Effect Sizes

Null Hypothesis Significance Testing (NHST)

When you read an empirical paper, the first question you should ask is 'how important is the effect obtained'. When carrying out research we collect data, carry out some form of statistical analysis on the data (for example, a *t*-test or ANOVA) which gives us a value known as a *test statistic*. This test statistic is then compared to a known distribution of values of that statistic that enables us to work out how likely it is to get the value we have *if* there were no effect in the population (i.e. if the null hypothesis were true). If it is very unlikely that we would get a test statistic of the magnitude we have (typically, if the probability of getting the observed test statistic is less than .05) then we attribute this unlikely event to an effect in our data (see Field, 2005). We say the effect is 'statistically significant'. This is known as Null Hypothesis Significance Testing (NHST for short).

NHST is used throughout psychology (and most other sciences) and is what you have been taught for the past 18 months (and for many psychology undergraduates it is all they are ever taught). It may, therefore, surprise you to know that it is a deeply flawed process for many reasons. Here are what some much respected statistics experts have to say about NHST (thanks to Dan Wright for some of these quotes – see web/acknowledgements sections of this handout):

Schmidt & Hunter (2002):

> "Significance testing almost invariably retards the search for knowledge by producing false conclusions about research literature" (p. 65).
>
> "Significance tests are a disastrous method for testing hypotheses" (p. 65)

Meehl (1978):

> "The almost universal reliance on merely refuting the null hypothesis is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" (p. 817).

Cohen (1994):

> "NHST; I resisted the temptation to call it Statistical Hypothesis Inference Testing". (p. 997)

Reason 1: NHST is Misunderstood

Many social scientists (not just students) misunderstand what the *p* value in NHST actually represents. If I were to ask you what *p* actually means which answer would you pick:

a) *p* is the probability that the results are due to chance, the probability that the null hypothesis (H0) is true.

b) *p* is the probability that the results are not due to chance, the probability that the null hypothesis (H0) is false.

c) *p* is the probability of observing results as extreme (or more) as observed, if the null hypothesis (H0) is true.

d) *p* is the probability that the results would be replicated if the experiment was conducted a second time.

e) None of these.

Someone did actually ask undergraduates at a UK university (not Sussex, but I won't name and shame) this question on a questionnaire and 80% chose (a) although the correct answer is

(c) – you will have known this if you read the first paragraph of this handout☺ Only 5% correctly chose (c)[1]. As such, many people who use NHST are not testing what they think they're testing and consequently the conclusions they draw are incorrect (because they are based on erroneous beliefs about what *p* means).

Reason 2: The Null Hypothesis is Never True

Now we know the correct interpretation of *p* we can think about the consequences of it. As we have seen *p* is the probability of observing results as extreme (or more) as observed, if the null hypothesis (H0) is true. There is one very important problem with *p* which is that for social science data the null hypothesis is never true (see Cohen, 1990). As such, *p* is completely meaningless because it's based on an assumption that can never be true!

Remember that the null hypothesis is that there is no effect in the population. Cohen points out that the null hypothesis is never true because we know from sampling distributions (see Field, 2005a, section 1.6) that two random samples will have slightly different means, and even though these differences can be very small (e.g. one mean might be 10 and another might be 10.00001) they are nevertheless different. In fact, even such a small difference would be deemed as statistically significant if a big enough sample were used (see Reason 3).

As such, a non-significant result should never be interpreted (despite the fact it often is) as 'no difference between means' or 'no relationship between variables'. So, significance testing can never tell us that the null hypothesis is true, because it never is! As such the idea of 'accepting the null hypothesis' is just plain wrong.

Reason 3: NHST depends upon Sample Size

Imagine we were interested in whether listening to Cradle of Filth (CoF) turns people into Granny-murdering devil-worshippers. We could (assuming ethics committees didn't exist and our moral values were very low indeed) expose unborn children to Cradle of Filth (or not) and see how they turn out years later. So, we have two groups (exposed to CoF in the womb vs. unexposed control group) and some kind of outcome variable (number of pentagrams drawn on the wall in blood). We could subject these to a simple *t*-test.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pentagrams Drawn	Equal variances assumed	1.046	.308	-6.222	198	.000	-2.20699	.35472	-2.90650	-1.50748
	Equal variances not assumed			-6.222	195.274	.000	-2.20699	.35472	-2.90656	-1.50742

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pentagrams Drawn	Equal variances assumed	2.744	.136	-1.510	8	.169	-2.20736	1.46173	-5.57810	1.16339
	Equal variances not assumed			-1.510	5.225	.189	-2.20736	1.46173	-5.91660	1.50188

SPSS Output 1

[1] Thanks to Dan Wright for these data.

