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

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 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

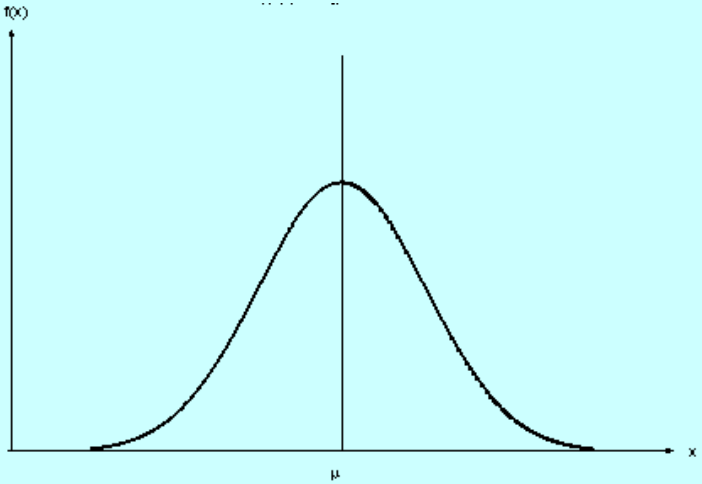
Law of Large Numbers

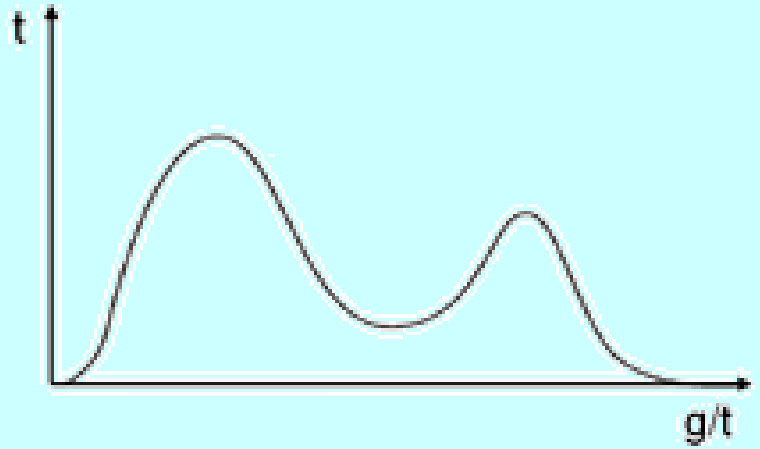
- Large random samples are representative of the population.
- So, if we take enough samples, the sample mean, *m,* will approach the population mean, μ.
- As sample size increases, the Gaussian describing the sample mean is more closely clustered around the population mean.

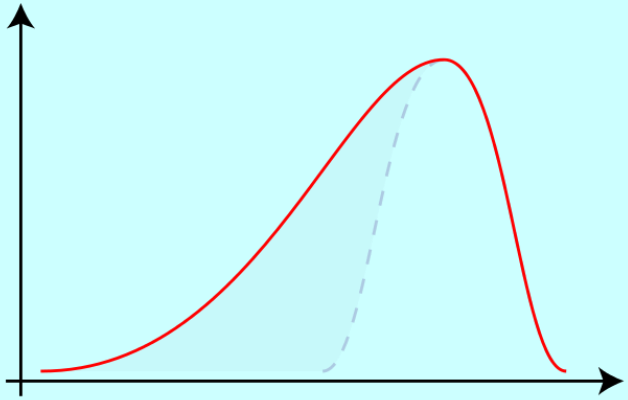
Central Limit Theorem

- 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.

