PSYC7112

Advanced Assessment Techniques

BASIC AND EXTREME psychometrics

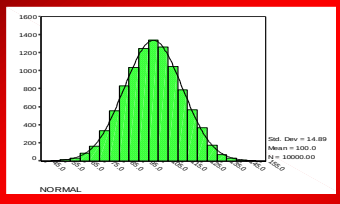
Dr Mark Horswill
School of Psychology, University of Queensland
m.horswill@psy.uq.edu.au

Contents:

- Test standardisation and the normal distribution
- Reliability & validity
- Individual score interpretation – SEM, Sediff, Reliable Change Index
- Evaluating diagnostic tests – 2x2 tables, likelihood ratios, ROC curves
- Signal detection theory
- Free extras (not testable in exam – only covered if time): power analysis and meta-analysis primers

The Normal Curve

Also known as the Laplace-Gauss curve as in "this curve is Gaussian in nature".



Karl Gauss

Pierre Laplace

IF we can assume something has a normal distribution THEN knowing just the mean and standard deviation can tell us how someone's score compares with everyone else.

Normal Curve in Psychological tests

Scores on many psychological tests tend to be approximately normally distributed .

The larger the sample and the wider the range of things measured, the closer to a normal curve the distribution usually becomes.

Because the normal curve is mathematically defined, if we can assume something is normally distributed then it means we can do more sensitive (parametric) statistical tests on it.

If something is normally distributed we know:

1. Mean = median = mode therefore 50% of people are below/above the mean
2. 68% of scores +/- 1 s.d. around mean.
3. 95% of scores +/- 2 s.d. around mean.
4. Tails of distribution are 2 to 3 s.d. from the mean.

e.g. 'Mentally retarded' has been defined as an I.Q. of below 2 s.d. below the mean (mean = 100, s.d. = 15, I.Q.<70). So - 'mentally retarded' is always in comparison with the rest of the population (it's not absolute).

Standard scores

Raw scores can be converted into standard scores to make interpretation simpler. That is, we anchor the mean and standard deviation of the scale - and therefore we know what any particular score means without having to explain the original scale. We can also compare performance across different scales.

z scores - mean of 0 and s.d. of 1

So - z score of 1 = you're 1 s.d. above the mean
z score of -.9 = nearly 1 s.d. below the mean
z score of 1.4 = between 1 & 2 s.d. above mean

$$z = \frac{X - \overline{X}}{s_X}$$

z = z score
X = raw score
$\overline{X}$ = mean of raw scores
s_x = s.d. of raw scores

T scores - mean of 50 and s.d. of 10

- **Avoids negative numbers (need < -5 s.d. to get negative) unlike z scores. To convert a z score into a T score, multiply by 10 and add 50.**
- **T scores are used by the Minnesota Multiphasic Personality Inventory (MMPI) - with some tweaking. Also used by the Stroop test.**

Note that this sort of transformation DOES NOT change the shape of the distribution. It is a *linear transformation* .

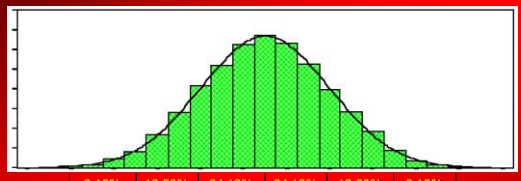
IQ scores - mean of 100 and s.d. of 15

IQ - 'deviation intelligence quotient'

So - because IQ is normally distributed:

- **IQ 100 = 50% smarter; 50% more stupid**
- **IQ 115 (+1 s.d.) = 16% smarter; 84% more stupid**
- **IQ 130 (+2 s.d.) = 2% smarter; 98% more stupid**
- **IQ 85 (-1 s.d.) = 84% smarter; 16% more stupid**
- **IQ 70 (-2 s.d.) = 98% smarter; 2% more stupid**

Percentage chart for normal distribution



	2.15%	13.59%	34.13%	34.13%	13.59%	2.15%
	-3 s.d.	-2 s.d.	-1 s.d.	Mean	+1 s.d.	+2 s.d.	+3 s.d.
Z score	-3	-2	-1	0	1	2	3
T score	20	30	40	50	60	70	80
IQ score	55	70	85	100	115	130	145

See Appendix of handout for a full conversion table.

Stanine scale

Used in school tests – 9 divisions, each .5 s.d. wide, with the middle band (5) from -.25 to +.25 s.d.

