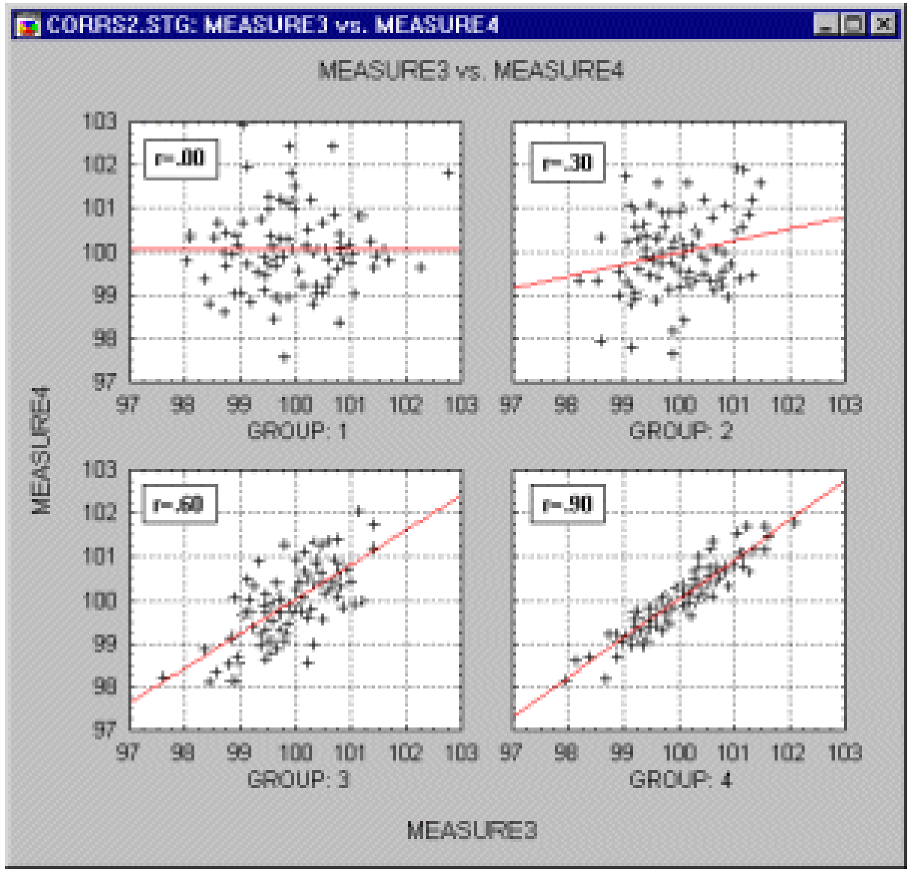




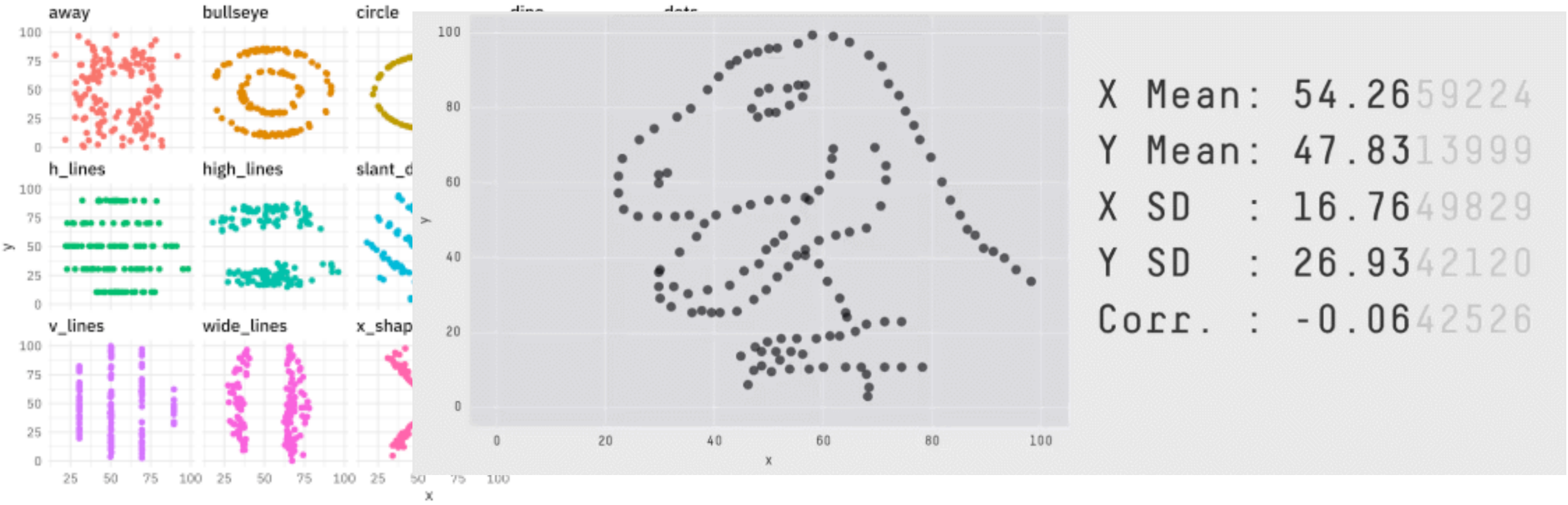


Correlation

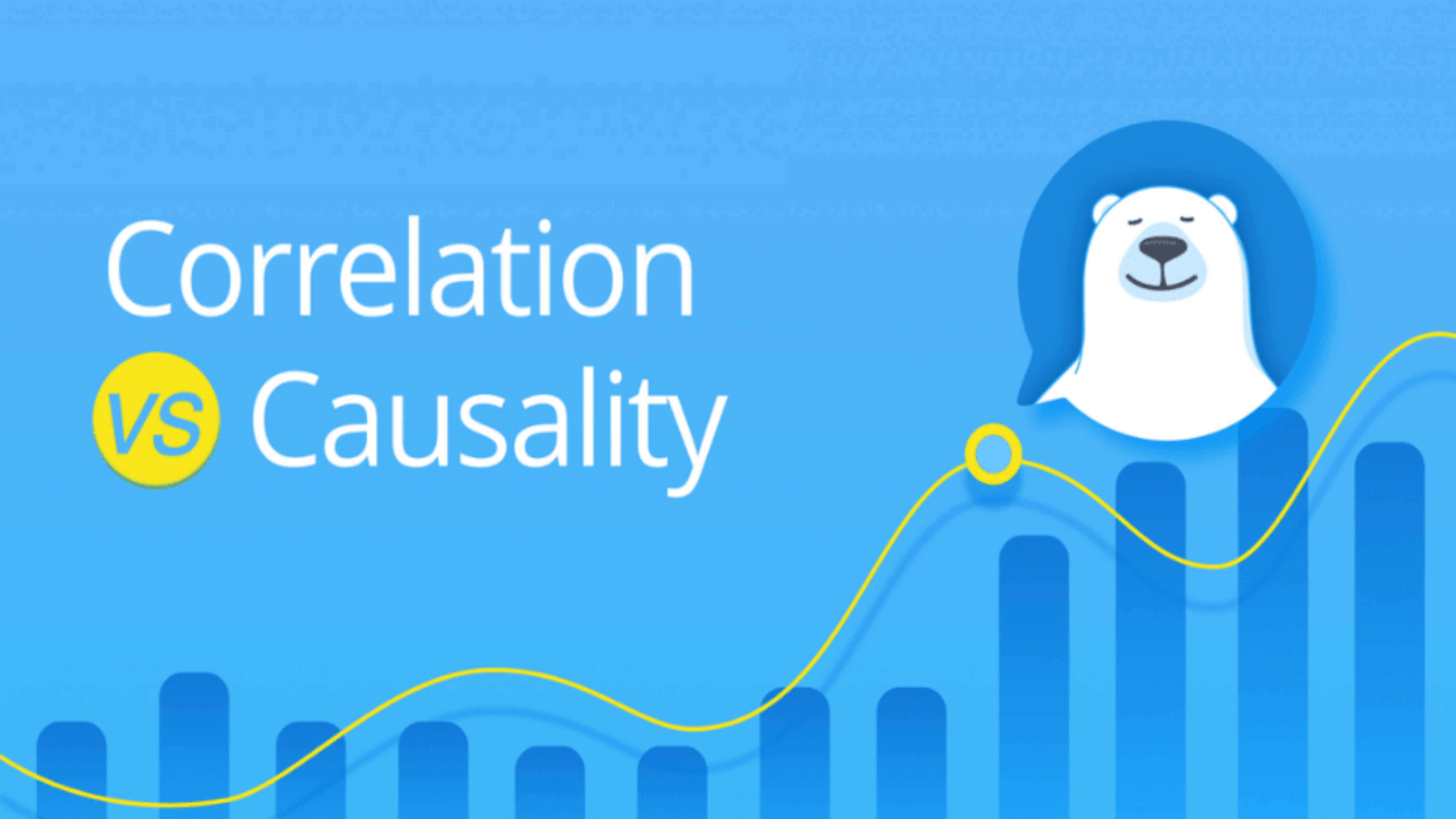
- Measure of the relation between two variables:
  - -1 variables perfect inverses (negative correlation)
  - 0 no correlation at all
  - +1 variables are perfectly correlated (they appear on a straight line with positive slope)
- Pearson's r
  - Computed as: $$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$
    - x and y are the sample means
    - s_x and s_y are the sample standard deviations
    - n is the sample size
  - Assumes variables are interval or ratio scale
  - Is independent of the measurement unit

