UNIVERSITY OF
BIRMINGHAM

School of Computer Science

EVALUATION METHODS AND STATISTICS

Prof. Chris Baber & Prof. Andrew Howes

Chair of Pervasive and Ubiquitous Computing

Aims of this Module

- This module provides an introduction to the use of empirical, scientific methods, including experimental design and statistics.
- This module is targeted at computer scientists with an interest in:
  - Developing systems that support human activity (Human-Computer Interaction)
  - Building computational models of human behaviour
  - Understanding human behaviour as an inspiration for Robotics, Machine Learning and Artificial Intelligence
  - Designing and analysing experiments to evaluate system performance

Outcomes of this Module

- On successful completion of this module, you will be expected to be able to:
  - identify and apply research methodologies for investigating human behaviour;
  - recognise the appropriateness of statistical techniques in data analysis;
  - conduct and report statistical tests;
  - interpret and critique research findings that are supported by statistical tests;
  - demonstrate understanding of experimental design, including sampling, participant selection, task design and research ethics.

Useful resources

- https://rcompanion.org/handbook/index.html
  - This is a well-written web handbook that explains statistics with lots of examples in R
- http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf
  - A very nice primer by Howard Stelman on Experimental Design and Analysis,
- Howell, D.C., 2002, *Statistical Methods for Psychology,* Pacific Grove, CA: Duxbury [5th edition]
  - This is a standard textbook on statistics

UNIVERSITY OF
BIRMINGHAM

Module Assessment and Student Effort

- Class tests:
  - There will be 2 class tests, accessed through Canvas (week 5 and week 8)
  - These are not meant to be difficult – the aim is to encourage reflection on the lecture notes, opportunity to think about design of experiments, and to practice Hierarchical Task Analysis as a method
  - The class tests are open book and you are expected to consult the lecture notes
- Examination
  - This is the primary assessment mode for this module
  - There will be a 2 hour exam
  - The exam is **closed book**
- Student Effort
  - This is a 10 credit module. That should equate to 100 hours of student effort.
    - c.20 hours will be in lectures – 2 hours per week
    - c.30 hours will be in lab classes and practical exercises – 3 hours per week
    - c.2 hours will be for the class tests
    - c.48 hours for background reading, revision for class tests and exam

UNIVERSITY OF
BIRMINGHAM

What do we mean by human behaviour?

- □ Perceptual-motor control
  - – Selecting an object on a visual display, controlling a prosthetic limb, flying a Drone, driving a car…
- □ Cognition
  - – Writing a program, navigating the world wide web, choosing what to eat in a restaurant, thinking and writing an essay, reading a textbook…
- □ Social, economic and collaborative activity
  - – Making and keeping friends, working in groups, Air Traffic Control, managing a business…
- □ Cultural, aesthetic, leisure activity
  - – Enjoying a concert or a visit to an art gallery, playing sports, exercise…

Q. How can we ‘measure’ these different types of behaviour?

Q. How can we demonstrate differences between examples of these types?

UNIVERSITY OF
BIRMINGHAM

Lecture 1: Introducing concepts

- □ In the first lecture, we won’t look at mathematical formulae but will consider…
  - what we mean by ‘evaluation’
  - how evlauation uses different scales of measurement
  - what is meant by data and by a normal distribution of data
  - how science makes arguments

UNIVERSITY OF
BIRMINGHAM

What do we mean by ‘evaluation’?

- Evaluation involves making a claim
  - that X is ‘fit for purpose’
  - that X is better than Y
- A claim is supported by evidence
  - that some aspect of X can be measured
- A claim can be proven or disproven
  - or, at least, the claim can be tested
  - that the measurement can be compared with other measurements

What do we mean making a claim?

- Science is not about taking for granted what scientists say
- Science is about doubting what science says
  - Is the claim based on ‘good’ measurement?
  - Is the claim based on ‘good’ analysis?
  - Are the data sufficient to support generalisation of the claim?
  - Does the claim rest on a ‘good’ theory?
  - Do the data and the theory match each other?
- Science is about argument

How can we test a claim?

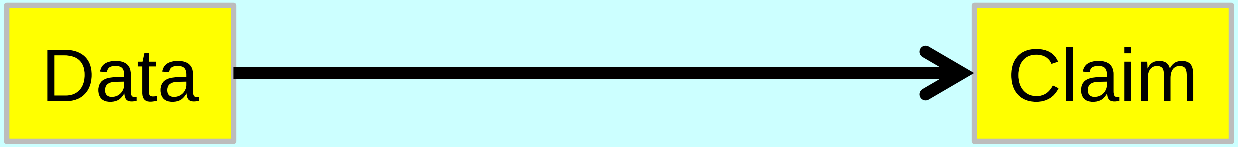
“Ready, text, go: typing speeds on mobiles rival keyboard users”

https://www.theguardian.com/science/2019/oct/02/ready-text-go-typing-speeds-mobiles-rival-keyboard-users

"People who tapped out messages with a single finger managed on average only 29 words per minute (wpm), but those who mastered the two-thumb technique hit a blistering 38wpm, only 25% slower than an average typer on a full-sized Qwerty keyboard. One volunteer thumbed out sentences on their mobile phone at a blur-inducing 85wpm, far exceeding the 52 wpm that people typically reach on a standard keyboard."