For example, the Neale Analysis of Reading.

STA(ndard) NINE

	4%	7%	12%	17%	20%	17%	12%	7%	4%
Stanine	1	2	3	4	5	6	7	8	9
s.d.	-1.75 to -2.25	-1.25 to -1.75	-.75 to -1.25	-.25 to -.75	-.25 to +.25	+.25 to +.75	+.75 to +1.25	+1.25 to +1.75	+1.75 to +2.25

Note that this scale can be mapped straight onto all the others on the previous slide (normal distribution)

Reliability:

"The extent to which measurements are consistent or repeatable; also, the extent to which measurements differ from occasion to occasion as a function of measurement error" (Cohen and Swerdlik, 2002, page 660)

- Lower measurement error = higher reliability.
- In Classical Test Theory, reliability is true variance (hypothetical variation of scores in a sample if no measurement error) divided by total variance (actual variation in data - including error)

We can **ESTIMATE** test reliability via:

Internal consistency

How much the item scores in a test correlate with one another on average (e.g. Cronbach's alpha, KR-20, Kappa).

Test-retest reliability

If people sit the same test twice, how much do their scores correlate between the two sittings?

Alternate-forms reliability

If people do two different versions of the same test, how much do their scores on the two versions correlate?

Inter-rater reliability

If a test involves an examiner making a rating - get two of them to do the rating independently and see how much their ratings correlate.

Different types of validity (traditional taxonomy)

- **Content validity** (are items in test judged to be representative of the domain?)
  - **Face validity** (how valid does a test intuitively APPEAR to be?)
- **Construct validity** (is the test judged to be an adequate measure of the theoretical construct that it was supposed to measure?)
- **Criterion-related validity** (test score correlated with other measures of interest?)
  - **Concurrent validity** (other measure of interest measured AT THE SAME TIME as the test)
  - **Predictive validity** (other measure of interest measured AFTER the test)

Why is it important for both the test *and* its criterion to have decent reliability?

Because the reliability of each ***limits*** the size of the validity coefficient (the correlation between test score and the criterion). That is:

- If the reliability of the test itself is low, then the validity coefficient may be lower than it should be.
- If the reliability of the criterion measure is low, then the validity coefficient may be lower than it should be.

"The validity coefficient is (always) less than or equal to the square root of the test's reliability coefficient multiplied by the square root of the criterion's reliability coefficient" (C&S, p.161)

Individual score interpretation

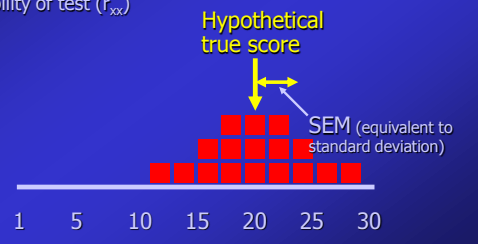
- In theory, if an individual repeatedly takes a test (losing their memory of the test after each go), then their scores will form a normal distribution.
- In Classical Test Theory, this distribution should be centered around their hypothetical 'true' score.
- We can estimate the chance than someone's actual score in a test is close to their 'true' score.
- The Standard Error of Measurement (SEM) is a statistic that tells you how much an individual's measured score is likely to deviate from their true score (low if high reliability, high if low reliability).

Standard Error of Measurement (SEM)

SEM can be *estimated* using:

- Standard deviation of test-takers' scores (s_x)
- Reliability of test (r_{xx})

$$SEM = s_x\sqrt{(1-r_{xx})}$$



If we assume the normal distribution…

- 68% of an individual's scores will be within 1 SEM of the true score (+/-1 standard deviation).
- 95% of an individual's scores will be within 2 SEMs of the true score (+/- 2 s.d.).
- 99.7% of an individual's scores will be within 3 SEMs of the true score (+/- 3 s.d.).

If someone's only taken a test once, then our best guess of the true score is their actual score.

On that assumption, we can estimate the likely *margin of error* in someone's score.

Confidence interval:

the range of scores that is likely to contain a person's true score (margin of error)

e.g. 95% confidence interval: +/- 2 s.d. (1.96 to be precise) with a normal distribution (95% of scores fall within 2 s.d. of the 'mean' which is their actual score).

Therefore the 95% confidence interval is the *actual score* +/- (2 x the SEM).