SPSS Output 1 shows the results of two independent *t*-tests done on the same scenario. In both cases the difference between means is —2.21 so these tests are testing the same difference between means. Look at the associated *t*-values and significance though. In one case *t* is highly significant ($p < .001$), but in the bottom table it is not ($p = .169$). How can this be: in both cases the tests are testing the same mean difference of –2.21?

Before I answer that question, let's look at another scenario (again, with CoF and granny murdering!). SPSS Output 2 shows another *t*-test, which yields a significant result ($p < .05$). Nothing strange about that you might think, but have a look at the 'Mean Difference'. The value is 0. Now, the *t*-test tests the null hypothesis that the difference between means is 0, therefore, if the difference between means really is zero then the resulting *t* should not be significant! How is it possible that this test is telling us that there is a significant difference between means when we know the difference is 0?

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pentagrams Drawn	Equal variances assumed	.994	.319	-2.296	999998	.022	.00	.00200	-.00851	-.00067
	Equal variances not assumed			-2.296	999997.4	.022	.00	.00200	-.00851	-.00067

SPSS Output 2

The answer to these two anomalies is the same: NHST depends on the sample size of the observed data. In the first case, the reason why it is possible to have the same mean difference but a difference in the significance of the resulting test is because the two tests were based on different samples: the significant test was based on a sample of 200, whereas the non-significant test was based on a sample of only 10! So, significance depends on your sample size.

In the later case, the reason why a difference of 0 is significant is because this test is based on a sample of 1 million data points. The mean difference is very small (-0.0046 or .00 to 2 decimal places) but given a big enough sample, it is significantly different from zero.

So, a trivial difference can be significant (and interpreted as 'important') if the sample is big enough, conversely, a big and important difference can be non-significant (and interpreted as 'trivial') in a small sample (see Field & Hole, 2003).

Reason 4: NHST is Illogical

NHST is based on probabilistic reasoning, which severely limits what we can conclude. Cohen (1994), points out that formal reasoning relies on an initial statement of fact followed by a statement about the current state of affairs, and an inferred conclusion. This syllogism from Field (2005a) illustrates what I mean:

- If a man has no arms then he can't play guitar.
    - This man plays guitar.
    - Therefore, this man has arms.

The syllogism starts with a statement of fact that allows the end conclusion to be reached because you can deny the man has no arms by denying that he can't play guitar. However, the null hypothesis is not represented in this way because it is based on probabilities. Instead it should be stated as follows:

- If the null hypothesis is correct, then this test statistic is highly unlikely.

- This test statistic has occurred.
- Therefore, the null hypothesis is highly unlikely.

If we go back to Field's (2005a) guitar example a similar statement would be:

- If a man plays guitar then he probably doesn't play for Fugazi (this is true because there are thousands of people who play guitar but only two who play guitar in the band Fugazi!).
  - Guy Picciotto plays for Fugazi.
  - Therefore, Guy Picciotto probably doesn't play guitar.

This is illogical: the conclusion is wrong because Guy Picciotto does play guitar. By extension, it should be clear that NHST allows very little to be said about the null hypothesis (see Cohen, 1994; Field, 2005a for more detail).

Reason 5: $p < .05$ is completely arbitrary!

Why do we use a $p < .05$ as a criterion for accepting significance? Well, essentially it's because Fisher said so. There is no 'magic' behind .05, it's just a reasonable figure and what Fisher decided was appropriate to be sufficiently confident that a genuine effect exists. To some extent it's arbitrary (for more detail on how Fisher reached this value see Field, 2005a or Field & Hole, 2003): IF Fisher had woken up in a 10% kind of mood we would all be working with $p < .10$ as our criterion. As such, it can be tempting in NHST to attribute 'importance' to an effect with $p = .04$, but assume that an effect with $p = .06$ is unimportant: in equal sample sizes, these effects are actually likely to be very similar!

What is an effect size?

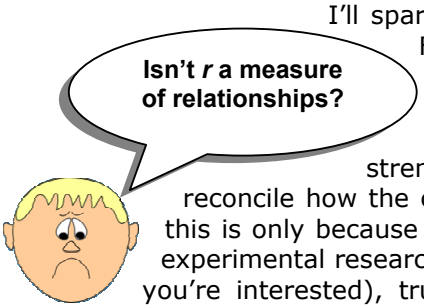
There is no magical answer to the problems with NHST (although see Cohen, 1994; Schmidt & Hunter, 2002 for some suggestions). However, one thing that can be used in conjunction with NHST (not necessarily as a replacement for) are effect sizes.