How do People Type on Mobile Devices? Observations from a Study with 37,000 Volunteers – Kseniia Palin , Anna Maria Feit , Sunjun Kim , Per Ola Kristensson , Antti Oulasvirta, *MobileHCI 2019*

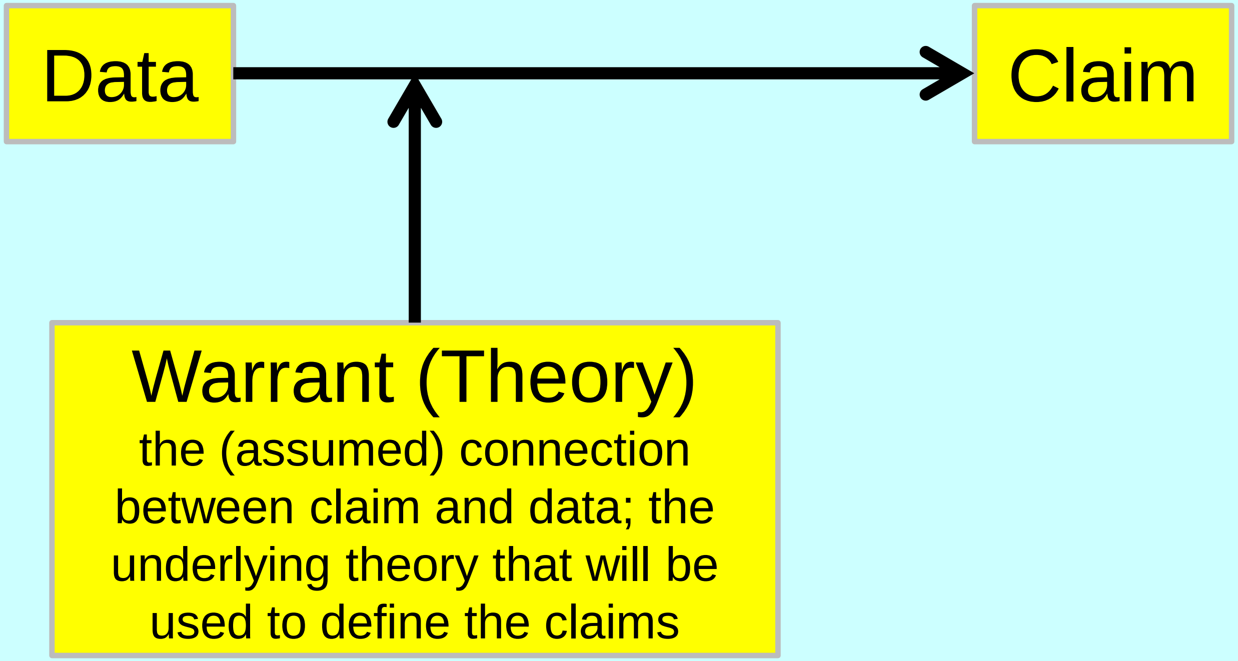
Toulmin’s (1958) model of Argument



EXERCISE#1a

- □ Working in groups of 3 or 4..
  - – what are claims that the newspaper article is making?
  - – do you believe the claim that texting is as fast as' typing?