EXAMPLE: the **WAIS IQ test** - reliability is .98; s.d. is 15; so SEM is: $15\sqrt{(1-.98)}$ = 2.12.

*If someone gets an IQ score of 105, their 95% confidence interval is [105 +/- (2*2.12)] from 101 to 109*

(i.e. their true IQ score is 95% likely to be in that range).

Standard error of the difference (SEdiff)

Use this to work out whether someone's score is significantly different from:

1. Their score on another test of the same thing
2. Someone else's score on the same test
3. Someone else's score on another test.

For 2 tests, we first transform them to the same scale e.g. a z score. Then:

$$SEdiff = \sqrt{SEM1^2 + SEM2^2}$$

SEdiff = standard error of the difference

SEM1 & SEM2 = Standard error of mean for tests 1 & 2

Standard error of the difference (SEdiff)

- Or, alternatively:

$$SEdiff = s.d.\sqrt{2 - r1 - r2}$$

SEdiff = standard error of the difference

s.d. = standard deviation of test 1 = standard deviation of test 2 (because they've been standardized)

r1 & r2 = reliability of tests 1 & 2

If you're comparing two scores on the same test, then SEM1=SEM2 and r1=r2 & just put the numbers in the formula as before.

Standard error of the difference between two scores - how to use it

- To be 95% confident that two individual scores are different then they would have to differ by at least *2 standard error of the differences* (1.96 to be precise) - because of the normal distribution.
- If the two scores differ by more than 2 SEdiff then we can say that they are significantly different at a 95% level of confidence.
- If the two scores differ by less than 2 SEdiff then we can say that they are not significantly different at a 95% level of confidence.

Standard error of the difference between two scores - an example

- A man has been on a program of treatment for depression.
- Before the treatment he scored 134 on a test of depression and 125 on the same test afterwards (high score, more depressed). How can you tell if this decrease in depression is down to chance or not?
- Reliability of the test is .92 and its standard deviation is 14 - so we calculate the standard error of the diff:

SEdiff = 14*sqrt(2-.92-.92) = 5.6

- The man's scores differ by 9 points which is 9/5.6 = 1.6 SEdiff. That's not enough to be 95% confident that they're different (the scores need to be 2 SEdiff apart) so we can't conclude the man is significantly less depressed after his treatment.

What if you expect scores to increase on repeated administrations of a test?

- Some tests are often applied more than once to the same person - and are associated with a practice effect. That is, you expect people to improve their scores even if the underlying trait being measured remains unchanged.
- So, if someone's score remains the same then this could actually indicate they have a problem.
- One way to address this is to re-standardise people's scores for their 2nd attempt against a sample of 2nd attempt scores (i.e. you correct for the expected improvement).
- Remember that any changes need to be significantly different to be considered meaningful - you'll always get some score fluctuation due to measurement error.

Evaluating diagnostic tests

Problem: Imagine a 55 year old man takes a test which indicates he has dementia. The probability of dementia is 1% in 55 year old men. If he has dementia, the probability is 80% that the test will detect it. If he does not have dementia, the probability is 10% that the test will incorrectly indicate he has dementia. What is the probability that this man actually has dementia?

Evaluating diagnostic tests

- We can answer this question using the following technique involving a 2x2 table:

	Disorder present	Disorder absent
Test positive	Correct positives	False positives
Test negative	False negatives	Correct negatives

Sensitivity: % of people with disorder who test positive

Specificity: % of people WITHOUT the disorder who test negative

Pre-test probability: % of people in the population who have the disorder

Using 2x2 tables to evaluate tests:

- **Problem:** Imagine a 55 year old man takes a test which indicates he has dementia. The probability of dementia is 1% in 55 year old men. If he has dementia, the probability is 80% that the test will detect it. If he does not have dementia, the probability is 10% that the test will incorrectly indicate he has dementia. What is the probability that this man actually has dementia?

So:

- Pre-test probability = .01
- Sensitivity (% correct positives) = .80
- Specificity (% correct negatives: 100%-10%) = .90.

Using 2x2 tables to evaluate tests:

	Disorder present	Disorder absent	
Test positive	Correct positives	False positives	Total positive
Test negative	False neg.	Correct neg.	Total negative
	Total with disorder	Total without disorder	Grand total

1. Choose arbitrary number of people (100000) and put in grand total box.
2. Multiply Grand Total by Pre-test probability to get Total with disorder.
3. Grand total minus Total with disorder = Total without disorder.
4. Multiply Total with disease by Sensitivity to get Correct hits.

