Effect Size, Sample Size, and Power
EDUR 9131: 22 April 2023
Bryan W Griffin

Topics

1. Review Hypothesis Testing Concepts
2. Effect Size d
3. Sample Size with d
4. Effect Size r; Sample Size with r
5. A Priori Power (Sensitivity Analysis)
6. Effect Size f; Sample Size with f
7. Effect Size f and ANCOVA
8. Sample size with categorical variables (to be added)

1. Review of Hypothesis Testing Concepts

1.1 Typical Null and Alternative Hypotheses

Group Comparisons (IV is Categorical, DV is Quantitative)

Null Hypothesis

Written

There is no difference in mean achievement between female and male students.

Symbolic

Ho: $\mu_f = \mu_m$

Alternative Hypothesis

Written

There is a difference in mean achievement between female and male students.

Symbolic

Ha: $\mu_f \neq \mu_m$

Relationships (IV is Quantitative, DV is Quantitative)

Null Hypothesis

Written

There is no correlation between hours studied and achievement.

Symbolic

Ho: $\rho = 0.00$

Alternative Hypothesis

Written

There is a correlation between hours studied and achievement.

Symbolic

Ha: $\rho \neq 0.00$

1.2 Types 1 and 2 Errors

Type 1 error

- incorrectly rejecting a true null hypothesis;
- claiming an effect using sample results when there is not an effect in the population;
- a false positive

Example

Using sample data, one concludes a new teaching strategy produces higher achievement than traditional practices when, in fact, it does not show this difference in the population

Type 2 error

- failure to reject a false null hypothesis;
- failure to identify an effect in the sample when there is an effect in the population;
- a false negative

Example

Based upon a sample one claims a new teaching strategy does not produce higher achievement than traditional practices when, in fact, it does in the population

Type 1 and 2 Errors Illustrated

Assume the following null hypothesis:

Ho: patient is not pregnant.

Question 1

Which of the following diagnoses would be Type 1 and Type 2 errors?

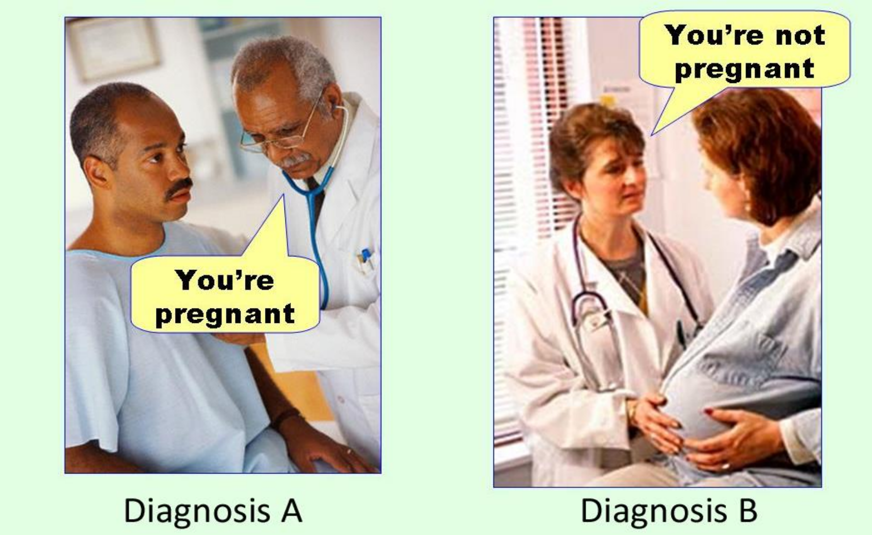
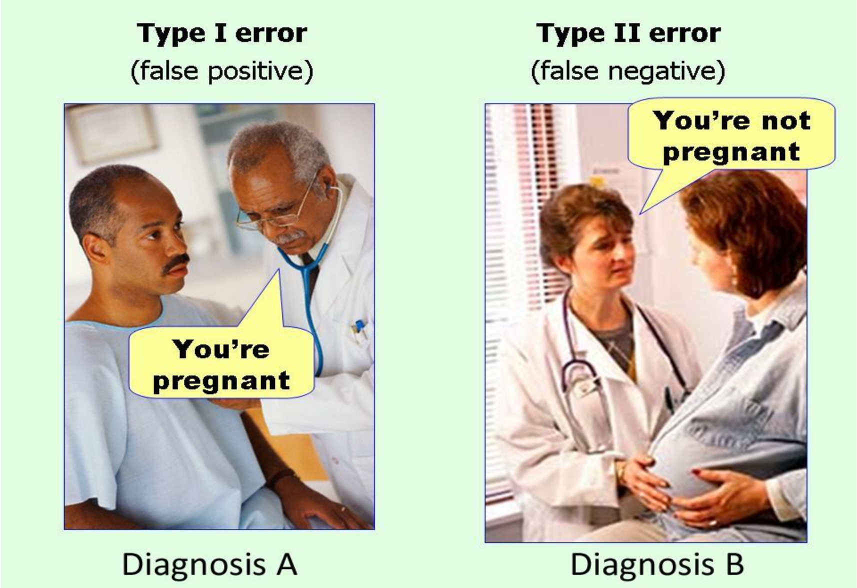


Image source

http://effectsizefaq.com/2010/05/31/i-always-get-confused-about-type-i-and-ii-errors-can-you-show-me-something-to-help-me-remember-the-difference/

Question 1 Answer



1.2 Alpha, Beta, Power, and Others

alpha (α)

- probability of a Type 1 error;
- normally .05 or .01, values .10 and .001 occasionally used;
- also called significance level

Example

There's a 5% chance we will find a relationship between hours studied and achievement in our sample when there isn't a relationship in the population

Based upon our sample evidence, there's a 1% chance we will claim our new teaching strategy is more effective when it is not more effective in the population

beta (β)

- probability of a Type 2 error;
- researchers often strive to set this rate .20 or less
-

Example

There's a 20% chance we will claim there is no relationship between hours studied and achievement when there is a relationship in the population

There's a 10% chance we will claim our new teaching strategy is not more effective when it is more effective in the population

power $(\mathbf{1-\beta})$

- probability of identifying an effect if one exists;
- probability of rejecting a false null hypothesis;
- complement of beta $(\mathbf{1-\beta})$ so it is the probability of not committing a Type 2 error

Example

There's an 80% chance we will detect a relationship between hours studied and achievement in our sample if there is a relationship in the population

With this small sample there is only a 40% chance we will correctly claim our new teaching strategy is more effective when it is more effective

n is study sample size

Example

In our pregnancy study we tested 15 doctors to determine if they could correctly identify pregnancy status of patients

To test our new teaching strategy, we sampled 108,637 students; 46,714 students received the new teaching strategy and 61,923 received current standard instructional approaches

effect size

- effect size denotes the magnitude of difference if comparing means or magnitude of relationship between variables
- may be unstandardized or standardized
- four commonly used ESs

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.

f^2: Ratio of variance predicted to variance not predicted

ES f^2 is the ratio of variance predicted to variance not predicted. In regression the index of variance predicted is R^2. If a regression model predicts 25% of the variance in the DV, then $R^2 = .25$. The amount of variance not predicted by the model is $1 - R^2$ or $1 - .25 = .75$. For this example, f^2 is .25 / .75 or 0.33; could also define f^2 as the signal to noise ratio. Traditionally researchers associate R^2 with regression, so f^2 has been associated with regression-type models.

Significant result

- simply means Ho rejected;
- does not mean results are important;
- larger samples are likely to lead to significant results even when effect sizes are trivial

Example

There is a statistically significant difference in student achievement between those exposed to the new teaching strategy and those exposed to traditional instruction (mean achievement scores: 83.35 vs. 83.23, SD = 7.00, d = .017).

(Note: To have an 80% chance of declaring this small difference “significant” at the .05 level requires a sample of 108,637 students)

Insignificant result

- simply means Ho was not rejected;
- does not mean results are unimportant

Example

There is not a statistically significant difference in red blood cell destruction counts between those taking Soliris (nonfictional drug) and those taking the rival experimental OxygenCell (fictional drug).

For this study why would failure to reject the null be important?

> “Alexion Pharmaceutical’s Soliris, at $409,500 a year, is the world’s single most expensive drug. This monoclonal antibody drug treats a rare disorder in which the immune system destroys red blood cells at night. The disorder, paroxysymal nocturnal hemoglobinuria (PNH), hits 8,000 Americans. Last year Soliris sales were $295 million.”
>
> Source: www.forbes.com/sites/matthewherper/2012/09/05/how-a-440000-drug-is-turning-alexion-into-biotechs-new-innovation-powerhouse

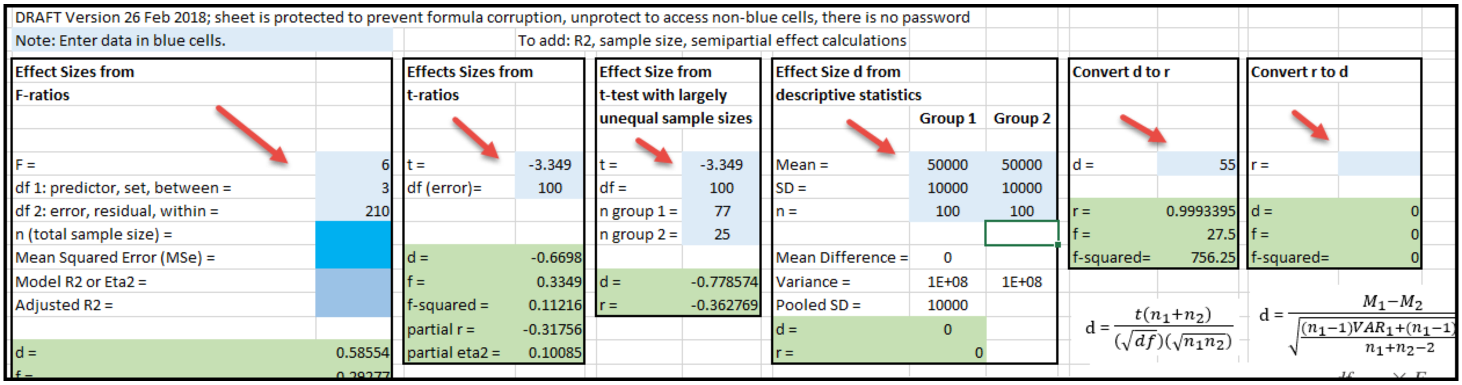
If OxygenCell costs only $120 per year then failure to find a difference (i.e., failure to reject the null) in red blood cell destruction counts between OxygenCell and Soliris means equivalent effectiveness at a huge personal cost savings.