Toulmin's model of Argument



Data

Claim

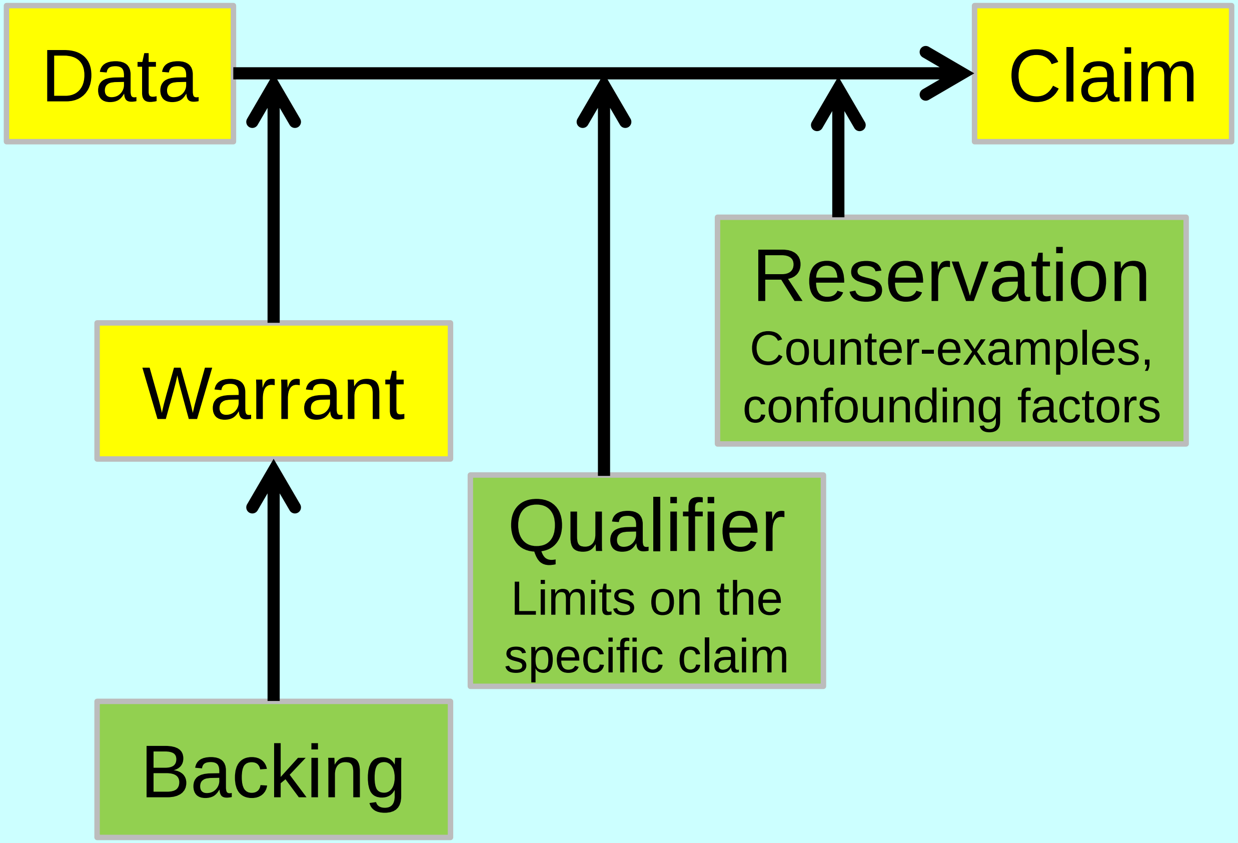
Warrant (Theory)

the (assumed) connection between claim and data; the underlying theory that will be used to define the claims

Exercise #1b

- □ What sort of theories do you believe would be needed to support the claims?
- □ What type of evidence would be presented to support these theories?
- □ How would you measure the evidence (i.e., how could this be defined as ‘data’)?

Toulmin’s model of Argument



Data
Claim
Warrant
Backing
Qualifier
Limits on the specific claim
Reservation
Counter-examples, confounding factors

UNIVERSITY OF
BIRMINGHAM

Exercise#1c

- What qualifiers (limits) might be placed on the claims?
- What arguments could be made against the claims?
- What evidence would be used to support the counter-arguments?
- How would you know which argument is correct?

What do we mean by measurement?

- Scales of Measurement
  - Nominal: data used to distinguish between categories, e.g., male = 1, female = 2.
  - Ordinal: data in rank (or other) order, e.g., Likert-type scales
  - Interval: data as quantity with equal, positive or negative units, where zero is simply another point on the scale
  - Ratio: an interval scale with an absolute zero.

Comparing Data from different Scales of Measurement

- Parametric
  - Interval and Ratio
  - Parametric data can be assumed to follow a **normal distribution**
  - Gaps between values are relative and meaningful.
- Non-parametric
  - Ordinal and nominal
  - Data do not follow normal distribution.
  - Data cannot be assumed to have equal magnitude intervals between data points

How do we know that data are valid?

- CONTENT VALIDITY
  - does the measurement describe what is being evaluated?
    - From ‘typing vs texting’ study, we could ask whether we believe that the activities are comparable, i.e., does it make sense compare typing with one finger vs two thumbs vs ten fingers?
- CONSTRUCT VALIDITY
  - do the measurements relate to the concept?
  - do the measurements permit generalization?
  - do the measurements perform similarly in related situations, i.e., do they have convergent validity?
- CRITERION-RELATED VALIDITY
  - do the measurements predict the outcome?
  - or, rather, can we show that the measures correlate with the outcomes in expected ways?

