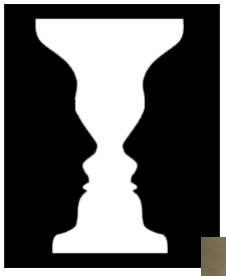
Numbers & Statistics—Damned Lies or Principled Arguments?

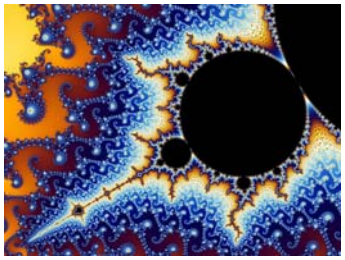
Gavin TL Brown, PhD
Presentation to FoEd PGSA
25 October, 2012

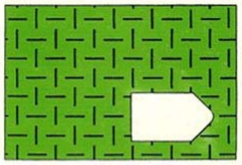
Basic problem

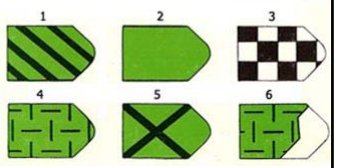
- Observations are used to generate and test theories
- They lead to inferences about the nature of reality and theories
- Actions are based on those ideas and observations
- If observations are wrong, then……
- Can you trust the data you are working with?

Humans are Pattern detectors









But not all patterns are attributable to a 'real' phenomenon

- Reality has patterns partly because random chance has patterns in it
  - Toss a coin enough times and you will get patterns like this:

T H T T H H T T T H H H T T T T H H H H

OR

T T T T T T T H H H H H H H

  - If you measure something often enough, one of your results will appear to be non-chance, even though it actually is a chance event
  - And this applies to every judgement made by a human

Test your intuition….

- In a hospital the sex of the last 6 children born was recorded. Which of the following results is attributable to chance?

1. G G G G G G
2. B G B B G G
3. B B B G G G

Role of Chance

- Chance plays a significant part in the results we observe
  - Sometimes the result we get could occur without our doing anything; just because something happens in my study doesn't mean it would **not** have occurred anyway
  - Only if it is unlikely to occur by chance can we say it is "*Statistically significant*"
    - It isn't likely to occur by chance (but it still might have!)
  - We can estimate the probability (*p*) of something happening by chance and use this to determine whether our result is likely to be real or a chance artefact

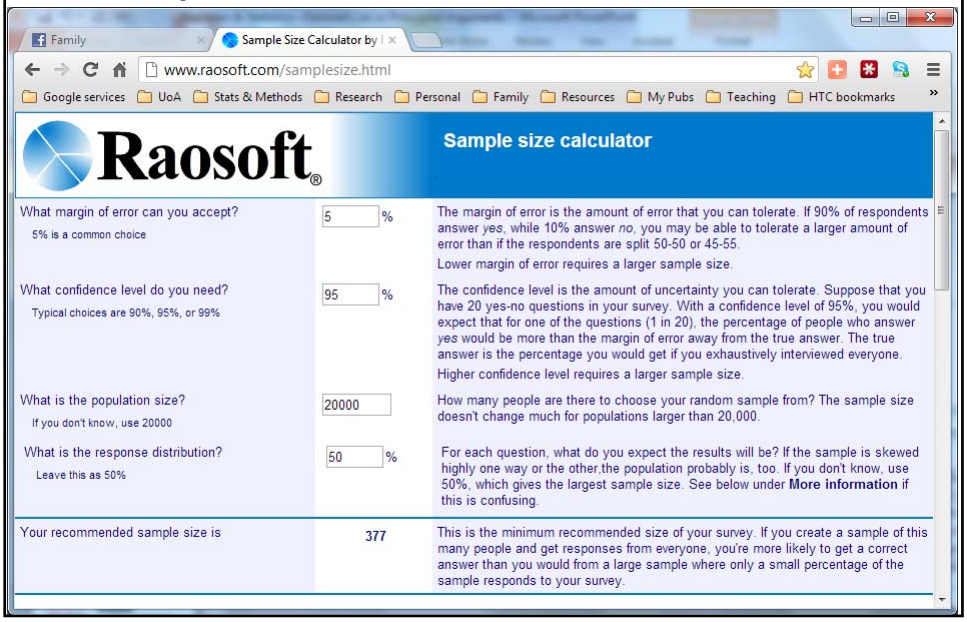
Properties of Chance

- Large samples are more accurate than small samples
- Small samples yield extreme results more often than large samples
  - Chance of all objects chosen from a population being the same is 12.5% if choosing only 4 but it is 1.56% if choosing 7
- So BEWARE of all results based on small samples
  - How big is big enough?