2. Effect Size d

Effect Size in Excel

This is a draft sheet designed to calculate various effect sizes. Enter numbers in Blue cells only.

http://bwgriffin.com/gsu/courses/edur9131/stat-data/effect-size.xlsx



DRAFT Version 26 Feb 2018; sheet is protected to prevent formula corruption, unprotect to access non-blue cells, there is no password

Note: Enter data in blue cells.

To add: R2, sample size, semipartial effect calculations

Effect Sizes from F-ratios	
F =	6
df 1: predictor, set, between =	3
df 2: error, residual, within =	210
n (total sample size) =	
Mean Squared Error (MSe) =	
Model R2 or Eta2 =	
Adjusted R2 =	
d =	0.58554
f =	0.29277

Effects Sizes from t-ratios	
t =	-3.349
df (error)=	100
d =	-0.6698
f =	0.3349
f-squared =	0.11216
partial r =	-0.31756
partial eta2 =	0.10085

Effect Size from t-test with largely unequal sample sizes	
t =	-3.349
df =	100
n group 1 =	77
n group 2 =	25
d =	-0.778574
r =	-0.362769

Effect Size d from descriptive statistics	**Group 1**	**Group 2**
Mean =	50000	50000
SD =	10000	10000
n =	100	100
Mean Difference =	0	
Variance =	1E+08	1E+08
Pooled SD =	10000	
d =	0	
r =	0	

Convert d to r	
d =	55
r =	0.9993395
f =	27.5
f-squared=	756.25

Convert r to d	
r =	
d =	0
f =	0
f-squared=	0

$$d = \frac{t(n_1+n_2)}{(\sqrt{df})(\sqrt{n_1 n_2})}$$

$$d = \frac{M_1 - M_2}{\sqrt{\frac{(n_1-1)VAR_1+(n_1-1)}{n_1+n_2-2}}}$$

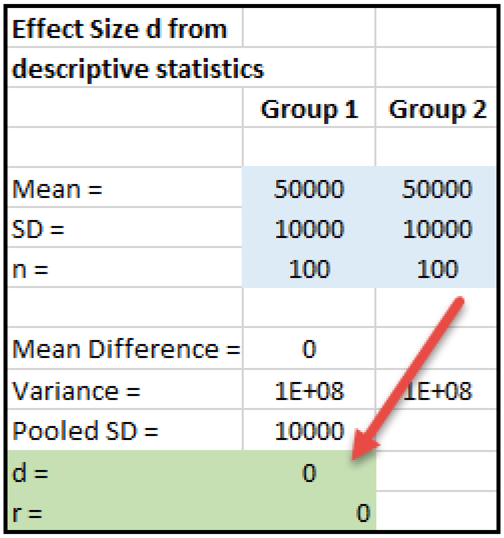
Cohen's d

Recall ES d (Cohen, 1988):

$$d = \frac{Mean_1 - Mean_2}{Standard\ Deviation\ Pooled}$$

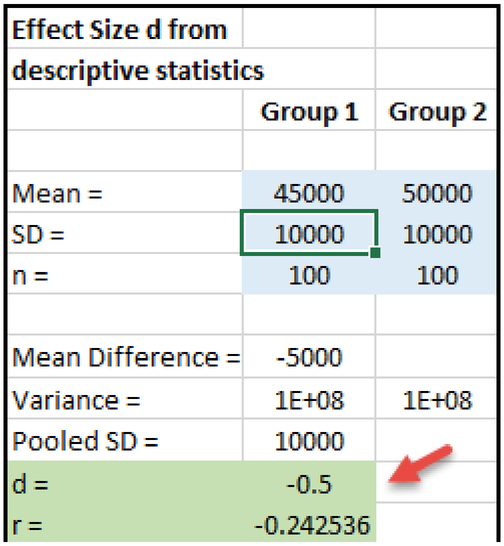
d tells us the following

- 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
    - $d = \frac{50{,}000 - 50{,}000}{10{,}000} = \frac{0}{10{,}000} = 0.00$



Effect Size d from		
descriptive statistics		
	Group 1	**Group 2**
Mean =	50000	50000
SD =	10000	10000
n =	100	100
Mean Difference =	0	
Variance =	1E+08	1E+08
Pooled SD =	10000	
d =	0	
r =	0	

- 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
  - $d = \dfrac{45{,}000 - 50{,}000}{10{,}000} = \dfrac{-5{,}000}{10{,}000} = -\,0.50$



Effect Size d from		
descriptive statistics		
	Group 1	**Group 2**
Mean =	45000	50000
SD =	10000	10000
n =	100	100
Mean Difference =	-5000	
Variance =	1E+08	1E+08
Pooled SD =	10000	
d =	-0.5	
r =	-0.242536	

- 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
  - $d = \dfrac{60{,}000 - 50{,}000}{10{,}000} = \dfrac{10{,}000}{10{,}000} = 1.00$

- 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
  - $d = \frac{70{,}000 - 50{,}000}{10{,}000} = \frac{20{,}000}{10{,}000} = 2.00$

Example

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

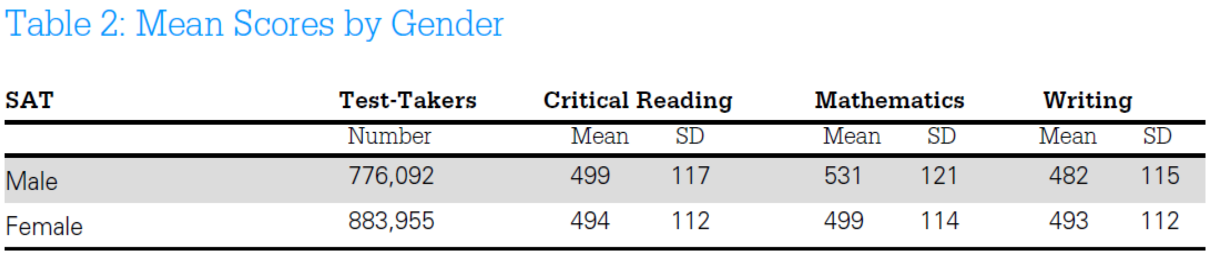


Table 2: Mean Scores by Gender

SAT	Test-Takers	Critical Reading		Mathematics		Writing	
	Number	Mean	SD	Mean	SD	Mean	SD
Male	776,092	499	117	531	121	482	115
Female	883,955	494	112	499	114	493	112

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

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

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

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

Effect size d can now be calculated using this SD:

$$d = \frac{M_1 - M_2}{\sqrt{\frac{(n_1-1)VAR_1+(n_1-1)VAR_2}{n_1+n_2-2}}}$$

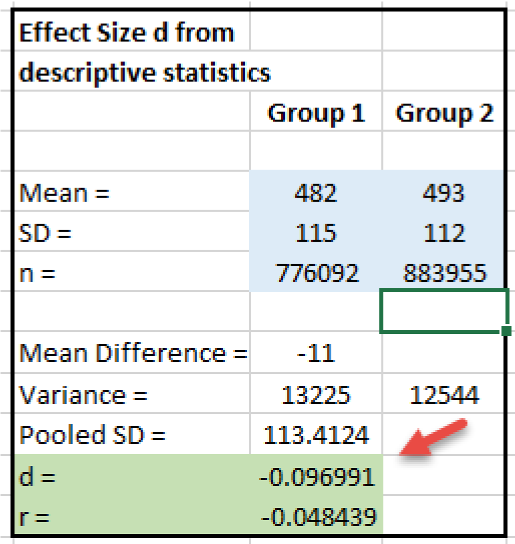
$$d = \frac{Mean_1 - Mean_2}{Standard\ Deviation\ Pooled}$$

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

Interpretation:

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

Use the Excel effect size calculators to estimate the effect size for writing score difference between female and males.



Effect Size d from descriptive statistics		
	Group 1	Group 2
Mean =	482	493
SD =	115	112
n =	776092	883955
Mean Difference =	-11	
Variance =	13225	12544
Pooled SD =	113.4124	
d =	-0.096991	
r =	-0.048439	

Example

Hall, J. M., & Ponton, M. K. (2005). Mathematics self-efficacy of college freshman. *Journal of developmental education*, *28*(3), 26.

Below is a table reported by Hall and Ponton (2005) comparing mathematics self-efficacy between students in calculus 1 and intermediate algebra.

