



Institute of
Data

2021



Data Science and AI

Module 2

Part 2:

Data Science Practices



Agenda: Module 2 Part 2

- Defining Data Science
- Hypothesising
- Statistical Evidence
- Statistical Proof
- Causation
- Statistical Inferences



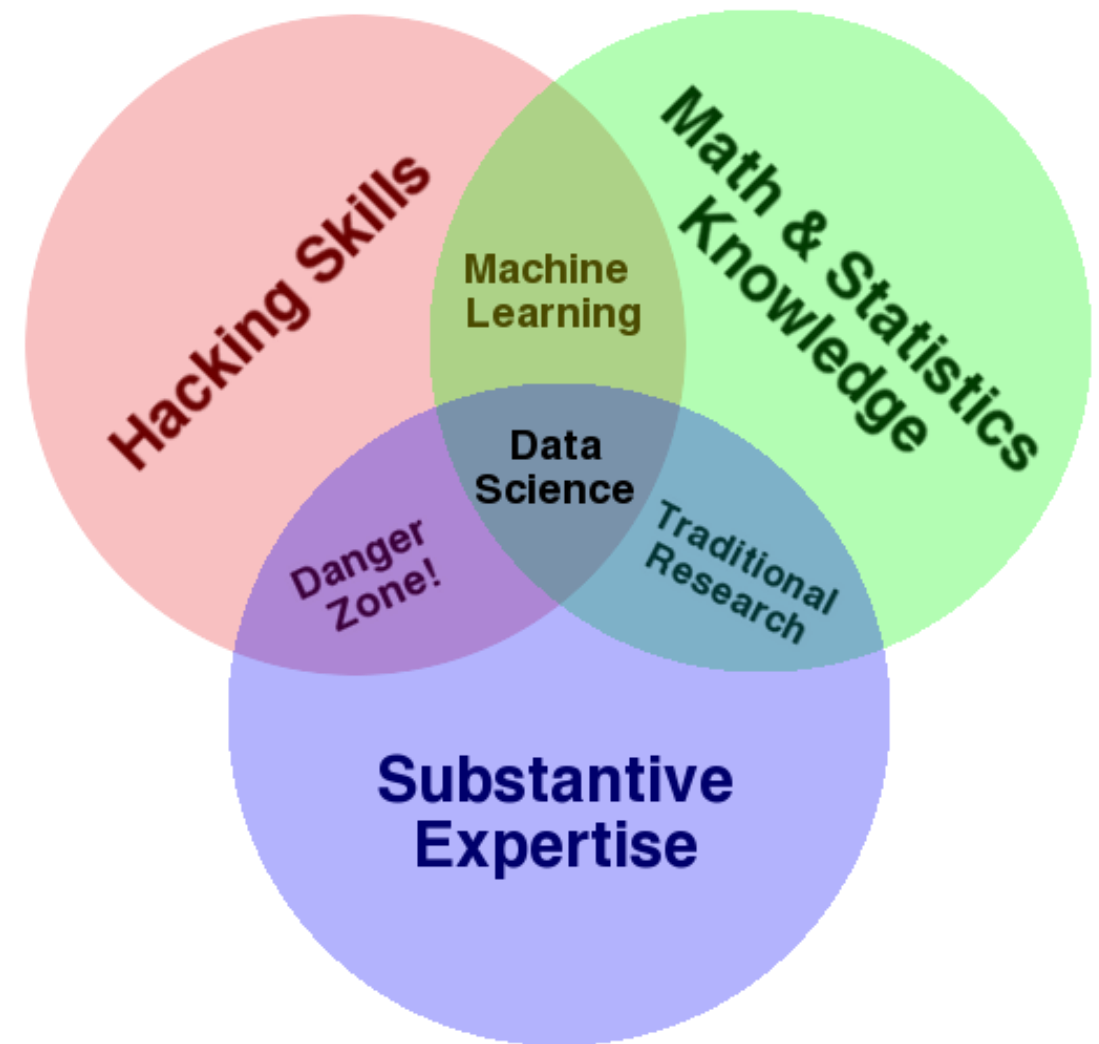
Defining Data Science

- What is data science?
- Users and use cases
- What makes a data scientist?
- The data science pipeline
- Testable hypotheses



What is Data Science?

- Cutting-edge techniques and tools for analysing data
- An interdisciplinary approach to problem-solving
- Business analysis on steroids
- The application of scientific method to practical problems



Drew Conway



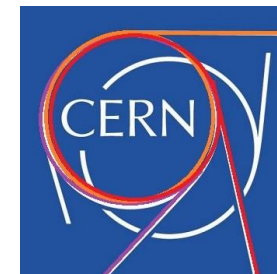
Who Uses Data Science?

NETFLIX

amazon.com[®]



Google





Where do data scientists come from?