Sample Size

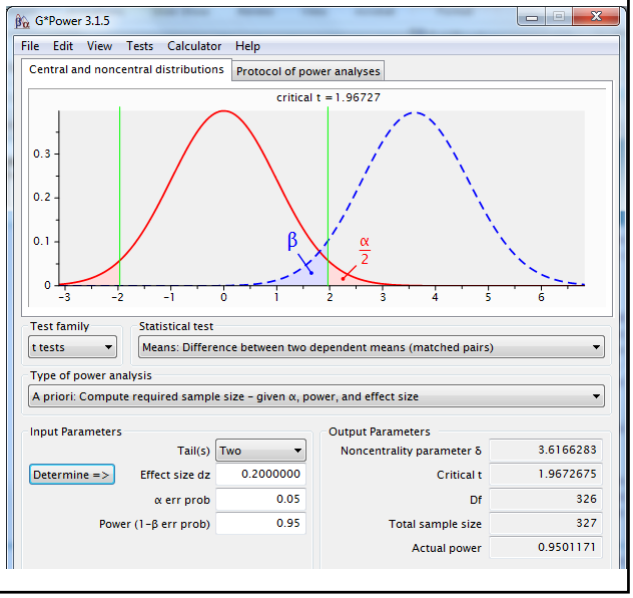
- Use Rao's free sample size calculator
  - http://www.raosoft.com/samplesize.html
- Key questions
  - Margin of error is acceptable—usually 5%
  - Confidence level—usually 95% (+/- 2 standard errors)
  - Population size—if over 20,000 it's large; but estimate it!
  - Response distribution—50%
    - If you ask a random sample of 10 people if they like donuts, and 9 of them say, "Yes", then the prediction that you make about the general population is different than it would be if 5 had said, "Yes", and 5 had said, "No". Setting the response distribution to 50% is the most conservative assumption. So just leave it at 50% unless you know what you're doing.

Example



Power Calculation

- G*Power software
  - http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/
- Helps determine how many people needed and what critical values needed to detect a difference of certain expected size



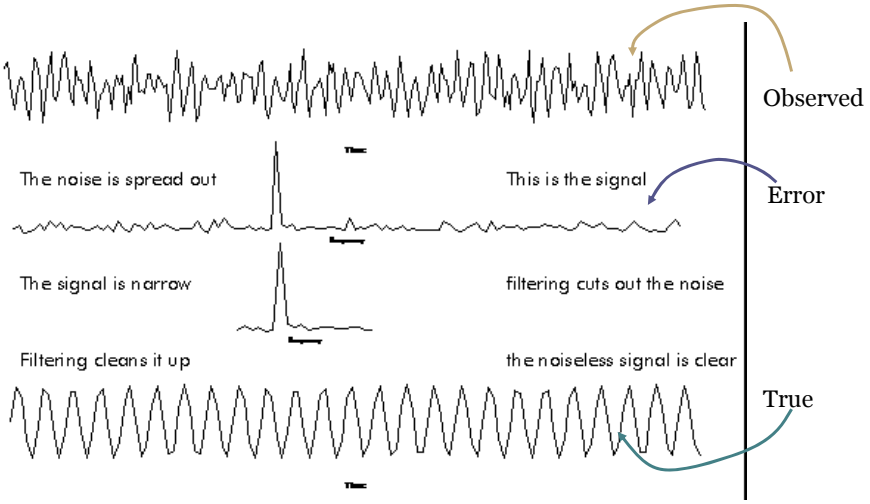
Sample size also depends on procedure

- Experiment: at least 30 in each condition
- *t*-test: up to 30 people
- *F*-test: similar numbers per group; at least 30 people in each group, but prefer more
- Multivariate statistics: at least 10 people per variable (30 questions= 300 people)
- CFA/SEM: at least 400 people