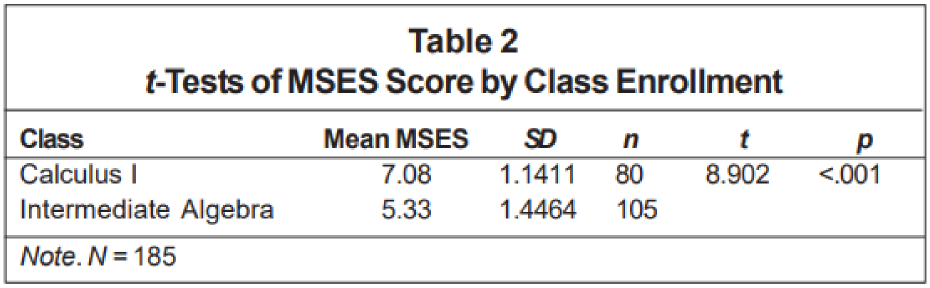


Table 2

t-Tests of MSES Score by Class Enrollment

Class	Mean MSES	*SD*	*n*	*t*	*p*
Calculus I	7.08	1.1411	80	8.902	<.001
Intermediate Algebra	5.33	1.4464	105		

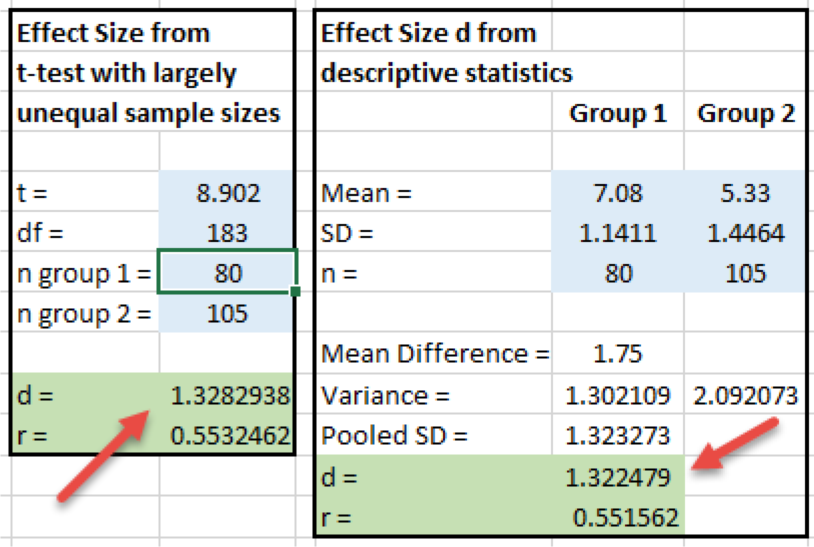
Note. N = 185

Using the t-ratio, what effect size d is obtained for self-efficacy differences between these two groups?

d = 1.328

Using the mean, SD, and group sample sizes, what effect size d is obtained for self-efficacy differences between these two groups?

d = 1.322



Effect Size from t-test with largely unequal sample sizes	
t =	8.902
df =	183
n group 1 =	80
n group 2 =	105
d =	1.3282938
r =	0.5532462

Effect Size d from descriptive statistics		
	Group 1	Group 2
Mean =	7.08	5.33
SD =	1.1411	1.4464
n =	80	105
Mean Difference =	1.75	
Variance =	1.302109	2.092073
Pooled SD =	1.323273	
d =	1.322479	
r =	0.551562	

Example

Bates, A. B., Latham, N., & Kim, J. A. (2011). Linking preservice teachers' mathematics Self-Efficacy and mathematics teaching efficacy to their mathematical performance. *School Science and Mathematics*, *111*(7), 325-333.

Bates et al (2011) provided the following table of M, SD, n, and t-values comparing preservice teachers who scored low vs high on basic mathematics skills.

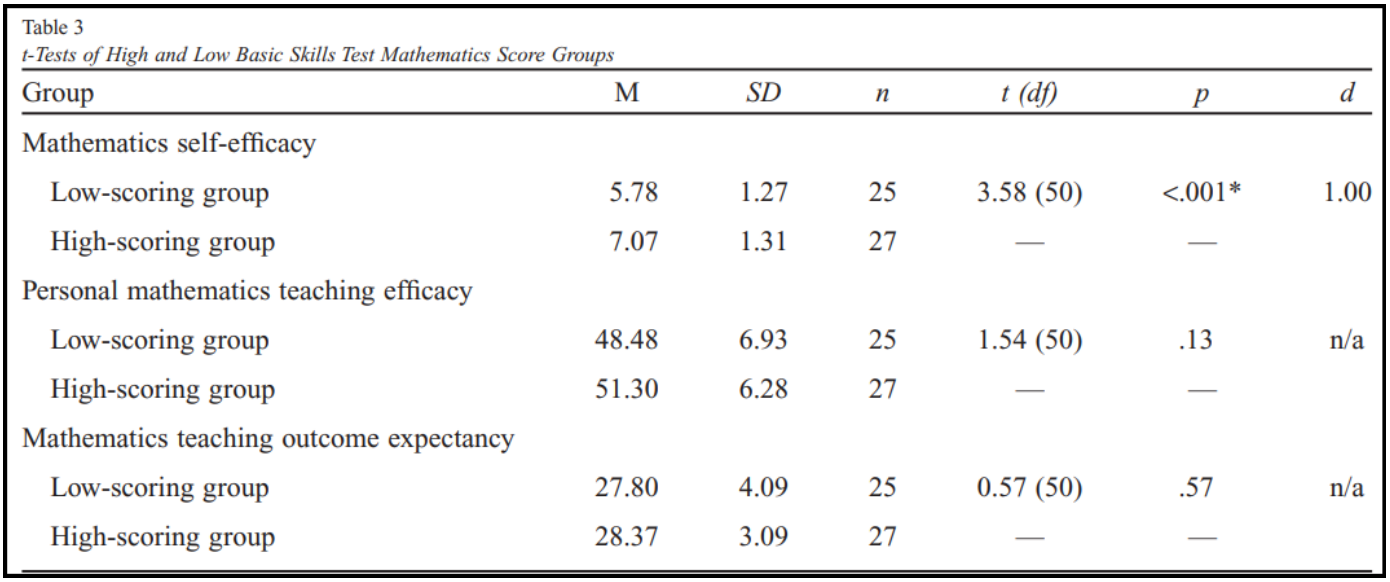
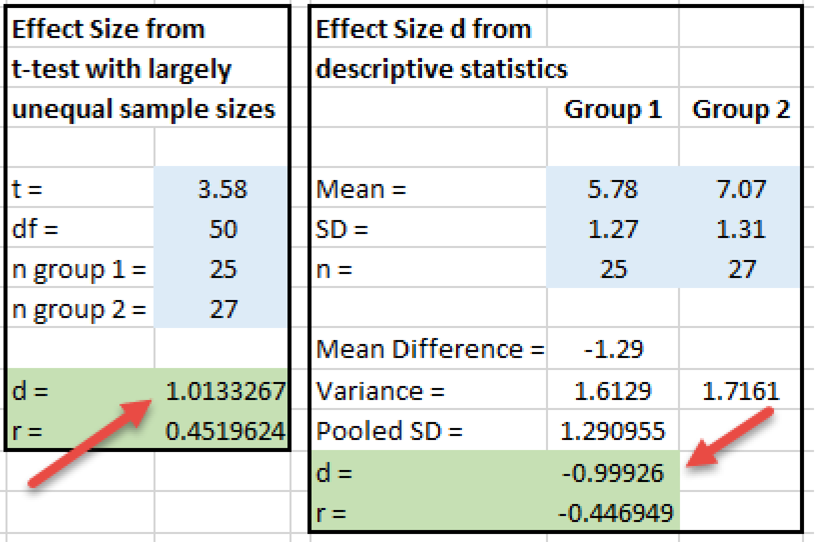


Table 3

t-Tests of High and Low Basic Skills Test Mathematics Score Groups

Group	M	*SD*	*n*	*t (df)*	*p*	*d*
Mathematics self-efficacy						
Low-scoring group	5.78	1.27	25	3.58 (50)	<.001*	1.00
High-scoring group	7.07	1.31	27	—	—	
Personal mathematics teaching efficacy						
Low-scoring group	48.48	6.93	25	1.54 (50)	.13	n/a
High-scoring group	51.30	6.28	27	—	—	
Mathematics teaching outcome expectancy						
Low-scoring group	27.80	4.09	25	0.57 (50)	.57	n/a
High-scoring group	28.37	3.09	27	—	—	

What effect size d do you get for the math self-efficacy mean differences between the groups? Bates et all reported a d of 1.00. Use the two procedures, M SD n for one d estimate, and t-ratio for the other d estimate.



Effect Size from t-test with largely unequal sample sizes	
t =	3.58
df =	50
n group 1 =	25
n group 2 =	27
d =	1.0133267
r =	0.4519624

Effect Size d from descriptive statistics		
	Group 1	Group 2
Mean =	5.78	7.07
SD =	1.27	1.31
n =	25	27
Mean Difference =	-1.29	
Variance =	1.6129	1.7161
Pooled SD =	1.290955	
d =	-0.99926	
r =	-0.446949	

Using descriptive statistics, the d = 1.01 and using the t-ratio d = -0.99. Why the difference?

Cohen's Guidelines/Recommendations/Suggestions for d

In the absence of prior research from which effect size d may be calculated, Cohen (1992; 1988, p. 24+) offered the following values of d as small, medium, and large. Many in social science researcher have adopted these guidelines.

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

3. Sample Size with d

Samples Sizes with d in Excel

Excel sheet designed to calculate samples sizes from effect sizes. Enter numbers in Blue cells only.

http://bwgriffin.com/gsu/courses/edur9131/stat-data/samplesizelinear_revision.28.Feb.2018.xlsm

Update April 2023 – Microsoft disables macros in downloaded Excel files, so this sheet won't work because it requires use of macros to function properly since sample size determination requires an iterative solution.