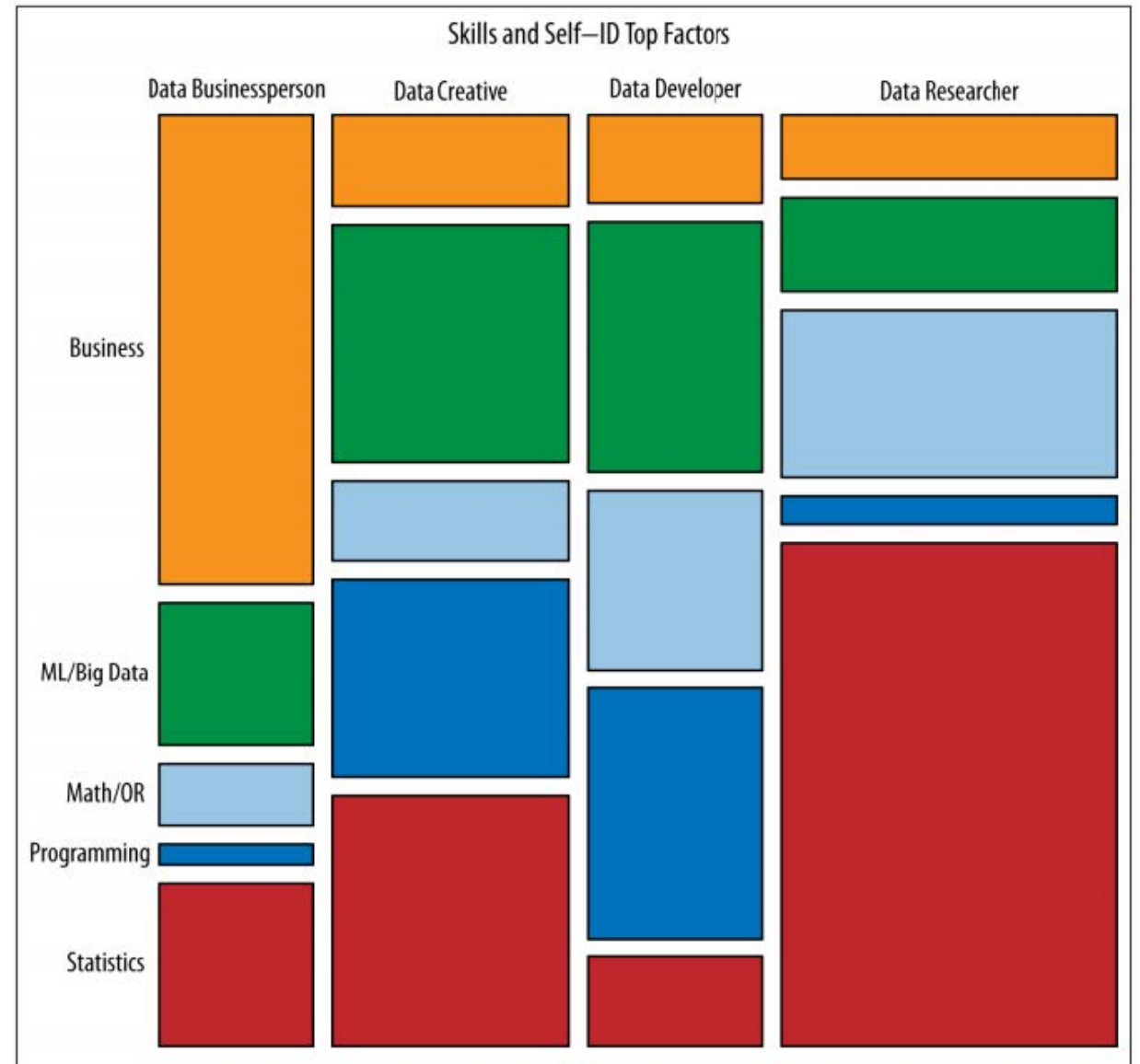
What are their typical strengths?

	Hacking Skills	Math & Stats	Substantive Expertise	Methodology	Abstraction	Communication
Data Science program graduates						
Scientists (especially physics)						
Statisticians						
Developers						
Business Analysts						