An effect size is simply an objective and standardized measure of the magnitude of observed effect (see Field, 2005a; 2005b). The fact that the measure is standardized just means that we can compare effect sizes across different studies that have measured different variables, or have used different scales of measurement. So, an effect size based on the Beck depression inventory could be compared to an effect size based on levels of serotonin in blood.

Effect Size Measures

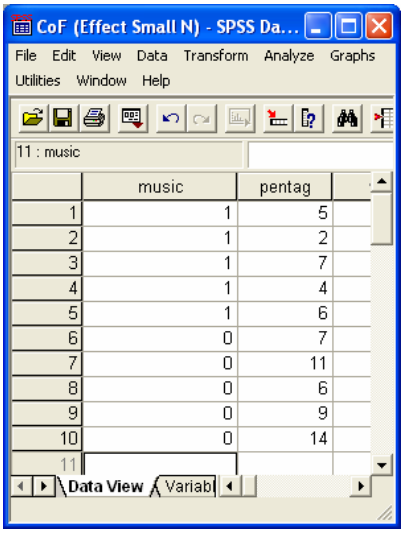
Many measures of effect size have been proposed, the most common of which are Cohen's, d, and Pearson's correlation coefficient, r (although there are others such as Hedges' g, Glass' Δ, odds ratios and risk rates: see Rosenthal, 1991).

For the purpose of this handout I'm going to stick with one of these: the correlation coefficient. There are three reasons for this: (1) it is probably the most common effect size measure (Field, 2001, in press; Rosenthal & DiMatteo, 2001); (2) you will be familiar with it already; and (3) it is incredibly versatile.



I'll spare you the details of the correlation coefficient (you can read Field, 2005a, chapter 4 if you're that interested). Many of you will be familiar with the correlation coefficient as a measure of the strength of relationship between two continuous variables; however, it is also a very versatile measure of the strength of an experimental effect. It may be difficult for you to reconcile how the correlation coefficient can also be used in this way; however, this is only because students are often taught about it within the context of non-experimental research. Although I don't want to get into it now (see Field, 2005a if you're interested), trust me that r can be used to express differences between

means and this is the measure that I prefer because it is constrained to lie between 0 (no effect) and 1 (a perfect effect) and is familiar to almost all students and researchers.



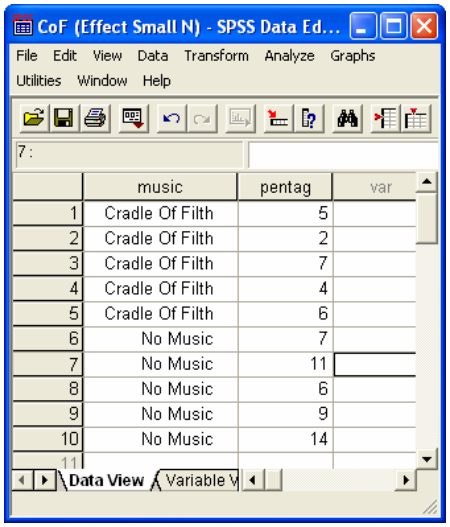


Figure 1

OK, so you don't believe me. Well, let's use our Cradle of Filth example again to look at a classic difference between means scenario. The data are in Figure 1. This shows the number of pentagrams drawn by children exposed in the womb to CoF compared to controls. First let's do an independent *t*-test on these means.

Group Statistics

	Music Listened To	N	Mean	Std. Deviation	Std. Error Mean
Pentagrams Drawn	No Music	5	9.4000	3.20936	1.43527
	Cradle Of Filth	5	4.8000	1.92354	.86023

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pentagrams Drawn	Equal variances assumed	1.454	.262	2.749	8	.025	4.6000	1.67332	.74132	8.45868
	Equal variances not assumed			2.749	6.545	.031	4.6000	1.67332	.58682	8.61318

SPSS Output 3

SPSS Output 3 shows the result of this *t*-test. We could conclude that listening to cradle of filth makes you less likely to draw pentagrams (the mean for the CoF condition is significantly lower than for controls). We could write that children exposed to CoF in the womb drew significantly fewer pentagrams than those exposed to no music, $t(8) = 2.75$, $p < .05$. Ok, let's now do a simple Pearson correlation on the same data (go, on try it for yourself). I know, it sounds crazy, but let's just see what happens.

SPSS Output 4 shows the results of this analysis. Note the significance of the correlation coefficient: it's .025. Now look back at the significance of *t* from SPSS Output 3: it's .025 also. That's odd isn't it? Well, no it's not, what we've just done is perfectly legitimate, the correlation expresses the difference between these two groups: in fact, provided you code your two groups with 0s and 1s, you can conduct a Pearson correlation on the data and the end result expresses the 'relationship' between the groups and the outcome variable. In fact, this

'relationship' is simply the difference between group means! This is known as a point-biserial correlation (see Field, 2005a, chapter 4).