Reliability

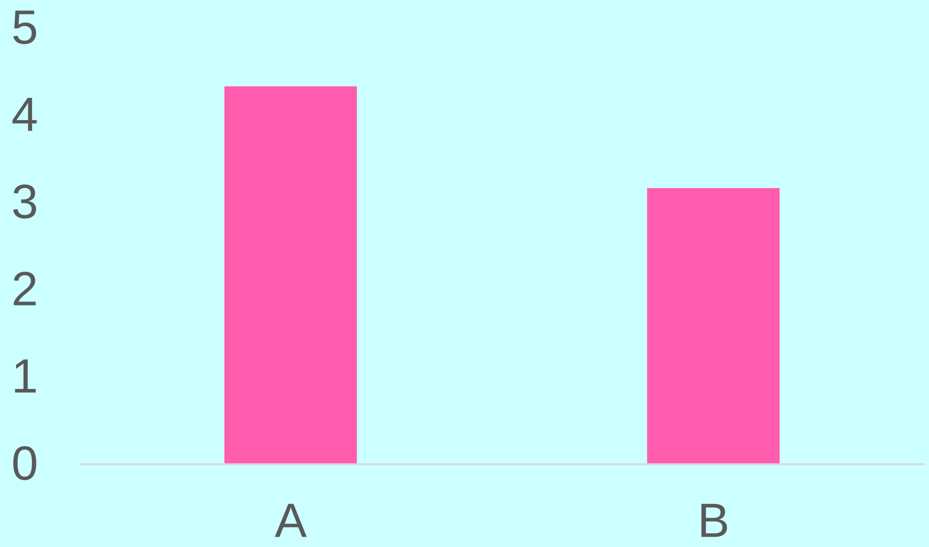
- RELIABILITY measures consistency of a measurement
  - Test-Retest
    - Take measures from same population on two occasions and correlate
  - Split-half
    - Take half the measures from same population and correlate
  - Alternative Form
    - Take measures from same population using two instruments and correlate
  - Inter-rater
    - Apply measures to same population by two experimenters
- Note, a measurement can be Reliable (i.e., consistent) but **not** Valid

UNIVERSITY OF BIRMINGHAM

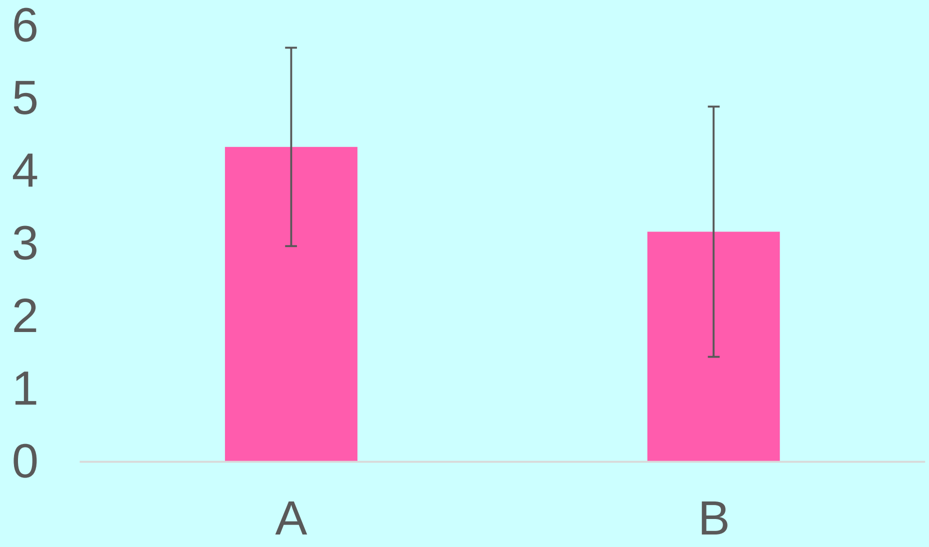
Data and Analysis

- All of the statistical tests we will consider in this module rely on assumptions:
  - we assume that we do not have access to the entire population of data, so we work with samples
    - we do not have access to N
  - we assume that the samples reflect the population
    - we do have access to n (as a sample of N)
  - we assume that the data can be described in terms of their distribution

Is there a difference?

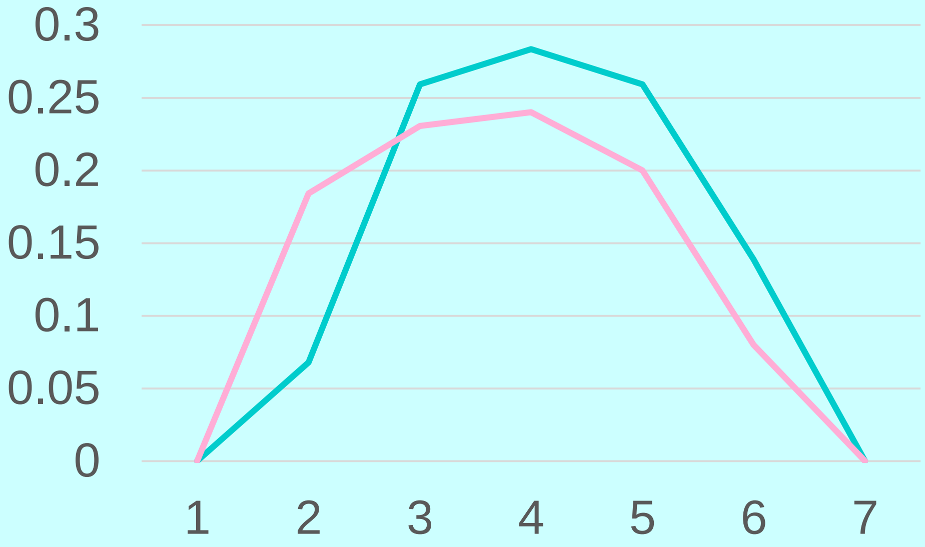


Is there a difference?



