**07 Effect Size (draft 26 Feb 2018)**

**Topics**

**1. Effect Size (ES): Unstandardized and Standardized**

**2. ES Use**

**3. Effect Size d**

**4. Effect Size r**

**5. Effect Size f**

**6. Which ES to Use**

add categorical data effect size

revise wording for prospective ancova

add interpretation of d, r, f, f2

discuss when to use d, r, f, f2

**1. Effect Size (ES): Unstandardized and Standardized**

The term effect size refers the

* size of the statistical relationship or difference observed in data
* denotes the magnitude of difference if comparing means
* magnitude of relationship between variables
* may be unstandardized or standardized
	+ unstandardized = females earn $2,354 more than males
	+ standardized = females earn 0.43 SD more than males

**1.1 Example 1 ES for Relationship**

Unstandardized

Regression of pay on hours worked

Pay’ = b0 + b1 Hours Worked

Pay’ = 0.00 + 12.50 (Hours Worked)

 ES = b1 = 12.50 (unstandardized ES for each hour worked, pay increases by $12.50)

Standardized

 Correlation between pay and hours worked

Pearson r = .90

or

Proportion of variance predicted for pay and hours worked (semi-partial correlation squared spr2, or ΔR2)

spr2 = .81

Unstandardized

The relationship between academic self-efficacy and test anxiety is b = -.16 (unstandardized regression coefficient)

Standardized

and r = -.45 (Pearson correlation)

**1.2 Example 2 ES for Mean Difference**

Unstandardized

Females earn $1.40 per hour more than men

Standardized

Females earn $1.40 per hour more than men, with a standard deviation of $5.60,

ES d = 1.40/5.60 = 0.25 (1/4 of a standard deviation more than men)

Unstandardized

Students exposed to the new teaching strategy scored, on average, 0.12 points (mean difference) higher on unit tests than students exposed to traditional instruction

Mean New Strategy = 83.35

Mean Traditional Teaching = 83.23

Mean Difference = 00.12

Standardized

With a SD = 7, the corresponding standardized effect size is

d = 0.12 / 7 = .017

**2. ES Use**

**2.1 Standardized vs Unstandardized**

Use unstandardized ES if measurement units are easily understood (e.g., temperature, dollars, pounds, time, percentage change), but often measurement in social science and behavioral research uses diverse scales that are not easily understood (e.g., anxiety scale A with range from 1 to 6 vs anxiety scale B with range 20 to 60). For these situations use standardized effect sizes.

**2.2 ES for Combining Studies vs Power Analysis and Sample Size Determination**

The ES calculations shown below are primarily for use in determining power analysis and sample size. Some are suitable for combining results from studies for statistical reviews (e.g., meta-analysis), but not all ES presented below will be suitable for meta-analyses without some stipulation (e.g., covariate adjusted ES).

**3. Effect Size d**

Much of what is understood about power and sample size can be attributed to Jacob Cohen, pictured below, author of Statistical Power Analysis for the Behavioral Sciences (1969; 1988).



**3.1 Cohen’s d**

$d =$ $\frac{Mean\_{1}- Mean\_{2}}{Standard Deviation Pooled}$

d tells us the following

* distance between two means in standard deviation units
* the larger d in absolute value, the greater separation between two means
* d = 0.00 indicates two means are the same
	+ Female Salary Mean = 50,000
	+ Male Salary Mean = 50,000
	+ Standard Deviation = 10,000
	+ $d =$ $\frac{50,000-50,000}{10,000}$ $=$ $\frac{0}{10,000}$ = 0.00
* d = -0.50 indicates the first mean is half a standard deviation below the second mean
	+ Female Salary Mean = 45,000
	+ Male Salary Mean = 50,000
	+ Standard Deviation = 10,000
	+ $d =$ $\frac{45,000-50,000}{10,000}=$ $\frac{-5,000}{10,000}$ = – 0.50
* d = 1.00 indicates the first mean is one standard deviation higher than the second mean
	+ Female Salary Mean = 60,000
	+ Male Salary Mean = 50,000
	+ Standard Deviation = 10,000
	+ $d =$ $\frac{60,000-50,000}{10,000}=$ $\frac{10,000}{10,000}$ = 1.00
* d = 2.00 indicates the first mean is two standard deviations higher than the second mean
	+ Female Salary Mean = 70,000
	+ Male Salary Mean = 50,000
	+ Standard Deviation = 10,000
	+ $d =$ $\frac{70,000-50,000}{10,000}=$ $\frac{20,000}{10,000}$ = 2.00

ES d graphically illustrated:

http://rpsychologist.com/d3/cohend/

**3.2 Example 1: IQ between College and High School Graduates**

College graduates have an estimated IQ of 115 while high school graduates have an estimated IQ of 105. The standard deviation of WAIS is about 15. (Source: <http://www.assessmentpsychology.com/iq.htm>)

$d $ $=\frac{Mean\_{1}- Mean\_{2}}{Standard Deviation Pooled}$

$d $ = $\frac{115- 105}{15}$ = $\frac{10}{15} $ = 0.667

Interpretation:

College graduates’ mean IQ is, on average, about 0.667 (or 2/3) standard deviations above the mean IQ of high school graduates.