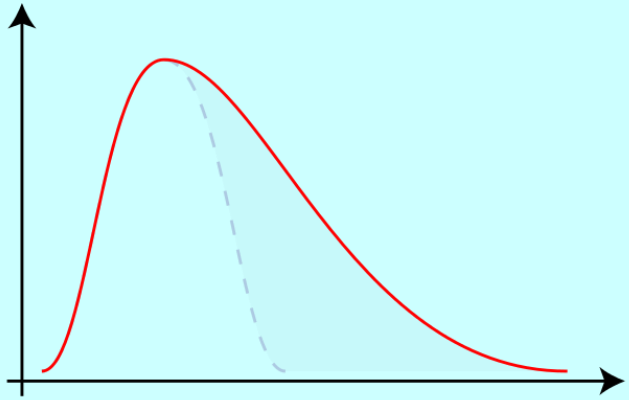
Distributions







Negative Skew



Positive Skew

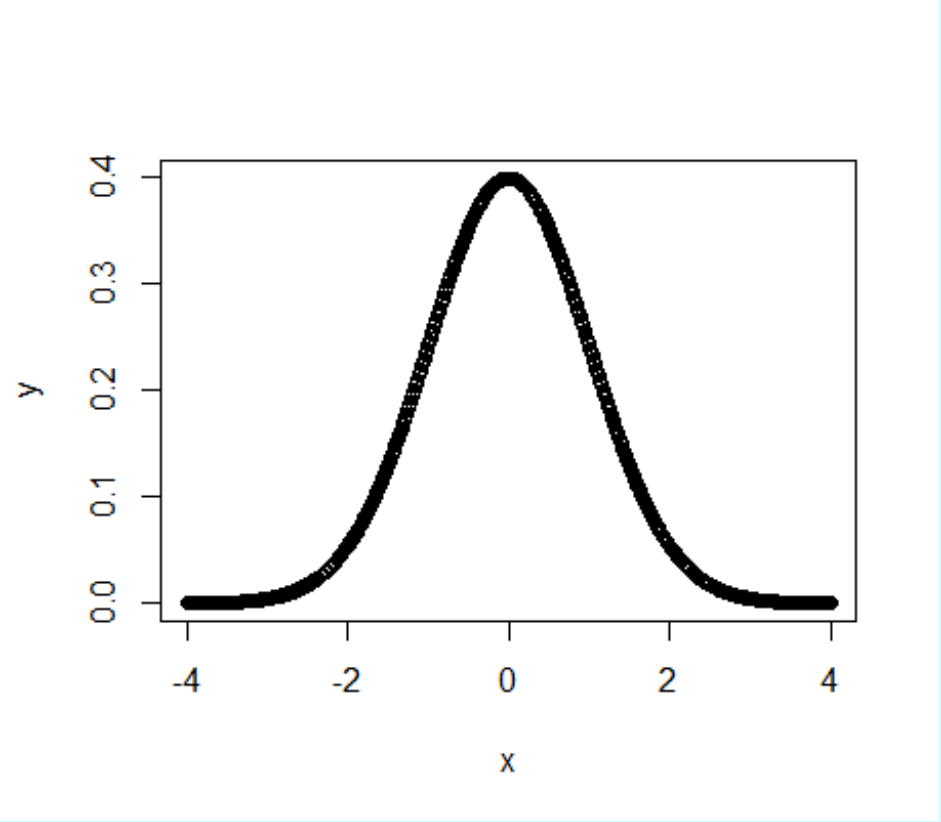
UNIVERSITY OF BIRMINGHAM

The Normal Distribution (again)

- Normal Distributions have the probability density:

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

- Rather than calculate this each time, we assume that all normal distributions can approximate this Standardized Normal Distribution which assume $\mu = 0$, $\sigma = 1$…



UNIVERSITY OF BIRMINGHAM

Standardized Normal Variable

- The z (from 'standardized') variable allows us to determine the probability that a value, x, falls within a normal distribution

$$z = \frac{x - \mu}{\sigma}$$

- Assume a data set with mean 918 and sd 180.17. What is the probability of obtaining a x<750?

First, sketch the distribution and mark where you think x lies…

Second, z = (750-918)/180.17 = -0.93

So…x is around 1 sd below the mean (we can check this as 918-180.17 = 737.83)

Z tables

- Rather than calculate z each time, we can use Normal Distribution tables to describe the 'standardized normal distribution' of $\mu = 0$, $\sigma = 1$
- To use this z-table when confidence interval, α, is 95%
  - Confidence level of 95% excludes 5% of distribution
  - Assuming a two-tailed test means 2.5% for each tail of the distribution
  - So, area under half of distribution is 50%-2.5% = 47.5%
  - In the z-table, 0.475 is z = 1.9 plus 0.06 = 1.96

The normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4

Confidence Level	α	$\alpha/2$	Z
85%	15%	7.5%	1.435*
90%	10%	5%	1.645*
95%	5%	2.5%	1.96

UNIVERSITY OF BIRMINGHAM

Cumulative Z table (same procedure)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Assume 95% confidence interval.
This is 5% off each side of the distribution, so 2.5% of one side. This is 97.5.
We look up .975 on the z-table.
This is a z-score of 1.96

(for 90% confidence, this would be .95 and a z-score of 1.64)