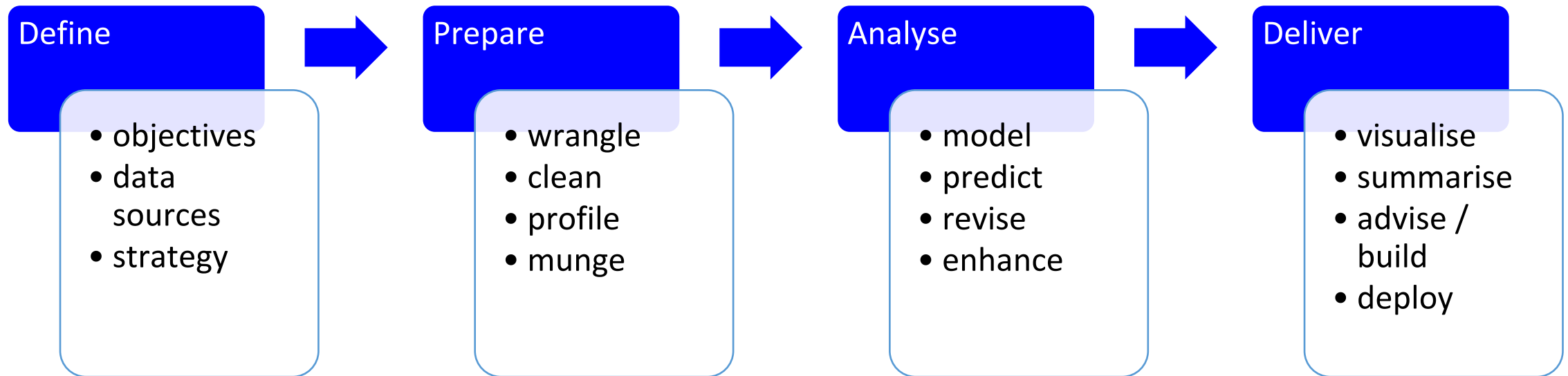
Relative Strengths

- These roles prioritise different skill sets.
- All roles involve some part of each skillset.
- *Where are your ambitions?*
- *Where are your strengths and weaknesses?*





The Data Science Pipeline





Defining the Problem

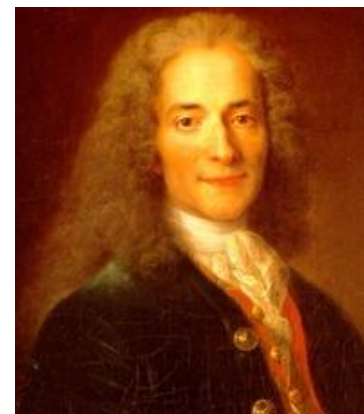
Every Solution Begins with a Question

- Any business problem, decision-support tool, or clever data product begins life with a well-defined need:
 - A set of questions that frame an analysis
- Sets up for a successful process
- Establishes the basis for reproducibility
- Creates scope for future expansion



“A problem well stated is half solved.”

— Charles Kettering



“Judge a man by his questions rather than by his answers.”

— Voltaire



What is your question?

- What is your name?
- What is your quest?
- What is the average airspeed of an unladen sparrow?





How to specify the question

A business challenge may be vague:

- “How can we grow our online market share?”

Data science questions need to be focused:

- “Is our website achieving sufficient user engagement?”
- “Are we presenting our products effectively to website visitors?”
- “Are our prices competitive?”
- “Is this market niche saturated?”

> Even these examples are a bit vague, but we could break each one down into a series of more granular questions with quantitative domains



The Elements of a Good Question

Specific

The dataset and key variables are clearly defined.

Measurable

The type of analysis and major assumptions are articulated.

Attainable

The available data are amenable to the question and unlikely to be biased.

Reproducible

The analysis can be repeated by another person or at another time.

Time-bound

The time period and population to which the analysis pertains is clearly stated.



Knowledge check

Does this question follow the SMART framework:

“Is there an association between number of passengers with carry-on luggage and delayed take-off time?”



Knowledge check

How about this (revised) question:

“Is there an association between the number of passengers (on JetBlue, Delta, and United domestic flights) with carry-on luggage and delayed take-off time in the data from flightstats.com between January 2015 and December 2015?”



Dataset Characteristics

- What would we look for if we wanted to be able to describe a dataset?
 - size, completeness
 - accuracy, precision
 - periodicity, stationarity
 - variance
 - bias
 - missing variables
 - correlated variables
 - due to causation or covariation
 - correlated samples
 - time series
 - contaminated or prejudiced sampling



Data Temporality

Cross-sectional

- 'static'
- treated as a snapshot in time
- causality is simultaneous

Longitudinal

- 'time series'
- treated as a series of snapshots with a temporal or serial dependence

Dynamic

- 'streaming'
- continuously accumulated or refreshed



Variables in Data Science

Features
Predictors

Independent variables
Inputs

A *predictor* is a *feature* that is useful in modelling the *response*. Specifically, its inclusion enables a *model* to account for more of the *variance* in the response.

Responses
Outcomes

Dependent variables
Outputs

A covariate is a variable that is possibly predictive of the response. It could also represent an interacting variable.

A confounding variable is one which influences the response but has not been measured (i.e. it introduces bias).



Data Preparation

def: Tidy data: the end goal of data cleaning and munging

- each variable should be in one column
- each observation should comprise one row
- each type of observational unit should form one table
- key columns for linking multiple tables
- top row contains (sensible) variable names
- in general, save data as one file per table



this is Codd's
3rd normal
form from
RDBMS theory

- search: “hadley wickham's tidy data paper”



Lab 2.2.1: Hypothesising

- Purpose:
 - To create a testable hypothesis
- Resources:
 - 'titanic.csv'
- Instructions:
 1. You should already be familiar with the 'titanic' dataset from the last module's homework. Now, think about what stories the data might tell, and devise a hypothesis that could be tested.
 2. Provide some data profiling results to support your assertion that this hypothesis is testable.





Statistical Evidence

- What is statistical proof?
- Revisiting the null hypothesis
- The Student's t -test



Statistical Proof

Can a hypothesis be proved?

- in science, no theory (or hypothesis) can actually be proved
 - must explain known phenomenon
 - must make testable predictions
 - *will gain acceptance if it survives rigorous testing*

How can a hypothesis be tested?