**3.3 Examples 2 and 3: SAT scores by Sex**

What are the effect sizes, d, by sex for Writing and Mathematics SAT performance? The College Board reports the following means for 2013 SAT results.



Source: http://media.collegeboard.com/digitalServices/pdf/research/2013/TotalGroup-2013.pdf

**Example 2: ES d for Writing SAT**

|  |  |  |
| --- | --- | --- |
|  | Female | Male |
| Mean | 493 | 482 |
| Standard Deviation | 112 | 115 |
| n (number of participants) | 883,955 | 776,092 |

To calculate d the pooled standard deviation is needed. Cohen provides the following formula for finding the pooled SD (denoted s below):



Pooled SD = $\sqrt{\frac{\left(883955-1\right)112^{2}+\left(776092-1\right)115^{2}}{882955+776092-2}}$ = 113.41

Effect size d can now be calculated using this SD:

(F1) d = $\frac{M\_{1}-M\_{2}}{\sqrt{\frac{\left(n\_{1}-1\right)VAR\_{1}+(n\_{1}-1)VAR\_{2}}{n\_{1}+n\_{2}-2}}}$

d = $\frac{Mean\_{1}- Mean\_{2}}{Standard Deviation Pooled}$

d = $\frac{493 - 482}{113.41}$ = $\frac{11}{113.41}$ = 0.097

Interpretation:

Females’ mean SAT writing score is about 0.097 standard deviations greater than males’ SAT writing score.

|  |
| --- |
| Screenshot from Excel Effect Size sheet |

**Example 3: ES d for Mathematics SAT**

|  |  |  |
| --- | --- | --- |
|  | Female | Male |
| Mean | 499 | 531 |
| Standard Deviation | 114 | 121 |
| n (number of participants) | 883,955 | 776,092 |

Find the pooled SD (denoted s below):



Pooled SD = $\sqrt{\frac{\left(883955-1\right)112^{2}+\left(776092-1\right)115^{2}}{882955+776092-2}}$ = 113.41

Effect size d can now be calculated using this SD:

d = $\frac{Mean\_{1}- Mean\_{2}}{Standard Deviation Pooled}$

d = $\frac{499 - 531}{113.41}$ = $\frac{-32}{113.41}$ = –0.272

Interpretation:

Females’ mean SAT mathematics score is about -0.272 standard deviations below the males’ SAT mathematics score (about 1/4th a SD below).

|  |
| --- |
| Screenshot from Excel Effect Size sheet |

**3.4 Cohen’s Guidelines/Recommendations/Suggestions for d**

In the absence of prior research from which effect size d may be calculated, Cohen (1988, p. 24+) offered the following values of d as small, medium, and large. Many in social science research have adopted these guidelines.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Small | Medium | Large |
| Effect Size d | .20 | .50 | .80 |

These guidelines are helpful for power analysis and sample size determination, the next topic following effect sizes.

**3.5 Cohen’s d from t-tests**

Cohen’s d can be computed from obtained t-ratios from independent samples t-tests. The formula is provided below where t = calculated t-ratio and df error is the degrees of freedom for the t-test (i.e., normally n – 2, where n = sample size):

(F2) d = $\frac{2t}{\sqrt{df}}$

The above formula is suitable when sample sizes are equal or nearly so for both groups, or if sample sizes for groups are unknown. If sample sizes are largely unequal, the formula below can be used:

(F3) d = $\frac{t(n\_{1}+n\_{2})}{(\sqrt{df})(\sqrt{n\_{1}n\_{2}})}$

In practice, the two will produce d values that are often similar, but sometimes they will differ. An example of different d values is provided below in the section on obtaining d from ANOVA models.

**Example 4: Dissertation Process Anxiety by Sex**

Doctoral students complete a questionnaire that included a scale to measure anxiety toward the dissertation process. The total sample was 19, but the sample by sex was female n = 14 and male n = 5. Descriptive statistics and t-test results are provided below; Group 1 = female, Group 2 = male.