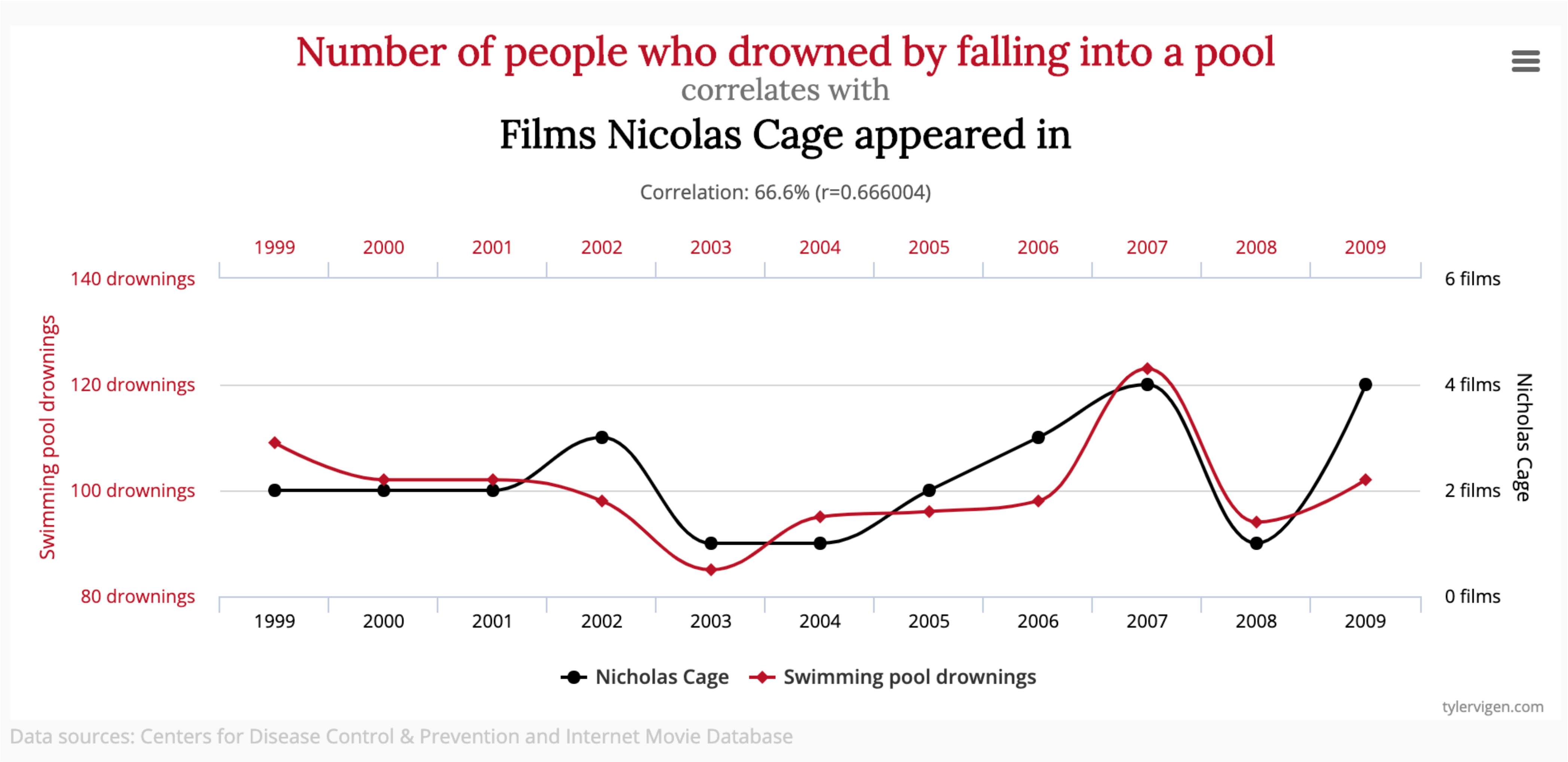


The Datasaurus Dozon

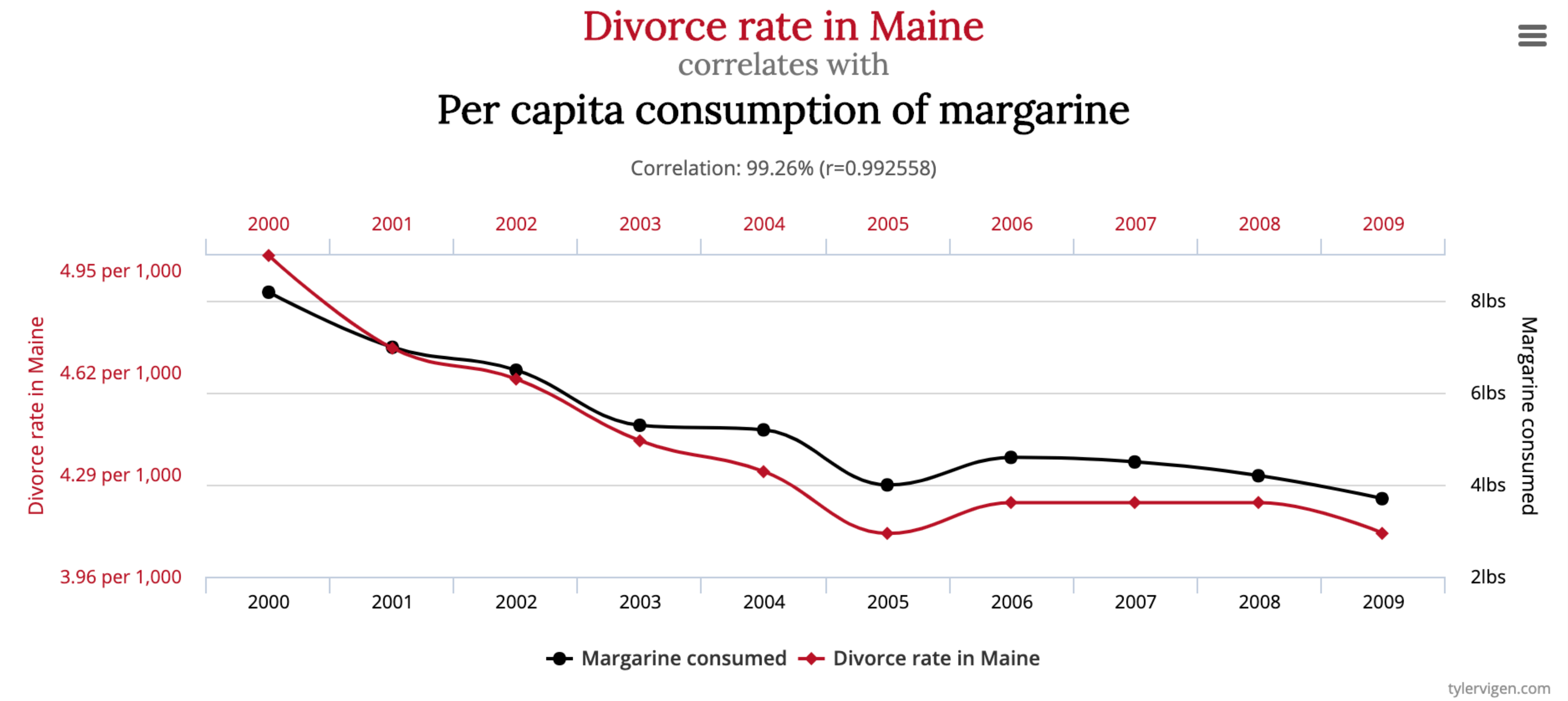


https://www.r-bloggers.com/2020/10/predicting-class-membership-for-the-tidytuesday-datasaurus-dozen/

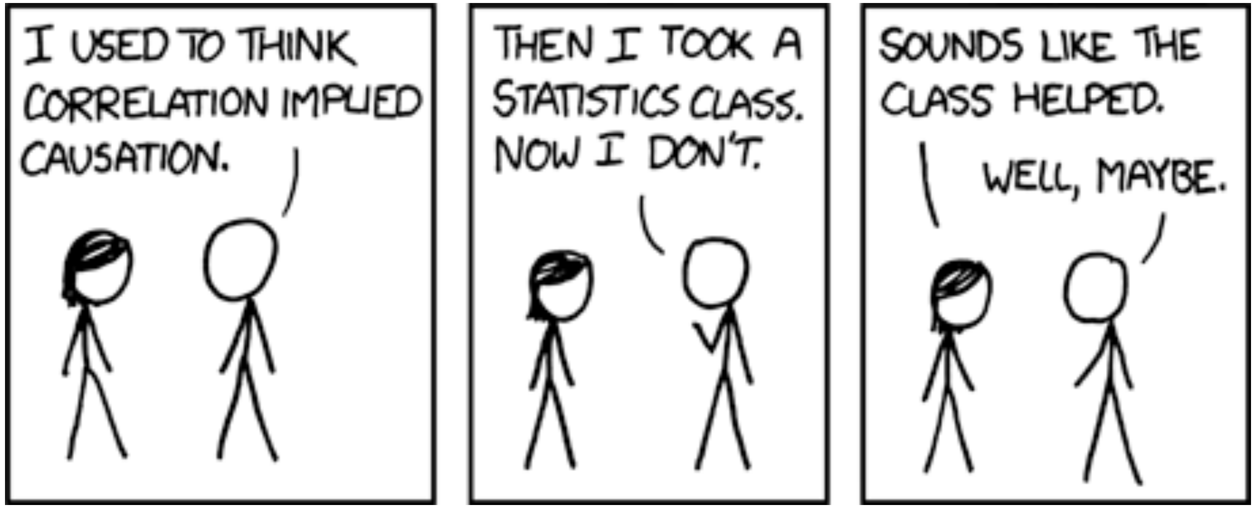

Correlation
vs Causality



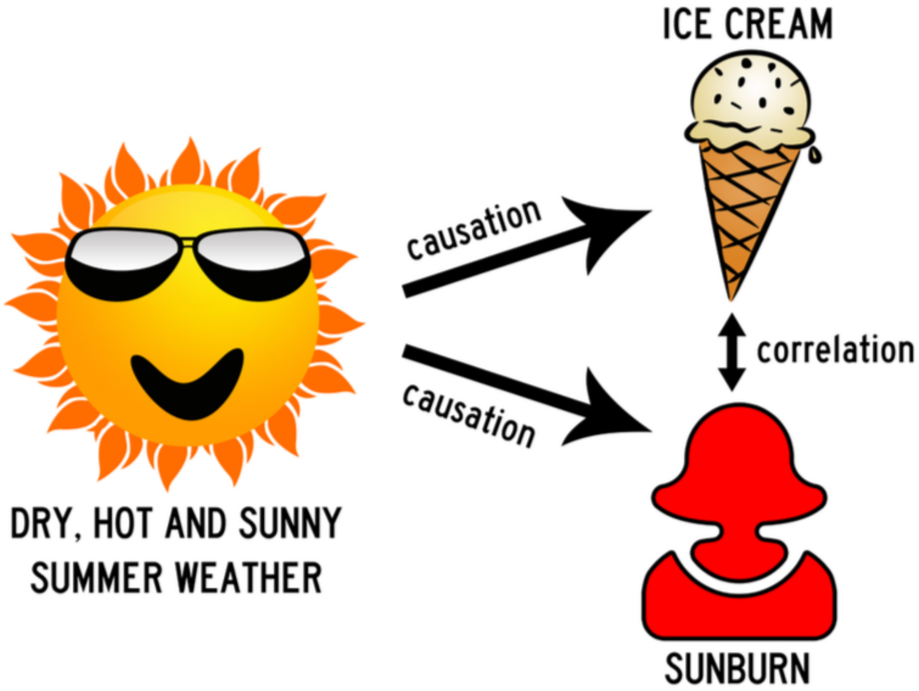
http://www.tylervigen.com/spurious-correlations



Data sources: National Vital Statistics Reports and U.S. Department of Agriculture



http://xkcd.com/552/



- For causation
  - Provide a theory (from domain knowledge, independent of data)
  - Show correlation
  - Demonstrate ability to predict new cases (replicate/validate)

The association between early career informal mentorship in academic collaborations and junior author performance

We also find that incr
only with a reduction
reduction
encourag
findings r
increase t
add a nev
women ir

“We find th
papers writ

Quantitative interpretation

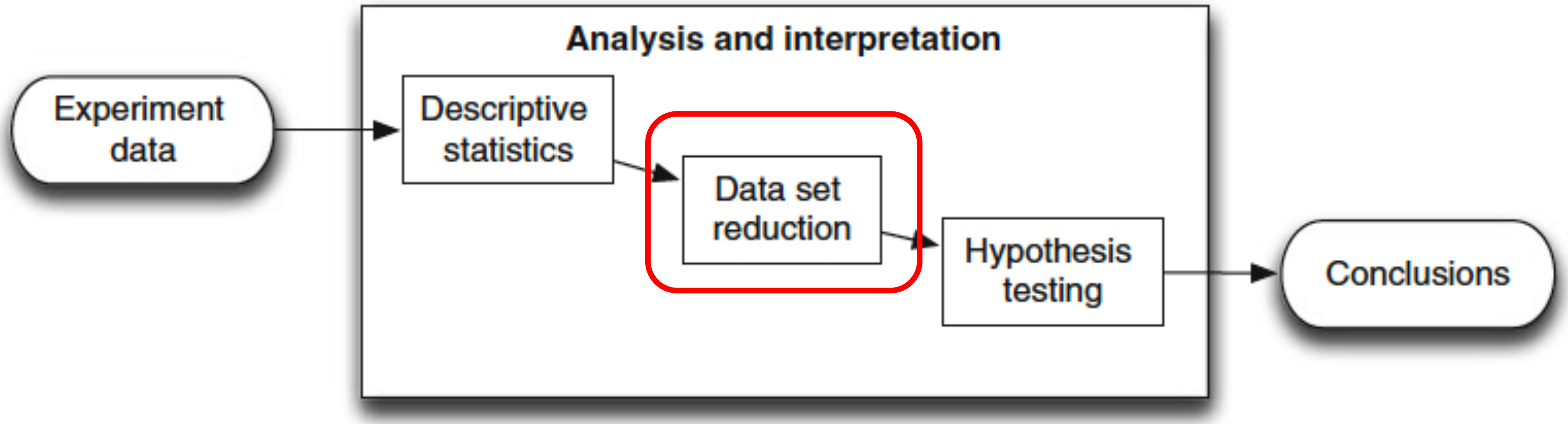
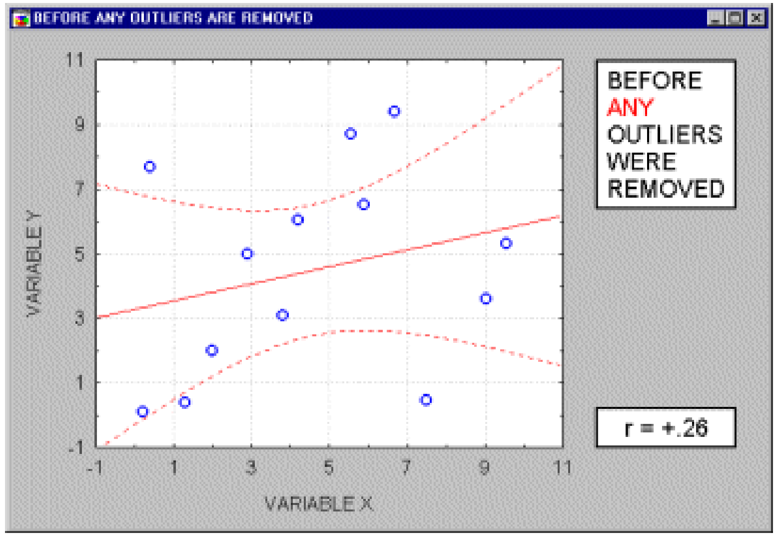
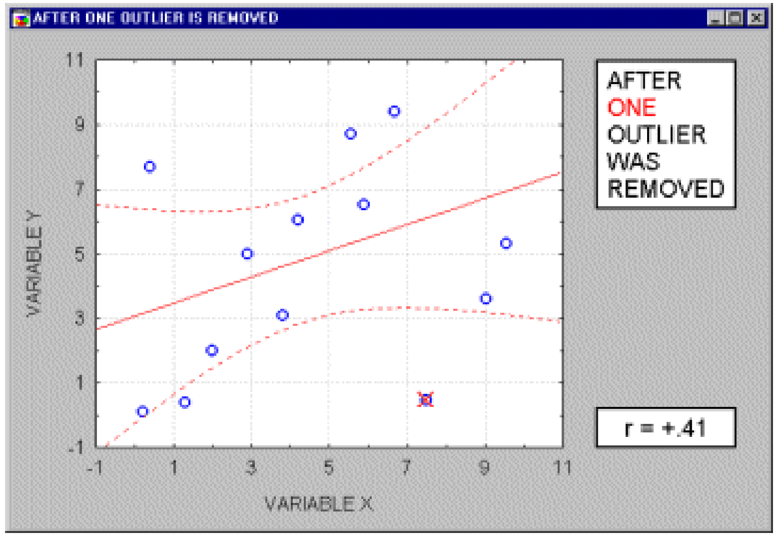