We will use Cohen's (1992) tables to estimate required sample sizes. Whatever effect size found or stated, use the Cohen effect size that is smaller. Example: d = .41, so use .20 in Cohen's table, not .50.

http://bwgriffin.com/gsu/courses/edur9131/stat-data/Cohen-1992-sampling-tables.pdf

http://bwgriffin.com/gsu/courses/edur9131/stat-data/Cohen-1997-sampling-tables.pdf

Cohen's 1992 Tables

Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychol. Bull.*, *112*, 1155-1159.

Cohen (1992) Table 1, p. 157, lists several effect sizes besides d and r, for example, f and f^2. Each of these can be viewed as equivalents to d and r for interpretation purposes. For each type of effect size, Cohen provides suggested standards for small, medium, and large based upon his research experience in psychology.

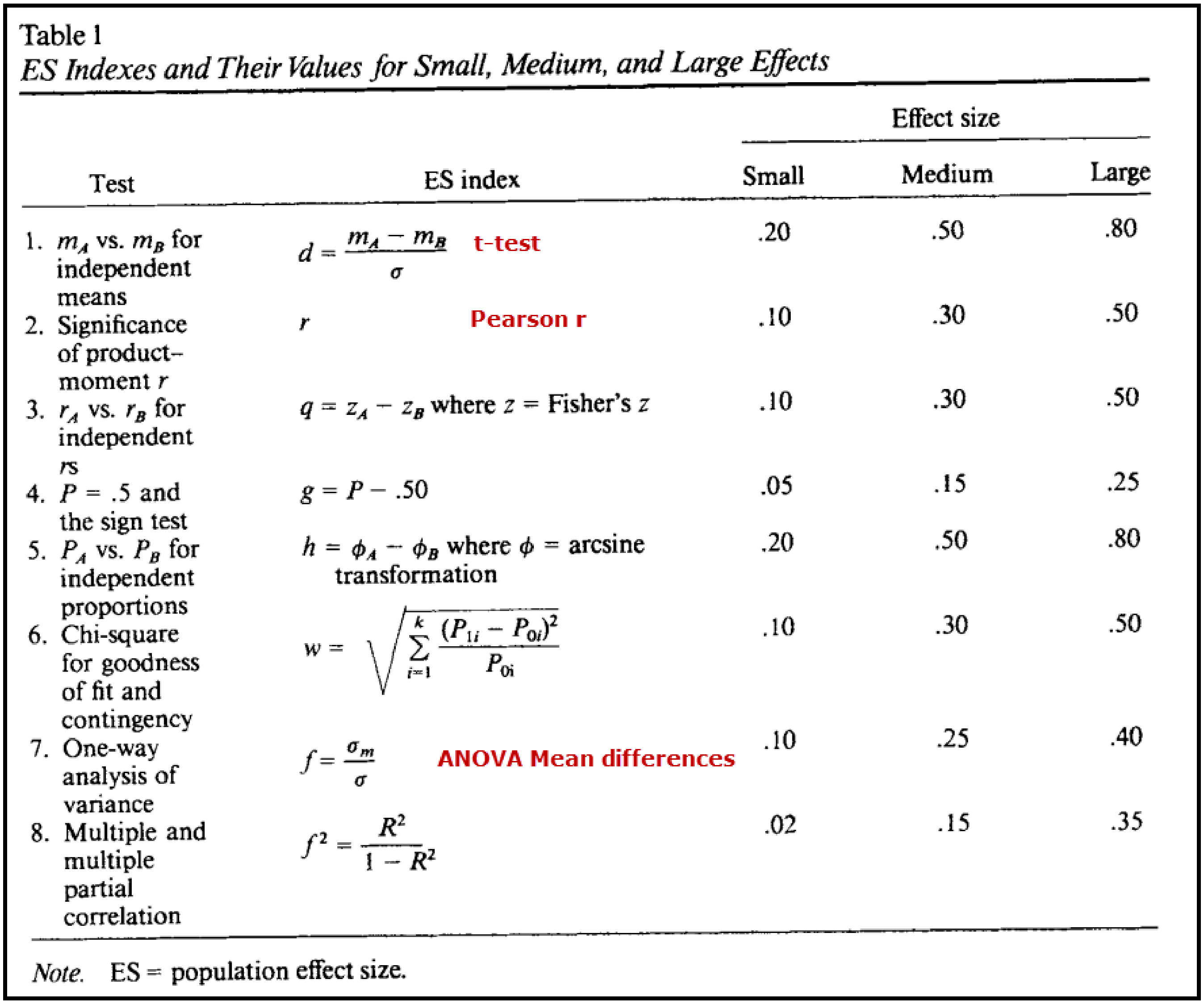


Table 1
ES Indexes and Their Values for Small, Medium, and Large Effects

Test	ES index	Effect size		
		Small	Medium	Large
1. m_A vs. m_B for independent means	$d = \frac{m_A - m_B}{\sigma}$ t-test	.20	.50	.80
2. Significance of product–moment r	r Pearson r	.10	.30	.50
3. r_A vs. r_B for independent rs	$q = z_A - z_B$ where z = Fisher's z	.10	.30	.50
4. P = .5 and the sign test	$g = P - .50$	.05	.15	.25
5. P_A vs. P_B for independent proportions	$h = \phi_A - \phi_B$ where ϕ = arcsine transformation	.20	.50	.80
6. Chi-square for goodness of fit and contingency	$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	.10	.30	.50
7. One-way analysis of variance	$f = \frac{\sigma_m}{\sigma}$ ANOVA Mean differences	.10	.25	.40
8. Multiple and multiple partial correlation	$f^2 = \frac{R^2}{1 - R^2}$	.02	.15	.35

Note. ES = population effect size.

Table 2 shows sample sizes for combinations of effect sizes (small, medium, large) and alpha levels (.01, .05, and .10). Note from the table title that only one power level is used, 0.80. Each additional power level (e.g., .85, .75) would require another table. Table from Cohen (1992) page 158.