The values of d using the three d formulas presented above (F1, F2, and F3):

(F1) d = $\frac{M\_{1}-M\_{2}}{\sqrt{\frac{\left(n\_{1}-1\right)VAR\_{1}+(n\_{1}-1)VAR\_{2}}{n\_{1}+n\_{2}-2}}}$ = $\frac{4.157-4.08}{\sqrt{\frac{\left(14-1\right)3.361+\left(5-1\right)0.092}{14+5-2}}}$ = 0.0478

(F2) d = $\frac{2t}{\sqrt{df}}$ = $\frac{20.09197}{\sqrt{17}}$ = 0.0446

(F3) d = $\frac{t(n\_{1}+n\_{2})}{(\sqrt{df})(\sqrt{n\_{1}n\_{2}})}$= $\frac{t(14+5)}{(\sqrt{14})(\sqrt{14\*5})}$ = 0.0506

|  |
| --- |
| Screenshot from Excel Effect Size sheet |

**3.6 Cohen’s d from one-way ANOVA**

Cohen’s d may also be derived from an ANOVA. The formula below works even when more than two groups are present, although interpretation of d in this situation is less clear. The reason it works stems from the fact that many effect sizes used to assess power and sample size are linear transformations, so one may easily convert among effect sizes d, r, r2 f, and f2, although interpretation can be difficult for some conversions. Additional, some ES do not maintain the directional sign that is present with ES d.

(F4) d = $2\left(\sqrt{\frac{F(df\_{predictors})}{df\_{error}}}\right)$

In formula F4 above, F is the calculated F ratio for the factor (variable) of interest, dfpredictors is the degrees of freedom for the factor, and dferror is the model error (within) degrees of freedom.

**Example 5: MPG between American and Japanese Vehicles**

SPSS provides a sample data file names cars.sav. Focus for this example will be a comparison of recorded MPG between vehicles of US and Japanese origin. Descriptive statistics, t-test results, and ANOVA results are presented below.

Descriptive Statistics



Two-group t-test



One-way ANOVA



Find Cohen’s d using formulas F1, F2, F3, and F4.

(F1) d = $\frac{M\_{1}-M\_{2}}{\sqrt{\frac{\left(n\_{1}-1\right)VAR\_{1}+(n\_{1}-1)VAR\_{2}}{n\_{1}+n\_{2}-2}}}$ = $\frac{20.13-30.45}{\sqrt{\frac{\left(248-1\right)40.666+\left(79-1\right)37.0881}{248+79-2}}}$ = –1.6356

(F2) d = $\frac{2t}{\sqrt{df}}$ = $\frac{-12.664}{\sqrt{325}}$ = –1.4049

(F3) d = $\frac{t(n\_{1}+n\_{2})}{(\sqrt{df})(\sqrt{n\_{1}n\_{2}})}$= $\frac{t(248+79)}{(\sqrt{325})(\sqrt{248\*79})}$ = –1.6411

(F4) d = $2\left(\sqrt{\frac{F(df\_{predictors})}{df\_{error}}}\right)$= $2\left(\sqrt{\frac{160.379(1)}{325}}\right)$ = 1.4049

The formulas indicate d as either -1.64 or -1.40 except for the value of d provided by the ANOVA F test which loses the negative sign since F tests are nondirectional.

|  |
| --- |
| Screenshot from Excel Effect Size sheet |

**3.7 Cohen’s d from Regression**

Mathematically simple regression produces the same test for mean differences as a two-group t-test assuming equal variances and a one-way ANOVA. A dummy variable was created coding American origin = 1 and Japanese origin = 0 using the cars.sav data above. Regression results are presented below.

Formula F2 can be used to find d. The t-ratio was obtained from the regression coefficient output. The model degrees of freedom are taken from the regression ANOVA table used to test the overall model fit Ho: R2 = 0.00 (i.e., df error are the residual degrees of freedom).

(F2) d = $\frac{2t}{\sqrt{df}}$ = $\frac{-12.664}{\sqrt{325}}$ = –1.4049

Regression Results for MPG and Origin



Note that results are identical to those provided by the two-group t-test and one-way ANOVA. If desired, one could adjust the effect size for unequal group sizes using formula F3 which results in the larger (in absolute value) estimate of d.

(F3) d = $\frac{t(n\_{1}+n\_{2})}{(\sqrt{df})(\sqrt{n\_{1}n\_{2}})}$= $\frac{t(248+79)}{(\sqrt{325})(\sqrt{248\*79})}$ = –1.6411

|  |
| --- |
| Screenshot from Excel Effect Size sheet |

**3.7 Other Standardized Measures of Mean Difference**

Two commonly referenced measures of standardized mean difference include Hedges’ g and Glass’$Δ$. Hedges’ g should be used if performing meta-analyses. For more information see: <http://en.wikipedia.org/wiki/Effect_size>

**3.8 Combining d**

Cohen’s d is often combined from multiple studies to obtain estimates of central tendency and variability in d, and this is especially common in meta-analytic studies. For estimating power and sample size, mean d can and should be used when information is available from multiple studies.

If different effect sizes estimates result from formulas F1, F2, F3, and F4 like above with the MPG data due to differences in sample size between groups, I recommend using the smaller d in absolute value for assessing power and sample size since it will result in a more conservative estimate thus requiring large n to achieve a specified power level.