- by formulating it in a way that makes its claims amenable to statistical analysis
 - must explain the data
 - must have a corresponding null hypothesis that can be rejected at a predefined level of confidence

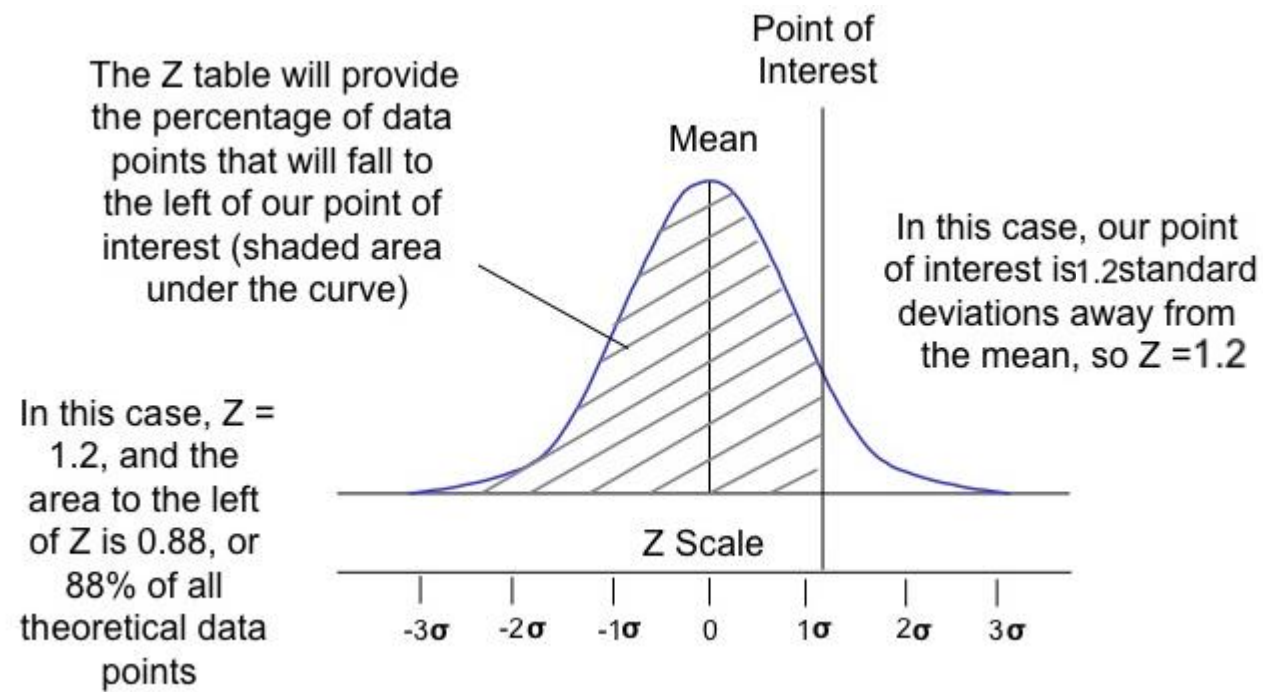


Statistical Proof – cont'd

Z-statistic

$$Z = \frac{\bar{X} - \mu}{\sigma}$$

- provides a measure of the likelihood that a data point belongs to a given population





The Null Hypothesis

Example:

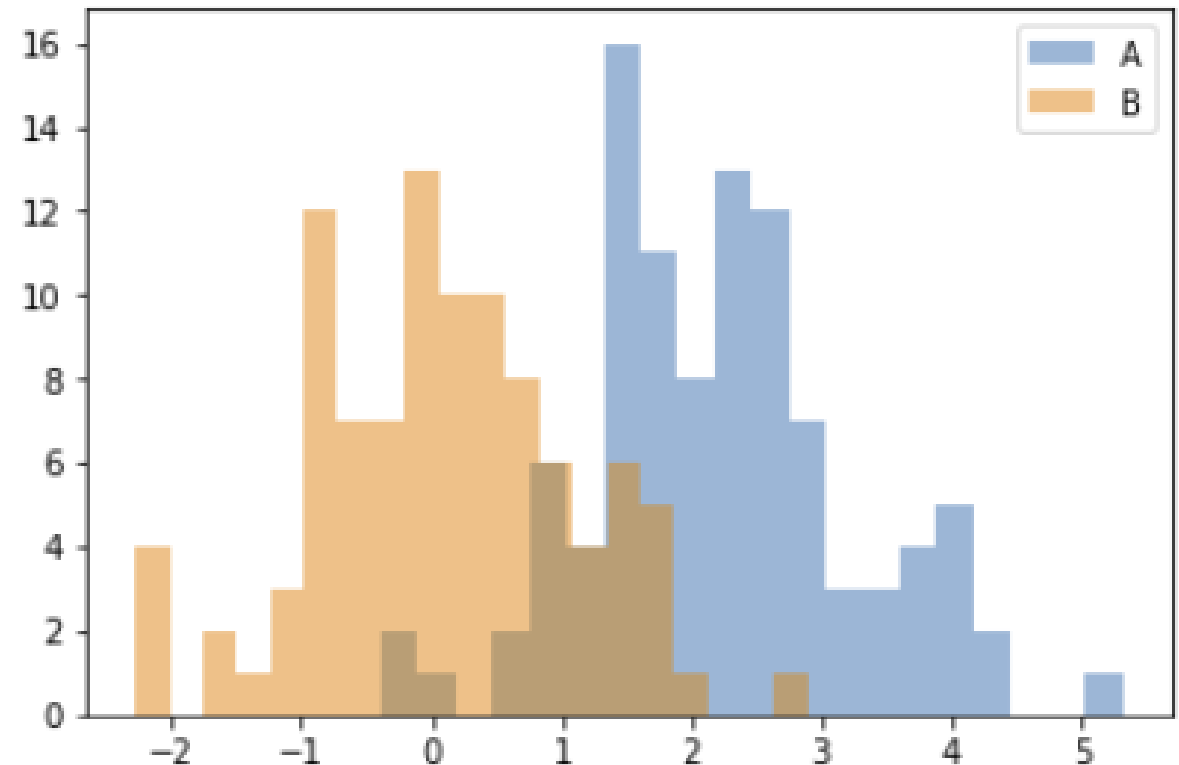
- dataset comprised of patients' responses to two different therapies:
 - drug A (the old drug, or 'control' treatment)
 - drug B (the new drug, or 'test' treatment).
- we are interested in testing the **alternative hypothesis H_a** :
 - A & B deliver significantly different outcomes
- but we do this by assuming (and then trying to reject) the **null hypothesis H_0** :
 - there is **no** significant difference between A & B
 - the distributions we get from the 'A' data and the 'B' data represent two sample sets from the same 'population'