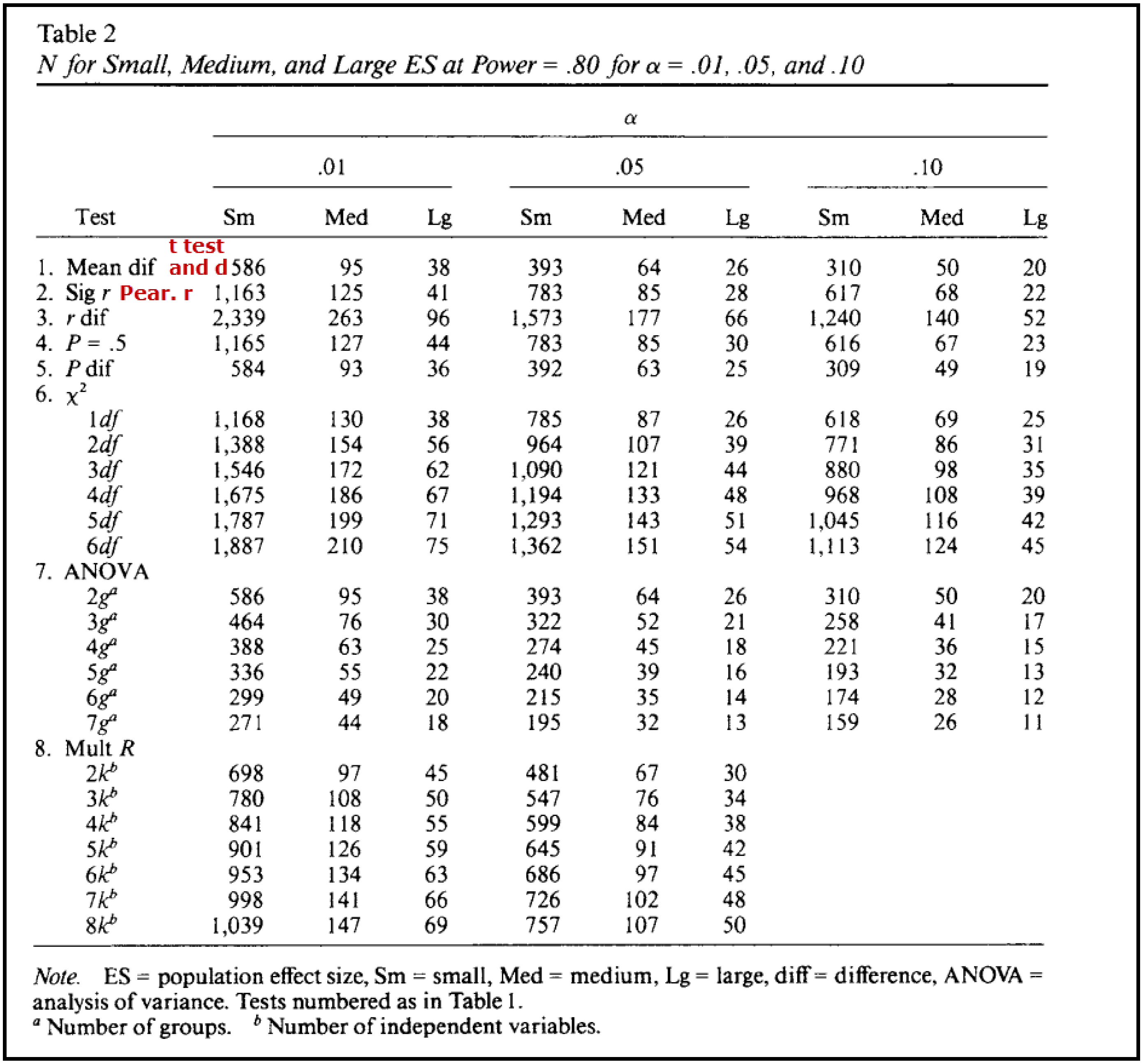


Table 2

N for Small, Medium, and Large ES at Power = .80 for α = .01, .05, and .10

	α								
	.01			.05			.10		
Test	Sm	Med	Lg	Sm	Med	Lg	Sm	Med	Lg
1. Mean dif t test and d	586	95	38	393	64	26	310	50	20
2. Sig *r* Pear. r	1,163	125	41	783	85	28	617	68	22
3. *r* dif	2,339	263	96	1,573	177	66	1,240	140	52
4. *P* = .5	1,165	127	44	783	85	30	616	67	23
5. *P* dif	584	93	36	392	63	25	309	49	19
6. χ^2									
1*df*	1,168	130	38	785	87	26	618	69	25
2*df*	1,388	154	56	964	107	39	771	86	31
3*df*	1,546	172	62	1,090	121	44	880	98	35
4*df*	1,675	186	67	1,194	133	48	968	108	39
5*df*	1,787	199	71	1,293	143	51	1,045	116	42
6*df*	1,887	210	75	1,362	151	54	1,113	124	45
7. ANOVA									
$2g^a$	586	95	38	393	64	26	310	50	20
$3g^a$	464	76	30	322	52	21	258	41	17
$4g^a$	388	63	25	274	45	18	221	36	15
$5g^a$	336	55	22	240	39	16	193	32	13
$6g^a$	299	49	20	215	35	14	174	28	12
$7g^a$	271	44	18	195	32	13	159	26	11
8. Mult *R*									
$2k^b$	698	97	45	481	67	30			
$3k^b$	780	108	50	547	76	34			
$4k^b$	841	118	55	599	84	38			
$5k^b$	901	126	59	645	91	42			
$6k^b$	953	134	63	686	97	45			
$7k^b$	998	141	66	726	102	48			
$8k^b$	1,039	147	69	757	107	50			

Note. ES = population effect size, Sm = small, Med = medium, Lg = large, diff = difference, ANOVA = analysis of variance. Tests numbered as in Table 1.

[a] Number of groups. [b] Number of independent variables.

Power = .80

This means a statistical test run with the sample sizes provided and with the effect size assumed (small, medium, or large) has an 80% chance of finding that effect.

α = .01, .05, or .10

Alpha (α) is the probability of making a Type 1 error in hypothesis testing (incorrectly concluding that you found an effect, difference, or relationship when there really is not effect, difference, or relationship in the population).

A value of .01 means this error is expected to occur 1 out of 100 tests.
A value of .05 means this error is expected to occur 5 out of 100 tests.
A value of .10 means this error is expected to occur 10 out of 100 tests.

Most researchers use .05 unless they have large samples and then .01 is used.

1. Mean dif (t-test)

Per-group sample size needed for two-group t-test or one-way ANOVA with two groups.

This means the total sample size will be twice the number presented in the table.

For example, alpha = .05 and large effect size (d) corresponds to a sample size of 26 for each group, thus total sample of 26 + 26 = 52 for both groups combined.

2. Sig r

Sample size needed for testing Pearson correlation.

For example, if we wanted to find a medium correlation of r = .30, with alpha = .01, we would need a sample of 125

7. ANOVA

This section provides sample sizes per group in ANOVA.

The numbers below the title ANOVA represent the number of groups, 2g = two groups, 3g = three groups, 4g = four groups, etc.

As with Mean Dif above, the sample sizes provided are per group, so to find the total sample size one must multiple the sample size provided by the number of groups.

For example, alpha = .05, effect size = medium, and there are 3 groups, the sample size reported is 52, so the total sample size would be 52 x 3 = 156.

Four Components Required to find n

α (alpha)

- probability of Type 1 error
- often set at .05, .01, or .001; sometimes may be as high as .10 for studies with small samples

Effect Size d

Standardized mean difference expressed in standard deviation units

Power $(1 - \boldsymbol{\beta})$

- probability of correctly rejecting a false null;
- probability of finding an effect if one exists in population;
- probability of not committing a Type 2 error;
- often researchers will set power to .80, .85, or .90.
- A value of .95 may be targeted if large samples are available.

degrees of freedom (df)

- number of parameters estimated in general linear model excluding the intercept
- df for common statistical tests
  - df = 1 for two-group t-test
  - df = 1 Pearson correlation
  - df = 2 if comparing 3 groups with ANOVA
  - df = 3 if comparing 4 groups with ANOVA
  - df = 4 if comparing 5 groups with ANOVA, etc.

Which Statistical Tests with d?

Any group comparison in which the

- independent or predictor variable is categorical with two groups (e.g., sex, experimental study of treatment vs. control)
- dependent or outcome variable is quantitative (e.g., test scores in percent correct, anxiety scale score, weight in lbs., distance in feet, etc.)

While effect size d can be associated with a variety of statistical tests, the focus of here will be d for an **independent samples t-test**.

Finding n with Excel

Note – be sure to enable content with the Excel file after downloading, otherwise it will be unable to assess iterative program to find solutions.

Important Limitations

This calculator is suitable only for designs with balanced data (equal or approximately equal sample sizes per group), non-directional alternative hypotheses, linear models with fixed predictors (although random predictors are discussed below), and for non-correlated linear models. For studies with large imbalances in sample size, directional tests, or correlated data, use the free software G*Power: http://www.gpower.hhu.de/en.html

Example

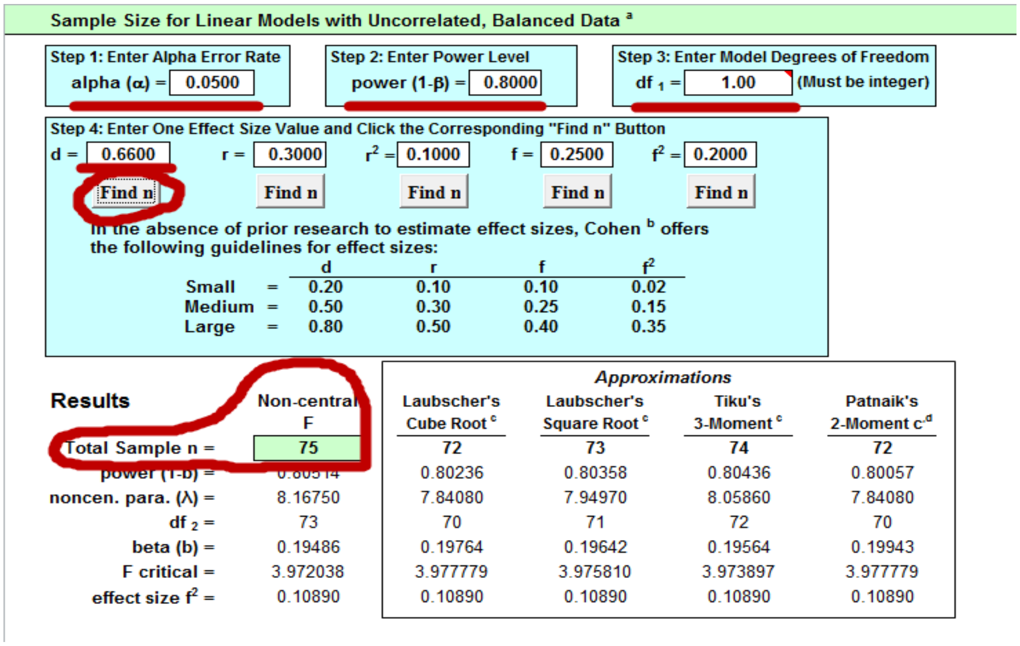
College graduates have an estimated IQ of 115 while high school graduates have an estimated IQ of 105. The standard deviation of WAIS is about 15. Source: http://www.assessmentpsychology.com/iq.htm

$$d = \frac{115 - 105}{15} = \frac{10}{15} = .667$$

Question 2

If comparing IQ between two groups, which size sample would be needed to

- detect an IQ difference of d = 0.66 when
- alpha = .05,
- power = .80, and
- df = 1 (standard value of df for finding n for t-test)



Sample Size for Linear Models with Uncorrelated, Balanced Data [a]

Step 1: Enter Alpha Error Rate
alpha (α) = 0.0500

Step 2: Enter Power Level
power (1-β) = 0.8000

Step 3: Enter Model Degrees of Freedom
df $_1$ = 1.00 (Must be integer)

Step 4: Enter One Effect Size Value and Click the Corresponding "Find n" Button
d = 0.6600 r = 0.3000 r^2 = 0.1000 f = 0.2500 f^2 = 0.2000

In the absence of prior research to estimate effect sizes, Cohen [b] offers the following guidelines for effect sizes:

		d	r	f	f^2
Small	=	0.20	0.10	0.10	0.02
Medium	=	0.50	0.30	0.25	0.15
Large	=	0.80	0.50	0.40	0.35

		Approximations			
Results	Non-central F	Laubscher's Cube Root [c]	Laubscher's Square Root [c]	Tiku's 3-Moment [c]	Patnaik's 2-Moment c [d]
Total Sample n =	75	72	73	74	72
power (1-b) =	0.80514	0.80236	0.80358	0.80436	0.80057
noncen. para. (λ) =	8.16750	7.84080	7.94970	8.05860	7.84080
df $_2$ =	73	70	71	72	70
beta (b) =	0.19486	0.19764	0.19642	0.19564	0.19943
F critical =	3.972038	3.977779	3.975810	3.973897	3.977779
effect size f^2 =	0.10890	0.10890	0.10890	0.10890	0.10890

Answer

Total n = 75 or 75 / 2 = 37.5 ~ 38 per group.