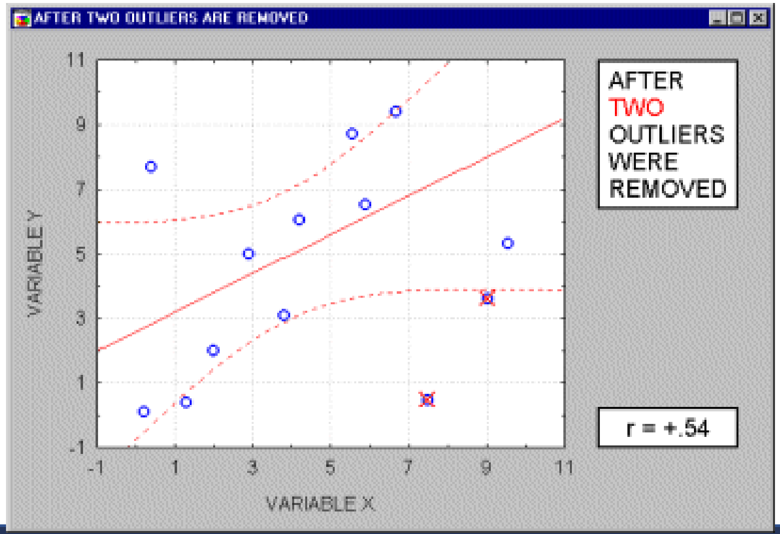
Fig. 10.1 Three steps in quantitative interpretation

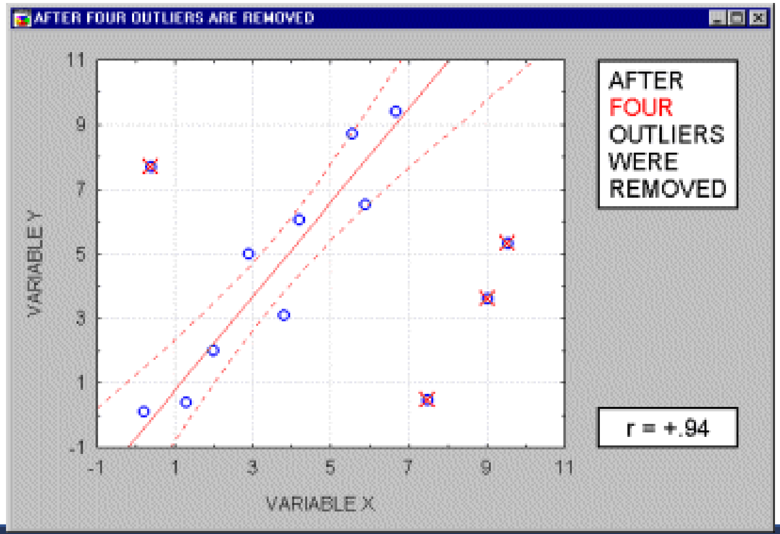
C. Wohlin et al., Experimentation in Software Engineering, Springer-Verlag Berlin Heidelberg 2012

Removal of outliers









Quantitative interpretation

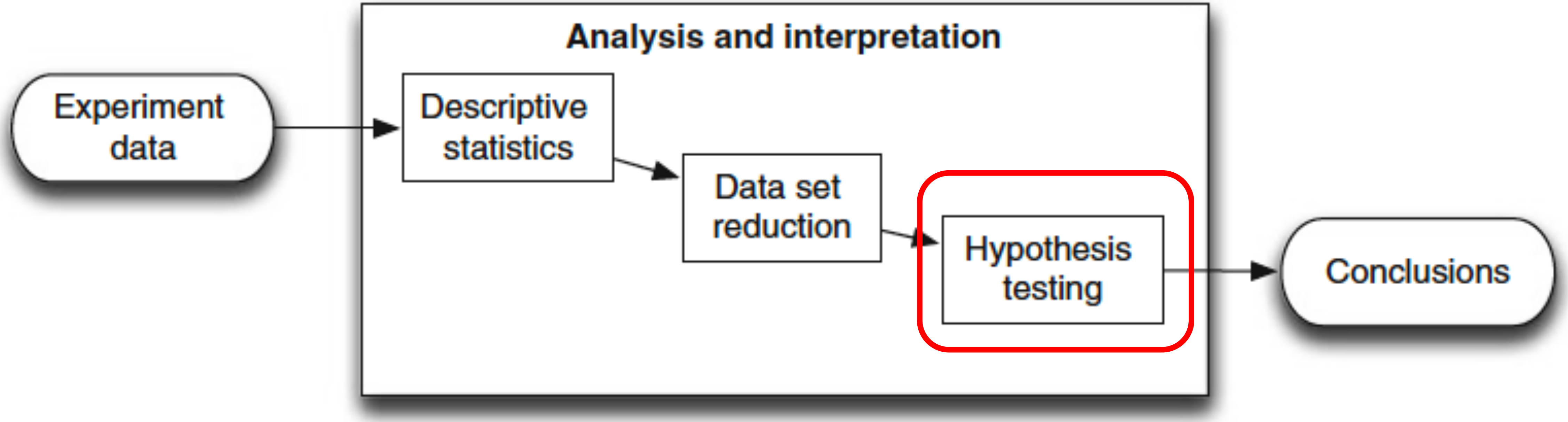


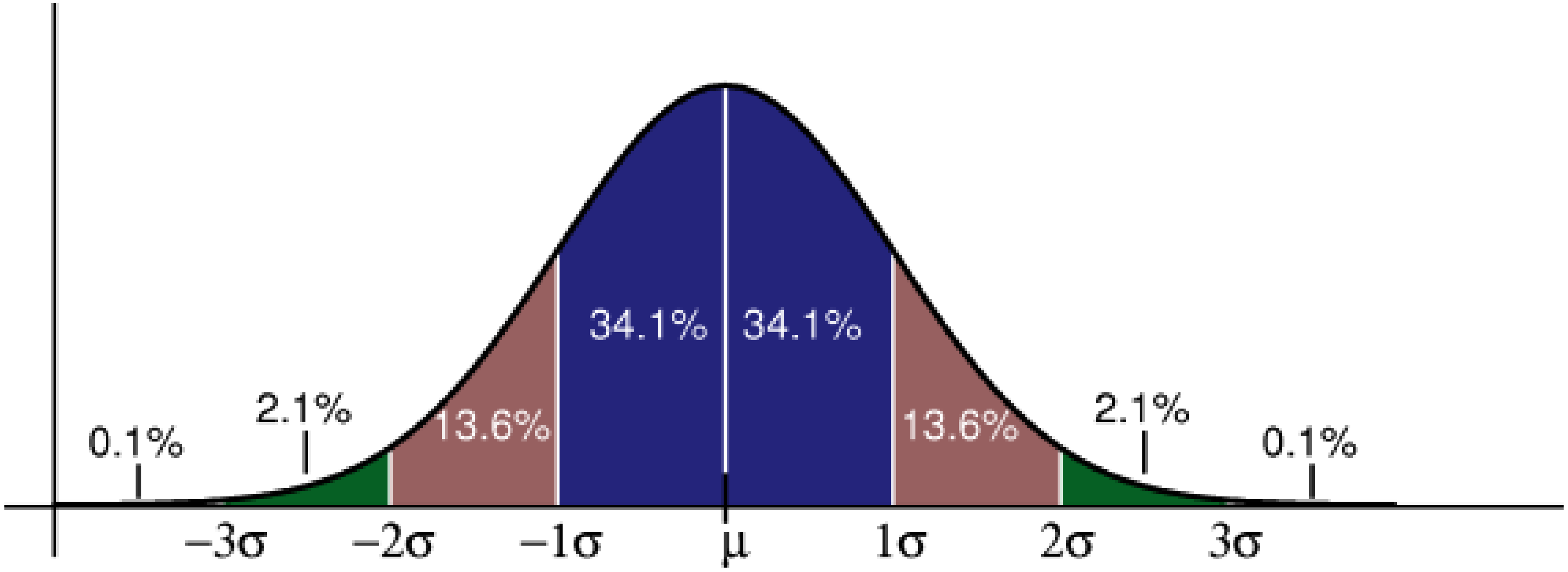
Fig. 10.1 Three steps in quantitative interpretation

C. Wohlin et al., Experimentation in Software Engineering, Springer-Verlag Berlin Heidelberg 2012

Two flavors of Hypothesis Testing

- Parametric tests -- operate on data from a probability distribution, such as the normal distribution or the t -distribution
- Non-parametric tests: distribution free

Normal Distribution



Source: wikipedia - http://en.wikipedia.org/wiki/Image:Standard_deviation_diagram.png

Checking your data is normal

- ➜Draw a Histogram
- ➜Compute the mean and standard deviation
- ➜Superimpose the expected normal curve over the histogram

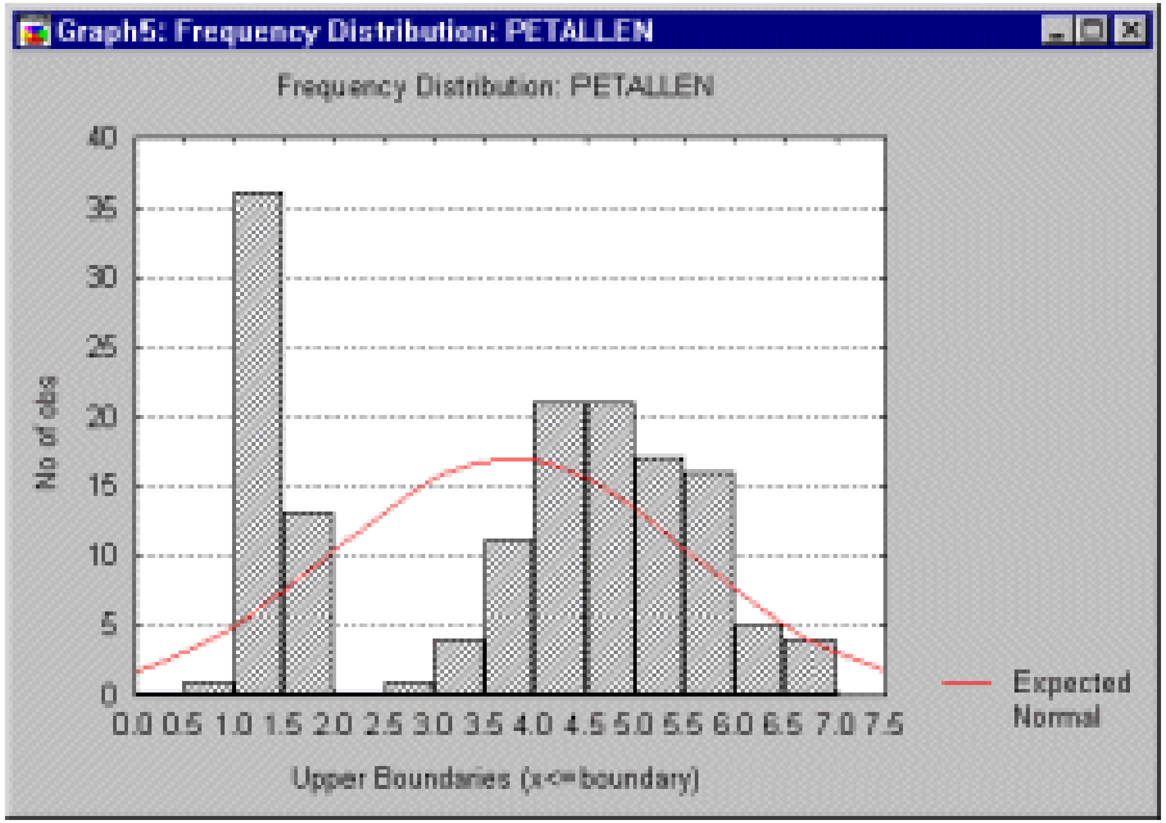


Image source:Statsoft (http://www.statsoft.com/textbook/stathome.html)