Testing the Null Hypothesis for Two Samples

Given two samples, A and B

- compute the means X_A, X_B
- compute the variances σ^2_A, σ^2_B
- calculate how close X_A is to X_B given the uncertainty implied by their variances
- calculate the likelihood that this value of our closeness parameter could be obtained at random





The Student's t -Test

The t -statistic for comparing two samples is:

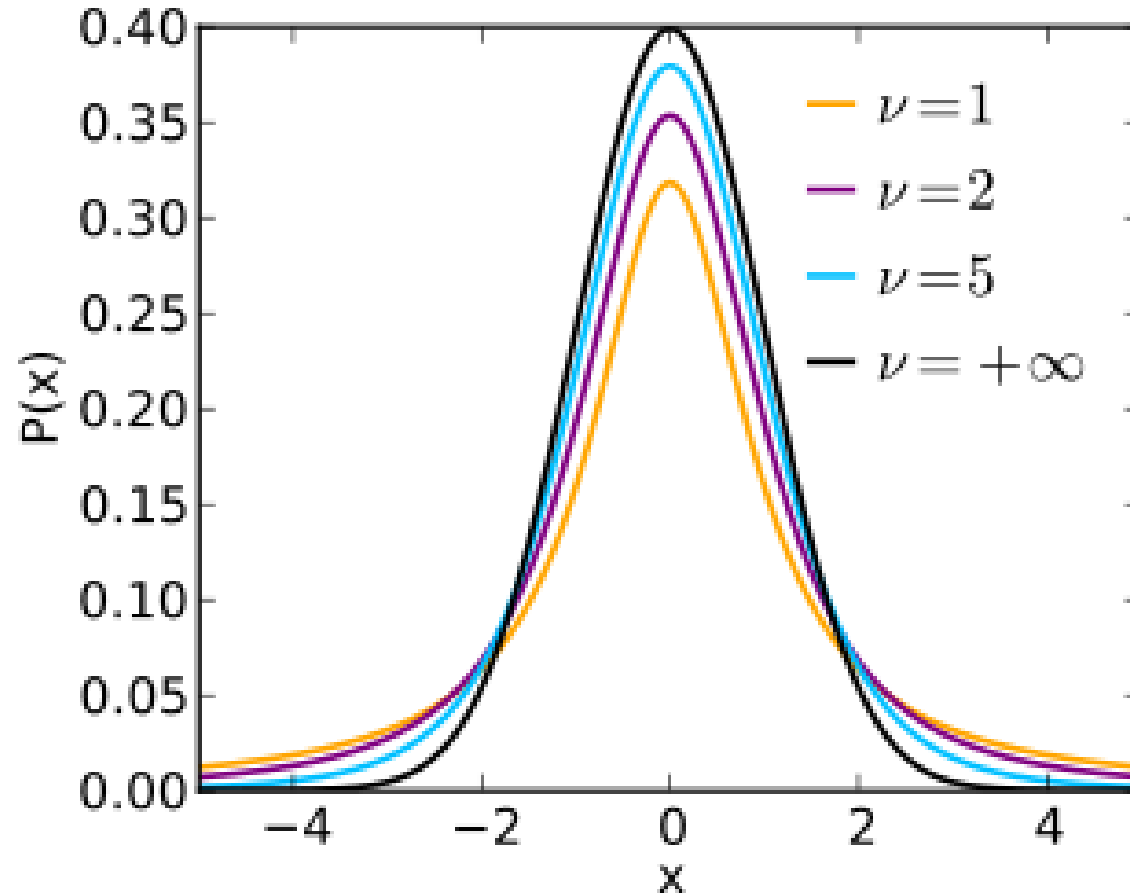
$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{1,2} \sqrt{2/N}}$$

where the *mutual* or *joint* standard deviation is given by:

$$s_{1,2} = \sqrt{\frac{\text{var}(X_1) + \text{var}(X_2)}{2}}$$



The t -Distribution



- ν is the number of degrees of freedom
- the distribution narrows (approaches normal distribution) as ν gets larger