Using 2x2 tables to evaluate tests:

	Disorder present	Disorder absent	
Test positive	Correct positives	False positives	Total positive
Test negative	False neg.	Correct neg.	Total negative
	Total with disorder	Total without disorder	Grand total

5. Multiply Total without disorder by Specificity to get Correct misses
6. Compute False positives and False negatives by subtracting Correct hits/Correct misses from column totals
7. Compute Total positive and Total negative by adding up rows

- Predictive value of a positive test is Correct positives ÷ Total positives.
- Predictive value of a negative test is Correct negatives ÷ Total negatives.

Example problem (old man with dementia):

	Disorder present	Disorder absent	
Test positive	800	9900	10700
Test negative	200	89100	89300
	1000	99000	100000

- Pre-test probability = .01
- Sensitivity = .80
- Specificity = .90.

- **Predictive value of positive test = 800/10700 = .07**
- **Predictive value of negative test = 89100/89300 = 1.00.**

- **So - what is the probability that this man actually has dementia given his positive test result? It's 7%.**

Likelihood ratios

- The 2x2 table method could only deal with tests that give a dichotomous result (positive/negative).
- However, most tests give you more than two outcomes (e.g. a number on a scale). Different numbers could indicate different severities - so by just having pass/fail you're throwing away information.
- Also, if you're using multiple diagnostic tests, it's possible but pain-staking to calculate their combined value using 2x2 table method.
- Both these problems are solved if you use Likelihood Ratios instead (& calculations are simpler).

Likelihood ratios

- The **post-test odds** (i.e. chance of positive test being correct) equals the **pre-test odds** (i.e. prevalence in the population) multiplied by the **likelihood ratio**.
- That is – the likelihood ratio is the ratio between people with a positive test who have the disease and people with a positive test who don't have the disease.
- **Likelihood Ratio = Sensitivity / (1 – Specificity)**
- Note that likelihood ratio calculations deal in **odds** (e.g. 2 to 1) rather than **probabilities** (e.g. 0.66).
- **Odds = probability/(1 - probability)**
- **Probability = odds/(1 + odds)**
  - Odds 3 (to 1) = .75 probability
  - Odds 9 (to 1) = .90 probability
  - Odds .5 (to 1) = .33 probability

Likelihood ratios

- So, if the likelihood ratio is 1 then the test has no predictive power (positive test doesn't tell you whether you've got the disease or not).
- A likelihood ratio of 2 means that if you have the disease then you're twice as likely to test positive for the disease as someone who doesn't have the disease.
- A likelihood ratio of .5 means that if you have the disease you LESS likely (half as likely) to test positive than someone without the disease.
- The bigger the ratio, the better the test can tell apart people with and without the disease.

Likelihood ratios

- So - to solve our previous problem using likelihood ratios, we'd convert pre-test probability into odds then multiply by the likelihood ratio for a positive test (then convert back to probabilities if we wanted).
- Pre-test probability = .01 (1%)
- Pre-test odds = .01/(1-.01) = 0.01 (to 1)
- Likelihood Ratio = Sensitivity / (1 – Specificity)
  = .8/(1-.9) = 8
- Post-test odds = LR x Pre-test probability
  = 8 x .01 = .08 (to 1)
- Post-test probability = .08/(1+.08) = .07 (7%)

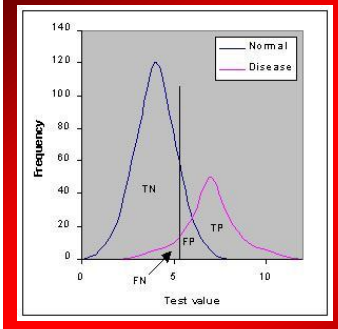
Likelihood ratios

- In the literature, you might well see likelihood ratios given for a test (positive/negative) instead of sensitivity/specificity.
- For a tutorial on how to use Likelihood ratios for multiple tests and for more than 2 possible outcomes - see: **gim.unmc.edu/dxtests/**