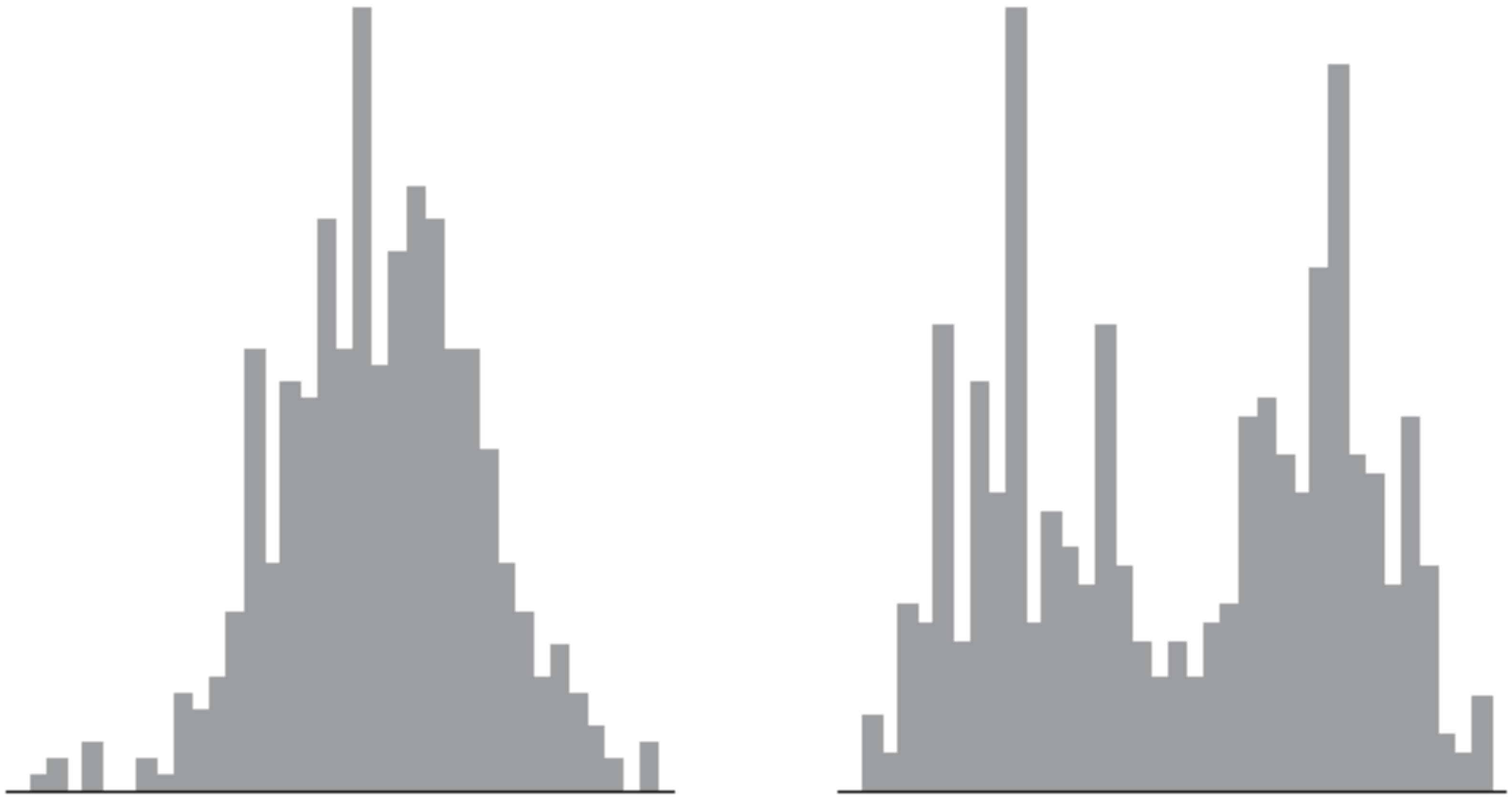


Fig. 2 Two very different samples with the same mean and standard deviation

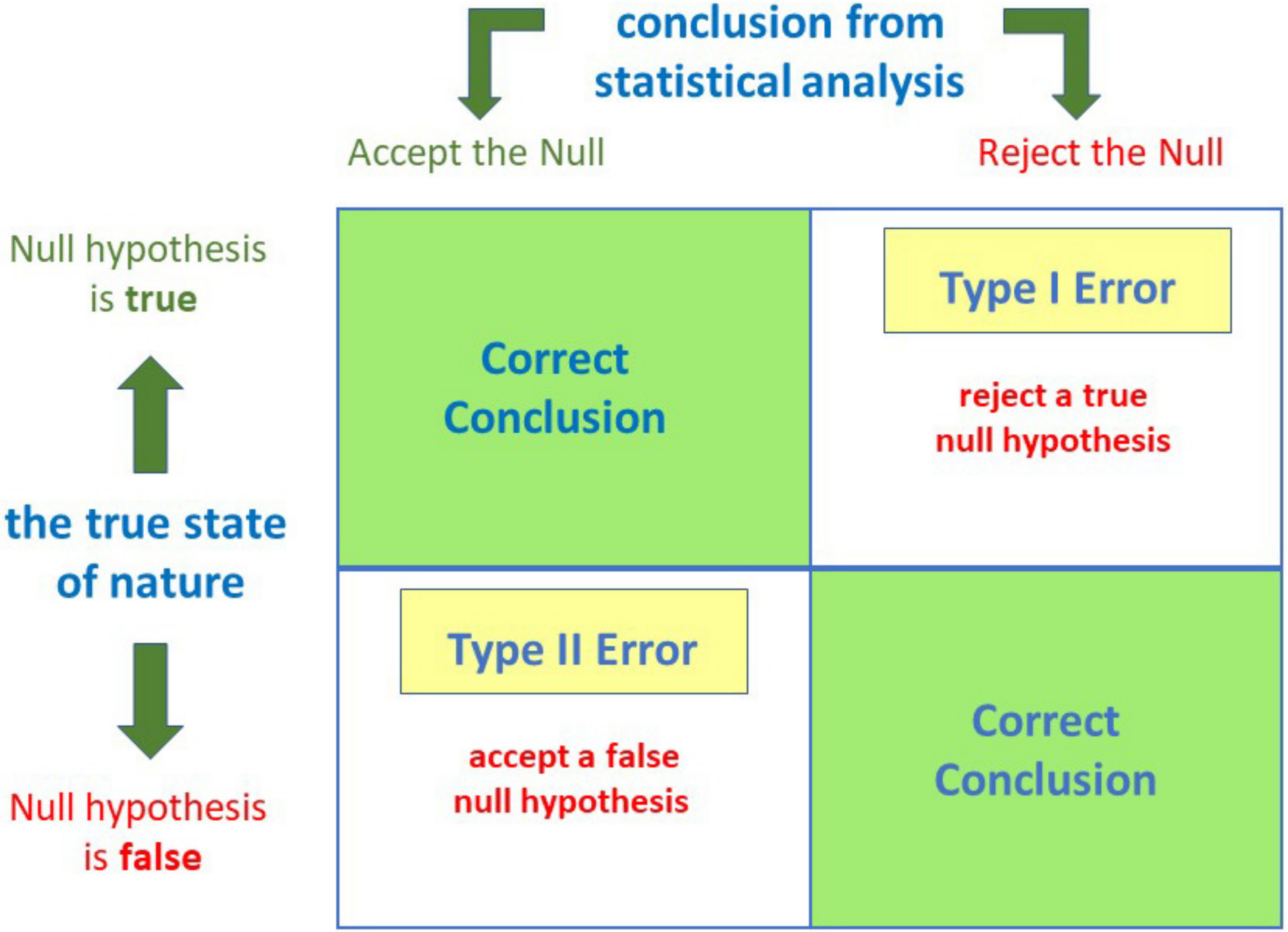
(Ch 6) Statistical Methods and Measurement . from F. Shull et al. (eds.), Guide to Advanced Empirical Software Engineering. Springer 2008

How will you measure things?

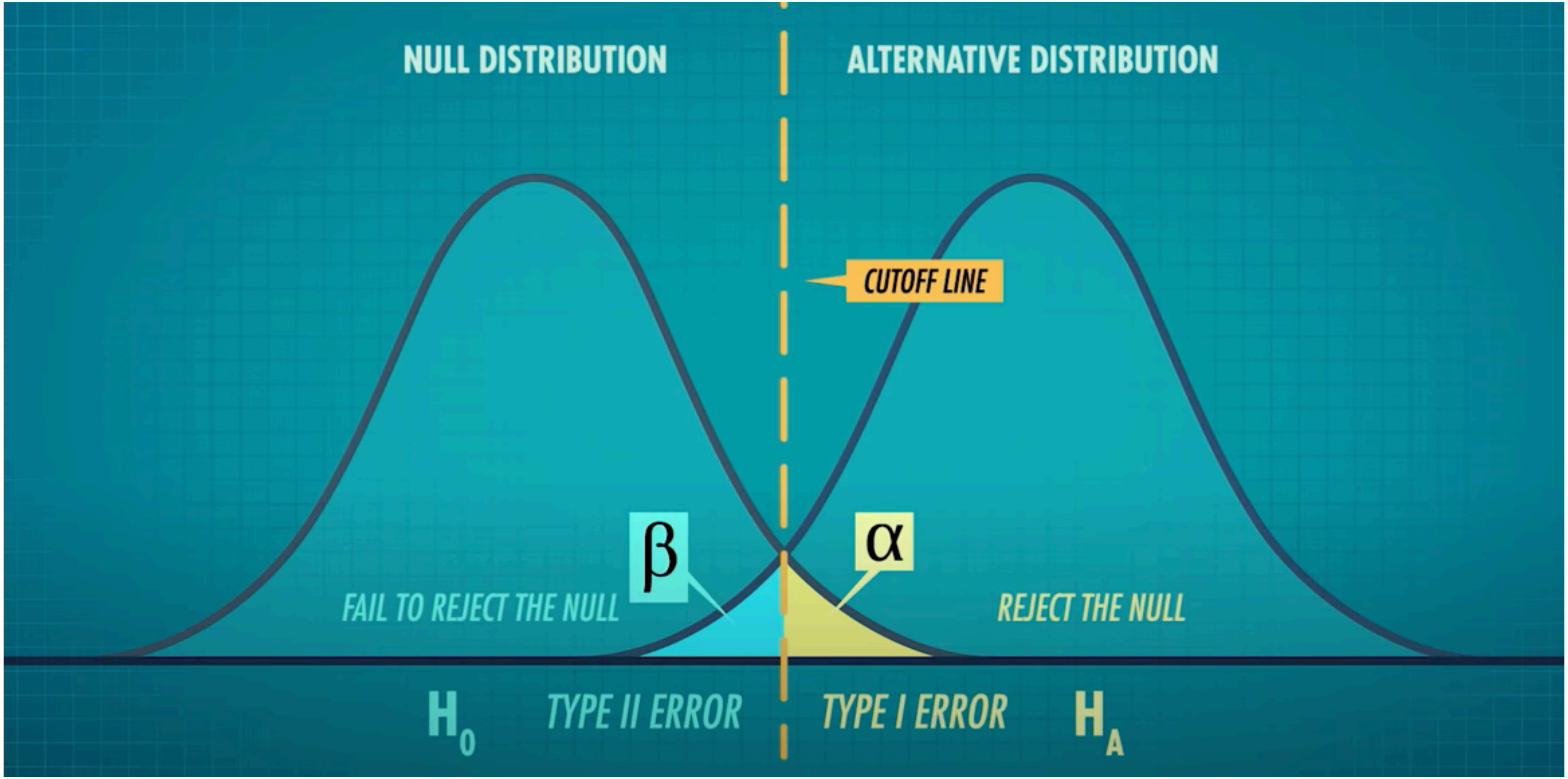
Type	Meaning	Example	Appropriate Stat. Test
Nominal Scale	Unordered categories	Gender, Political preferences, Place of residence	Non-parametric tests
Ordinal Scale	Ranking of objects into ordered categories (intervals between the values are not necessarily of the same size)	Satisfaction, Happiness, grades	Non-parametric tests
Interval Scale	Differences between points on the scale are meaningful (equal intervals)	Celsius, Fahrenheit Temperature, IQ (intelligence scale), SAT scores	Parametric tests Non-parametric tests
Ratio Scale	Ratios between points on the scale are meaningful (ordered, equal intervals with a zero point)	weight, height, sales figures, ruler measurements, number of children	Parametric tests Non-parametric tests