Statistical Errors

Type I errors

- false positives (FP)
- we erroneously rejected the null hypothesis

Type II errors

- false negatives (FN)
- we erroneously upheld the null hypothesis

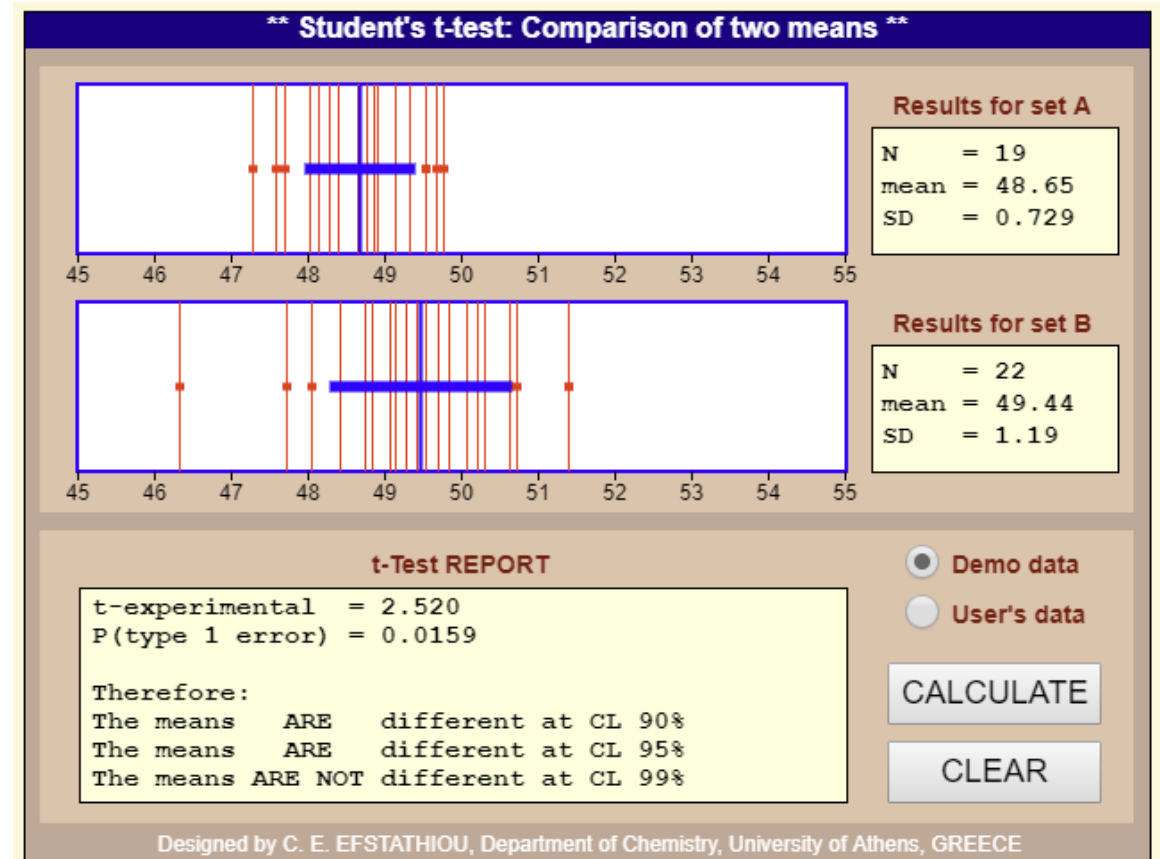
predicted positives $PP = TP + FP$ predicted negatives $PN = TN + FN$

actual positives $P = TP + FN$ actual negatives $N = TN + FP$



Lab 2.2.2: Statistical Proof

- Purpose:
 - To learn how to use the Student's *t*-test for comparing two samples
- Materials:
 - 'Lab 2.2.2.ipynb'
- Reference:
 - <http://195.134.76.37/applets/AppletTtest/AppletTtest2.html>





Discussion

- Is it sufficient to declare statistical significance at $p < 0.05$?
 - how much confidence is enough?
- Is it okay to mine for significance by testing each variable in turn?
 - how would we control the error estimate in multivariate testing?
- Resources:
 - Statistical Thinking for Managerial Decisions
<https://home.ubalt.edu/ntsbarsh/Business-stat/opre504.htm>
 - Statistics: The Art & Science of Learning from Data
<http://www.artofstat.com/webapps.html>



ANOVA

Analysis of variance

- generalises t -test to >2 samples (groups)
 - more conservative
 - reduces Type I errors
- decomposes data additively
 - compares mean squares, F -statistic
 - can test a nested sequence of models
- comprises a suite of methods
 - one-way, two-way, multiple