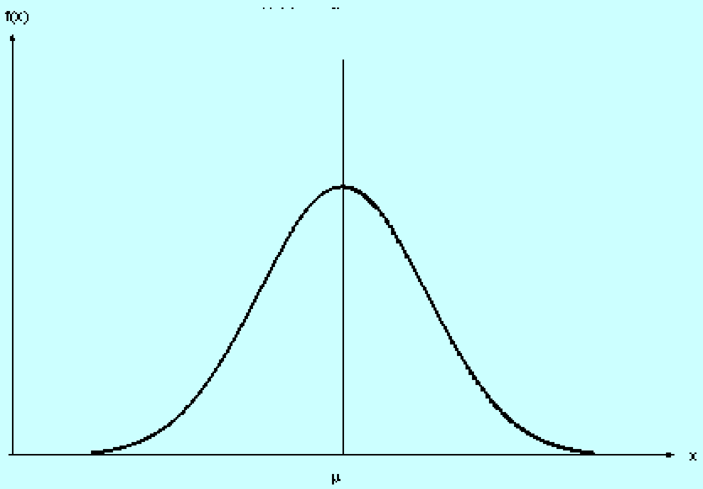
UNIVERSITY OF BIRMINGHAM

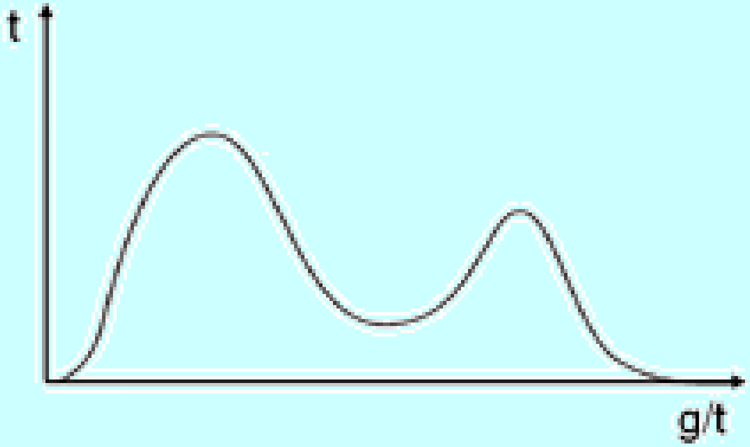
Is there a difference?

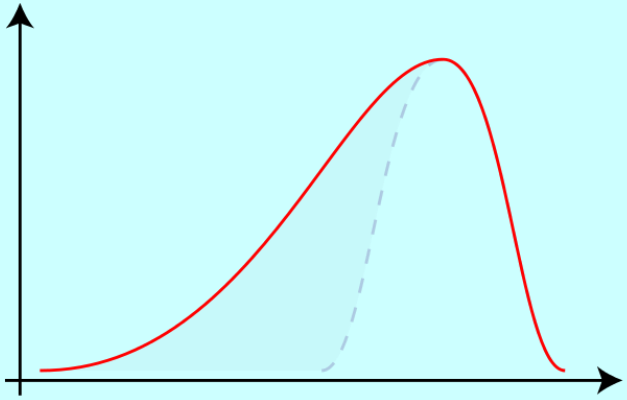


UNIVERSITY OF BIRMINGHAM

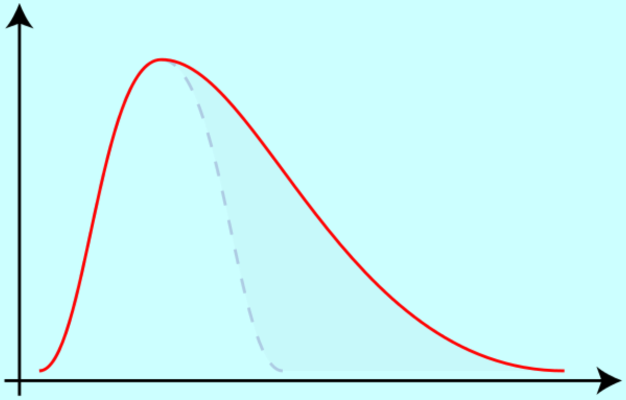
Distributions







Negative Skew



Positive Skew

UNIVERSITY OF
BIRMINGHAM

Measures of Central Tendency

Data	Mean	Median	Mode
{1, 1, 1, 2, 2, 3, 4, 5, 5, 6}	3	2.5	1
{1, 27, 28, 29, 30}	23	28	
{1, 1, 1, 1, 1, 1, 2, 2, 3, 3, 3, 3, 50}	5.5	2	1

- Mode (most common value) is actually in the data
- Median (middle value) less affected by extreme scores (so more appropriate if data are skewed, i.e., if data are non-parametric)
- Mean (average value) most commonly used (assumes data are parametric, i.e., follow a normal distribution)

