Research Methods

Effect Size

Andy Gillett

1

1

Statistical Significance

For many statistically oriented social scientists, quantitative analysis is virtually synonymous with significance testing. The whole point and purpose of the exercise is taken to be 'have we got a significant result?' Is $p < 0.05$?' This refers to *statistical significance.*

(Robson, 2011, p. 446)

2

2

Statistical Significance

If a particular result is marked as significant, we can utter a sigh of relief as this means that the observed phenomenon represents a significant departure from what might be expected by chance alone. That is, it can be assumed to be real and we can report it. However, if a particular result is not significant, we must not report it (for example, if the mean scores of two groups are not significantly different, we *must not* say that one is larger than the other even if the descriptive statistics do show some difference!).

(Dörnyei, 2007, p. 210)

3

3

Statistical Significance

Conventional levels of α in the social sciences are either .05 or .01. When other levels are specified, they tend to be even lower, sauch as α = 0.001. It is rare for researchers to specify α levels of higher than .05. The main reason is editorial policy: Manuscripts may be rejected for publication if α > .05.

(Kline, 2004, pp. 38-39)

4

4

Statistical Significance

It is important to stress here that statistical significance is *not* the final answer because even though a result may be statistically significant (i.e. reliable and true to the larger population), it may not be valuable. 'Significance' in the statistical sense only means 'probably true' rather than 'important'.

Dörnyei (2007, p. 210)

5

5

Statistical Significance

In one example we obtained a statistically significant result, but the actual difference in the mean scores of the groups was very small (21.36, 22.10, 22.96)....

With a large enough sample (in this case, N = 435) quite small differences can become statistically significant, *even* if the difference between the groups is of little practical importance..

(Pallant, 2016, pp. 258-260)

6

6

Effect Size

One way that you can assess the importance of your finding is to calculate the 'effect size' (also known as 'strength of association'). This is a set of statistics that indicates the relative magnitude of the differences between means, or the amount of the total variance in the dependent variable that is predictable from knowledge of the levels of the independent variable (Tabachnick & Fidell, 2013, p. 54).

(Pallant, 2016, p. 212)

7

7

Effect Size

Effect sizes either measure the sizes of associations between variables or the sizes of differences between group means.

Several measures have been proposed and used.

The more frequent effect size measures are:

- coefficient of determination – R^2 ;
- the standardized mean difference - Cohen's d - and
- η^2 - eta squared - the correlation ratio.

IBM SPSS Statistics calculates partial eta squared as part of the output from some techniques (e.g. analysis of variance), and Cohen's d for t-tests.

8

8

Describing Data using Statistics

Description

Inference

- ❑ Exploring Relationships
- ❑ Comparing Groups

9

9

Correlation

There is much interest in whether there is a relationship between pupils' ability at sport and their maths ability. Accordingly a study was carried out in which both sport and maths ability were rated on each of 10 pupils sampled at random from a class of first year pupils in a secondary school. Sport ability was rated on a scale of 1 to 20. Maths ability was measured by a pencil and paper test on a scale from 1 to 30.

10

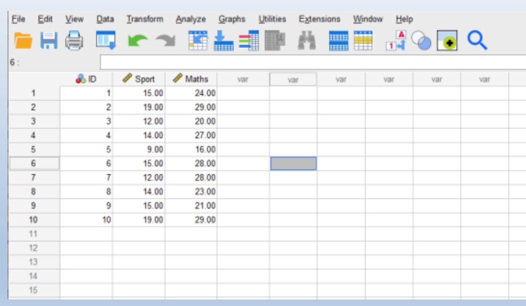
10

Correlation

Pupil No.	1	2	3	4	5	6	7	8	9	10
Sports score	15	19	12	14	9	16	20	14	15	19
Maths score	24	29	20	27	16	28	28	23	21	29