A sample of 38 participants per group (total n = 76) would provide an 80% chance of detecting an effect size d = .66 at a Type 1 error rate of .05.

Answer (April 2023 Update)

Using Cohen's Table 2

1. Convert d to closest listed d provided by Cohen. Since our calculated d is between two values listed by Cohen, use the smaller value. Since d = .66, Cohen's values are .20, .50, and .80. Since .66 is less than .80. we must use Cohen's .50 (medium effect size).

2. Find the tabled sample size for:

- alpha = .05
- power = .80 (only value provided by Cohen)
- d = .50 (1. Mean dif)

3. Tabled value is 64 per group, so total sample is n = 64*2 = 128.

Question 3

Using the sample specifications as above (d = .66, alpha =.05), what size sample would be needed if power were raised to .95 instead of .80?

Answer

Total n = 122 so 122 / 2 = 61 participants per group

Answer (April 2023 Update)

Cannot address this question with Cohen's Table 2 since only power = .80 is provided.

Example: Writing SAT d

Recall the calculated d for Writing SAT scores between females and males.

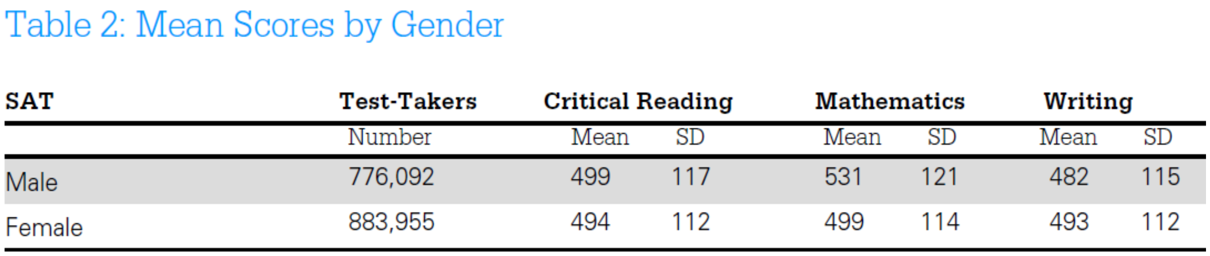


Table 2: Mean Scores by Gender

SAT	Test-Takers	Critical Reading		Mathematics		Writing	
	Number	Mean	SD	Mean	SD	Mean	SD
Male	776,092	499	117	531	121	482	115
Female	883,955	494	112	499	114	493	112

$$d = \frac{M_1 - M_2}{\sqrt{\frac{(n_1-1)VAR_1 + (n_1-1)VAR_2}{n_1+n_2-2}}}$$

d = $\frac{Mean_1 - Mean_2}{Standard\ Deviation\ Pooled}$

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

Question 4

What size sample would be needed to detect a mean difference in SAT writing scores of d = .097 when

- alpha = .01
- power = .90
- and df = 1 (as usual for two-group t-tests)

Answer

Total n = 6,329 so 6,329 / 2 = 3,164.5 or 3,165 participants per group. Large samples are required to detect small differences, especially with low Type 1 error rates and high power.

Answer (April 2023 Update)

Using Cohen's Table 2

1. Convert d to closest listed d provided by Cohen. The calculated d = .097 and the smallest d provided by Cohen is .20, we will use that. Note also we have to use power of .80.

2. Find the tabled sample size for:

- alpha = .01
- power = .80 (only value provided by Cohen)
- d = .20 (1. Mean dif)

3. Tabled value is 586 per group, so total sample is n = 586*2 = 1,172.

Question 5

What happens to n in Question 4 above if we are more willing to increase the Type 1 rate to .05 and reduce power to .80?

- d = .097
- alpha = .05
- power = .80
- and df = 1 (as usual for two-group t-tests)

Answer

Total n = 3,339 so 3,339 / 2 = 1,669.5 or 1,670 participants per group.

Answer (April 2023 Update)
Using Cohen's Table 2

1. Convert d to closest listed d provided by Cohen. The calculated d = .097 and the smallest d provided by Cohen is .20, we will use that.

2. Find the tabled sample size for:

- alpha = .05
- power = .80 (only value provided by Cohen)
- d = .20 (1. Mean dif)

3. Tabled value is 586 per group, so total sample is n = 393*2 = 786.

Example: Mathematics SAT d
What is ES d for the female vs male comparison of math SAT scores?

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

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

Effect size d can now be calculated using this SD:

$$d = \frac{Mean_1 - Mean_2}{Standard\ Deviation\ Pooled}$$

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

Question 6
What size sample would be needed to detect a mean difference in SAT mathematics scores of d = -.27 when

- alpha = .05
- power = .85
- and df = 1 (as usual for two-group t-tests)

Answer

Total n = 488 so 488 / 2 = 244 participants per group.

Answer (April 2023 Update)

Using Cohen's Table 2

1. Convert d to closest listed d provided by Cohen. The calculated d = -.272 and the smallest d provided by Cohen is -.20 (or .20 in absolute value), we will use that. Also have to use power of .80.

2. Find the tabled sample size for:

- alpha = .05
- power = .80 (only value provided by Cohen)
- d = -.20 (1. Mean dif)

3. Tabled value is 586 per group, so total sample is n = 393*2 = 786.

Question 7

Cohen (1992, p. 159) poses the following scenario

- detect medium difference between two means (d = .50),
- alpha = .05,
- power = .80, and
- df = 1 (note that for independent samples t-tests, df will be 1 for sample size calculations)

Answer

Cohen (1992) indicates the answer is n = 64 per group; since there are two groups, the total n = 128 which matches the result found with Excel.

Answer (April 2023 Update)

Using Cohen's Table 2. We will revise this example to have an alpha of .10.

1. Find the tabled sample size for:

- alpha = .10
- power = .80 (only value provided by Cohen)
- d = .50 (1. Mean dif)

2. Tabled value is 50 per group, so total sample is n = 50*2 = 100.

4. Effect Size r; Sample Size with 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 (1992) 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 Sample Size with r

Values Need to find n

- α **(alpha),** often set at .05, .01, or .001.
- **Power (1-β)** usually to .80, .85, or .90. A value of .95 may be targeted if large samples are available.
- **Effect Size r**
- **Degrees of Freedom (df)** which defaults to 1 for zero-order correlations.

4.4 Finding n for r with Excel

Example

I collected data to assess the association between **perceived autonomy support** and **student ratings of instruction**. To measure **perceived autonomy support** students were asked to respond to the following three items:

- "The instructor was willing to negotiate course requirements with students,"
- "Students had some choice in course requirements or activities that would affect their grade," and
- "The instructor made changes to course requirements or activities as a result of student comments or concerns."

Responses to these items ranged from 1 ("strongly disagree") to 5 ("Strongly agree"). Cronbach's alpha for these three items was α =.85.

Overall rating of the instructor was assessed by responses to this item:

10. Overall, how would you rate this instructor?

with response options ranging from 1 = "Poor" to 5 = "Excellent."

Results

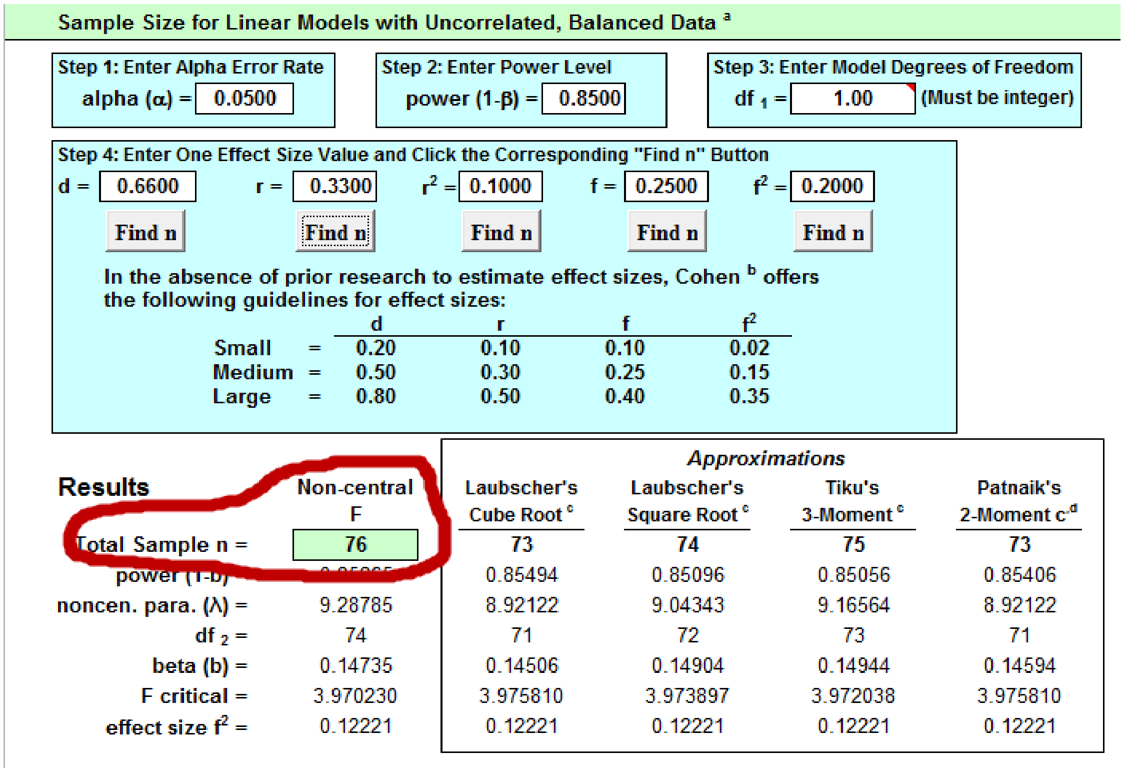
	Rating of Instructor	Autonomy Support
Mean	3.95	3.50
Standard Deviation	1.14	1.06

Pearson r = .33
n = 914

Question 8

What sample size is needed to detect an r = .33 for

- alpha = .05,
- power = .85, and
- df = 1 (default df for Pearson r when calculating sample size).