ROC curves

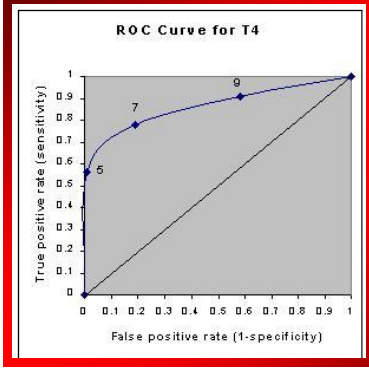
- The number of correct and false positive test results depends on (1) how accurate the test is but also (2) where you set the "pass mark" for the test (e.g. at what score are people labelled demented?).
- We may want to use different "pass marks" in different clinical situations depending on whether it's more important to minimise false positives or false negatives (or if you just want to maximize discrimination between groups).
- A ROC curve is a plot of Correct positive rate versus False positive rate - where each point on the curve is a different "pass mark" for the test.

ROC curves



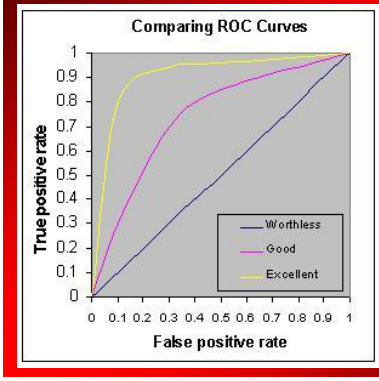
- Vertical line is "pass mark".
- As we move the "pass mark" to the right, the true positives decrease but so do false positives - due to overlap in distributions (because the test isn't perfect).

ROC curves



- This is the ROC curve for all possible "pass marks".
- As true positive rate increases with a changed pass mark, false positive rate does too.

ROC curves



- The more the line curves away from the diagonal, the better the test is at discriminating people with disorder from controls.
- The area under the curve is the accuracy of the test.

A real example of using a ROC curve

- Sara Olsen's honours project with Gina Geffen: to find out whether adding a particular neuropsych test (nonword repetition) to an existing battery improves the chance of detecting concussion.
- Concussion group vs controls completed batteries within 24 hours of injury.
- Sara used a discriminant analysis (like a multiple regression but where the DV is discrete – concussion versus no concussion) to come up with the best linear combination of tests that could detect concussion (with and without the new test).
- She then calculated a composite score of all the tests (with and without the new test).

Using ROC curve data to find the 'pass mark' for optimal discrimination of concussion vs no concussion

- Go to SPSS – graphs menu – ROC curve…
- Value of state variable: enter DV (e.g. concussion = 1 vs no concussion = 2)
- Test variable: enter composite test score from battery.
- Tick all boxes and check out 'options' sub menu.
- Note you can choose how to calculate the area under the curve (2 options). Non-parametric is 'safer'.

Using ROC curve data to find the 'pass mark' for optimal discrimination of concussion vs no concussion

- Cut and paste "co-ordinates of the curve" into Excel.
- Work out specificity (=1- (1-specificity)) for each row.
- Add together sensitivity & specificity for each row.
- Look to see where this sum is highest: that's your optimal pass mark (best discrimination) – read off sensitivity and specificity at that point.
- Repeat whole exercise with the new test in the battery.
- Go and work out your 2x2 tables (pre-test probability is % participants with concussion) – did adding the new test correctly classify more people? No…

ROC curves

- Rough guide to accuracy/discrimination between people with disorder and controls (area under ROC curve):
  - .90-1.0 excellent
  - .80-.90 good
  - .70-.80 fair
  - .60-.70 poor
  - .50.-60 fail
  - (.50 is a straight diagonal which is chance, i.e. you are getting the same rate of correct positives as false positives so it's pointless to perform the test).
- See **gim.unmc.edu/dxtests/** for further info.

Signal Detection Theory

- ROC curves are one application of signal detection theory - which should also be used in scoring certain tests.
- Use it whenever you've got a task that involves discriminating between stimuli:
- e.g. the recognition memory task in California Verbal Learning Test - have you seen this word previously or not? (YES or NO).
- If you use % correctly recalled as the score, people can maximise their score by saying they recognise every word.

Signal Detection Theory

- Four possible outcomes:
  - Correct hit
  - False positive/false alarm
  - Correct miss
  - False negative
- People have different criterion for how familiar a word has to feel before they say they recognise it.
- Liberal criterion - they'll say YES even if they only have a vague recollection (will get more correct).
- Conservative criterion - they'll say YES only if they're absolutely certain they remember it.