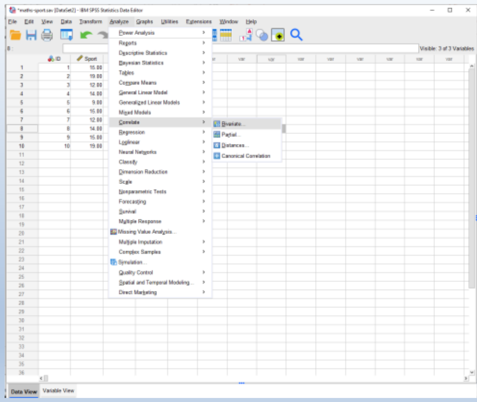
11

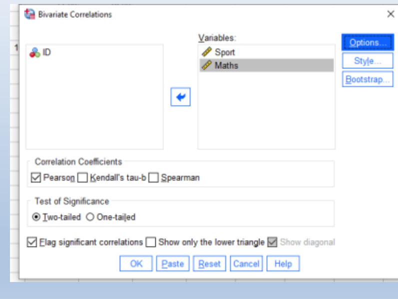
11



12

12





Descriptive Statistics

	Mean	Std. Deviation	N
Sport	14.4000	3.06232	10
Maths	24.5000	4.45346	10

Correlations

		Sport	Maths
Sport	Pearson Correlation	1	.725*
	Sig. (2-tailed)		.018
	N	10	10
Maths	Pearson Correlation	.725*	1
	Sig. (2-tailed)	.018	
	N	10	10

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

The correlation, r, has the value 0.725.

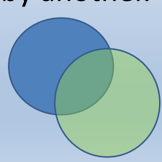
The *p*-value is the probability that the correlation found between the two variables could be a chance result. As *p* here is less than 0.05 the correlation is unlikely to be a chance result.

The conclusion is that there is a relationship between SPORT scores and MATHS scores in the wider population of first year secondary school pupils. The results can be generalised.

Coefficient of Determination (R^2)

$R^2 \equiv r^2$

A measure of how much variance is shared, how much of one variable can be explained by another.



Coefficient of Determination

Figure 15.3 **No common variance.**

If the correlation (r) between variable X and variable Y = 0, then the coefficient of determination = 0^2 = 0 or 0%. There is no common variability.

(Burns & Burns, 2008, p. 349)

Coefficient of Determination

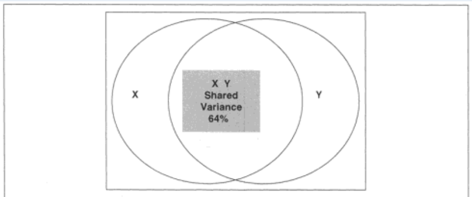


Figure 15.4 **64% shared variance with r = 0.8.**

If the correlation (r) between variable X and variable Y = 0.8, then the coefficient of determination = 0.8^2 = .64 or 64%. 64% of the variance in one variable is predictable from the variance in the other.

(Burns & Burns, 2008, p. 349)

Coefficient of Determination

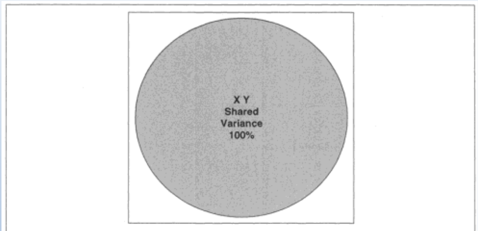


Figure 15.5 **All variability common.**

If the correlation (r) between variable X and variable Y = 1, then the coefficient of determination = 1^2 = 1 x 100 = 100%: 100% of the variability is common to both variables.

(Burns & Burns, 2008, p. 350)

Coefficient of Determination

Converted to a percentage, 0% indicates that the model explains **none** of the variability of the dependent variable. 100% indicates that the model explains **all** the variability.

Size of Effect	r	R^2
Small	.1 - .3	.01 – .09
Medium	-3 - .5	.09 – .25
Large	.5 – 1.0	.25 – 1.0

(Cohen, 1988, 1992)

Coefficient of Determination

R^2

A measure of how much variance is shared, how much of one variable can be explained by another.

Effect Size

Describing Data using Statistics

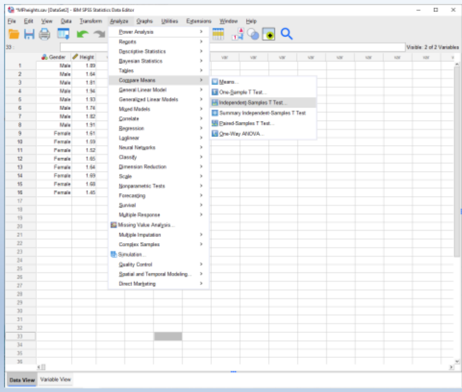
Description

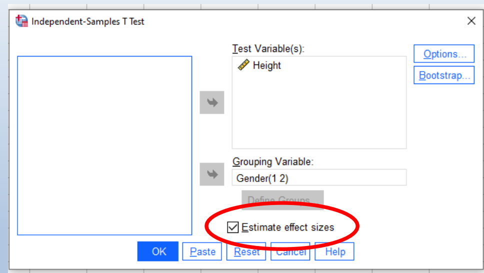
Inference

- ❑ Exploring Relationships
- ❑ Comparing Groups

t-test

	Gender	Height
1	Male	1.89
2	Male	1.64
3	Male	1.81
4	Male	1.94
5	Male	1.93
6	Male	1.74
7	Male	1.82
8	Male	1.91
9	Female	1.61
10	Female	1.59
11	Female	1.52
12	Female	1.65
13	Female	1.64
14	Female	1.69
15	Female	1.68
16	Female	1.45





Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Height	Male	8	1.8350	.10461	.03698
	Female	8	1.6038	.08245	.02915

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
Height	Equal variances assumed	.569	.463	4.911	14	.000	.23125	.04709	.13025	.33225
	Equal variances not assumed			4.911	13.275	.000	.23125	.04709	.12973	.33277

Independent Samples Effect Sizes

				95% Confidence Interval	
		Standardizer[a]	Point Estimate	Lower	Upper
Height	Cohen's d	.09418	2.455	1.101	3.763
	Hedges' correction	.09963	2.321	1.041	3.557
	Glass's delta	.08245	2.805	1.038	4.516

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control group.

Cohen's d - Effect Size for t-Tests

Cohen's d

Size of Effect	d
Small	.2
Medium	.5
Large	.8

(Cohen, 1988, 1992)

Comparing Groups (3 groups or more)

An experiment was carried out in which a sample of adults was selected and randomly allocated to four groups. All were asked to consume amounts of alcohol as follows:

Group 1: NO ALCOHOL
Group 2: ONE UNIT OF ALCOHOL
Group 3: TWO UNITS OF ALCOHOL
Group 4: THREE UNITS OF ALCOHOL

after which they were asked drive a car simulator and perform an emergency stop at 40 miles per hour. The distance to stopping the simulator in metres by each group member is shown below.

31

Comparing Groups (3 groups or more)

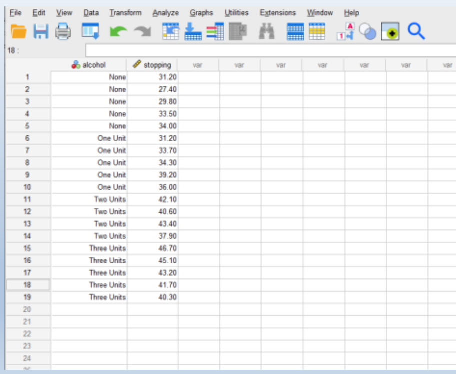
The results were as follows:

NO ALCOHOL	**31.2**	**27.4**	**29.8**	**33.5**	**34.0**
ONE UNIT	**31.2**	**33.7**	**34.3**	**39.2**	**36.0**
TWO UNITS	**42.1**	**40.6**	**43.4**	**37.9**	**39.0**
THREE UNITS	**46.7**	**45.1**	**43.2**	**41.7**	**40.3**

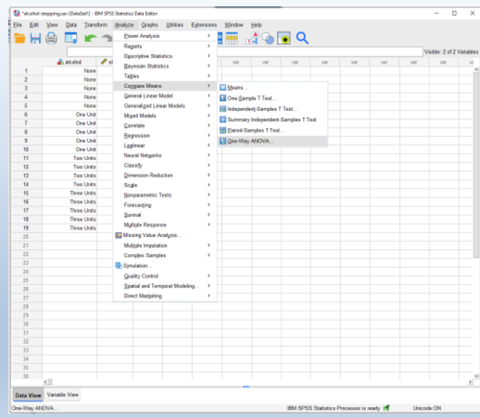
Stopping Distance in Metres

32

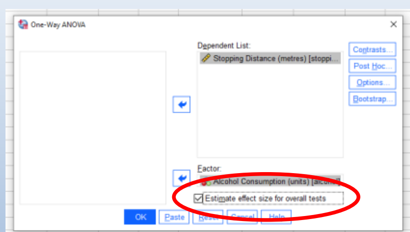
ANOVA



33



34



35

Descriptives

Stopping Distance (metres)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean: Lower Bound	95% Confidence Interval for Mean: Upper Bound	Minimum	Maximum
None	5	31.1800	2.71698	1.21507	27.8064	34.5536	27.40	34.00
One Unit	5	34.8800	2.96597	1.32642	31.1973	38.5627	31.20	39.20
Two Units	4	41.0000	2.36220	1.18110	37.2412	44.7588	37.90	43.40
Three Units	5	43.4000	2.56515	1.14717	40.2149	46.5851	40.30	46.70
Total	19	37.4368	5.60151	1.28507	34.7370	40.1367	27.40	46.70

36

Significance

ANOVA

Stopping Distance (metres)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	457.008	3	152.336	21.202	.000
Within Groups	107.776	15	7.185		
Total	564.784	18			

$p = < 0.05$
Therefore significant

The conclusion is that there is a significant difference between the distance it takes to stop a car simulator according to the amount of alcohol consumed.