Frequentist vs. Bayesian statistics

- This lecture course follows a frequentist approach
  - probabilities describe events with long-term frequencies
  - one can not relate probabilities to hypotheses
  - Data are sampled on the assumption of maximum likelihood estimate
  - MLE is tested by null hypothesis testing and assigning confidence intervals

- In contrast, Bayesian approaches use probabilities to reason about uncertainty of events or hypotheses
- Data collection updates a prior Bayesian probabilites distrtibution

UNIVERSITY OF
BIRMINGHAM

Law of Large Numbers

- □ Large random samples are representative of the population.
- □ So, if we take enough samples, the sample mean, *m*, will approach the population mean, μ.
- □ As sample size increases, the Gaussian describing the sample mean is more closely clustered around the population mean.

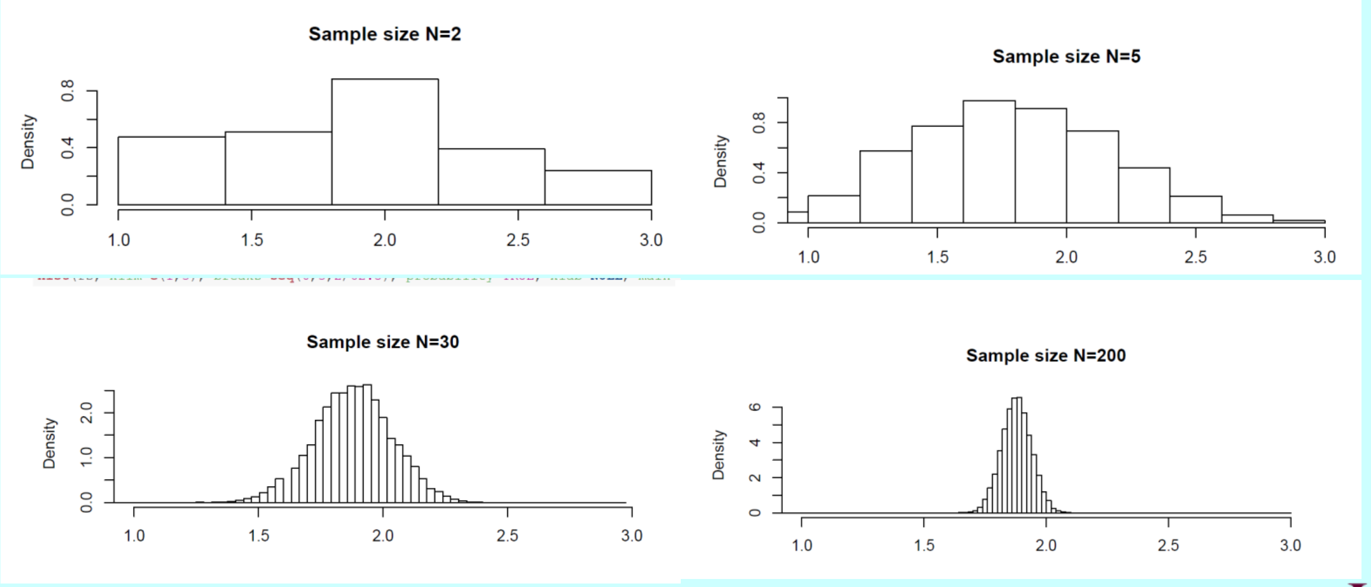
UNIVERSITY OF BIRMINGHAM

Central Limit Theorem

- The sample distribution will be normal if:
  - the population that is sampled is normal

  or

  - if the sample size is large enough



UNIVERSITY OF
BIRMINGHAM

Dispersion

- Range: between smallest and largest value
- Quartiles: 25%, 50%, 75%
- Variance: spread of values from their average (calculated as the averaged, squared difference between each data point and mean)
- Standard deviation: the amount of variation in a set of data (calculated as the square root of variance)

The Empirical Rule (68: 95: 99.7).