ANOVA – cont'd

One-way ANOVA

- F -statistic:

$$F = \frac{(\text{variance between groups})}{(\text{variance within groups})} = \frac{SS_T / (I - 1)}{SS_E / (n_T - I)}$$

I = number of groups

n_T = number of subjects

- compare this statistic to F -distribution for $I - 1, n_T - I$ degrees of freedom
- reject H_0 for $F \geq F_{\text{critical}}$

<https://www.marsja.se/four-ways-to-conduct-one-way-anovas-using-python/>



Controlled Trials

objectives:

- to evaluate an experimental cohort (*test group*) against a baseline (*control group*)
- to measure every factor that has the potential to influence the response variable



challenges / considerations:

- the control group must be representative of the test group in every way except for the influence of the effect that is under test
- if we have limited understanding of the phenomenon, we may neglect important variables
 - *this will lead to experimental bias*
- others?



Randomised Controlled Trials

objective:

- to minimise experimental bias by evenly distributing uncontrolled variables between the study cohorts

challenges / considerations:

- different classes of subjects should be evenly distributed between cohorts
 - e.g. age range, weight range, sex, medical status
 - requires data profiling of subjects prior to commencing experiment
- others?



Blind Randomised Controlled Trials

blind

- subjects do not know if they have been allocated to the test group or the control group



double blind

- experimenters do not know which individuals are test subjects or control subjects
- *only the analysts know!*



A/B Testing

def: a randomised experiment with two variants

examples

- evaluate / compare options for improving performance
 - marketing campaigns
 - website engagement
 - product variants
- conversion rate
 - proportion of sales resulting from all visits
- funnel
 - stages from visit through to conversion



Experimental Design for Big Data

- processing time (cost)
 - sample small subsets of the data
 - design the experiment, validate analytic methods before progressing to full dataset
 - for time-dependent data, need to sample many epochs so that periodicity is captured
- the curse of high-dimensionality
 - special methods required when number of features $\sim 10^3$
 - $O(n^2)$ algorithms too slow
 - exploit sparseness where possible
 - large number of features \rightarrow many spurious correlations
- *other issues?*



Causation

- Causation vs correlation
- Domain knowledge



Causation vs Correlation

example:

- a study finds that homicide correlates with ice cream consumption
 - what does this mean?

Headline #1: *'Ice Cream Linked to Murder'*

- scientists are desperately trying to discover which brands or flavours of ice cream are driving the murder rate

Headline #2: *'Heat Wave Pushes Murder Rate Up'*

- scientists suspect elevated brain temperatures increase mental instability
- meanwhile, ice cream sales are soaring



Causation vs Correlation – cont'd

A few cups of coffee may lower colon cancer risk

Posted: 01 August 2007 17:08 hrs

TOKYO : Drinking a few cups of coffee a day may lower the risk of advanced colon cancer, at least for women, Japanese researchers said Wednesday.

The study, supported by Japan's health ministry, showed women who drink more than three cups of coffee a day were 56 percent less likely to develop advanced colon cancer than those who drink no coffee at all.

"Drinking coffee sustains the secretion of bile acid and keeps down cholesterol levels, the mechanisms thought to prevent colon cancer," the report said.

But unfortunately the effect was not seen in men, the medical research team said.

Many men smoke and drink alcohol more than women, and those habits probably offset the effect of coffee, the study said.

The research team tracked down about 96,000 people in Japan aged from 40 to 69 between the early 1990s and 2002, of whom 726 men suffered colon cancer.



Photos

The screenshot shows the CancerConsultants.com website. The header includes the site name and 'oncology resource center'. A navigation bar has links for Patient Home, Professional Home, Newsletters, and Feedback Survey. A search bar is present with a search button. On the left, there is a sidebar with a 'Quick Links' section containing a dropdown menu for 'Information by Disease' set to 'All', and a 'Cancer News' link. The main content area features a banner for 'Cancer News: Rectal Cancer: Article' with a 'Printable Version' link. Below this is the section 'Rectal Cancer News' with the article title 'Coffee Does Not Decrease Risk of Colorectal Cancer'. The article text states that researchers from the Harvard School of Public Health reported that coffee consumption does not appear to reduce the risk of colorectal cancer, contrary to previous studies. It mentions a review of 12 studies with 646,846 participants and 5,403 cases of colorectal cancer, finding no significant effect of coffee consumption on colorectal cancer risk.



Causation vs Correlation – cont'd

Simpson's paradox

- a trend appears in different groups of data but disappears or reverses when these groups are combined
 - common in social-science and medical-science statistics
 - https://en.wikipedia.org/wiki/Simpson%27s_paradox
- caused by experimental bias
- results in H_0 rejected despite insufficient statistical power
 - difference in means is too small
 - variances are too large
 - number of samples is too small



Can't we just use 'common sense'?