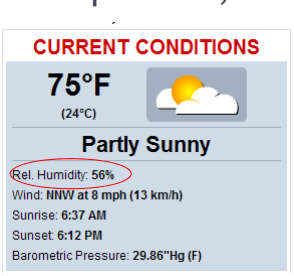
Error in Measurement

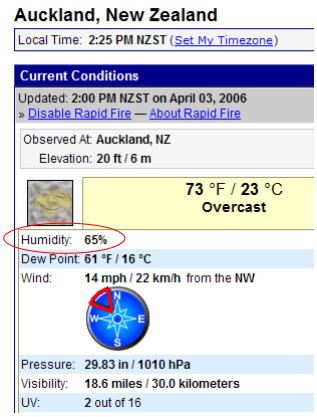
- In addition to sampling, all measures have some error
  - The length of a certain platinum bar in Paris is a metre
  - It is supposed to be 1/10,000,000th of the distance from equator to pole
  - **but** it is short by 0.2 mm according to satellite surveys
- Less error in measures of physical phenomena and more in social phenomena

Signal to Noise Ratio—a way to conceive of what we are trying to do



Compare Temperatures, Wind, Pressure, Cloud, Precipitation, Humidity



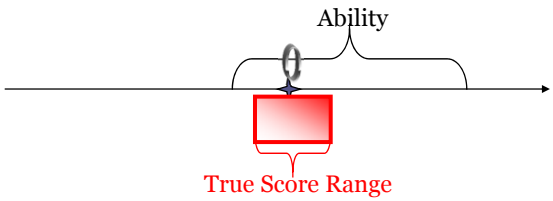


Most important things

- Are complex, subtle, dynamic
- Difficult to quantify
- Hence, our estimations of how much, how many, how often, etc. contain error: So more is better!
- Thus, classical test theory posits
  - O= *t* + *e*
  - What we observe is equal to 'true' (real) value plus error
  - Error includes random and systematic faults

Different measurements

- *Observed*—What you actually get on a measurement (e.g., test/ inventory/ observation)
- *True*—What you should get if the measure were perfect, bearing in mind measure is a sample of domain
- *Ability*—What you really are able to do or know of a domain independent of what's in any one measure



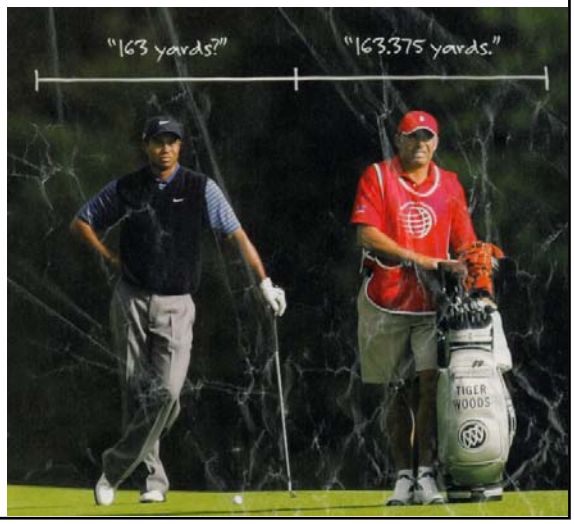
Because the observed value contains error we can only ESTIMATE what the real performance value is

Why Care?

- If measures are inaccurate or inconsistent then the quality of Interpretations and Actions based on those measures is threatened.
- Thus, an indication of how accurate or consistent a measurement is MUST be attempted to give credibility to the researcher's opinions & actions

How accurate does it really need to be?

- Close enough might be good enough
- Level of accuracy required depends on level of risk you can live with



Finding a balance between risk and accuracy

Accuracy

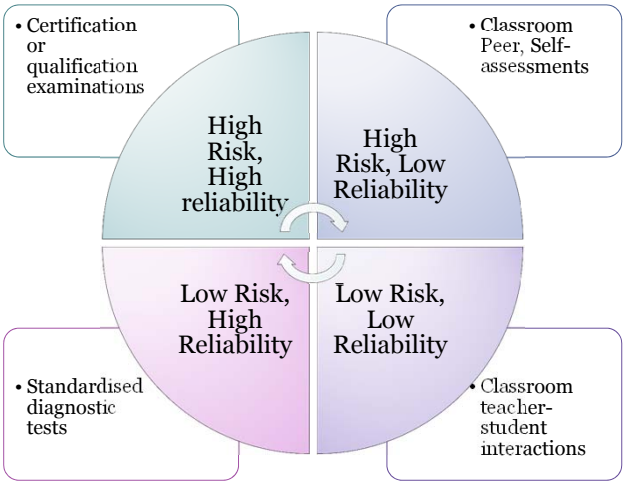
Importance of getting it right—risk of being wrong!