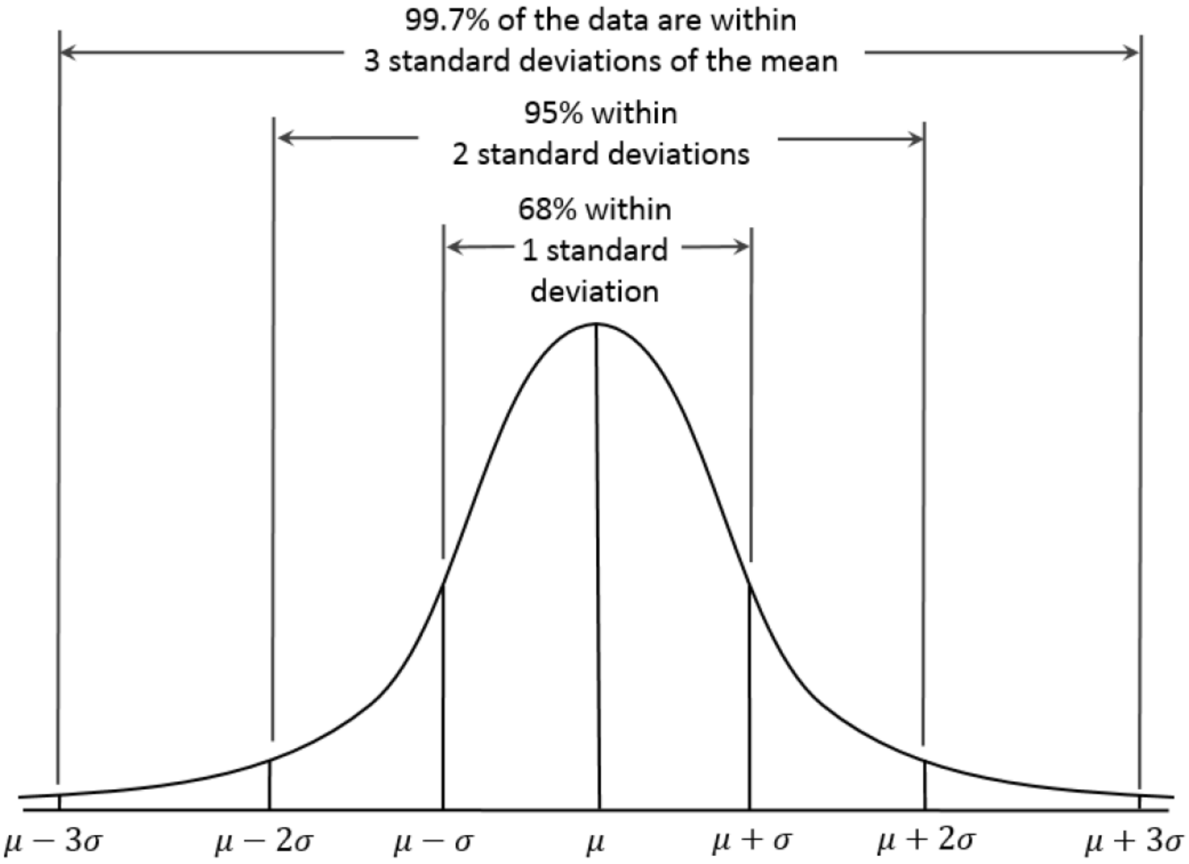


Figure 1: The general Gaussian distribution

UNIVERSITY OF
BIRMINGHAM

Law of Large Numbers (again)

- Given a distribution with mean μ and variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean (μ) and a variance σ^2/n as sample size, n, increases.

Central Limit Theorem (again)

- Also known as the Law of Small Numbers because the approximation (that the mean of randomly chosen samples will follow a normal distribution) works even for small sample sizes.
- If know the distribution of sample means, we can say how confident we are that the true mean is within a given interval of the sample mean.
- We can express the closeness of a statistic to the mean as a standard deviation (error from the mean).
- We can express a confidence interval in terms of the number of standard deviations from the mean we can go to maintain the confidence level.

Data Collection Exercise

- I have designed a ‘video game’
- Each of you will download and play several versions of the game
- Performance will be measured in terms of response time and signal detection measures

Downloading and Playing the Drones Games

UNIVERSITY OF BIRMINGHAM

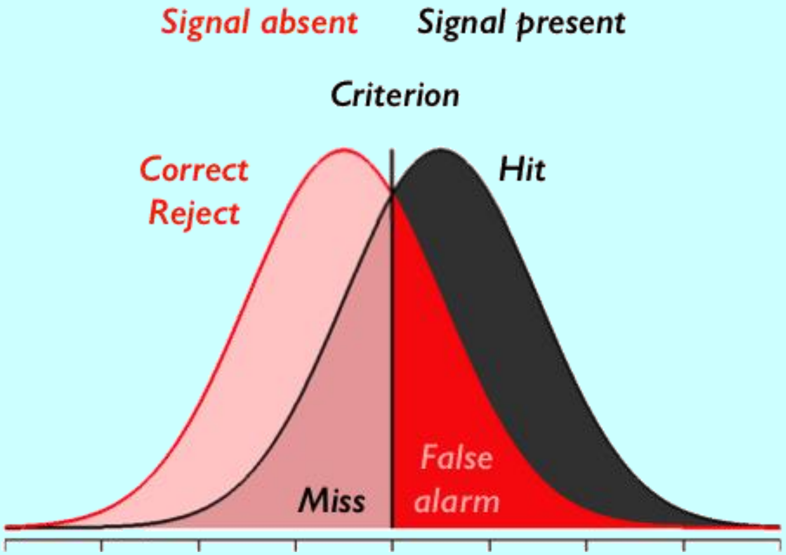
Signal Detection Theory [1]

- □ Tasks can involve spotting a ‘signal’ against a background of ‘noise’

		STATE OF THE WORLD	
		SIGNAL	NOISE
RESPONSE	YES	Hit	False alarm
	NO	Miss	Correct Rejection

Signal Detection Theory [2]

- We can assume that signals and noise occur with defined (normal) distributions.