Hypothesis Testing

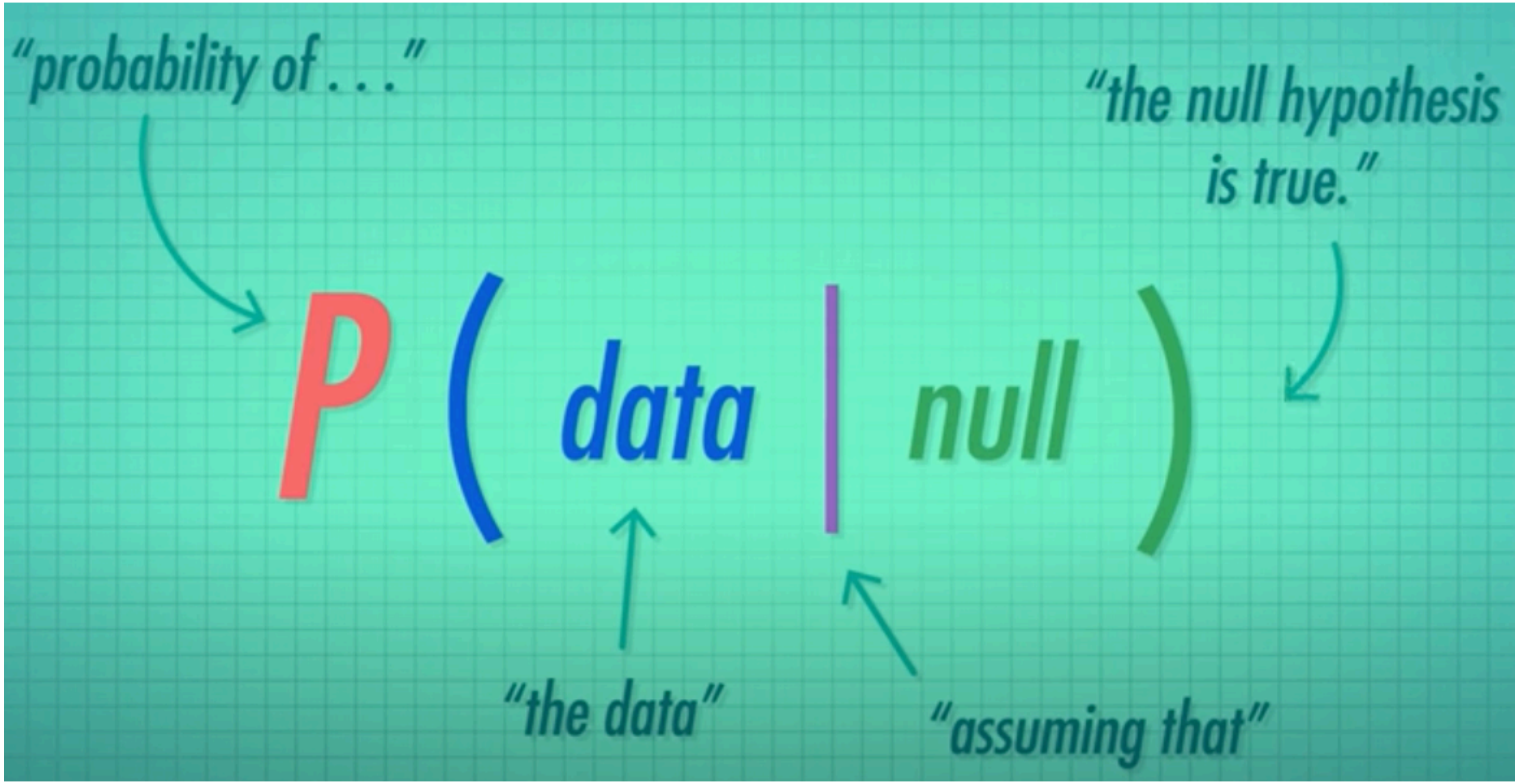
- Set up some hypotheses
  - Null hypothesis (H_0) - asserts that a relationship does not hold
    - In many cases, this is the same as saying there is no difference in the the means of two different treatment groups
  - Alternative hypotheses (H_1, …) - each asserts a specific relationship
  - Type I error: A false positive (rejecting H_0 when it's true)
  - Type II error: A false negative (accepting H_0 when it's false)
- For the statistical tests
  - P value (we calculate this) - probability that a relationship observed in the sample happened by chance
  - Alpha level (selected a priori) - a threshold for p at which we will accept that a relationship did not happen by chance (typically 0.01 or 0.05)
    - This allows us to fix the probability of a type I error in advance
    - If $p < \alpha$, we say the result was significant



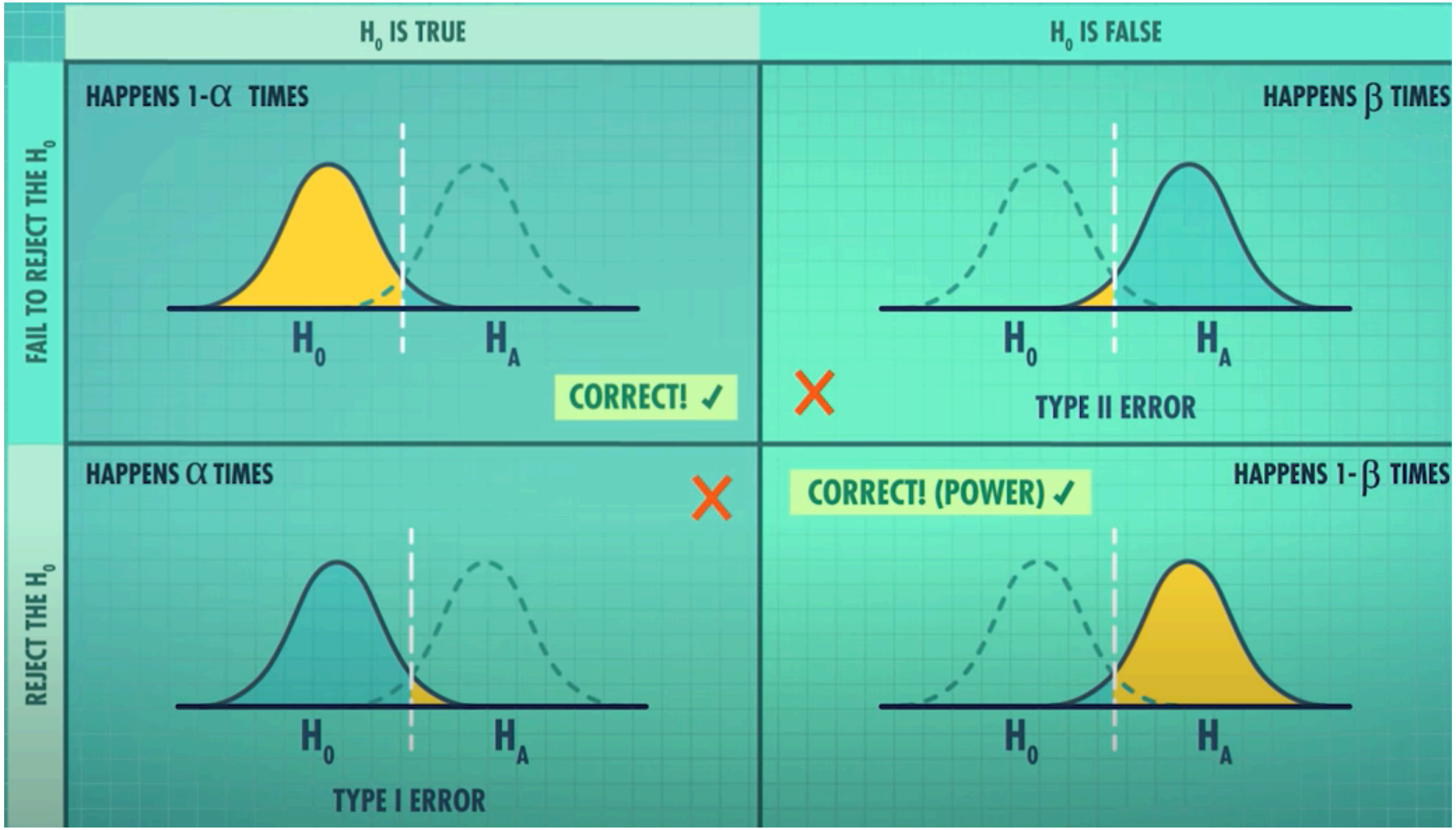
https://www.simplypsychology.org/type_I_and_type_II_errors.html

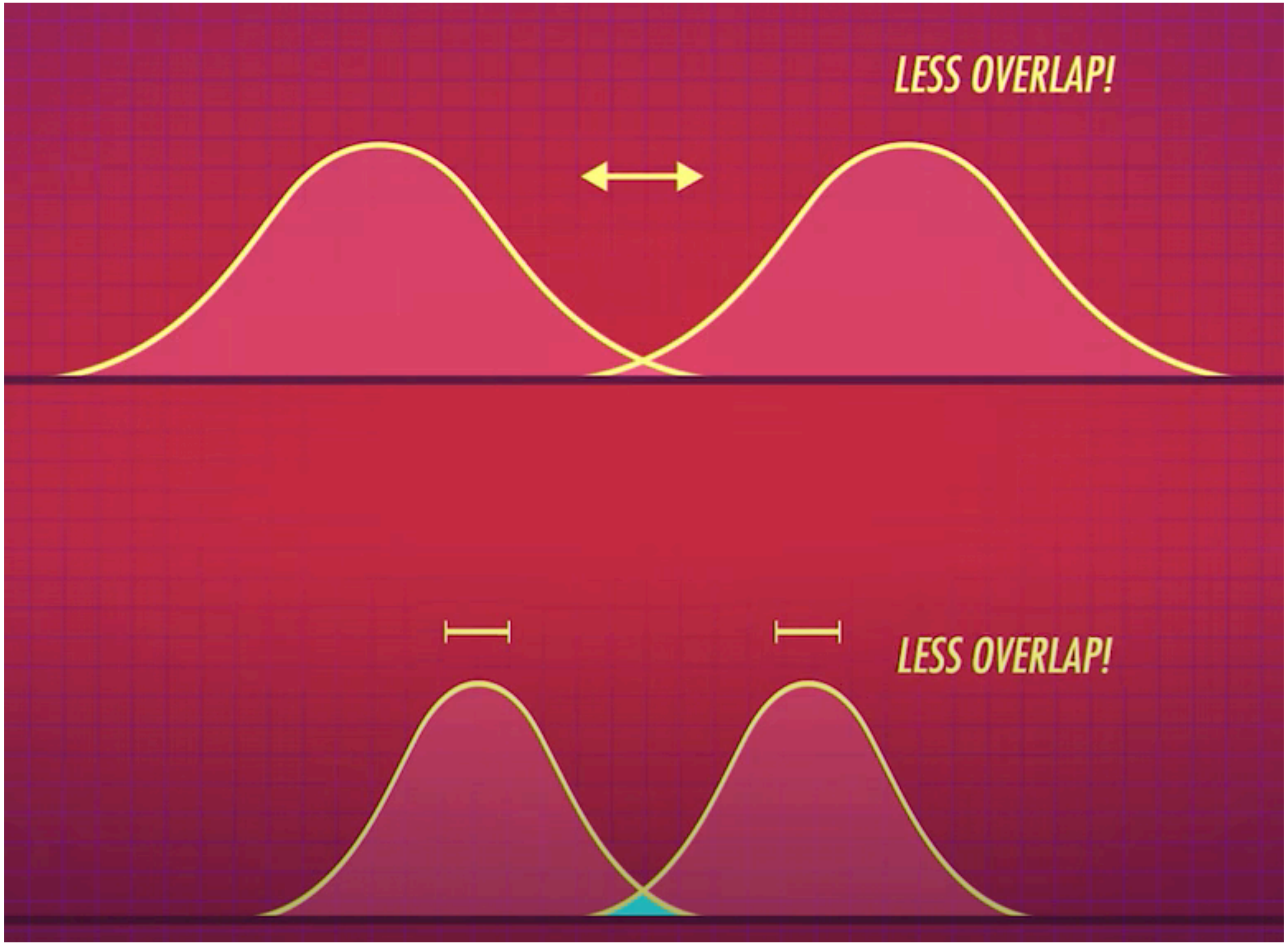

NULL DISTRIBUTION
ALTERNATIVE DISTRIBUTION
CUTOFF LINE
β
α
FAIL TO REJECT THE NULL
REJECT THE NULL
H_0
TYPE II ERROR
TYPE I ERROR
H_A

P-value


"probability of . . ."
"the null hypothesis
is true."
P (data | null)
"the data"
"assuming that"