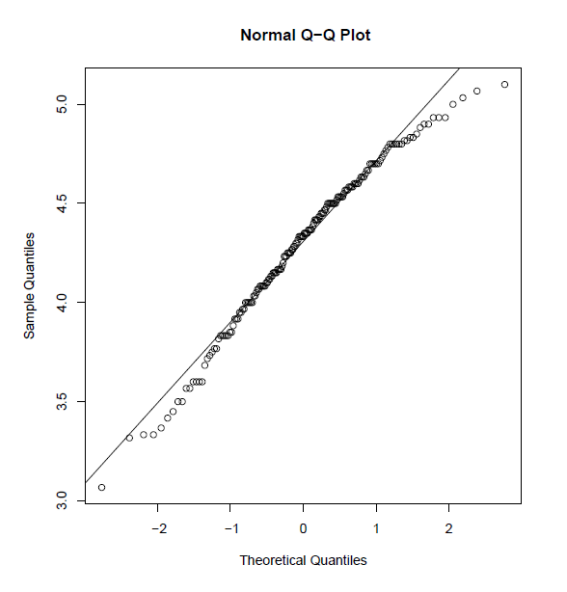
UNIVERSITY OF BIRMINGHAM

Testing for Normality

- We can plot the data in a histogram and visually check if this looks like a ‘bell curve’
- We can convert the data to a Q-Q plot
- We can apply a Shapiro-Wilk test

Q-Q plot

- Quantile-Quantile plot: sorts all data from experiment and plots one sample against the other. R as qqplot() to do this
- More usefully, you want to check thet relationship between samples follows a normal distribution. So, would apply the qqnorm() command in R.
- This plots samples to a hypothetical line. The closer the points are to the line, the more likely the data are normally distributed



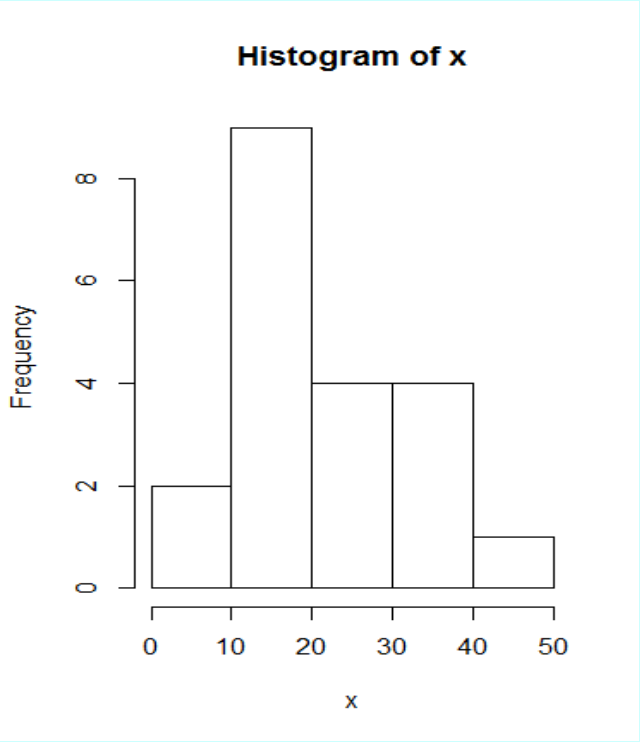
UNIVERSITY OF BIRMINGHAM

Shapiro-Wilk test

- Checks for correlation between samples and assumed normal distribution
- Compares the slope of the observed data against expected values normalised to the sum of square
- Values > 0.5 show normality

Trimming and Winsorization

- Suppose your data has many samples in the tails of the distribution…
- …this could lead to a skewed distribution.
- If you trimmed (i.e., removed samples from) the tails, this could make the distribution normal.
- Trimming
  - Remove a fixed percentage (say, 5%) of samples.
  - E.g., 20 samples, 5% trimming would 1 extreme sample from *each* tail
- Winsorization
  - Trim (as above) and then replace these with most extreme values remaining in each tail



3	**7**	**12**	**15**	**17**	**17**	**18**	**19**	**19**	**19**	**Mean = 22.55**	**Mean.wins. = 22.05**
20	**22**	**24**	**26**	**30**	**32**	**32**	**33**	**36**	**50**		