Evidence strength, i.e., how strong the signal needs to be for the operator to say ‘yes’

Source: http://gru.stanford.edu/doku.php/tutorials/sdt

Response Criterion, β

- Response criterion is the ratio of evidence, given a signal, to evidence, given noise:

$$\beta = \frac{p(X|S)}{p(X|N)} = \frac{p(H)}{p(FA)}$$

- When the probability of signal or noise is equal, the distributions will overlap sufficiently for β to be at their intercept (as in the previous slide)

UNIVERSITY OF BIRMINGHAM

Response Criterion, β

- When the probability of signal is expected to be higher (because of additional information) then the operator shifts the response criteria to accept a signal with lower evidence strength (i.e., shift from ‘conservative’ to ‘risky’ response)

- This suggests an optimal setting of criterion:

$$\beta_{opt} = \frac{p(N)}{p(S)}$$

UNIVERSITY OF BIRMINGHAM

Response Criterion, C

$$C = -\frac{[z(H) + z(FA)]}{2}$$

When C = O, the observer is ‘unbiased’ (equally likely to respond to signal or noise); when C>0 the criterion moves to the left and the observer uses a more strict criterion to accept a signal; when C <0 the criterion moves to the right and the observer accepts more instances as signals

Tends to be more commonly used in contemporary reports

How well can people set Response Criteria?

- Given acceptable signal : noise ratio, people are able to adjust response criteria…
- …and to adjust response criteria taking into account changes in costs and values.
- BUT changes can be ‘sluggish’
  - Operators might introduce additional ‘yes’ responses
  - Operators might not be able to perceive probability distribution of data (over estimating rate events)
  - Operators might be happier lowering response criteria as uncertainty increases

UNIVERSITY OF BIRMINGHAM

Sensitivity

- Sensitivity relates the separation to the spread of the response curves
- Response criterion describes the response bias of the operator
- Sensitivity describes the ability of the operator to resolve (or detect) a signal against the noise:

$$d' = z(H) - z\,(FA)$$

- When H or FA is 1 or 0, the z scores are at infinity. This could be addressed through correction to adjust the p-value used to calculate z, where *p = 1 / N* (N is number of trails).
- Or we could use a different measure of sensitivity, such as A':

$$A' = 1 - \frac{1}{4}\left\{\frac{p(FA)}{p(H)} + \frac{[1 - p(H)]}{[1 - p(FA)]}\right\}$$

UNIVERSITY OF BIRMINGHAM

Sensitivity and Specificity
(as used in testing algorithm performance)

□ $\text{Sensitivity} = \dfrac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$

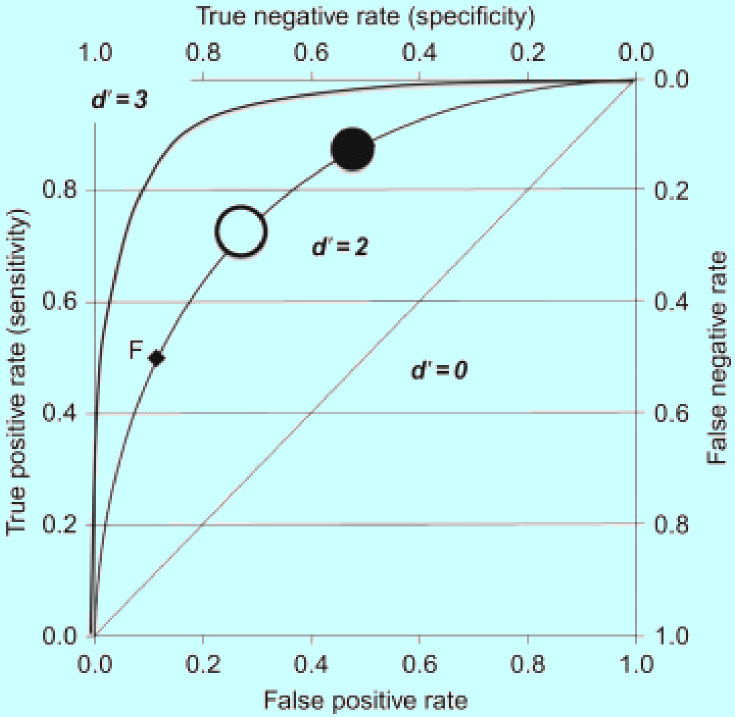
How well can you detect positive cases

□ $\text{Specificity} = \dfrac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$

How often do you mistake a negative case

Receiver Operating Characteristic

- Detection improves after a threshold
- Perceived cost of missed detection influences response criterion



UNIVERSITY OF BIRMINGHAM

Dorneich et al., 2018: https://www.sciencedirect.com/science/article/pii/B9780081018699000091