**Common sense is the collection of
prejudices acquired by age eighteen.**

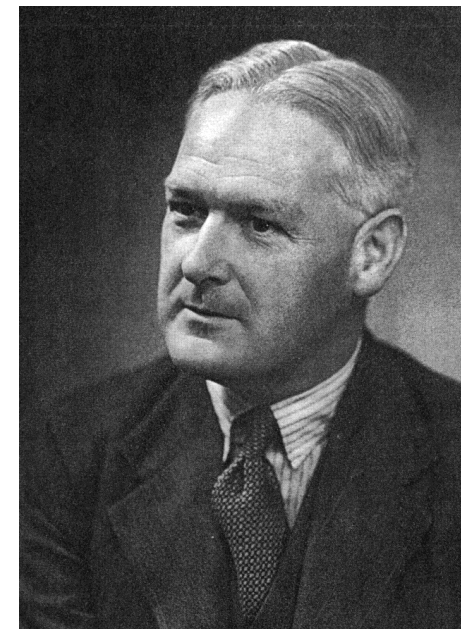
Albert Einstein



Criteria for Evaluating Causation

- Strength of association
- Consistency
- Specificity
- Temporality
- Biological gradient
- Plausibility
- Coherence
- Experiment
- Analogy

> **subject matter expertise + statistics + reasoning**



Bradford Hill



Lab 2.2.3: Statistical Inference

- Purpose:
 - To consolidate the basic concepts of sampling and distributions.
- Materials:
 - 'Lab 2.2.3.ipynb'



Appendix



Statistical Power

def: the probability that the test correctly rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_1) is true

example

- let A, B be the control & test cohorts:

$$D(N) = \frac{1}{N} \sum_{i=1}^N B_i - A_i$$

- define test statistic:

$$T(N) = \frac{D(N) - \mu_D}{\sigma_D / N}, \quad \mu_D = 0 \quad (H_0)$$



Statistical Power – cont'd

- specify $p < 0.05$ for significance
- from the t -distribution, $p = 0.05$ corresponds to $t = 1.64$
- therefore, to reject H_0 we require:

$$T(N) > 1.64$$

- specify power > 0.9 to detect $\mu_D > 1$
- after a few more steps, we obtain this requirement:

$$N > 8.56 \sigma_D$$



Statistical Power – cont'd

Errors and Power in Significance Testing

Select null hypothesis value p_0 :

0.05

Type of alternative hypothesis:

☒ greater ☐ not equal ☐ less

Show:

☒ Type I error ☒ Type II error ☒ Power

True value of p :

0.05 0.15 0.95

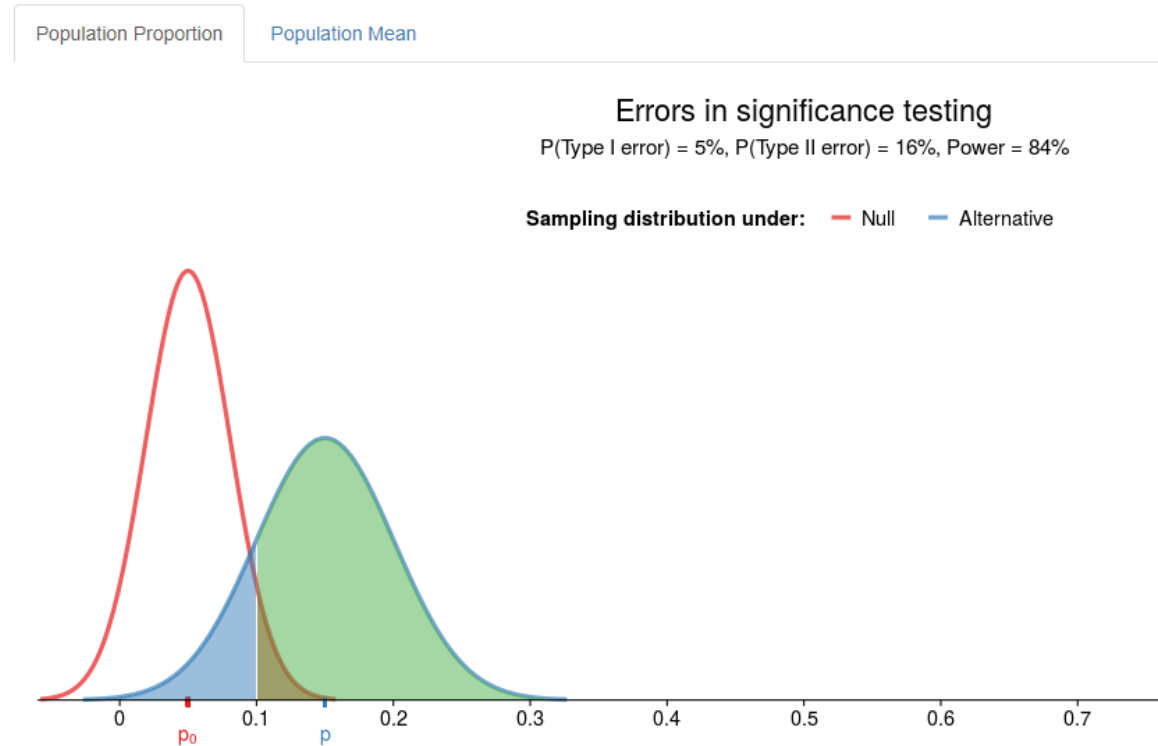
Sample size n :

30 50 200

Type I Error α :

0 0.05 0.15

[Download Graph](#)



<https://istats.shinyapps.io/power/>