Effect Size

ANOVA Effect Sizes[a]

		Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Stopping Distance (metres)	Eta-squared	.809	.492	.870
	Epsilon-squared	.771	.390	.844
	Omega-squared Fixed-effect	.761	.377	.837
	Omega-squared Random-effect	.515	.168	.631

a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

η^2 - Effect Size for ANOVA

η^2 - eta squared - the correlation ratio

Size of Effect	η^2
Small	.2
Medium	.5
Large	.8

(Cohen, 1988, 1992)

A Large Effect

ANOVA Effect Sizes[a]

		Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Stopping Distance (metres)	Eta-squared	.809	.492	.870
	Epsilon-squared	.771	.390	.844
	Omega-squared Fixed-effect	.761	.377	.837
	Omega-squared Random-effect	.515	.168	.631

a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

ANCOVA

Tests of Between-Subjects Effects

Dependent Variable: Final Test Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	470.422[a]	4	117.606	17.449	.000	.833	69.794	1.000
Intercept	212.601	1	212.601	31.543	.000	.693	31.543	.999
pretest	13.414	1	13.414	1.990	.180	.124	1.990	.260
method	430.354	3	143.451	21.283	.000	.820	63.849	1.000
Error	94.362	14	6.740					
Total	27193.610	19						
Corrected Total	564.784	18						

a. R Squared = .833 (Adjusted R Squared = .785)

b. Computed using alpha = .05

Significant Effect For Method ($p<0.001$)

η^2 (Eta Squared) = 0.82

Pearson *r* or correlation coefficient

Pearson's correlation, often denoted *r* and introduced by Karl Pearson, is widely used as an *effect size* when paired quantitative data are available.

For instance if one were studying the relationship between birth weight and longevity. The correlation coefficient can also be used when the data are binary. Pearson's *r* can vary in magnitude from −1 to 1, with −1 indicating a perfect negative linear relation, 1 indicating a perfect positive linear relation, and 0 indicating no linear relation between two variables.

Coefficient of determination (R^2)

A related *effect size* is r^2, the coefficient of determination (also referred to as R^2 or "*r*-squared"), calculated as the square of the Pearson correlation *r*. In the case of paired data, this is a measure of the proportion of variance shared by the two variables, and varies from 0 to 1.

For example, with an *r* of 0.21 the coefficient of determination is 0.0441, meaning that 4.4% of the variance of either variable is shared with the other variable. The r^2 is always positive, so does not convey the direction of the correlation between the two variables.

43

43

(Partial) Eta-squared (η^2)

(Partial) eta squared effect size statistics indicate the proportion of variance of the dependent variable that is explained by the independent variable.
Values can range from 0 to 1.
It describes the ratio of variance explained in the dependent variable by a predictor while controlling for other predictors, making it analogous to the R^2.

44

44

Effect Size: Cohen's d

Cohen's d presents the difference between groups in terms of standard deviation units.

$$\textbf{Effect Size} = \frac{\textbf{[Mean of experimental group]} - \textbf{[Mean of control group]}}{\textbf{Standard Deviation}}$$

Cohen (1988, pp. 24-27) proposed that *d* = 0.2 be considered a 'small' effect size, 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size.
This means that if two groups' means don't differ by 0.2 standard deviations or more, the difference is trivial, even if it is statistically significant.

45

45

Effect Size: Chi-squared (χ^2)

Chi-squared looks at the relationships between two or more categorical variables.

Commonly quoted effect sizes are Phi (for 2 x 2 studies) or Cramer's V (for larger studies).

Cohen (1988, pp. 24-27) proposed that 0.1 be considered a 'small' effect size, 0.3 represents a 'medium' effect size and 0.5 a 'large' effect size.

46

46

Effect Size: Chi-squared (χ^2)

What's your main course? * What's your gender? Crosstabulation

Count

		What's your gender?		
		female	male	Total
What's your main course?	Education	54	8	62
	LInguistics	7	28	35
	Economics	12	21	33
	Social Sciences	15	22	37
	Other	4	12	16
Total		92	91	183

sex	major
female	Economics
female	Social Sciences
male	Other
male	Education
female	Social Sciences
female	Economics
male	LInguistics
male	Social Sciences
female	Economics
female	Social Sciences
female	Social Sciences
male	Education
male	Education
female	Social Sciences
male	Social Sciences
male	Social Sciences
male	LInguistics
male	Social Sciences
male	LInguistics
female	Education
male	LInguistics
female	Other
male	Economics
male	Economics
male	LInguistics
female	Education
male	Economics
female	Education
female	Education

47

Effect Size: Chi-squared – A large effect

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	54.504[a]	4	<.001
Likelihood Ratio	59.758	4	<.001
Linear-by-Linear Association	25.597	1	<.001
N of Valid Cases	183		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.96.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.546	<.001
	Cramer's V	.546	<.001
N of Valid Cases		183	

48

48

Effect Size

A statistically significant tells us that the independent variable has a *significant* effect on the dependent variable.