**4. Effect Size r**

**4.1 Pearson Correlation Coefficient, r**

Pearson’s correlation coefficient, r, is a standardized effect size measure. The reason for this can be seen in the following formula for r; both variables are first standardized with Z scores which have a mean of 0.00 and standard deviation of 1.00.

r = $\frac{Z\_{x}Z\_{y}}{n-1}$

Pearson r

* is a measure of linear relationship;
* ranges from -1.00 to 1.00;
* indicates no linear relationship when r = 0.00; and
* is typically used to identify associations between quantitative variables (e.g., test grades and hours studied; level of motivation and level of persistence; number of publications and annual merit pay increase)
* r can be misleading as an effect size (discussion to be added, compared with unstandardized regression slope)

**4.2 Cohen’s Guidelines/Suggestions for Pearson r**

In the absence of prior research from which Pearson’s r may be found, Cohen (1988, p. 24+) offered the following values of r as small, medium, and large. As with effect size d, many in social science researchers have adopted these guidelines.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Small | Medium | Large |
| Pearson r | .10 | .30 | .50 |

**4.3 Combining r values**

Like Cohen’s d, Pearson r from multiple studies can be combined to obtain a mean correlation for power analysis and sample size determination. Those familiar with r argue that Fisher’s Z transformation should be used before calculating the mean value of r. For our purposes, a basic mean of r without Fisher’s Z transformation will work well. If your goal is to conduct a meta-analysis, a more rigorous approach should be used.

**Example 6: Combining Mathematics Self-efficacy and Mathematics Anxiety Correlations**

Below are four studies that report the Pearson correlation between mathematics self-efficacy and anxiety for high school or college students. The values are:

-.53 (college)

 .58 (college, anxiety was reverse scored so high scores indicate low anxiety)

-.53 (high school)

-.24 (high school, international sample)

Since the r of .58 is based upon reversed scored anxiety, a negative sign will be added to that correlation to make the interpretation consistent with the other correlations, i.e., negative relation (higher anxiety, lower efficacy).

The table below shows the unadjusted mean of the four correlations and the mean based upon the Fisher Z transformation. For this example, there is little difference in vales, -.47 vs -.47915.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Study | Pearson r |  |  | Converted to Fisher Z |
| 1 | -.53 |  |  | -0.59015 |
| 2 | -.58 |  |  | -0.66246 |
| 3 | -.53 |  |  | -0.59015 |
| 4 | -.24 |  |  | -0.24477 |
|  |  |  | Mean Fisher Z | -.52188 |
| Mean r | -.47 |  | Converted to r | -.47915 |

Akin, A., & Kurbanoglu, N. (2011). The relationships between math anxiety, math attitude and self-efficacy: A structural equation model. Studia Psychologica, 53(3), 263-273.



Hackett, G. (1985). The role of mathematics self-efficacy in the choice of math-related majors of college men and women: A path analysis. Journal of Counseling Psychology, 32, 47-56.



Pajares, F., & Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in mathematical problem-solving. Contemporary Educational Psychology, 20, 426-443.



Ferla J, Valcke M, Cai Y. (2009). Academic self-efficacy and academic self-concept: reconsidering structural relationships. Learn. Individ. Dif. 19:499–505.



**4.4 Conversion between d and r**

Effect sizes d and r can be converted one from the other:

d = $\frac{2r}{\sqrt{1-r^{2}}}$

r = $\sqrt{\frac{d^{2}}{d^{2}+4}}$

These formulas assume equal group sample sizes. For two-group t-test results converting from d to r results in a point-biserial correlation. If either r or d are calculated from partial tests (i.e. partial F test or t-test from variable contribution in regression or ANOVA), then the resulting d and r values are partial effects (i.e., partial r rather than Pearson r).

|  |
| --- |
| Screenshot from Excel Effect Size sheet |

(Note – material to add: Converting to r from d using formulas that take into account differing sample sizes, F1 and F3, will produce a correlation that differs from the correlation obtained using Pearson r formula.)

**5. Effect Size f**

**5.1 Analysis of Variance: Effect Size f,** $η^{2}$ **and Multiple R2**

**5.1.1 One-way ANOVA**

One-way ANOVA models are like independent samples t-tests except that ANOVA may be used to compare two or more groups.

For example, ANOVA may be used to compare mean

* science test scores among three different teachers;
* SAT scores by college;
* motivation scores between treatment and control groups; or
* salary between females and males.

**Effect Size f,** $η^{2}$**and Multiple R2**

Cohen (1988) uses **effect size f** for sample size determination and power analysis for ANOVA and ANCOVA models.

f = $\sqrt{\frac{η^{2}}{1-η^{2}}}$

The term $η^{2}$ (eta-squared) is a measure of

* model fit; how well the ANOVA model is able to predict or explain variance on the dependent variable (DV);
* how much variance in the DV can be predicted by knowing group membership on the independent variable (IV);
* proportional reduction in error: how much error reduction in prediction can be expected by using the IV to assist in predicting the DV.