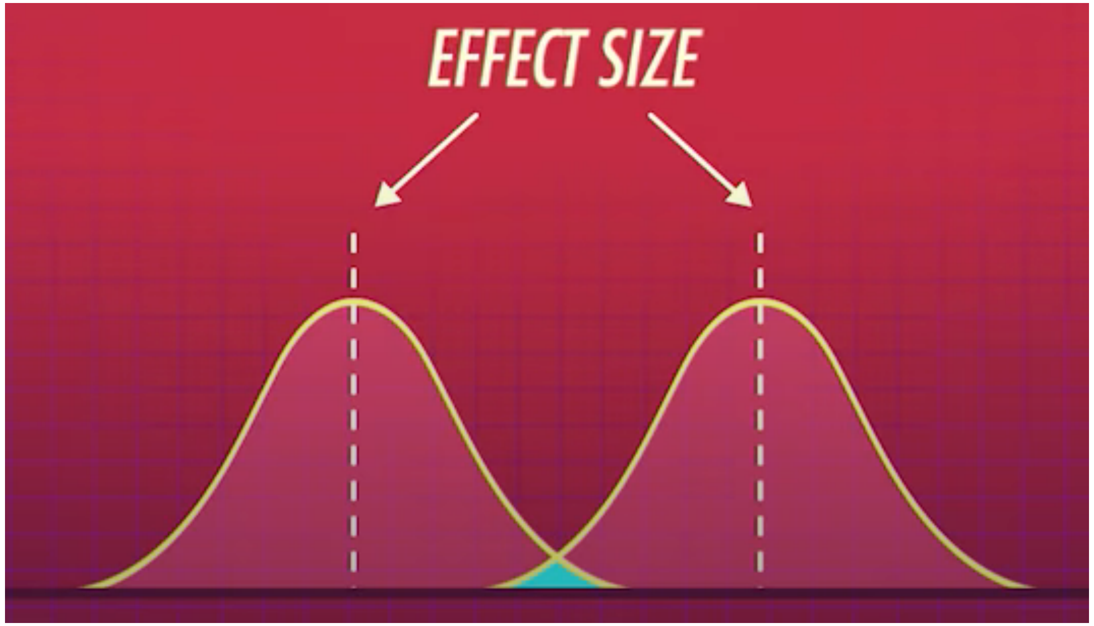
Statistical Power



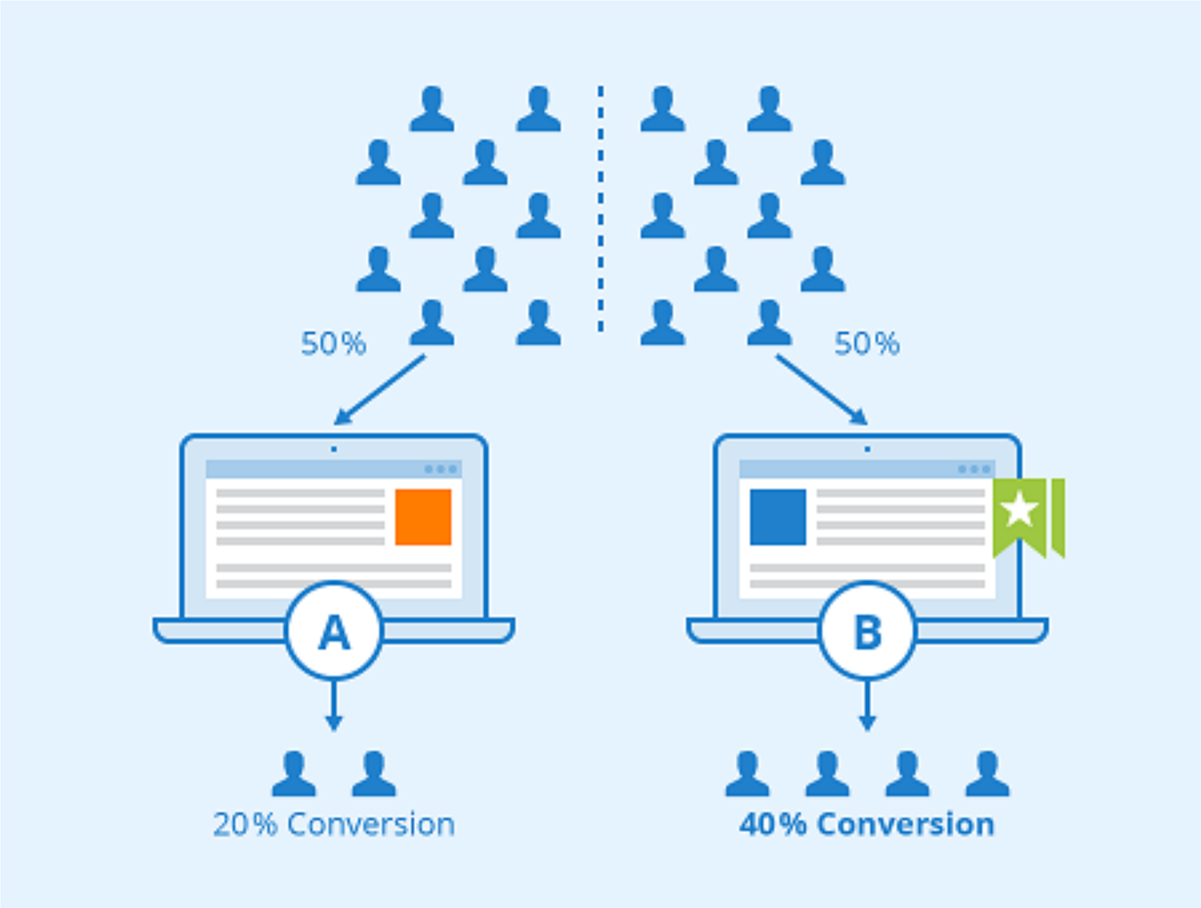

LESS OVERLAP!
LESS OVERLAP!

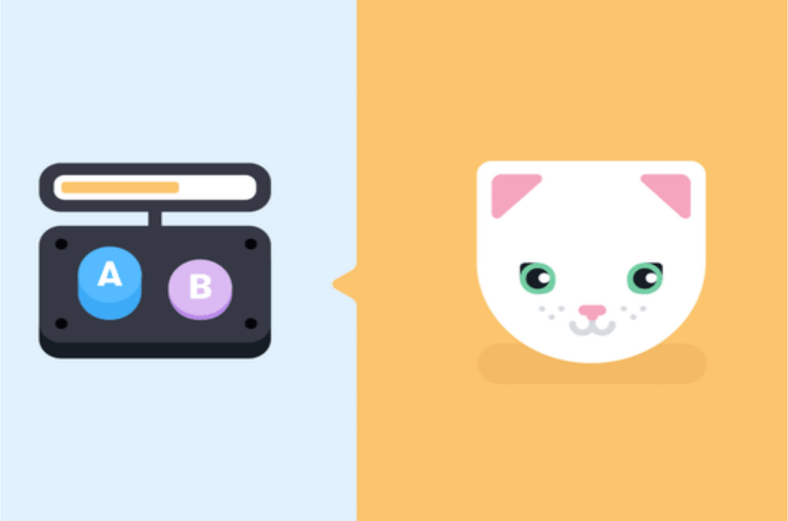
Effect size

- increase sample size
- 80%
- Estimate how many subjects they need based on estimates of effect size and power



A/B Testing





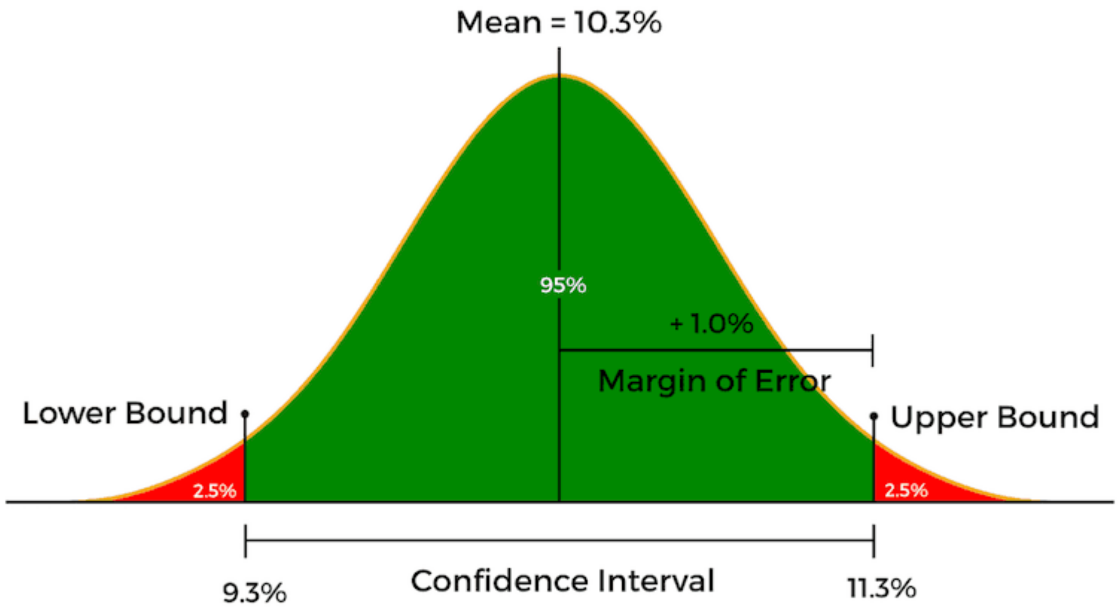
A/B testing

- Controlled randomized experiment with two variants, A and B, which are the control and treatment.
- One group of users given A (current system); another random group presented with B; outcomes compared.
- Often used in web or GUI-based applications, especially to test advertising or GUI element placement or design decisions.

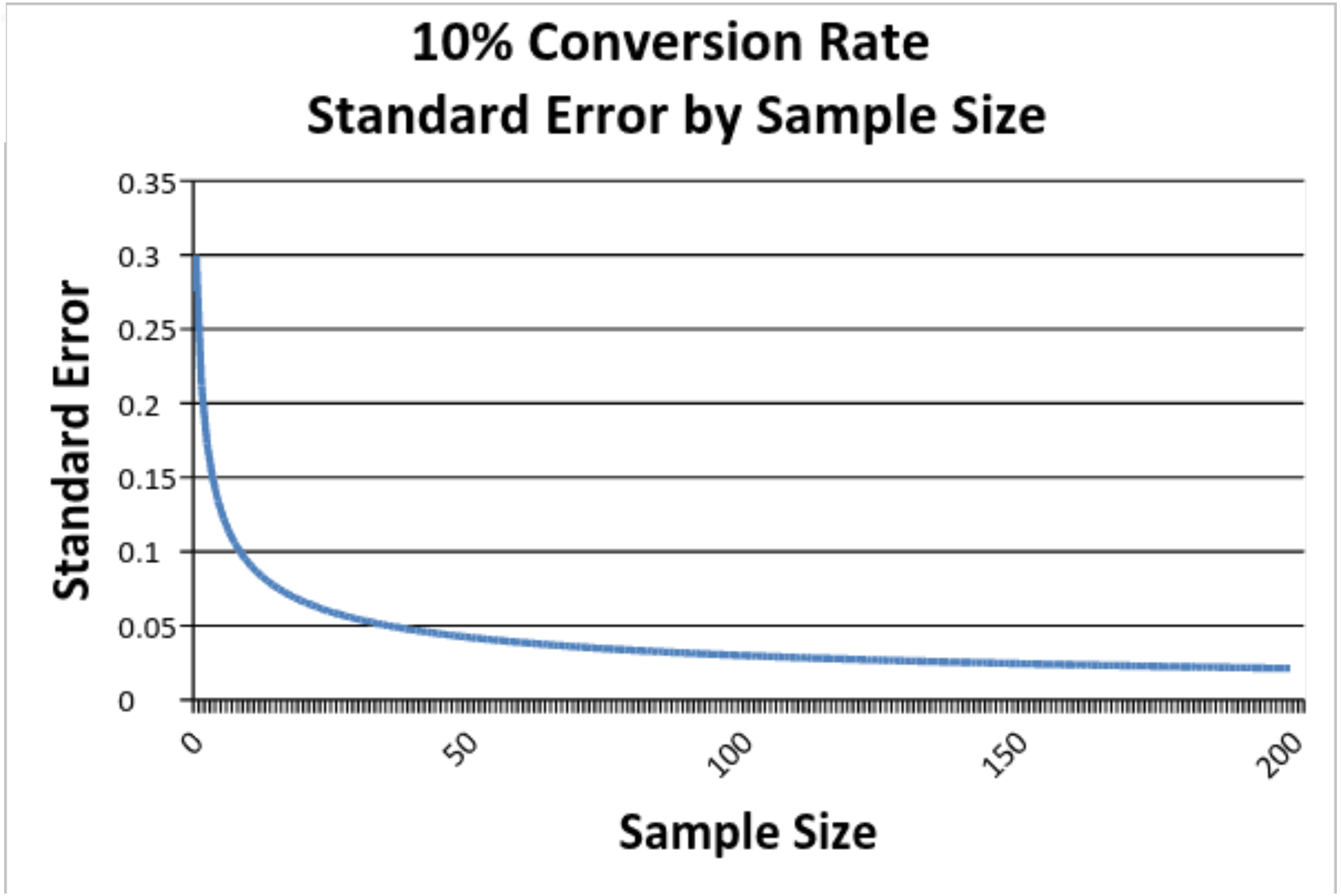
EXAMPLE

- 3,662 visitors and get 378 conversions:

  conversion rate = 378/3,663 = 10.3%   ± 1.0 %



EXAMPLE

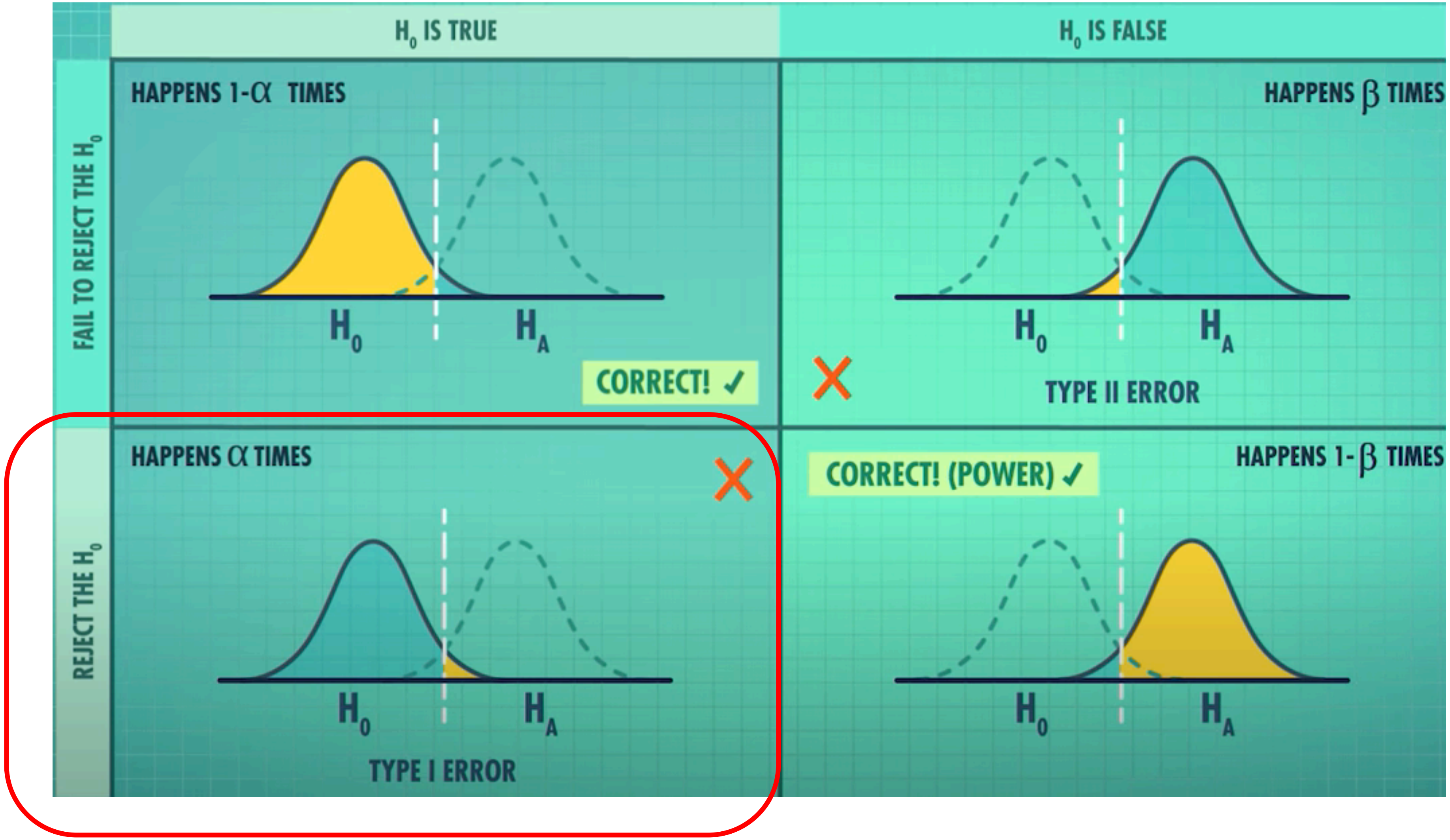


10% Conversion Rate
Standard Error by Sample Size
Standard Error
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
50
100
150
200
Sample Size

Hypothesis testing

- H0: Variation B is NOT a meaningful improvement over Variation A.
- H1: Variation B will convert at a higher rate for our overall population than Variation A will.

1. Test says Variation B is better & Variation B is actually better
2. Test says Variation B is better & Variation B is not actually better (**type I error**)
3. Test says Variation B is not better & Variation B is actually better (**type II error)**
4. Test says Variation B is not better & Variation B is not actually better

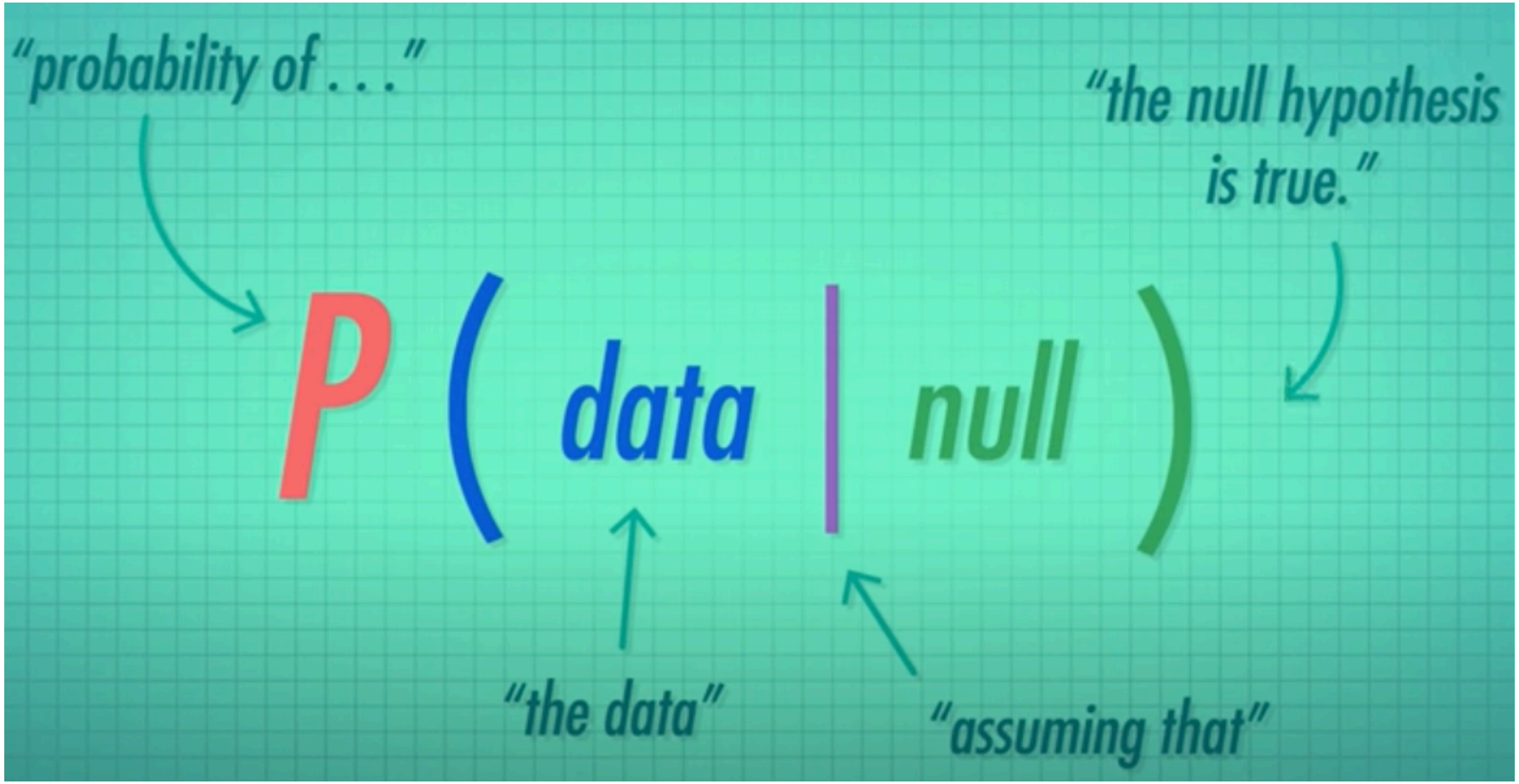

H$_0$ IS TRUE
H$_0$ IS FALSE
FAIL TO REJECT THE H$_0$
REJECT THE H$_0$
HAPPENS 1-α TIMES
H$_0$
H$_A$
CORRECT! ✓
HAPPENS β TIMES
TYPE II ERROR
HAPPENS α TIMES
TYPE I ERROR
CORRECT! (POWER) ✓
HAPPENS 1-β TIMES

EXAMPLE

Type I error

If Variation B performs better in our sample. How do we know whether or not that improvement will translate to the overall population? How do we avoid making a type I error?

P-value

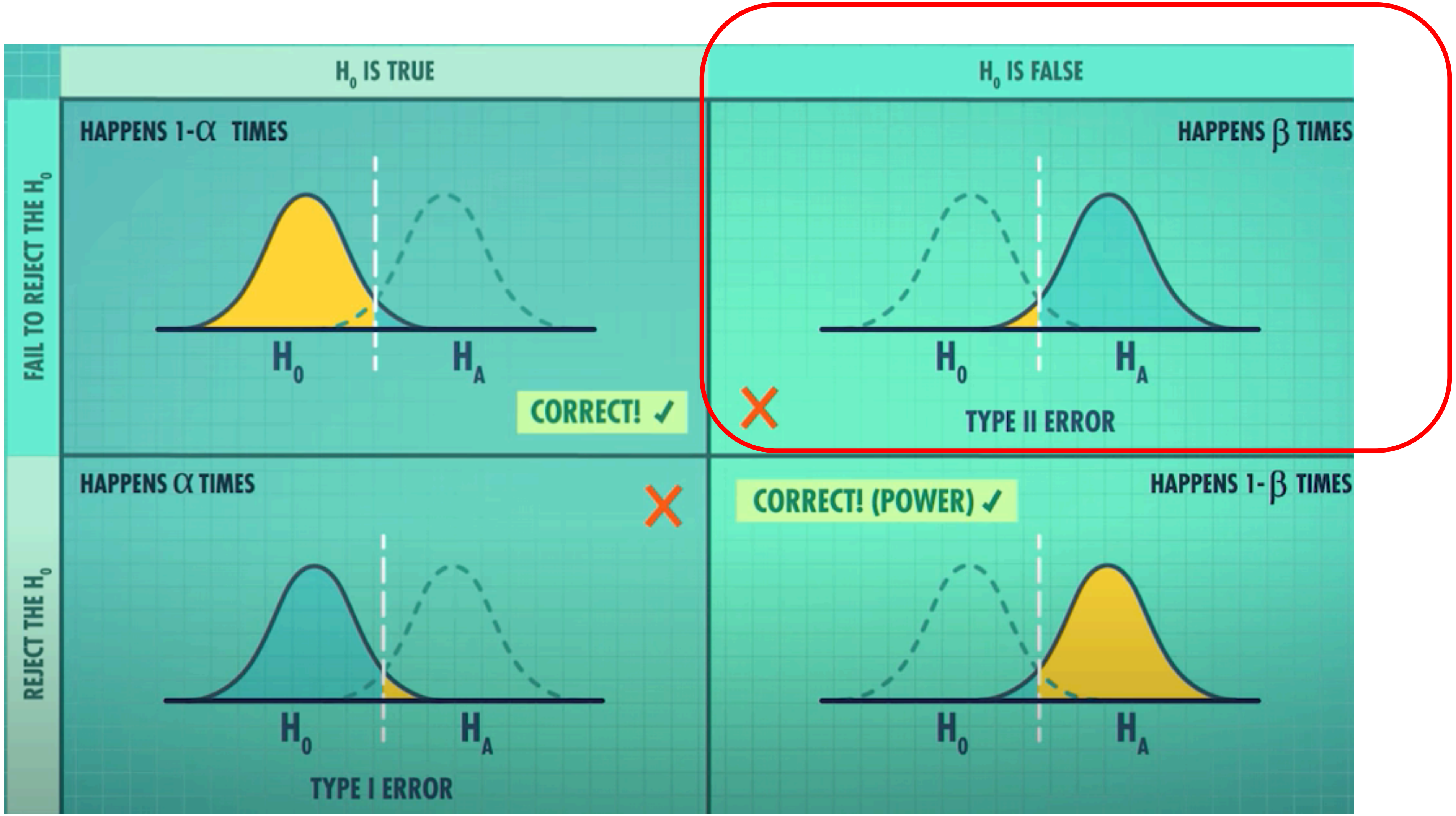

"probability of . . ."
"the null hypothesis is true."
P (data | null)
"the data"
"assuming that"

EXAMPLE

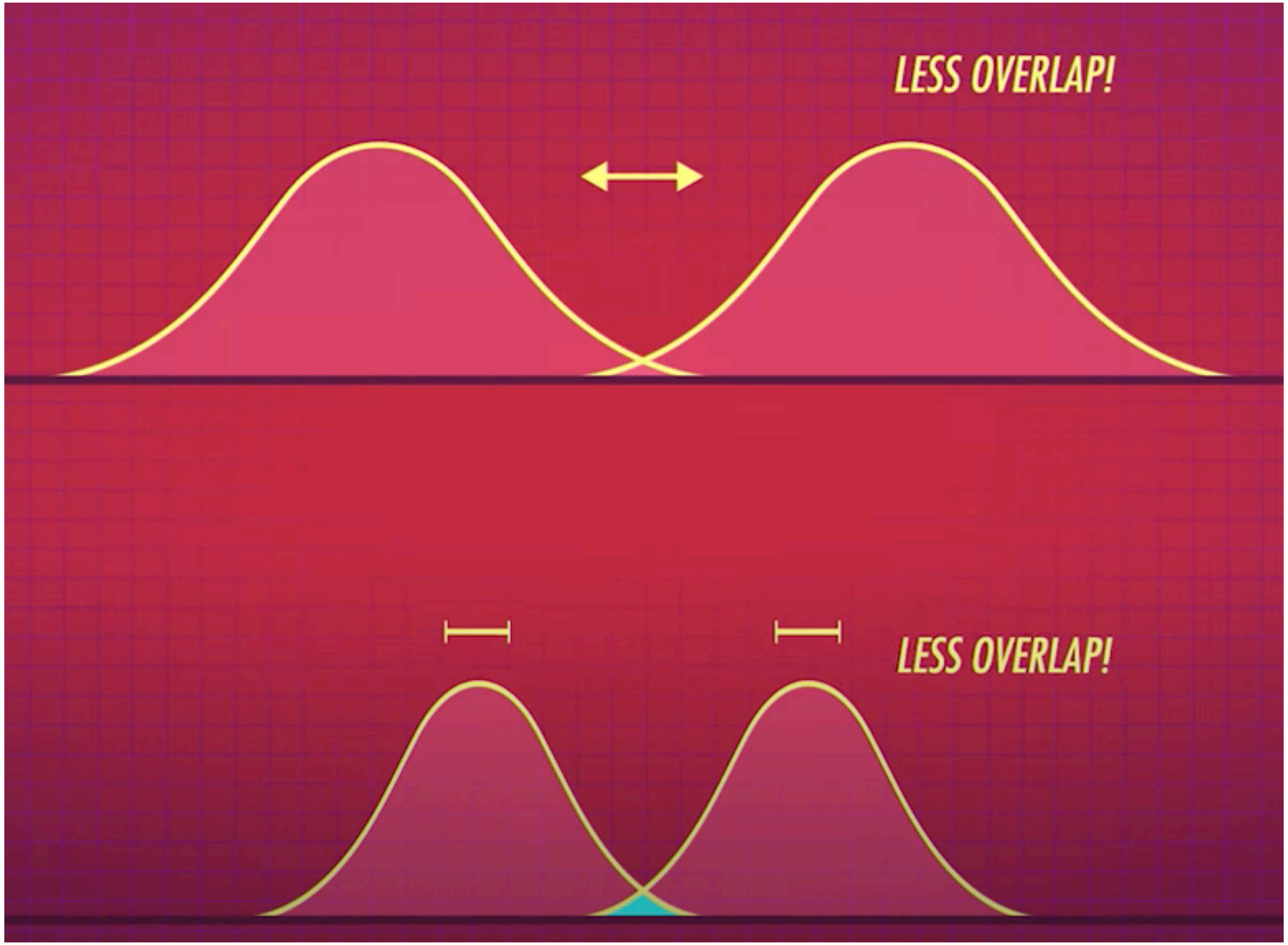
Hypothesis testing

- H0: Variation B is NOT a meaningful improvement over Variation A.
- H1: Variation B will convert at a higher rate for our overall population than Variation A will.