Sample Size for Linear Models with Uncorrelated, Balanced Data [a]

Step 1: Enter Alpha Error Rate
alpha (α) = 0.0500

Step 2: Enter Power Level
power (1-β) = 0.8500

Step 3: Enter Model Degrees of Freedom
df $_1$ = 1.00 (Must be integer)

Step 4: Enter One Effect Size Value and Click the Corresponding "Find n" Button

d = 0.6600 r = 0.3300 r^2 = 0.1000 f = 0.2500 f^2 = 0.2000

In the absence of prior research to estimate effect sizes, Cohen [b] offers the following guidelines for effect sizes:

		d	r	f	f^2
Small	=	0.20	0.10	0.10	0.02
Medium	=	0.50	0.30	0.25	0.15
Large	=	0.80	0.50	0.40	0.35

		Approximations			
Results	Non-central F	Laubscher's Cube Root [c]	Laubscher's Square Root [c]	Tiku's 3-Moment [c]	Patnaik's 2-Moment c [d]
Total Sample n =	76	73	74	75	73
power (1-b)	0.85265	0.85494	0.85096	0.85056	0.85406
noncen. para. (λ) =	9.28785	8.92122	9.04343	9.16564	8.92122
df $_2$ =	74	71	72	73	71
beta (b) =	0.14735	0.14506	0.14904	0.14944	0.14594
F critical =	3.970230	3.975810	3.973897	3.972038	3.975810
effect size f^2 =	0.12221	0.12221	0.12221	0.12221	0.12221

Answer

n = 76

Answer (April 2023 Update)

Using Cohen's Table 2

1. Convert r to closest listed r provided by Cohen. The calculated r = .33 and values for r suggested by Cohen are .10, .30, and .50. The r of .33 is between .30 and .50, so use the smaller of the two, .30. Also we have to use power of .80.

2. Find the tabled sample size for:

- alpha = .05
- power = .80 (only value provided by Cohen)
- r = .30 (2. Sig r)

3. Tabled value is 85.

Important Note About N for Pearson r
(Technical note; does not apply to EDUR 9131 activities or assessments)

If one uses a sample size source that is developed specially Pearson r (e.g., G*Power, Cohen's 1988 or 1992 correlation tables), then there will be minor differences from the sample size reported by the Excel spreadsheet. Normally the difference will be between 2 and 6 cases. For example, using specifications for Question 8 above, G*Power reports n = 79 which is 3 cases more than reported by the Excel spreadsheet.

To compensate, add 4 cases to the sample size listed in Excel (i.e., n + 4 = adjusted n; 76 + 4 = 80). This correction will provide a power level that closely matches the desired power in most research situations (i.e., alpha = .01 or .05, power = .80 to .95).

The reason that a sample size discrepancy exists results from the assumption with the Excel sheet that linear models with fixed predictors are specified. This means the dependent variable (or criterion variable) is assumed to be a random variable, but that all the predictors in the model are assumed to be fixed, not random. This is the same assumption Cohen (1988; 1992) made when developing his power and sample size tables for effect sizes d, f, and f^2, so the power levels and sample size values provided by the Excel sheet and by Cohen will match with the exception for Pearson r.

In situations where both predictors and criterion are assumed to be random, the same size values provided by Cohen and the Excel sheet using effect sizes d, f, or f^2 will be slightly underestimated. Through simulations I have found that adding the model degrees of freedom to the Excel or Cohen specified sample size (i.e, n + model df = adjusted n) creates an adjusted sample size that meets or exceeds specified power in most research situations. Model degrees of freedom represents the number of parameters estimated through regression or ANOVA. For example, for a regression model with three quantitative predictors, the model df = 3; for an ANOVA with 10 groups, the model df = 9; for an ANCOVA or regression with two covariates (quantitative predictors) and two categorical predictors, one with three groups and one with four groups, and no interaction specified, the model df would be: 2 for covariates, 2 for three groups, and 3 for four groups, so df = 7; the adjusted sample size would be n + 7.

I simulated regression with 6 predictors, so model df = 6, and one of predictors has an effect size of $f^2 = 0.05357$. With alpha = .05, power = .90, and df = 1, the Excel computed n = 199. Random samples were taken from these data 250,000 times and the calculated power (number times Ho was rejected) was .8925 (close to the specified value of .90). This was repeated a second time and the empirical power level obtained was .8901 (again, close to the specified value of .90).

The simulations were performed again, but this time with model df added to the sample size, so the adjusted sample size was 199 + 6 = 205. The first run of 250,000 samples produced a power level of .9017 and the second run of 250,000 produced a power level of .8989, both closer to the targeted power level of .90 than the previous estimates. This shows that adding model df to the Excel sample size estimate for models with random predictors provides an easy way to determine sample size for random variate models.

Question 9

Sridevi (2013) reports a correlation of -.22 between test anxiety and midterm examination scores among secondary education students in Pakistan. Source: Sridevi, K. V. (2013). A Study of Relationship among General Anxiety, Test Anxiety and Academic Achievement of Higher Secondary Students. *Journal of Education and Practice*, *4*(1).

What sample size is needed to detect an r = -.22 for

- alpha = .01,
- power = .90, and
- df = 1 (default df for Pearson r when calculating sample size).

Answer

n = 296

Answer (April 2023 Update)

Using Cohen's Table 2

1. Convert r to closest listed r provided by Cohen. The calculated r = -.22 is closest, without exceeding in absolute value, is -.10. Also we have to use power of .80.

2. Find the tabled sample size for:

- alpha = .01
- power = .80 (only value provided by Cohen)
- r = -.10 (2. Sig r)

3. Tabled value is 1,163.

Question 10

Cohen (1992, p. 159) poses the following scenario, what n is needed?

- large r, so r = .50,
- alpha = .01,
- power = .80, and
- df = 1 (note that for independent samples t-tests, df will be 1 for sample size calculations)

Answer

Excel sheet indicates n = 39; Cohen (1992) indicates the answer is n = 41 (again, note the discrepancy in sample size of 2 due to the assumption of fixed vs random variables).

Example

Four studies were found that reported the Pearson correlation between mathematics self-efficacy and anxiety for high school or college students (see presentation on Effect Sizes for these studies). The Pearson r 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 (a recommended approach when averaging correlations). 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

Question 11

What size sample of college students is needed to have 90% chance to detect a correlation of -.47 with alpha at .01?

Answer

Excel sheet reports n = 56.

Answer (April 2023 Update)

Using Cohen's Table 2

1. Convert r to closest listed r provided by Cohen. The calculated r = -.47 is between the values of -.30 and -.50, so we use -.30 for Cohen's table. We have to reduce the 90% chance of detecting the correlation (i.e., power level) to .80 for Cohen's table.

2. Find the tabled sample size for:

- alpha = .01
- power = .80 (only value provided by Cohen)
- r = -.30 (2. Sig r)

3. Tabled value is 125.

5. A Priori Power Analysis (Sensitivity Analysis)

April 2023: This section is not possible with Cohen's tables.

A priori power analysis can be used to determine whether a study will be sensitive enough to detect a given sized effect.

Example

An EdS student wishes to investigate which of two instructional strategies may result in greater student motivation to learn science. Results can be analyzed via a two-group t-test.

The EdS student has access to two classes for this proposed study. The first class has 23 students and the second has 25 students for total sample size of 48.

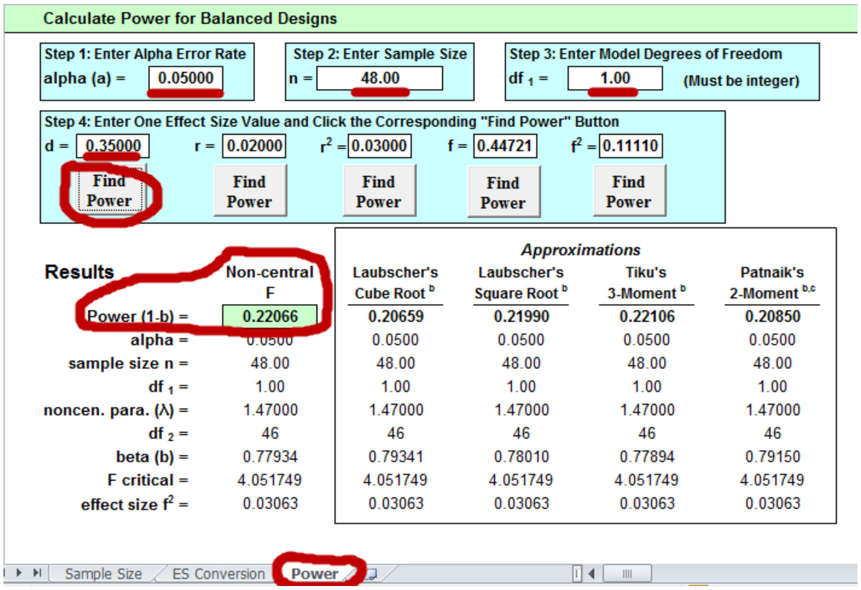
The EdS student also expects a standardized mean difference of about d = .35 given prior published research.