ANOVA can be performed via linear regression, so Multiple R2, a measure of model fit in regression, can be used to estimate f:

(F5) f = $\sqrt{\frac{η^{2}}{1-η^{2}}}$ = $\sqrt{\frac{R^{2}}{1-R^{2}}}$

Multiple R2 has the same interpretation as $η^{2}$; they are the same measure.

**Correspondence among** $η^{2}$, **R2, d, and f**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| $η^{2}$ |  **R2** | **d** | **f** | **Cohen’s View of f** |
| .90 | .90 | 6.00 | 3.00 |  |
| .75 | .75 | 3.46 | 1.73 |  |
| .50 | .50 | 2.00 | 1.00 |  |
| .25 | .25 | 1.15 | 0.58 |  |
| .14 | .14 | 0.80 | 0.40 | Large |
| .10 | .10 | 0.67 | 0.33 |  |
| .06 | .06 | 0.50 | 0.25 | Medium |
| .05 | .05 | 0.46 | 0.23 |  |
| .01 | .01 | 0.20 | 0.10 | Small |

Note: To show correspondence among d and the other entries, one must assume only two groups are compared; there is no direct correspondence (I think) between d and the other indices when more than two groups are compared.

Loosely explained, **effect size f** is an index of how well an ANOVA model is able to explain or predict variance on the DV. The better ANOVA can predict scores on the DV, the greater will be **effect size f**.

**Example 7: Instructor Reputation Effect size**

I conducted and published a study on student ratings of instruction with a focus on instructor reputation. A sample of n = 920 students participated.

**Instructor Reputation**

39. Before taking this course, what did you hear about this instructor?

Responses were coded into three general categories:

* Negative Information
* No Information
* Positive Information

**Overall Instructor Rating**

30. Overall, how would you rate this instructor?

Response options ranged from 1 = “Poor” to 5 = “Excellent”

**ANOVA Results**

Descriptive Statistics

|  |  |  |  |
| --- | --- | --- | --- |
| Instruction Reputation | Mean | Std. Deviation | N |
| Negative Info | 3.0765 | 1.06033 | 170 |
| No Info | 4.0632 | 1.12970 | 522 |
| Positive Info | 4.3465 | .84901 | 228 |
| Total | 3.9511 | 1.13831 | 920 |

Tests of Between-Subjects Effects

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | Type IIISum of Squares | df | Mean Square | F | Sig. |
| Reputation | 172.252 | 2 | 86.126 | 77.539 | .000 |
| Error | 1018.547 | 917 | 1.111 |  |  |
| Corrected Total | 1190.799 | 919 |  |  |  |

R Squared = .145 (Adjusted R Squared = .143)

How typically reported in published studies: F(2, 917) = 77.539

**Effect Size f Calculation based upon R2**

f = $\sqrt{\frac{R^{2}}{1-R^{2}}}$

f =$\sqrt{\frac{.145}{1-.145}}$ =$\sqrt{\frac{.145}{.855}}$ =$\sqrt{.1695}$ = .41

If R2 is not provided, it can be calculated from the ratio of the sums of squares (SS) due to the factor (predictor, or between) to the total sums of squares:

R2 =$\frac{SS factor}{SS total}$ = $\frac{SS between}{SS total}$ = $\frac{172.252}{1190.799}$ = 0.1447

**Effect Size f Calculation based upon F-ratio**

Another option to compute effect size f uses the calculated F ratio from ANOVA. The F ratio from the above ANOVA is 77.539 with degrees of freedom, df, of 2 and 917.

The formula F6 contains three numbers:

* F = the F ratio calculated in ANOVA or regression;
* df predictors (or df 1 or df between)) = degrees of freedom for the independent variable (or set of variables); and
* df error (or df 2 or df within) = degrees of freedom that remain and pooled in the error term.

For one-way ANOVA models, the df for the predictor (only one predictor in one-way ANOVA) is

df1 = J – 1 (where J is the number of groups)

Since there are three groups (Negative Information, No Information, and Positive Information), this leads to 2 degrees of freedom:

df1 = J – 1 = 3 – 1 = 2

For this ANOVA model F = 77.539, df1 = 2, and df2 = 917 (sample size – df1 – 1 = 920 – 2 – 1 = 917)

(F6) f = $\sqrt{\frac{F(df\_{predictors})}{df\_{error}}}$= $\sqrt{\frac{77.539(2)}{917}}$ = 0.411

|  |
| --- |
| Screenshot from Excel Effect Size sheet |

**Example 8: Math Achievement and Student Group**

Watt, Huerta, and Lozano (2007) studied several student outcomes for those participating in two groups, AVID and GEAR UP, and those participating in none. A total of four groups were studied: AVID, GEAR UP, AVID and GEAR UP (students enrolled in both), and a control group.

Below are the ANOVA results comparing math achievement across the four study groups. What is the effect size f for this study?