Tolerance for error

Ability to recover from or cope with errors

Risk

Thinking about risk and accuracy

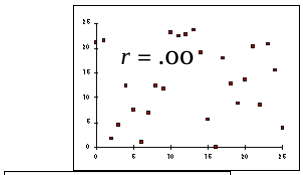


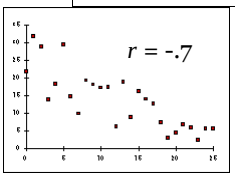
Estimating accuracy

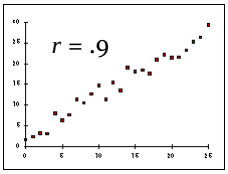
- Multiple indicators of a construct reduce the influence of randomness
- Correlation of variables to other variables that are supposed to be related indicates consistency
  - Pearson: continuous variable to another
  - Point biserial: dichotomous variable to total without the variable
  - Cronbach's alpha: median inter-correlation of all items being evaluated
  - Cohen's kappa: chance adjusted agreement of judges to each other
- Conventional but indicative only methods

Correlation: Measure of Agreement

- As one variable changes, how does the other one change?
  - Range -1 to +1;
  - 1 = perfect agreement
  - 0 = no pattern, thus independent
  - Can be negative (one goes up, the other down)—very bad
  - Notice the closer the value is to 1.0 the less spread out or variable the data is. Agreement or consistency comes when scores are close to each other.
  - Without variance in the data correlation values will be reduced
  - r^2= proportion of variance explained





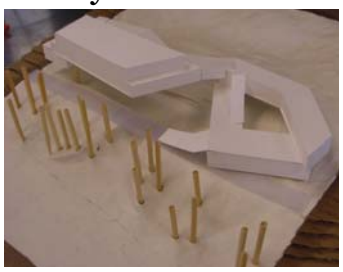


Critical values for Pearson Correlations

- > .9 ➔ use for highest stakes possible (consistency relationship explains 81% or more of variance)
- > .8 ➔ publish assessment (consistency relationship explains 64% or more of variance)
- > .7 ➔ classroom teacher use only (consistency relationship explains 49% or more of variance)
- < .6 ➔ random noise (consistency relationship explains less than 36% of variance)

Confirming quality of measurements

- A model is a simplification of reality
- The quality of the model depends on the degree to which it fits to relevant theory and data collected to test the theory
- If the model differs from the data by only chance, then we have a basis for accepting that the simplification represents reality



Models

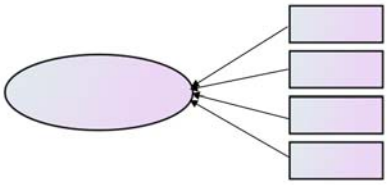
- Everything is connected to everything in the real world
  - It's messy and hard to make sense of
- **BUT**
  - in a model we select for theoretical reasons the important connections that we THINK explain most of what is going on in the phenomenon of interest
  - It is not the real thing, but a simplification
- The arrangement of the connections between and among variables of interest constitute testable expressions of our theories about how things go together

Model frameworks

- Manifest models
  - Formative analysis
  - Use only directly measured variables
  - Presume variables are sufficiently accurate indicators or proxies
    - E.g., sex, age,
  - predictors create the domain;
    - E.g., socio-economic status which is an aggregation of contributing variables such as income, education, and wealth.
- Latent trait theory
  - Reflective analysis
  - Theoretically proposed constructs explain behaviour as obtained from responses to direct measures
  - manifest variables are indicators of a latent trait that causes the responses given to each indicator
  - Variables are samples of latent trait domain

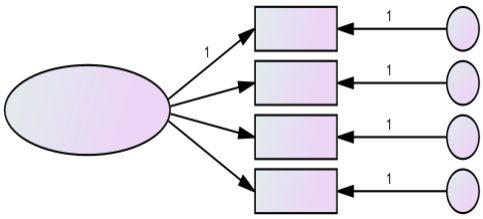
Model Frameworks

- Reflexive



- More common in sociology
  - Trait is what the indicators are and nothing more
  - Trait is a composite index of the sum of predictors