Signal Detection Theory

- E.g. Older people are less likely to guess (conservative criterion) so they'll get a lower % correct BUT you don't know whether that's because their memory is actually worse or because of their response style from this % alone.
- That is, there's a confound between SENSITIVITY (ability to discriminate between words that you heard previously and those you didn't – ***not*** the same as sensitivity in the context of diagnostic tests) and RESPONSE BIAS (criterion for saying yes).
- You want to measure SENSITIVITY - but if you just look at correct hits then sensitivity is contaminated by RESPONSE BIAS.
- We can disentangle sensitivity and response bias by looking at false positives as well as correct hits.

Signal Detection Theory

- This is what **signal detection theory** does for us (we enter hit rate and false alarm rate into a formula and get out sensitivity and response bias as two separate scores).
- We can then use the sensitivity score (often called d prime or d') instead of hit rate as our raw score on the test.
- See the following for further info (including details of how to do this in SPSS, Excel etc) – this article is available free on the web (search on Google):
  - Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. Behavior Research Methods, Instruments, & Computers, 31(1), 137-149.

PSYC7112 - EXTRAS

POWER ANALYSIS & META ANALYSIS

Power analysis should be done for any empirical project

- Traditionally psychologists use statistical significance as the benchmark for quantitative results.
- Significant results (5% level) are seen as describing effects that are important and substantial.
- Non-significant results are seen as either indicating trivially small effects or even no effect at all.
- This is wrong - but power analysis addresses this issue.
- **These days, papers without power analysis are regularly rejected by journals as a matter of course.**

Sample size/effect size trade off

- Statistical significance just tells you the probability that something is due to chance.
- This depends on the magnitude of the effect and the sensitivity of the experiment (mainly how many people you've tested).
- If you've got a really big effect then it's easier to detect so you don't need so much sensitivity (i.e. you can get away with smaller samples).
- If you've got a small effect then you'll need more sensitivity to detect it as statistically significant (i.e. you'll need to test a larger sample).

Power analysis

- Power analysis tells us how many people we need to detect a certain size of effect as significantly difference from zero.
- It's especially useful for interpreting non-significant results (very typical for student projects!).
- You can do it either by using very simple equations.
- Or by consulting tables (I'll email you a copy of Cohen 1992 which has these tables).
- Or by using free software downloadable from the internet.

Ideal world

- Before running any study, we predict what magnitude of effect we would consider "substantial"/worth knowing about in whatever domain we're looking at (e.g. see literature).
- Then we work out how many people we need to test to stand a decent chance (usually 80%) of finding such a effect size statistically significant.
- Then we design our study to test that many people.
- In practice, the number of people we test is constrained by logistic factors - but it's still worth knowing if you're setting yourself an impossible task or not.

Effect sizes

- If you're doing a study where you're looking at correlations, you can use the correlation coefficient itself as the effect size.
- Cohen defined small, medium, and large effect sizes to act as a rule of thumb.
- You traditionally square correlation coefficients to get a measure of what they actually mean in terms of % of the variance accounted for - though some argue r itself is a better estimate of this (e.g. Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97*(2), 307-315).
- The less people you test, the greater the likely margin of error in the correlation coefficient you get out (likely to be further from the actual population correlation).

To give an 80% chance of detecting these correlations as significant at the 5% level, we need the following numbers of people:

	Correlation (r)	% of shared variance (r^2)	Minimum number of people needed
'Large'	.50	25%	28
'Medium'	.30	9%	85
'Small'	.10	1%	783

So - if you get a non-significant result, it could be because the correlation is non-existent or it could be that you haven't tested enough people.

Effect sizes - differences between 2 means (t-test)

- Consider an experiment with 2 groups where you want to see if there's a "substantial" difference between those two groups (say in reaction time).
- The effect size would be the difference in reaction time between the two groups (in whatever units you measure it in - e.g. milliseconds).
- BUT - it's more useful to standardise it.
- One standardised measure of effect size is Cohen's d (the difference between the groups measured in standard deviations).