(F6) f = $\sqrt{\frac{F(df\_{predictors})}{df\_{error}}}$= $\sqrt{\frac{1.98(3)}{132}}$ = 0.212

|  |
| --- |
| Screenshot from Excel Effect Size sheet |

**5.2 Analysis of Variance with Several Factors**

Multi-way ANOVAs contain more than one factor (i.e., predictor, independent variable). As an example, one may model student reading comprehension (DV) by type of phonics program (factor 1), sex (factor 2), and grade level (factor 3).

Calculated effect sizes for factors in such models represent partial effect sizes which are effect sizes that partial out, or control, for the statistical effects of other predictors. When calculating and reporting such effect sizes one should be clear that these effect sizes are adjusted based upon the statistical control of other variables so readers won’t confuse these effect sizes with those calculated from unadjusted models (e.g., the effect sizes from all examples shown above which represent simple bivariate associations).

**Example 9: Blood Pressure by Disease and Drug**

Stata, a statistical analysis company, illustrates two-way ANOVA using change in systolic blood pressure (DV) by drug assignment (Factor 1) and patient disease (Factor 2). The data are reported below in table form and can be read into Stata with the following command.

“use http://www.stata-press.com/data/r13/systolic”



ANOVA results are reported below. What are effect size f for drug, disease, and the interaction between drug and disease?

Drug f

(F6) f = $\sqrt{\frac{F(df\_{predictors})}{df\_{error}}}$= $\sqrt{\frac{9.05(3)}{46}}$ = 0.768

Disease f

(F6) f = $\sqrt{\frac{F(df\_{predictors})}{df\_{error}}}$= $\sqrt{\frac{1.88(2)}{46}}$ = 0.285

Drug by Disease Interaction

(F6) f = $\sqrt{\frac{F(df\_{predictors})}{df\_{error}}}$= $\sqrt{\frac{1.07(6)}{46}}$ = 0.373

|  |
| --- |
| Screenshot from Excel Effect Size sheet   |

 

These effect sizes, 0.768 for drug, 0.285 for disease, and 0.373 for the interaction are partialed effects that take into account the other predictors in the model. These effect sizes can be used for assessing power and sample size determination for those who wish to replicate or perform a similar study.

**5.3 Analysis of Covariance from Existing Studies**

The calculating of effect size f can be performed on most any linear modeling analysis that uses the F test, this includes multiple regression and analysis of covariance.

**Example 10: Car MPG by Origin Controlling for Car Weight**

Using the SPSS data file cars.sav, an ANCOVA is performed comparing adjusted MPG by origin after taking into account vehicle weight. For these data origin consists of three locations: US, Japan, and Europe. Vehicle weight is measured in pounds.

Results are presented below. What are the effect size f for origin and weight?

Origin f

(F6) f = $\sqrt{\frac{F(df\_{predictors})}{df\_{error}}}$= $\sqrt{\frac{6.426(2)}{393}}$ = 0.18

Weight f

(F6) f = $\sqrt{\frac{F(df\_{predictors})}{df\_{error}}}$= $\sqrt{\frac{478.914(1)}{393}}$ = 1.10

|  |
| --- |
| Screenshot from Excel Effect Size sheet  |



**5.4 Multiple Regression**

Regression and ANOVA are part of the general linear model and therefore share the same underlying linear model, so the effect sizes learned above with ANOVA also apply to regression. One differences is that ANOVA traditionally uses F-tests while regression incorporates t-tests of parameter estimates. If the degrees of freedom for the predictor is 1.00, the t-test is equivalent to an F-test, i.e., F = t2. In situations where predictor degrees of freedom do not equal 1.00 (e.g., sets of predictors are tested, or a factor with more than two categories is modeled), one uses an F-test in regression.

Given this equivalence, formulas posted above continue to work for regression, and the t-test can be converted to f by squaring t and using formula F6.

(F6) f = $\sqrt{\frac{F(df\_{predictors})}{df\_{error}}}$

(F7) f = $\sqrt{\frac{t^{2}}{df\_{error}}}$

**Example 11: Car MPG Regression on Weight, Horsepower, and Engine Displacement**

Using the SPSS data file cars.sav, and MPG was regressed upon three predictors: vehicle weight, horsepower, and engine displacement in cubic inches.

SPSS results are presented below. Two tables of information are needed, the ANOVA summary so df error can be found (df2 = 388), and the coefficient table so t-ratios can be obtained.





What are the effect size f for the three predictors?

Horsepower

(F7) f = $\sqrt{\frac{t^{2}}{df\_{error}}}$ = $\sqrt{\frac{-4.153^{2}}{388}}$ = .21

Weight

(F7) f = $\sqrt{\frac{t^{2}}{df\_{error}}}$ = $\sqrt{\frac{-6.186^{2}}{388}}$ = .31

Displacement

(F7) f = $\sqrt{\frac{t^{2}}{df\_{error}}}$ = $\sqrt{\frac{-0.786^{2}}{388}}$ = .039

|  |
| --- |
| Screenshot from Excel Effect Size sheet  |

**Example 12: Car MPG by Origin Controlling for Car Weight and Horsepower**

Using the cars.sav data, below is a regression analysis of MPG with predictors origin, weight, and horsepower. Origin is recoded into two dummy variables for the regression analysis. Recall that dummy variables are coded 1, 0 to represent one group vs. rest. One must have J – 1 dummy variables for groups represented (e.g., if there are three Origins – US, Japan, and Europe – then two dummy variables are needed).