UNIVERSITY OF BIRMINGHAM

Experimental Design

Hypothesis: Reaction time to congruent words will be faster than reaction time to incongruent words	
Independent Variable: Congruent Words (colour of ink = name of word), Incongruent Words (colour of ink ≠ name of word)	
Control Condition: Congruent Words	**Experimental Condition:** Incongruent Words
Dependent Variable(s): Reaction Time	
Task: participants will be asked to read, as quickly as possible, single words on a display. The words will be the names of colours and will be presented either in the same colour as the word's name or in a different colour	
Confounding Variables: performance could be affected by ability to perceive colour ('colour-blindedness') and knowledge of the names of colour ('language skills')	

UNIVERSITY OF BIRMINGHAM

False Claims and Errors

- **Type I error**
  - We could accept the Alternative hypothesis when it is false (false positive).
  - Many statistics tests are designed to minimise this error.
  - Type I errors define the significance level (α) that the experimenter will accept (conventionally 5%)

False Claims and Errors

- **Type II error**
  - We could accept the Null hypothesis (fail to reject it) when it is false (false negative).
  - The probability of a Type II error is defined as β
  - The probability of correctly rejecting a false null hypothesis is defined as 1- β, which called Power.

Participants (Subjects)

	Source of Unsystematic DV variation	Control
Between Subjects	Individual differences	Match participants on key characteristics
		Random allocation to condition
Within Subjects	Practice / Order effects	Modify order of stimuli
	Boredom / fatigue effects	Design in breaks
	Asymmetric transfer effects	Counterbalance conditions

Sample Size calculation

- $$n = \frac{2(za + z1 - \beta)^2 . \sigma^2}{\Delta^2}$$

za		
α-error	5%	1%
2-sided	1.96	2.5758
1-sided	1.65	2.33

$z1 - \beta$				
Power	80%	85%	90%	95%
	0.8416	1.0363	1.2816	1.6449

Δ = estimated effect size, i.e., difference between conditions
σ = standard deviation (of data set)

UNIVERSITY OF BIRMINGHAM

Types of t-test

- Dependent means / Paired Samples
  - Each participant completes all conditions
- Independent Samples
  - Different participants are allocated to each condition

t-statistic (paired samples)

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

D = difference between trials;
N = number of samples

UNIVERSITY OF BIRMINGHAM

Independent t-test

- Participants are assigned to *only* one trial
- If you have the same number of participants per trial (equal sample sizes):

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\left(\frac{{s_1}^2}{N_1} + \frac{{s_2}^2}{N_2}\right)}}$$

Is this is a ‘good’ result?

- We consider the generalisability of the result by its Effect Size
- Effect size indicates the size of the difference between samples independent of sample size
  - This could relate to proportion of explained variance, e.g., using R^2 or (partial) eta squared
  - Or could relate to difference in averages, e.g., using Cohen’s d

	Small	Medium	Large	Very large
R^2	1%	9%	25%	
Partial eta^2	0.02	0.13	0.26	
d	0.2	0.5	0.8	>1

Here is a nice paper about Effect Size:
https://www.leeds.ac.uk/educol/documents/00002182.htm

Here is an online calculator for Effect size:
https://www.ai-therapy.com/psychology-statistics/effect-size-calculator

UNIVERSITY OF BIRMINGHAM

Paired t-test effect size

- There are several formulae, I prefer Cohen's *d*…
  - For a paired t-test…

$$d = \frac{|m_1 - m_2|}{\sqrt{{s_1}^2 + {s_2}^2 - (2rs_1s_2)}}$$

where r is the correlation coefficient between groups

For this dataset, r = 0.3277 (we'll calculate this in a later lecture)

So, the Effect size for this result is 1.001

(meaning that the difference is larger than 1 standard deviation)

Here is an online calculator for Effect size:
https://www.ai-therapy.com/psychology-statistics/effect-size-calculator

UNIVERSITY OF BIRMINGHAM

Cohen's d for independent samples

$$d = \frac{|m_1 - m_2|}{pooled\ variance}$$

$$pooled\ variance = \sqrt{\left(\frac{({s_1}^2 + {s_2}^2)}{2}\right)}$$

- Why you don't use this for paired samples
  - The effect size assumes a differences between sets of independent scores
  - The paired t-test corrects for correlation, r, into account to correct for pooled variance

Comparing three conditions

	No Communication	Texting	Telephoning
Driving	Control (C)	A	B

- We *could* conduct 3 t-tests { C x A, C x B, A x B}
- This is a *bad idea* because it will inflate the likelihood of Type I error…
  - …because we are accumulating error through repeating the tests
  - i.e., α = 95%, then accumulated t-tests will produce .95 * .95 * .95 = .857
  - Which means that, rather than the p = 0.05 we believe we are testing for, we are actually at 1 – 0.857 = 0.143
- So, rather than multiple t-tests, we employ ANOVA