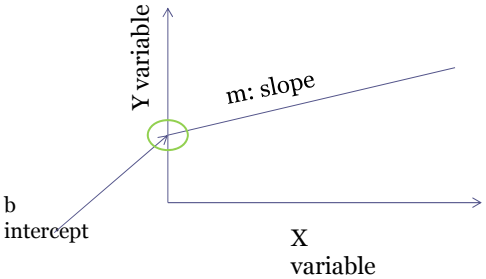
- Formative



- More common in psychology & education
  - Trait contains variables plus others not measured at this time; but indicators are also influenced by random

Understanding linear models

- linear regression
  - Changes in XXX cause a linear change (increase or decrease) in YYY
  - Formula: Y= m*X + b
    - m=slope [standardised beta = a proportion of standard deviation]
    - b=intercept [starting point of equation; represents all the unknown stuff]



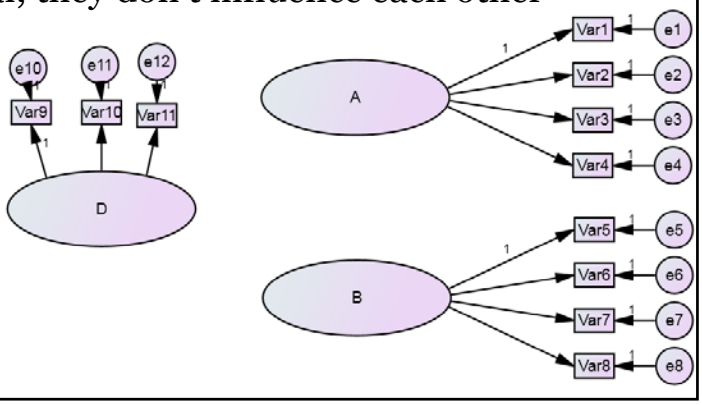
Interpretations:
1. For every 1 *SD* change in X, you will get m**SD* change in Y.
2. This relationship explains *x*% of variance in Y

Prediction, Causation, Association

- Most common models assume linear (i.e., correlations and regressions) relationships (paths) exist among constructs.
- And these relations can be diagrammed and statistically calculated provided enough data exists –*does it work?*
- And then the quality of the model to the data can be estimated---*it works but is it worth keeping?*

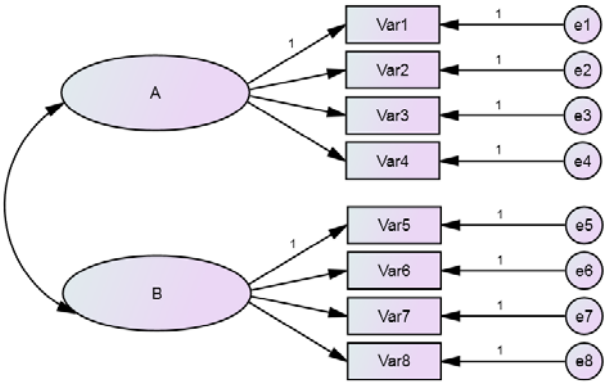
All things are independent

- The groups are uncorrelated (*r*=.00) or orthogonal; they don't influence each other