Regression output for the MPG model is presented below.



Error degrees of freedom = 386

Note that in this example it is not possible to provide an overall effect size for Origin since it is modeled in regression by two dummy variables. We could calculate an effect size for each dummy variable comparison, but this does not assess Origin as a whole. One must have the F test for the Origin contribution to calculate effect size using formula F6 (or have the semi-partial correlation and R2).

In situations where a dummy variable is used to represent only two groups, such as sex (female vs. male) or only two Origins are included (e.g., US vs Japan), then effect size f from the t-ratio would be appropriate. It is only when more than one dummy is required to represent group membership that an F test or other relevant statistics are needed to calculate an effect size.

Effect size estimates are possible for the other predictors, Weight and Horsepower.

Horsepower

(F7) f = $\sqrt{\frac{t^{2}}{df\_{error}}}$ = $\sqrt{\frac{-4.857^{2}}{386}}$ = .247

Weight

(F7) f = $\sqrt{\frac{t^{2}}{df\_{error}}}$ = $\sqrt{\frac{-8.863^{2}}{386}}$ = .451

**5.6. Effect Size f and ANCOVA**

Note – to be revised, formula not yet added to Excel spreadsheet

The effect size estimates for ANCOVA explained above are post-hoc estimates – done after the fact on existing analyses. In situations where one wishes to perform a true experimental study, with randomly formed groups – this is an important point – and analyze data with ANCOVA, this one can take into account the proportion of variance explained given the covariates to compute an adjusted effect size for power and sample size determination.

Recall that an ANCOVA is formed when covariates are added to an ANOVA model. The addition of covariates leads to increased power and precision of model estimates due to a reduction in the model error term. The addition of covariates may also result in adjusted group means on the dependent variable (DV) to reflect statistical control, or partialing effects, of added covariates.

Cohen (1988) writes that sample size and power analysis procedures for ANCOVA proceeds like that for ANOVA except that the adjusted means – adjusted for the covariates – are used for calculating effect sizes of interest.

One approach to incorporating covariate influence on the ANOVA model for determining n or power is to adjust effect size estimates based upon anticipated explanatory power added by covariates.

For example, if one believes that a covariate, or set of covariates, will correlate with the DV at the r = .30 level (or multiple R = .30 in case of multiple covariates), then this correlation can be employed to adjust the effect size used in determining sample size for an ANOVA model. The formula for adjustment follows:

adjusted effect size = $\frac{effect size}{\sqrt{(1-r^{2})}}$= $\frac{effect size}{\sqrt{(1-R^{2})}}$

where effect size refers to d or f; neither squared effect sizes such as f2 or r2, nor Pearson r, should be adjusted using the above formula since the adjustment for these values is nonlinear.

Below are two illustrations of this process.

**Illustration A**

Sample size for an ANOVA model with four groups is estimated using the following criteria

* alpha = .05
* power = .80
* df = 3
* f = .20

Resulting ANOVA total n = 277 so the per-group n would be 277 / 4 = 69.25 or about 70 per group.

A covariate will be added to this ANOVA model and prior research suggests the correlation between this covariate and the DV is r = .35. With this value of r, the ANCOVA adjusted effect size, f, would be:

adjusted f $=\frac{f}{\sqrt{(1-r^{2})}}=\frac{.20}{\sqrt{(1-.35^{2})}}=\frac{.20}{\sqrt{(1-.1225)}}$

$=\frac{.20}{\sqrt{.8775}}=\frac{.20}{.9367}=$.2135

One would then use this adjusted f of .2135 as the effect size for determining sample size. There will be no need to adjust degrees of freedom from the original ANOVA model sample size calculation; only the effect size estimate must be adjusted.

Entering the following criteria

* alpha = .05
* power = .80
* df = 3
* adjusted f = ~~.20~~ .2135

results in a total n of 244 or about 61 participants per group. Adding the covariate has reduced the required sample size from 277 to 244.

**Illustration B**

We wish to compare mathematics SAT means between females and males. Given College Board results cited earlier, the anticipated effect size d is -0.27.

Using these criteria

* alpha = .05
* power = .85
* df = 1
* d = -0.27

the sample size for this t-test is n = 495 or about 248 per group.

We wish to add three covariates to this study: mathematics self-efficacy, mathematics anxiety, and IQ scores. While it is difficult to know how much predictive power these three measures will bring to the model of mathematics SAT scores, we know from prior research that each variable correlates between .20 and .40, in absolute value, with various academic assessments. Together we anticipate these three variables would contribute to a total model R2 of about .25.