To give an 80% chance of detecting these differences between 2 groups as significant at the 5% level, we need the following numbers of people:

	Cohen's d (s.d. between means)	% of group 1 lower/higher than the average of group 2	Minimum number of people needed IN EACH GROUP
'Large'	.80	79%	26
'Medium'	.50	69%	64
'Small'	.20	58%	393

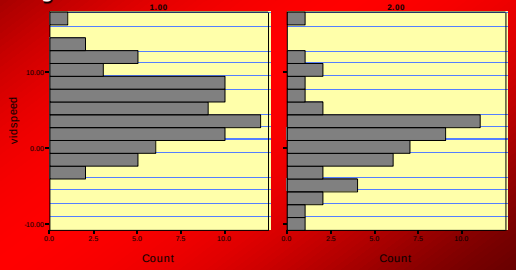
Medium effect size = "apparent to an intelligent viewer" = the difference in heights between men and women at age 18.

Special treat for PSYC7112 students: Cohen's d calculator

- I'll email you an Excel file I've written that calculates Cohen's d when given the means, standard deviations, and group sizes of 2 groups.
- Enter the data into the columns – Cohen's d (and the pooled standard deviation) should appear.

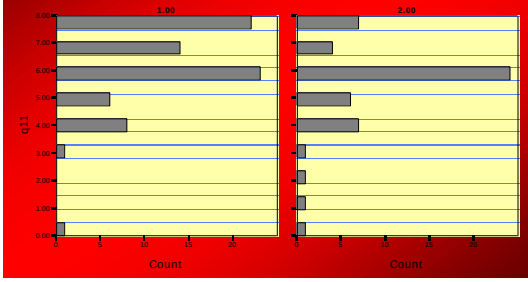
What do different Cohen d sizes actually look like?

- Males (1) versus females (2) – video speed test (miles per hour compared with car in video).
- Large effect size: Cohen's d = .8



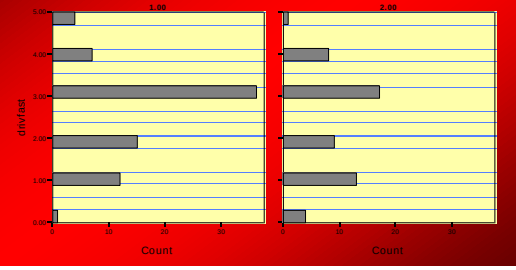
What do different Cohen d sizes actually look like?

- Males (1) versus females (2) – 'I enjoy driving'
- Medium effect size: Cohen's d = .5



What do different Cohen d sizes actually look like?

- Males (1) versus females (2) – 'How often do you drive fast?'
- Smallish effect size: Cohen's d = .25



POWER ANALYSIS

- So - if we expect only a trivially small or zero correlation between two variables then I just need to test **enough people** to give a **good chance** (say, 80%) to detect a **medium-sized correlation** (where I argue that a smaller correlation than this is trivial/unimportant in the circumstances - whether it's due to chance or not).
- THEN if our correlation does indeed turn out to be non-significant, we can say, *"despite having enough people to stand a good chance of detecting a medium correlation, the correlation was not significant."*

POWER ANALYSIS

"There was no significant difference between the mean biases of experts and novices, t(78) = -.149, p = .882, Cohen's d = 0.03. Power analysis revealed that in order for an effect of this size to be detected (80% chance) as significant at the 5% level, a sample of 34886 participants would be required."

Taken from results section of Waylen, Horswill, Alexander, & McKenna (2004), "Do expert drivers have a reduced illusion of superiority?"

See this article for further information: Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

See here for a free power analysis guide and software: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

G*Power 3 guide:

- Click 'test family' for a list of different tests
  - Exact = use this for determining if a correlation is significantly greater than 0.
  - F-tests (= ANOVAs)
  - T-tests
  - Chi² tests
  - z tests
- If you want to work out how many subjects to run in order to find a certain effect size then click on the 'a priori' option in "type of power analysis".
- If you want to find out how much chance you've got of detecting a certain effect size given a certain number of participants, click 'post hoc' instead.

G*Power 3 guide:

- **E.g. what's the chance I'll get a significant result in my 2 group between subject experiment when I've tested 20 people in each group and I'm looking for a medium effect size (2 tailed test)?** (i.e. Independent samples t-test with a Post-Hoc power analysis)
  - Effect size d = .5 (half a SD between the group means)
  - Alpha = .05 (level of significance)
  - Sample size n1 (group 1) = 20 people
  - Sample size n2 (group 2) = 20 people
  - Click 'Calculate'
- Power is .38 (you've got a 38% chance of getting a significant results under these conditions).