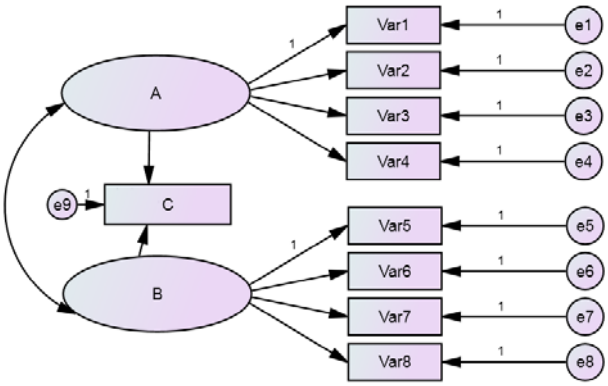
Correlations

- 2 things exist simultaneously and behave in a coordinated fashion



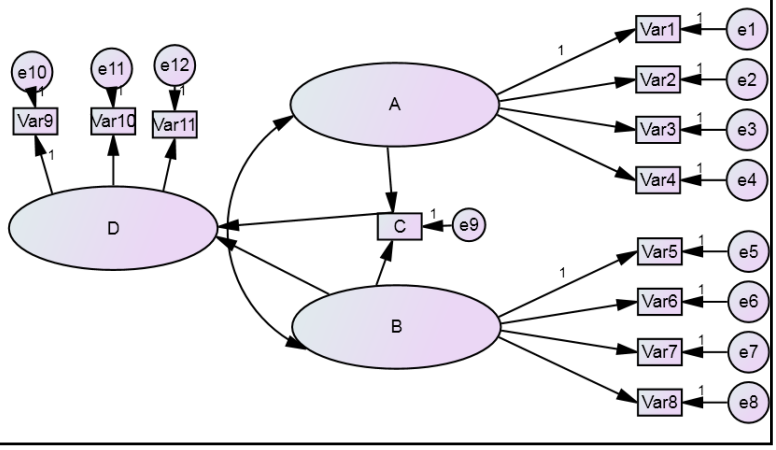
Correlation + Regression

- correlated things may jointly influence a 3rd thing



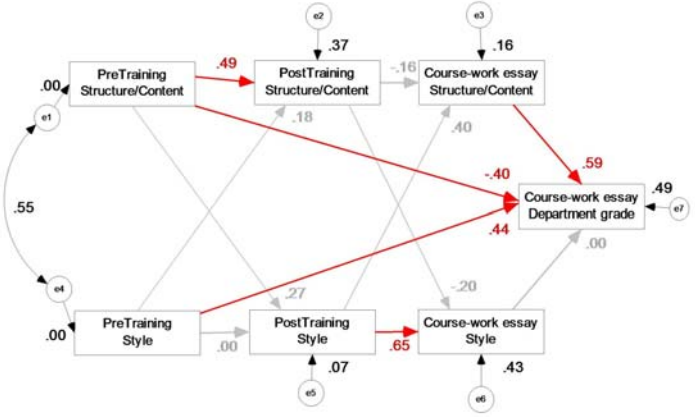
MIMIC SEM

Multiple Indicators & Multiple Causes Structural Equation Model



Linear Path Model

- Markov chain with cross-lagging



Mathematical Testing of Formative Models

- How close are they? Does the model fit the data?
  - Models are rejected if they do NOT have close fit to the data
    - the data can't be wrong—it's the reality we are trying to model…so make sure the data are representative & robust
  - Models are NOT accepted if they have close fit to the data
    - They are NOT YET DISCONFIRMED—Popper
    - Multiple models can fit equally well the same data
    - Fit could be attributable to chance factors in the data we collected
    - So proceed with caution

Latent Trait Theory

- Multiple manifest indicators are required to have stable estimation of the latent trait's existence, strength, and direction
  - Hence, factor analysis expects 3 to 6 items per factor
  - Hence, test scores rely on 5 to 30 test questions
- WHY?
  - CHANCE….ERROR….DEFICIENCIES IN STIMULI
  - Observed behaviour is not perfectly controlled or reflective of our TRUE intelligence, attitude, belief, ability, etc.
    - *I chose B but I meant A; I chose response 3 but I meant 4*
  - Our response mechanism interferes
    - *I want 3.4 but I had to choose 3 or 4*
- Hence, all values are ESTIMATES
  - A range of most likely values exists
  - Multiple indicators reduces error/chance effects

Using statistics to answer questions you might want to ask

- Chi-square test (χ^2)
  - Is my sample different from the population? Does my model differ from the data? Do the characteristics of the drop-outs differ by from the stay-ins?
- Regression
  - Does this predictor have a meaningful effect on the dependent variable?
- *t*-test, *F*-test, Cohen's *d* effect size
  - Are the differences in group means due to chance and are they big enough to care about?

x^2 : Ratio of Observed to Expected—a way to tell if a model differs from data

- Contingency Table—Frequency of Categories: Observed (Expected)

Sex	CatA	CatB
M	5 (10)	15 (10)
F	15 (10)	5 (10)

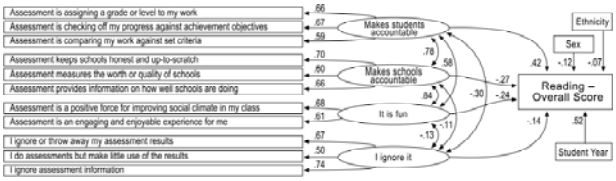
- Sum of all (Expected-Observed) compared to critical values contingent on *df* and *N*
- good for large samples, $n>30$; but cells must have minimum number of cases

X^2 Which party do you belong to & what is your attitude?