Correlations

		Music Listened To	Pentagrams Drawn
Music Listened To	Pearson Correlation	1	-.697*
	Sig. (2-tailed)	.	.025
	N	10	10
Pentagrams Drawn	Pearson Correlation	-.697*	1
	Sig. (2-tailed)	.025	.
	N	10	10

*. Correlation is significant at the 0.05 level (2-tailed).

SPSS Output 4

Hopefully, I've demonstrated that *r* can be used to express the difference between two groups, but why did *r* and *t* have the same significance value. The answer is that the two statistics are directly related: *r* can be easily obtained from several common test statistics. For example, when a *t*-test has been used *r* is a function of the observed *t*-value and the degrees of freedom on which it is based:

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

With the data above, we get:

$$r = \sqrt{\frac{2.749^2}{2.749^2 + 8}} = .697 ,$$

which is exactly the same as calculating a Pearson *r* for the original data (see SPSS Output 4)! The reason why the minus sign has gone is because when we calculate Pearson *r* for these kind of data, the sign of the correlation entirely depends on which way you code the groups (try running a correlation on the same data but change it so that Cradle of Filth are represented by a code of 1 and 'No Music' by a code of 0: the resulting *r* will be positive).

In fact, *r* can also be obtained from a variety of other test statistics. When ANOVA has been used and an *F*-ratio is the test statistic, then when there is 1 degree of freedom for the effect, the following conversion can be used:

$$r = \sqrt{\frac{F(1,-)}{F(1,-) + df_R}}$$

In which F(1,-) is simply the *F*-ratio for the effect (which must have 1 degree of freedom) and df_R is the degrees of freedom for the error term on which the *F*-ratio is based.

The reason the degrees of freedom for the effect need to be 1 is simply because this means that 2 things are being compared. It's difficult to interpret effect sizes for complex effects involving lots of groups because you have no idea which groups contribute to the effect. So, it's best to calculate effect sizes for focussed comparisons such as comparisons of two groups or interactions with only 1 degree of freedom.

"What about categorical data?" I hear you ask. No problem, *r* can also be used to express relationships in categorical data because it is directly related to the chi-square statistic (again, provided this chi-square statistic has only 1 degree of freedom):

$$r = \sqrt{\frac{\chi^2(1)}{N}}$$

Finally, *r* can be calculated from the probability value of a test-statistic. First, you must convert the probability into a *z*-score using tabulated values of the normal distribution (see Field, 2005a), and then simply divide the resulting *z* by the square root of the total sample size on which it is based:

$$r = \frac{Z}{\sqrt{N}}$$

Why are effect sizes useful?

Effect sizes are useful because they provide an objective measure of the importance of an effect. It doesn't matter what effect you're looking for, what variables have been measured, or how those variables have been measured we know that a correlation coefficient of 0 means there is no effect, and a value of 1 means that there is a perfect effect. Cohen (1992, 1988) has made some widely accepted suggestions about what constitutes a large or small effect:

- r = 0.10 (small effect): in this case, the effect explains 1% of the total variance.
- r = 0.30 (medium effect): the effect accounts for 9% of the total variance.
- r = 0.50 (large effect): the effect accounts for 25% of the variance.

We can use these guidelines to assess the importance of our effects (regardless of the significance of the test statistic). However, r is not measured on a linear scale so an effect with r = 0.4 isn't twice as big as one with r = 0.2.

For example, in our earlier Cradle of Filth example, we had three sets of data on which we conducted *t*-tests. Figure 2, shows these data and the significance of these tests. You'll remember we had two samples with similar differences between means, but different significances and one data set with a near zero difference between means that was highly significant. If we calculate the effect sizes using the values of *t* in SPSS outputs 1 and 2 we find that our two samples with similar mean difference yield similar effect sizes (.40 and .47); also our sample with a near zero difference between means produces an effect size of zero. As such the effect sizes better reflect what's going on than the significance of the test statistic!