G*Power 3 guide:

- **E.g. how many people do I want to test in order to stand a reasonable chance (80%) of detecting a medium effect size when comparing two between-subject groups? (i.e. a priori)**
  - Effect size d = .5 (half a SD between the group means)
  - Alpha = .05 (level of significance)
  - Power = .80 (80%)
  - Let the allocation ratio be 1 (equal numbers in both groups).
  - Click 'Calculate'
- Total sample size needed is 128 people (64 in each group).

META ANALYSIS

- Some have argued that progress in psychological research is slower than it ought to be despite thousands of studies.
- One reason could be how we analyse our data.
- Imagine we have a drug that affects learning - and this drug always increases learning by half a standard deviation (medium effect size, Cohen's d = .5, difference in heights of men and women).
- We do a typical drug study and give 15 people the drug and 15 people get a placebo.

Optimizing cumulative scientific knowledge in psychology

- This study is replicated 100 times.
- However, doing a power analysis tells us that power is 37% (n = 30, d = .5).
- That means out of 100 studies, only 37 will show a significant effect (alpha = 5%, 1 tailed).
- So - when we do our literature review, we review these 100 studies and conclude that (1) the evidence that the drug works is contradictory but (2) on balance, most studies show that the drug doesn't work (voting system).

Optimizing cumulative scientific knowledge in psychology

- Another common interpretation is that we have to determine what differed between the experiments that showed an effect and those that did not.
- The reality is that the drug has worked exactly the same every time (differences are just sampling error).
- Schmidt (1992) & others argue that this is a pervasive problem with psychological studies: they tend to be chronically underpowered due to logistic constraints - and psychologists confuse statistical significance with 'ecological' significance.

Sampling error example:

- Imagine we had a test to predict job performance.
- We try it out on 1428 people and then measure their job performance.
- The correlation between the two (criterion validity coefficient) is .22 (Schmidt et al., 1985).
- In organisational psychology, the average study size is n = 68.
- What happens if we take random samples of 68 people from the overall 1,428 sample to simulate a number of replication studies from organisational psychological research?

21 validity studies (n = 68):

.04	.20	.26*
.14	.02	.17
.31*	.23	.39*
.12	.11	.22
.38*	.21	.21
.27*	.37*	.36*
.15	.14	.29*

- These are the range of correlation coefficients you get picking random samples of 68 people from the overall sample of 1428 people.
- Traditional lit review would conclude 38% studies show effects - only use test in those organisations…

Meta analysis

- Meta-analysis can solve these problems.
- It involves combining the results from a number of similar studies to get a much bigger sample size - which means a more accurate estimate of the true effect size.
- Here's a simple example…

Relationships between self-assessed skill and risk taking

- I've looked at the relationship between self-assessed skill and risk-taking in a number of studies in the past.
- These studies have yielded contradictory findings.
- So – we did a meta-analysis to resolve this discrepancy… (Horswill et al., 2004 - Journal of Applied Social Psychology)

Correlations between self-report speed and self-assessed skill

Study:	**r**	**n**	**p**	**Type of study**
Present study	.11	163	.162	Internet
Horswill, 1994	.18	995	<.001	Postal survey & lab data
McKenna & Horswill, 2002 (Study 1)	.31	126	<.0005	Lab data
McKenna & Horswill, 2002 (Study 2)	.23	400	<.0005	Postal survey & lab data
Meta-analysis	.20		**<.0001**	

How to do that meta-analysis:

To get the 'grand' r:

- Step 1: convert r into r' (Fisher transformation) using tables (e.g. Howell, 2002, p.746).
- Step 2: multiply each r' by its n and then sum them.
- Step 3: divide this by the sum of all the n's.
- Step 4: convert the 'grand' r' back to r (tables).

How to do that meta-analysis:

To get the overall significance:

- Step 1: Convert p into z scores (e.g. Howell, p.759, where p is the 'smaller portion' - divide p by 2 first if two tailed).
- Step 2: multiply each z by its n and then sum them.
- Step 3: sum the squares of each n and square root the result.
- Step 4: divide the number from step 2 by the number from step 3 (this gives you an overall z score).
- Step 5: Convert the z score back to p using table (remember to multiply by 2 if it's 2-tailed).

What if the original results don't report r?

- You can convert any sort of output statistic (Cohen's d, t, F, etc.) into r and back again (see handout for formulae).
- Just convert all the study's results into r and proceed.
- Note that there can be plenty more sophistication in a meta-analysis compared with what I've just shown - this is just a very simple example to give you an idea.