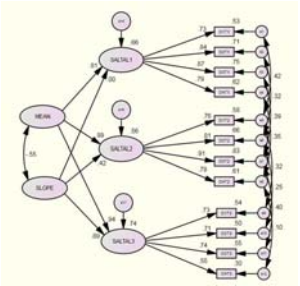
	Party A	Party B	Party C
In favour	30%	90%	80%
Opposed	70%	10%	20%

- $\chi^2_{(2)} = 93.0$; $p < .05$
  - What is the χ^2 really asking?
  - What do the statistics mean?
  - Where is the difference most likely to be?

x^2 in latent trait theory



- Do the models differ from the data?
  - This statistic is used in
    - Confirmatory factor analysis
    - Structural equation modeling
    - Latent curve modeling

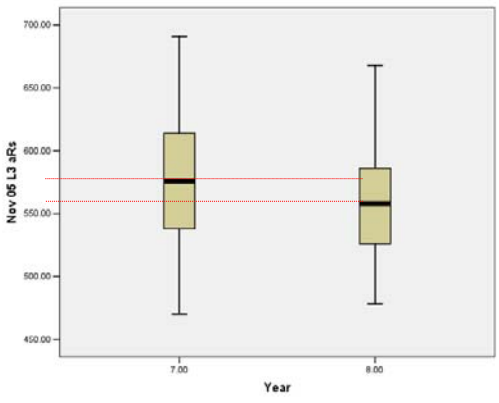


Differences in Group Means

- Assuming the measurement of the construct is sufficiently accurate so as to merit usage, how can we compare groups or times?
- Statistic depends on structure of design
  - Repeated measures with matched groups
  - Contrasting groups with cross-sectional data
- Two evaluations
  - Differ from chance➔statistical significance
  - Big enough to care about➔practical significance

Differences in Means

- How different do means have to be to be real differences?
  - Statistical significance: greater than chance
  - Practical significance: large enough to care about



Conclusion?

Student's *t*-test

- Observed Value1 (+/- 2 std errors) minus Observed Value2 (+/- 2 std errors)
- Compare result to critical values of that result appearing by chance given number of cases
- good for small sample $n<30$
- Uses
  - Compare sub-group to overall mean
  - Compare two different groups
  - Compare same group in repeated measures
- Prone to interpretive error if multiple tests conducted

L3 Nov 05: Difference to expected value

- Is my sample different to the Norm group?
  - Expected $M = 555$ (average for Y7)
  - Observed $M = 566$

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
L3Nov05aRs Nov 05 L3 aRs	82	566.4390	48.83282	5.39268

One-Sample Test

	Test Value = 555					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference Lower	Upper
L3Nov05aRs Nov 05 L3 aRs	2.121	81	.037	11.43902	.7093	22.1688

Conclusion?

Fisher's *F*: Analysis of Variance (ANOVA)

- Ratio of variance within a set of scores to variation between two different scores

$$\frac{\text{(Mean Square Between Groups)}}{\text{(Mean Square Within Groups)}}$$

- Compare to critical value taking into account number of cells between groups and number of cases within groups
- More robust than multiple *t* tests

L3 Nov 05: Univariate ANOVA

- Comparison of Years 7—8

Report

L3Nov05aRs Nov 05 L3 aRs

Year	Mean	N	Std. Deviation
7.00	575.5161	31	54.45418
8.00	560.9216	51	44.74186
Total	566.4390	82	48.83282

ANOVA

Conclusion?

L3Nov05aRs Nov 05 L3 aRs

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4106.767	1	4106.767	1.738	.191
Within Groups	189049.4	80	2363.118		
Total	193156.2	81			