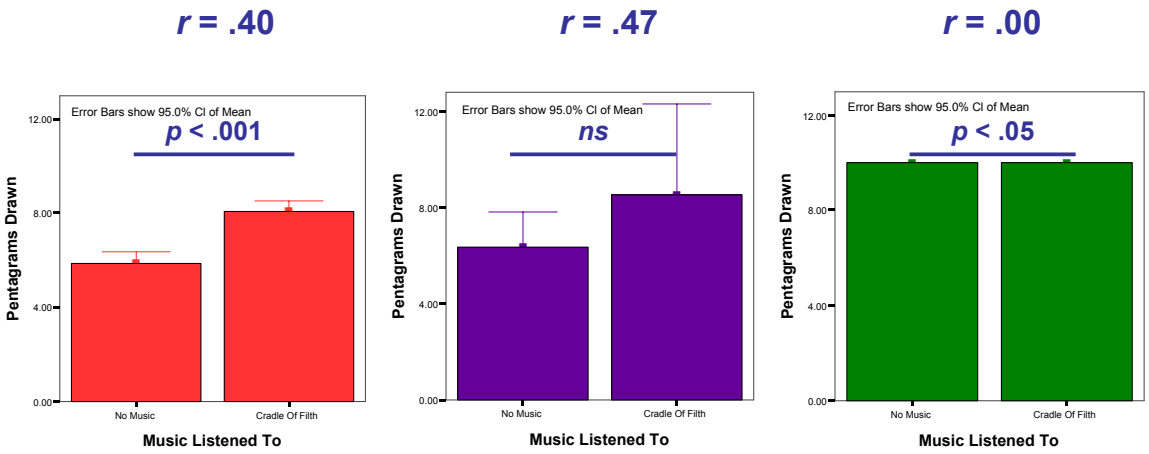


Figure 2

Finally, the effect size in the sample is not actually that interesting: it is the size of the effect in the population that is important. However, because we don't have access to this value, we use the effect size in the sample to estimate the likely size of the effect in the population (see Field, 2001). This can be done by assimilating effect sizes from similar studies using a technique called meta-analysis (see Field, 2001, 2005b, 2003; Rosenthal & DiMatteo, 2001), but that will have to wait for another time!

Further Reading

Clark-Carter, D. (2003). Effect size: The missing piece of the jigsaw. *The Psychologist*, 16 (12), 636-638.

Field, A. P. (2005). *Discovering statistics using SPSS* (2nd edition). London: Sage.

Wright, D. B. (2003). Making friends with your data: improving how statistics are conducted and reported. *British Journal of Educational Psychology*, *73*, 123–136.

References

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Field, A. P. (in press). Is the meta-analysis of correlation coefficients accurate when population effect sizes vary? *Psychological Methods*.

Field, A. P. (2005a). *Discovering statistics using SPSS* (2nd edition). London: Sage.

Field, A. P. (2005b). *Meta-analysis*. In J. Miles & P. Gilbert (eds.) *A handbook of research methods in clinical and health psychology* (pp. 295-308). Oxford: Oxford University Press.

Field, A. P. (2001). Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, *6* (2), 161–180.

Field, A. P., & Hole, G. (2003). *How to design and report experiments*. London: Sage.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806-834.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (revised). Newbury Park, CA: Sage.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods and literature reviews. *Annual Review of Psychology*, *52*, 59-82

Schmidt, F., & Hunter, J. (2002). Are there benefits from NHST? *American Psychologist*, *57*(1), 65-66.

Web

See Dan Wright's pages on NHST (and a very useful list of links under the 'extra bits'):

http://www.sussex.ac.uk/Users/danw/masters/qual%20and%20quant/ills_of_nhst.htm

Acknowledgement

Some of the NHST material was directly 'borrowed' from Dan's website (above), for which I'm very grateful. More indirectly he has been very influential over the years in encouraging me to read various things about NHST to which I would have remained oblivious had I not known

him. So, although he cannot be blamed for my obsession with Cradle of Filth, his influence permeates much of this handout.

Task

Lotze et al (2001) did a study to see what areas of the brain were activated during anal stimulation: they inserted balloons (not party ones) into people's rectums and inflated them while the person was in an fMRI scanner. Then they sang happy birthday and … OK, they didn't, but they really did do the balloon thing (see NeuroImage, *14*, 1027-1034 if you don't believe me). One of the areas of the brain in which they were interested was the secondary somatosensory cortex (S2). Let's imagine we replicated this study, but looked at two levels of balloon inflation: 200ml and 400ml. These inflations were done on the *same* participants. The data below show the level of S2 activation when the anal balloon was inflated to 200ml and 400ml.

200ml	400ml
3.87	6.01
4.05	8.93
5.78	9.24
6.89	7.28
6.98	8.72
7.06	7.86
7.54	8.98
10.92	12.50

- ✓ Enter the data into SPSS.
- ✓ Conduct a *t*-test on the data to compare the means of the two conditions.
- ✓ Use the *t*-statistic to compute an effect size
- ✓ Write a mini-results section reporting the results and effect size in APA format.