However, it does *not* tell us is how large an effect the independent variable has.

Measures of effect size are used to determine the strength of the relationship between the independent and dependent variables. Measures of effect size reflect how large the effect of an independent variable is.

49

49

Statistical Significance

There is one last issue I would like to address here. That is the case where the p-value is greater than .05 but the effect size is large. A p-value greater than .05 usually results in the author stating that the results are not statistically significant, but this can be a misleading practice if the effect size is, in fact, large.

(Larson-Hall, 2012, p. 249)

50

50

Statistical Significance

One problem is that the a =.05 cut-off point is an arbitrary dividing line. The biggest problem, however, is the fact that sample sizes play an outsized role in the determination of p-values (Cohen, 1994; Kline, 2004).

If a study has a small sample size but a large effect size then we would expect that if the study were repeated with more participants the p-value would become smaller. In such cases it is perfectly legitimate for an author to put forth their case that with larger sample sizes, the p-value would likely be lower than the cut-off and thus there is at least a case to be made for the importance of the result.

Further research needs to be done but a large effect size means that the result should not be summarily dismissed. Effect size is a much better indicator of the importance of a result than the p-value since it does not depend on sample size (Kline, 2004).

(Larson-Hall, 2012, p. 249)

51

51

Why report effect sizes

Effect size statistics provide an indication of the magnitude of the differences between your groups (not just whether the difference could have occurred by chance).

A significant *p*-value tells us that an intervention works.

Whereas an effect size tells us how much it works.

Effect size is independent of sample size, so effect sizes can be used to quantitatively compare the results of studies done in a different setting. It is widely used in meta-analysis.

52

52

Why report effect sizes

Effect Sizes. For readers to appreciate the magnitude or importance of a study's findings, it is recommended to include some measure of effect size in the Results section. Effect sizes are statistical estimates; therefore, whenever possible, provide a confidence interval for each effect size reported to indicate the precision of estimation of the effect size. Effect sizes may be expressed in the original units (e.g., mean number of questions answered correctly, kilograms per month for a regression slope) and are most easily understood when reported as such. It is valuable to also report an effect size in some standardized or units-free or scale-free unit (e.g., Cohen's *d* value) or a standardized regression weight. Multiple degree-of-freedom effect-size indicators are less useful than effect-size indicators that decompose multiple degree-of-freedom tests into meaningful one degree-of-freedom effects, particularly when the latter are the results that inform the discussion. The general principle to follow is to provide readers with enough information to assess the magnitude of the observed effect.

(American Psychological Association, 2020, p. 89).

53

53

Effect Sizes

Effect Size	Small	Medium	Large
r	.1 - .3	.3 - .5	.5 – 1.0
R^2	.01 - .09	.09 - .25	.25 – 1.0
η^2	.2 - .5	.5 - .8	.8 – 1.0
d	.2 - .5	.5 - .8	.8 – 1.0
Cramer's V	.1 - .3	.3 - .5	.5 – 1.0

(Cohen, 1988, 1992)

But, as Cohen (1988, p. 532) pointed out, his conventions were devised "with much diffidence, qualification, and invitations not to employ them if possible."

54

54

References

American Psychological Association (2020). *Publication manual of the American Psychological Association* (7th ed.). Washington, DC: American Psychological Association.

Burns, R. B. & Burns, R. A. (2008). *Business research methods and statistics using SPSS*. London: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Routledge.

Cohen, J. (1992). A power primer. *Psychological Bulletin,* 112, 155-159;

Dörnyei, Z. (2007). *Research methods in applied linguistics.* Oxford: Oxford University Press.

Kline, R. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Larson-Hall, J. (2012). How to run statistical analyses. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition* (pp. 245-274). London: Wiley-Blackwell.

Pallant, J. (2016). *SPSS survival manual* (6th ed.). London: McGraw Hill.

Robson, C. (2011). *Real world research* (3rd ed.). London: Wiley.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics.* (5th ed.). Boston: Allyn & Bacon.