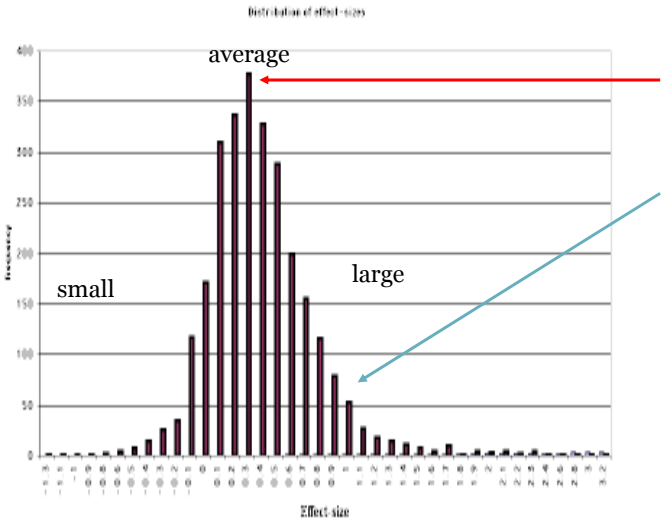
Practical Significance

- Small differences can be non-chance with large samples
- We have limited resources so need to look for magnitude of effect
- Best, constant scale for measuring effect is proportion of *SD*; thus, effect size
- Most things have an effect in education

Effect Size

- *d* = 1.0
  - average students receiving treatment would exceed 84% of students not receiving that treatment.
  - large, blatantly obvious, grossly perceptible
  - difference between mean IQ of PhD graduates and high school students
- *d* = .31
  - not perceptible to the naked observational eye
  - approximately equivalent to the difference between the height of a 5'11" and a 6'0" person.

Hattie meta-analysis of effects



M effect size is .40

Seek effects >.40

Effect Size

- The difference between two means divided by their spread (usually *SD*); Cohen's *d*
- $(M_{group1} - M_{group2})/((SD_1+SD_2)/2)$

Group	Reading	Writing	Mathematics
Female	514	512	505
Male	478	472	508
SD	100	100	100
d	.36	.42	-.03

Effect Size: Change scores

- The difference between post-pre scores

Problems

1. Unreliable
2. Are you measuring same thing both times
3. Regression to the mean

So: use effect size to estimate difference in time

Effect sizes apply to between groups and across time

- Show if change too large to be explained by maturation
- Difference between *M* scores /spread ➔ Effect-size

$$\frac{M_2 - M_1}{SD_{(1+2)2}}$$

Example. $M1$ (pre-test) = 12,
$M2$ (post-test) = 15
average SD = 6
effect size (d) ➔ $\frac{15-12}{6} = .5$

Summary: Quantitative methods

- A systematic way of looking at phenomena and attempting to determine if it is beyond chance and large enough to matter
- Testing theories about how things ought to go together; not just throwing things in the air and hoping to find something
- Strong concern for the credibility of a set of arguments concerning the observations

Big issues to keep in mind

- This approach to research still leaves a large number of problems that are still in contention
  1. Are psychological attributes actually measureable?
     - Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept. New York: Cambridge University Press.*
     - Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics. Cambridge, UK: Cambridge University Press.*
  2. Do regression-based models actually depict causal relationships in the real world?
     - Pearl, J. (2000). *Causality: Models, reasoning, and inference. Cambridge, UK: Cambridge University Press.*

Where to begin reading?

- Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: LEA.
- Blaikie, N. (2003). *Analyzing quantitative data*. London: Sage
- Coladarci, T., Cobb, C. D., Minium, E. W., & Clarke, R. B. (2008). *Fundamentals of statistical reasoning in education* (2nd ed.). Danvers, MA: John Wiley & Sons.
- Field, A. (2005). *Discovering statistics using SPSS (and sex, drugs and rock 'n' roll)* (2nd ed.). Thousand Oaks, CA: SAGE.
- Jaeger, R. M. (1983). *Statistics: A Spectator Sport.* Beverly Hills, CA: Sage.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. London: Penguin Books.
- Popham, W. J. (1967). *Educational statistics: Use and interpretation.* New York: Harper & Row.
- StatSoft. (2007). *Electronic Statistics Textbook* (Web: http://www.statsoft.com/textbook/stathome.html ed.). Tulsa, OK: StatSoft.

Fun stuff

- Alder, K. (2002). *The measure of all things: The seven-year odyssey that transformed the world.* London: Abacus.
- McGrayne, S. B. (2011). *The theory that would not die: How Bayes' Rule cracked the Enigma Code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy.* New Haven, CN: Yale University Press.
- Gould, S. J. (1996). *The mismeasure of man* (2nd ed.). New York: W. W. Norton & Co.
- Taleb, N. N. (2004). *Fooled by randomness: The hidden role of chance in the markets and in life* (2nd ed.). New York: Texere.
- Wainer, H. (1997). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot.* New York: Copernicus Books.