The resulting adjusted d value can be found as follows:

adjusted d$=\frac{d}{\sqrt{(1-R^{2})}}=\frac{-.27}{\sqrt{(1-.25)}}=\frac{-.27}{\sqrt{.75}}$

$=\frac{-.27}{.866}=$-.3118

What sample size is needed to detect SAT mean differences once these covariates are taken into account?

Using these criteria

* alpha = .05
* power = .85
* df = 1
* adjusted d = ~~-.27~~ -0.3118

adjusted f$=\frac{f}{\sqrt{(1-R^{2})}}=\frac{.412}{\sqrt{(1-.06)}}=\frac{.412}{\sqrt{.94}}$

$=\frac{.412}{.9695}=$ 0.425

**Cautions**

The procedures outlined above assume no correlation between covariates and group membership, i.e., no differences in covariate means among groups of the factor/independent variable of interest.

If group membership is correlated with covariate scores (e.g., group A has a lower covariate mean than group B), then the correlation employed to adjust the effect size will likely be overlarge.

In most cases the result of a correlation between covariate and group membership is a reduction of power for detecting mean differences on the DV; thus estimated sample sizes for the ANCOVA will be too small.

The ANCOVA steps explained above work best for experimental designs in which groups are randomly formed since this likely produces no correlation between group membership and covariates. Stated differently, random formation of groups from a common population of participants assures that groups, over the long run, are equivalent on possible confounding variables. This equivalence results in no differences, or only trivial differences, in covariate means between groups. In this situation there is no correlation, or only a trivially small correlation, between group membership and covariate scores.

In studies with intact, pre-existing groups that were not randomly formed, it is very likely that covariates will correlate, perhaps substantially, with group membership variables. For sample size determination, partial correlations to identify the unique contribution of covariates partialed from group membership variables are needed for effect size adjustment.

Sources for ANCOVA adjustment approach outlined above:

* Lipsey, M. (1990). Design sensitivity: Statistical Power for Experimental Research. p 131. (2 group comparison)
* Cohen (1988) p. 432. (4 groups)
* Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analyzing data: A model comparison perspective (2nd ed.). Belmont, CA: Wadsworth. p. 442. (3 group comparison)
* Shan G, Ma C (2014) A Comment on Sample Size Calculation for Analysis of Covariance in Parallel Arm Studies. J Biomet Biostat 5: 184.
* Beyene, N. & Lui, K. (2001). Sample Size Determination for Analysis of Covariance. Proceedings of the Annual Meeting of the American Statistical Association, August 5-9.
* PASS Sample Size Software, Chapter 551 Analysis of Covariance, NCSS.com (note their ES is f, not d, for output; p. 10)

**6. Which ES to Use**

Four ESs have been presented: d, r, f, and f2. Which should be used and when? One answer is that it does not matter which is used because each is an algebraic transformation of the other; if you know d, then you can find r, f, and f2; if you know f2, you can find d, r, and f; and so on.

In practice, use of specific ES follows some tradition, described below.

**d: mean difference of two groups**

ES d is defined as the mean difference between two groups divided by a standard deviation which is commonly the pooled SD, but could be other SDs, so ES d is the standardized mean difference between two groups. ES d is suitable for any situations in which two groups are compared by a mean difference.

**r: Pearson correlation**

The ES r is naturally associated with Pearson correlation, since it is the Pearson correlation, so any situation for which Pearson r is suitable is also suitable for use of ES r.

**f: Mean differences among groups**

Cohen (1988) defined ES f as the ratio of standard deviation of means to the standard deviation of raw scores; this is like ES d, and f is equal to d/2. ES f is traditionally used for ANOVA-type models since f is defined in terms of the standard deviation of means.

**f2: Ratio of variance predicted to variance not predicted**

ES f2 is the ratio of variance predicted to variance not predicted. In regression the index of variance predicted is R2. If a regression model predicts 25% of the variance in the DV, then R2 = .25. The amount of variance not predicted by the model is 1 - R2 or 1 - .25 = .75. For this example, f2 is .25 / .75 or 0.33; could also define f2 as the signal to noise ratio. Traditionally researchers associate R2 with regression, so f2 has been associated with regression-type models.

**References**

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd). Erlbaum: Hillsdale, NJ.

Watt, Huerta, and Lozano (2007) A comparison study of AVID and GEAR UP 10th grade students in two high schools in the Rio Grande Valley of Texas. Journal of Education for Students Placed At Risk, 12, 185-212.

7. Effect sizes (Cohen's d and r): interpretation and calculations

* <http://www.bwgriffin.com/gsu/courses/edur9131/content/Hallam_RR_backgroundmusic.pdf> (table 4, 118) t-test, ANOVA
* <http://www.bwgriffin.com/gsu/courses/edur9131/activities/Menon_ST_2001_employee_empowerment_Applied_Psychology.pdf> (table 1 Menon correlations)