ANOVA and the F ratio

- ANOVA calculates an F ratio to explain variation in the data
- The F ratio is derived from the Sum of Squares (which we used as a measure of variance for t-tests last week)
- F ratio compares the variation in the data explained by our model (defined by the experimental conditions) against the variation of data not explained by our model…

Calculating Degrees of Freedom

- To reject the null hypothesis (that our data come from the same distribution), we need to compare our calculated value against an expected value for an experiment of this design, using degrees of freedom

  - We have three experimental conditions, so the Between group d.f. = 3 -1 = 2

  - For the Within group d.f., assume we have 21 people in total…

    d.f. = Number of samples – number of conditions = 21 – 3 = 18

  - For the Total d.f., we have 21 – 1 = 20

Can we reject the null hypothesis?

- MS(between) = SS(between) / d.f. (between)
  = 204.7 / 2 = 102.35

- MS (within) = SS (within) / d.f. (within)
  = 486.6 / 18 = 27

- F ratio = MS (between) / MS (within)
  = 102.35 / 27 = 3.79

This is greater than Fcrit, so we can reject the null hypothesis with 95% confidence

Power in ANOVA

- Eta-squared (η^2)
  - Also called Correlation Ratio
  - Assume that correlation between dependent variable and levels of independent variable can be plotted as a scatterplot and a line of best fit applied
  - Assume that, rather than a linear (r) regression we can fit a *k*th-order polynomial regression
  - We can represent this simply as the ratio between calculated Sum of Squares for treatment and total:

$$\eta^2 = \frac{SS_{treatment}}{SS_{total}}$$

  - This calculates the maximum squared correlation between IV and DV, and (usefully) can be read as the percentage variation explained by the IV
    - E.g., if SStreatment = 352.52 and SStotal = 786.82, η^2 = 0.447
    - Meaning that SStreatment explains 44.7% of the variance in the data

Friedman Analysis of Variance

- Non-parametric tests make no assumption about the underlying distribution of data
- Developed by Milton Friedman (economics Nobel laureate)
- Ranks the set of data and then asks whether the ranks are more likely to be binned in the independent variable categories

Friedman Statistic

$$Q = \frac{12}{mk(k+1)} \sum_{j=1}^{k} R_j^2 - 3m(k+1)$$

- k = levels of independent variable (columns)
- m = discrete measures (rows)
- R = rank across each row

Human Modelling & Simulation

- Model: human behaviour can be represented by heuristics and algorithms
- Simulation: an application, in software (and hardware) that can be used to run a model

Pew and Mavor, 1998, *Modeling Human and Organizational Behavior*

Performance vs. Competence

- Performance Models
  - Make statements and predictions about the time, effort or likelihood of error when performing specific tasks;
- Competence Models
  - Make statements about what a given user knows and how this knowledge might be organised.

Keystroke Level Models

- Related to Time and Motion studies
- Human information processor as linear executor of specified tasks
- Unit-tasks have defined times
- Prediction = summing of times for sequence of unit-tasks