1. Test says Variation B is better & Variation B is actually better
2. Test says Variation B is better & Variation B is not actually better (**type I error**)
3. Test says Variation B is not better & Variation B is actually better (**type II error)**
4. Test says Variation B is not better & Variation B is not actually better



H_0 IS TRUE
H_0 IS FALSE
FAIL TO REJECT THE H_0
HAPPENS $1-\alpha$ TIMES
H_0
H_A
CORRECT! ✓
HAPPENS β TIMES
H_0
H_A
TYPE II ERROR
REJECT THE H_0
HAPPENS α TIMES
H_0
H_A
TYPE I ERROR
CORRECT! (POWER) ✓
HAPPENS $1-\beta$ TIMES
H_0
H_A

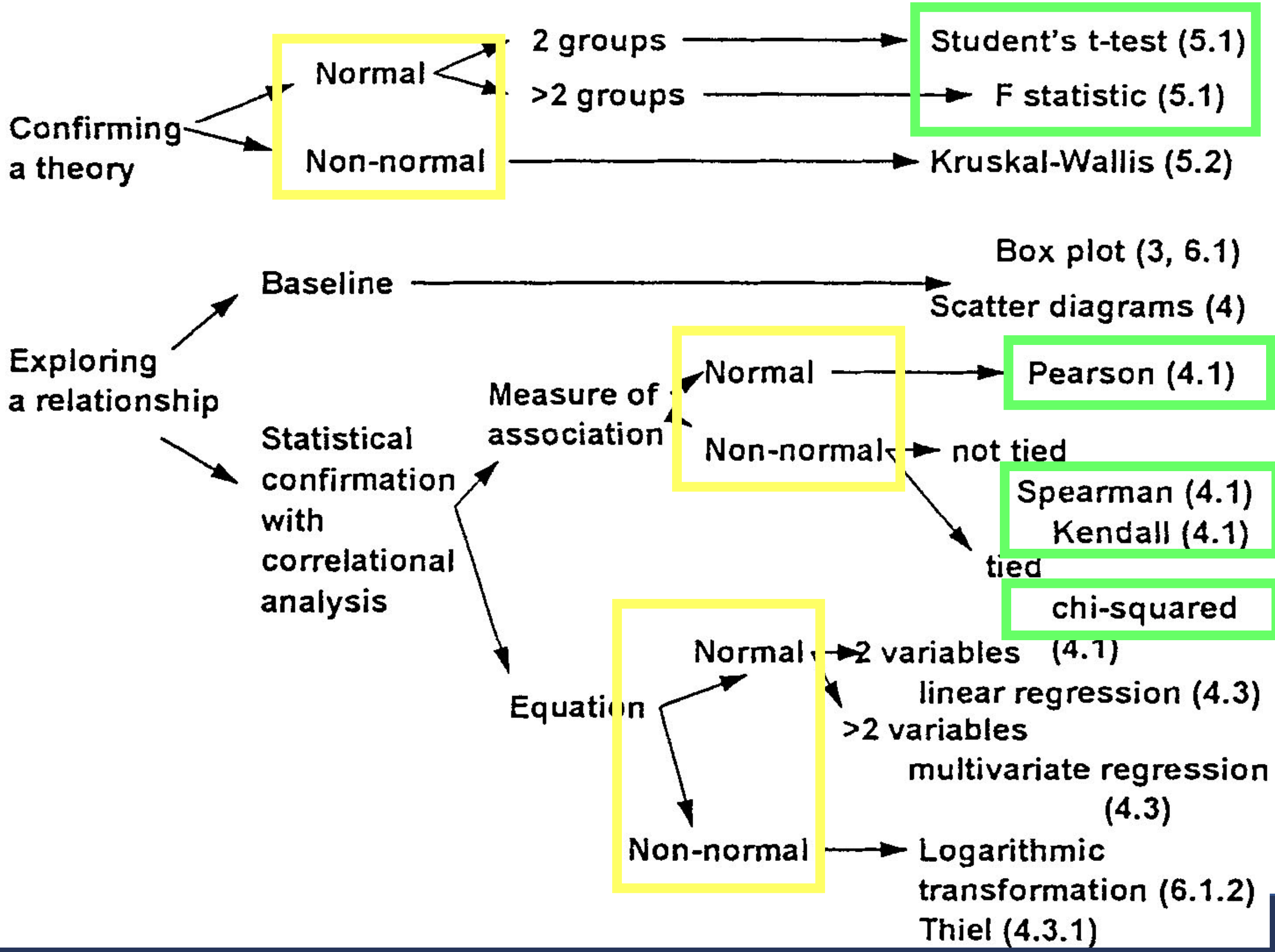

LESS OVERLAP!
LESS OVERLAP!

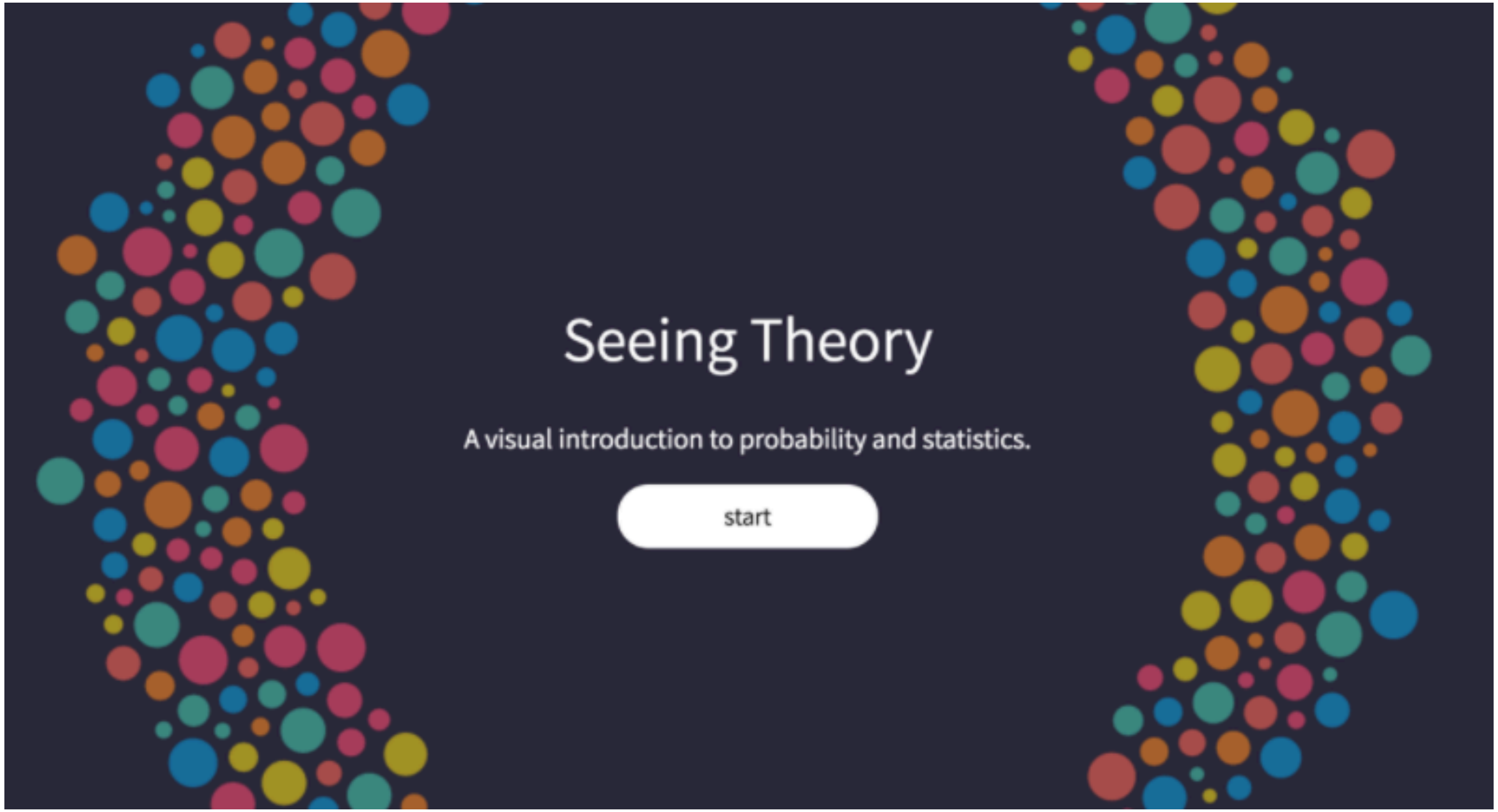
Statistical significance cannot be used as a stopping point

Usability: A/B testing

- However, it cannot..
  - Tell you why
  - Let you test drastic redesigns of your website or app.
  - Tell you if you're solving the right/wrong problem.

Which Statistical Test?



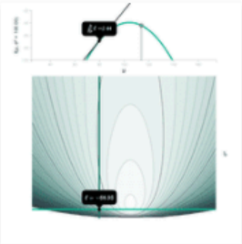


https://seeing-theory.brown.edu/

Visualizations

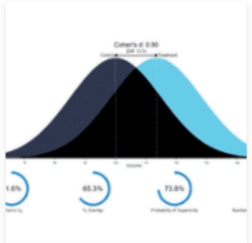
Share

Since 2014 I've tried to illustrate various statistical concepts using interactive visualizations. This page contains an updated overview of my visualizations.



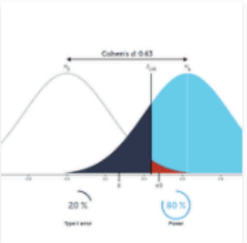
Maximum Likelihood

An interactive post covering various aspects of maximum likelihood estimation.



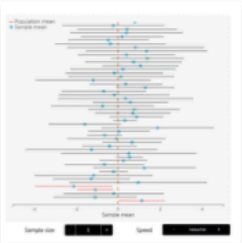
Cohen's d

An interactive app to visualize and understand standardized effect sizes.



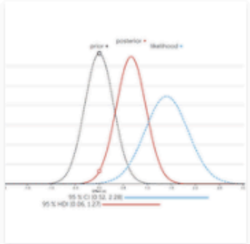
Statistical Power and Significance Testing

An interactive version of the traditional Type I and II error illustration.



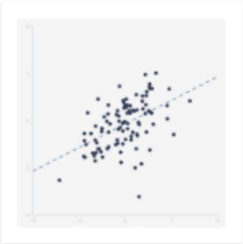
Confidence Intervals

An interactive simulation of confidence intervals



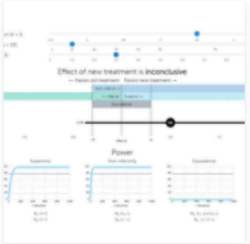
Bayesian Inference

An interactive illustration of prior, likelihood, and posterior.



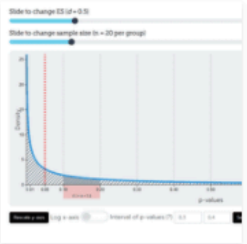
Correlations

Interactive scatterplot that let's you visualize correlations of various magnitudes.



Equivalence and Non-Inferiority Testing

Explore how superiority, non-inferiority, and equivalence testing relates to a confidence interval



P-value distribution

Explore the expected distribution of p-values under varying alternative hypothesises.

https://rpsychologist.com/viz

References

- (Ch 10) Analysis and Interpretation . from C. Wohlin et al., Experimentation in Software Engineering, Springer-Verlag Berlin Heidelberg 2012

- (Ch 6) Statistical Methods and Measurement . from F. Shull et al. (eds.), Guide to Advanced Empirical Software Engineering. Springer 2008

- (Ch 6) Hypothesis Testing. from MacKenzie. Human-Computer Interaction. Elsevier 2013