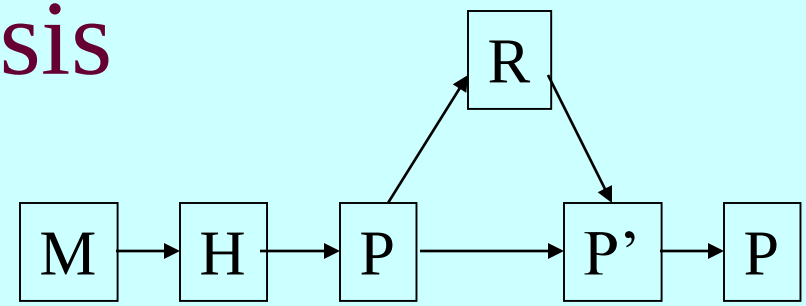
Example: cut and paste

Task Model: Select line – Cut – Select insertion point – paste

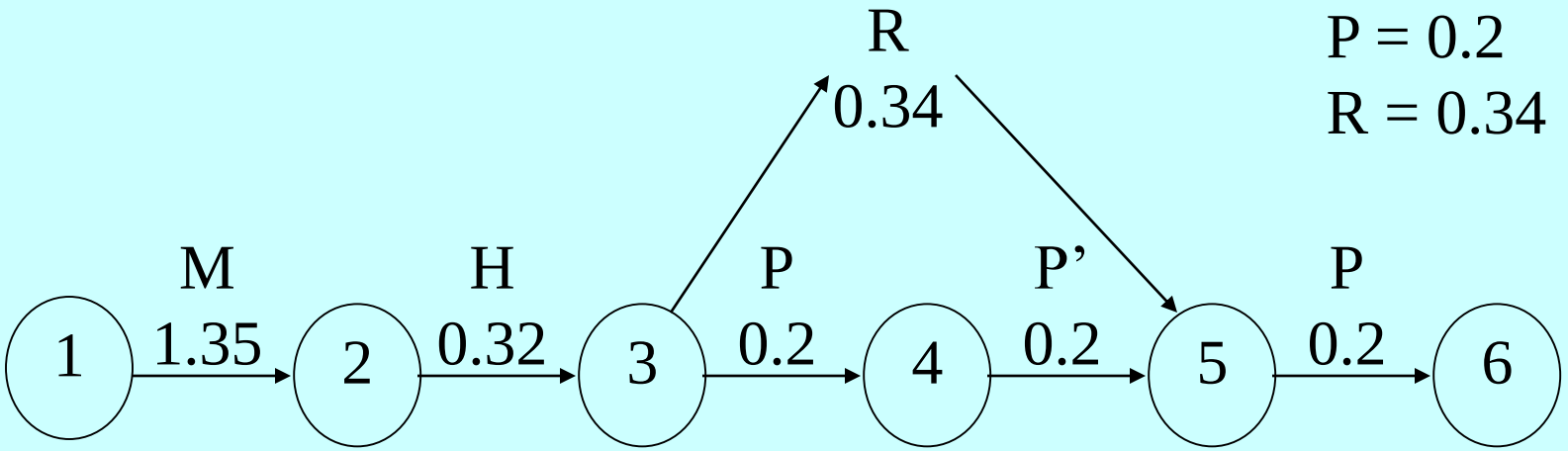
Task One: select line

- move cursor to
- start of line
- press (hold) button
- drag cursor to
- end of line
- release button

Critical Path Analysis

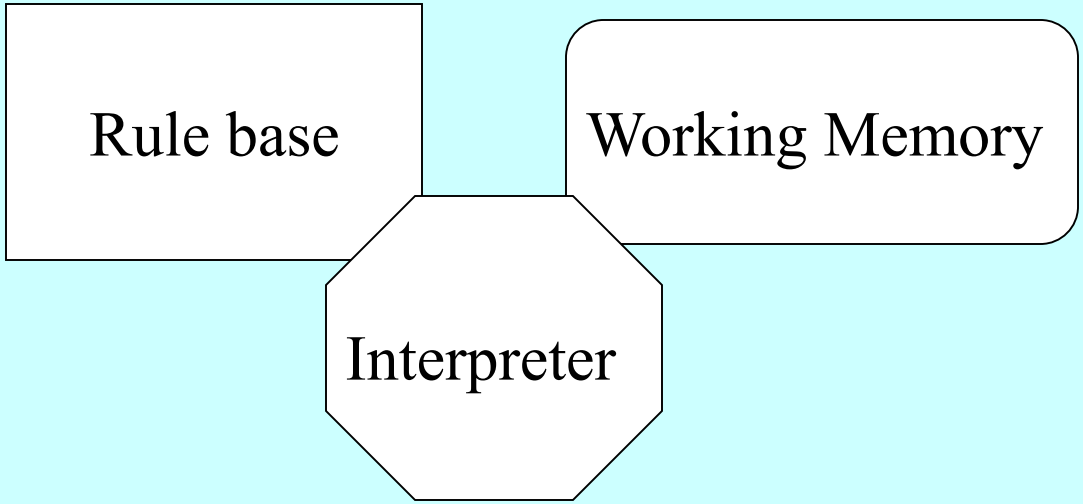


M = 1.35
H = 0.32
P = 0.2
R = 0.34



Production Systems

Architecture of a production system:



Production Systems

- *If* Condition *Then* Action
- Condition
  - Event – external (to model, e.g., light on)
  - State – internal (to model, e.g., selection)
- Action – operation associated with condition

Adaptive Control of Thought, Rational (ACT-R)

- ACT-R symbolic aspect realised over subsymbolic mechanism
- Symbolic aspect in two parts:
  - Production memory
  - Symbolic memory (declarative memory)
- Theory of rational analysis

ACT*