Question 12

What would be the expected power level for this study?

- d = .35
- alpha = .05
- n = 48
- df = 1

Note on the Excel sheet the bottom tab named "Power" – select that tab to access the power calculator.



Calculate Power for Balanced Designs

Step 1: Enter Alpha Error Rate
alpha (a) = 0.05000

Step 2: Enter Sample Size
n = 48.00

Step 3: Enter Model Degrees of Freedom
df $_1$ = 1.00 (Must be integer)

Step 4: Enter One Effect Size Value and Click the Corresponding "Find Power" Button
d = 0.35000 r = 0.02000 r^2 = 0.03000 f = 0.44721 f^2 = 0.11110

Results

		Approximations			
	Non-central F	Laubscher's Cube Root [b]	Laubscher's Square Root [b]	Tiku's 3-Moment [b]	Patnaik's 2-Moment [b,c]
Power (1-b) =	0.22066	0.20659	0.21990	0.22106	0.20850
alpha =	0.0500	0.0500	0.0500	0.0500	0.0500
sample size n =	48.00	48.00	48.00	48.00	48.00
df $_1$ =	1.00	1.00	1.00	1.00	1.00
noncen. para. (λ) =	1.47000	1.47000	1.47000	1.47000	1.47000
df $_2$ =	46	46	46	46	46
beta (b) =	0.77934	0.79341	0.78010	0.77894	0.79150
F critical =	4.051749	4.051749	4.051749	4.051749	4.051749
effect size f^2 =	0.03063	0.03063	0.03063	0.03063	0.03063

Answer

Note that the Excel power analysis worksheet assumes equal group sizes and departures from balanced samples will lead to incorrect power estimates. However, when samples are approximately equal, power estimates should be close to actual power values as demonstrated by this example, Excel power = .22066 and while actual power = .2203 (G*Power estimate of power for imbalanced sample sizes).

Question 13

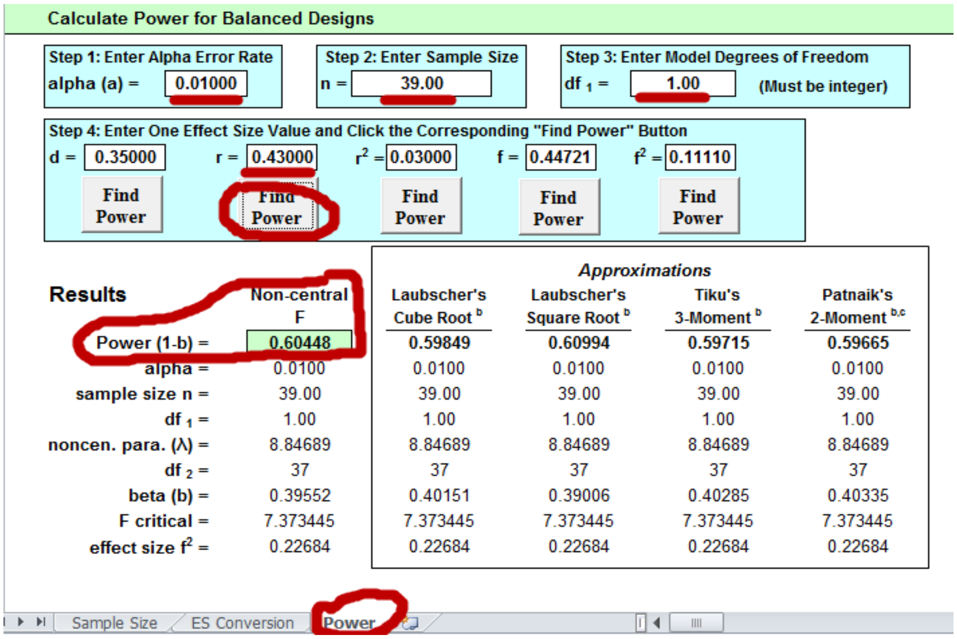
We plan to investigate the correlation between classroom technology usage self-efficacy and classroom technology integration among teachers. The anticipated correlation is r = .43 (as reported in a prior study) and the total number of possible teachers for participation is 39.

What would be the expected power level for this study?

- r = .43
- alpha = .01
- n = 39
- df = 1

Answer

Excel sheet power = .60488 (exact power = .579, discrepancy due to correlation assumption of random variables)



Calculate Power for Balanced Designs

Step 1: Enter Alpha Error Rate
alpha (a) = 0.01000

Step 2: Enter Sample Size
n = 39.00

Step 3: Enter Model Degrees of Freedom
df $_1$ = 1.00 (Must be integer)

Step 4: Enter One Effect Size Value and Click the Corresponding "Find Power" Button
d = 0.35000 r = 0.43000 r^2 = 0.03000 f = 0.44721 f^2 = 0.11110

		Approximations			
Results	Non-central F	Laubscher's Cube Root b	Laubscher's Square Root b	Tiku's 3-Moment b	Patnaik's 2-Moment b,c
Power (1-b) =	0.60448	0.59849	0.60994	0.59715	0.59665
alpha =	0.0100	0.0100	0.0100	0.0100	0.0100
sample size n =	39.00	39.00	39.00	39.00	39.00
df $_1$ =	1.00	1.00	1.00	1.00	1.00
noncen. para. (λ) =	8.84689	8.84689	8.84689	8.84689	8.84689
df $_2$ =	37	37	37	37	37
beta (b) =	0.39552	0.40151	0.39006	0.40285	0.40335
F critical =	7.373445	7.373445	7.373445	7.373445	7.373445
effect size f^2 =	0.22684	0.22684	0.22684	0.22684	0.22684

Question 14

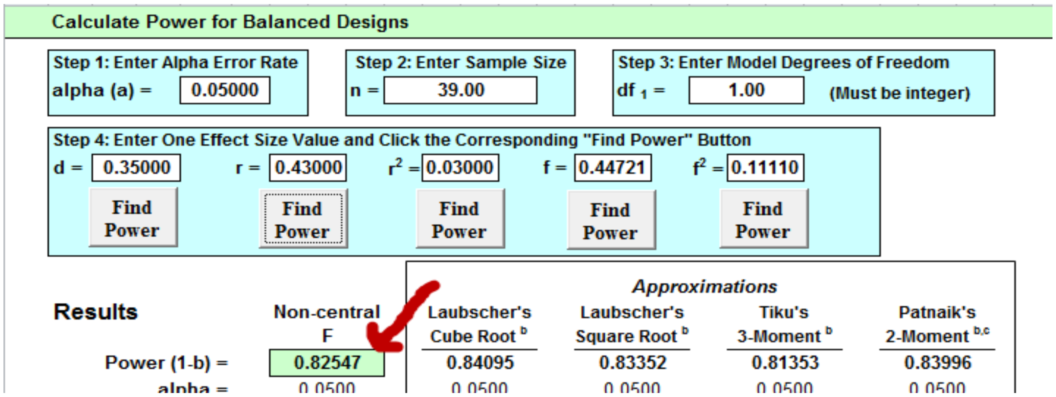
The study outlined in Question 13 has low power (Excel power = .604 and G*Power power = .579). What could be done to increase power for this study if sample size is fixed? What new power level is obtained?

Recall the original values for Question 13:

- r = .43
- alpha = .01
- n = 39
- df = 1

Answer

One option is to raise Type 1 error rate from .01 to .05. This will provide more power to detect a correlation of .43. According to the Excel power calculator, power would now be about .82 (or about .799 according to G*Power).



Calculate Power for Balanced Designs

Step 1: Enter Alpha Error Rate
alpha (a) = 0.05000

Step 2: Enter Sample Size
n = 39.00

Step 3: Enter Model Degrees of Freedom
df $_1$ = 1.00 (Must be integer)

Step 4: Enter One Effect Size Value and Click the Corresponding "Find Power" Button
d = 0.35000 r = 0.43000 r^2 = 0.03000 f = 0.44721 f^2 = 0.11110

Results	Non-central F	*Approximations* Laubscher's Cube Root [b]	Laubscher's Square Root [b]	Tiku's 3-Moment [b]	Patnaik's 2-Moment [b,c]
Power (1-b) =	0.82547	0.84095	0.83352	0.81353	0.83996
alpha =	0.0500	0.0500	0.0500	0.0500	0.0500

Question 15

Consider the study in Question 13 again. If the response rate from participating teachers is only 1/3, this will result in a total sample of 13 teachers. What would be the power level with only 13 teachers with the following values?

- r = .43
- alpha = .10
- n = 13
- df = 1

Answer

With alpha = .05 power is about .347 (Excel results), and with alpha = .10 power is about .486 (Excel results). With G*Power the corresponding power levels would be .322 (for alpha = .05) and .452 (for alpha = .10).

6. Effect Size f

6.1 Analysis of Variance, Effect Size f, η^2 and Multiple R^2

One-way ANOVA

Simple one-way ANOVA models are like independent samples t-tests, except while the t-test is limited to two groups comparisons, 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, $\eta^2$and Multiple R^2

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

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

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 R^2, a measure of model fit in regression, can be used to estimate f:

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

Multiple R^2 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: 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 III Sum 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:

$$R^2 = \frac{SS\ factor}{SS\ total} = \frac{SS\ between}{SS\ total} = \frac{172.252}{1190.799} = 0.1447$$

Degrees of Freedom Due to ANOVA

Lastly, before sample size can be estimated, model df must be calculated. For a one-way ANOVA df will be the number of groups minus one:

df = 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:

df = J - 1 = 3 - 1 = 2

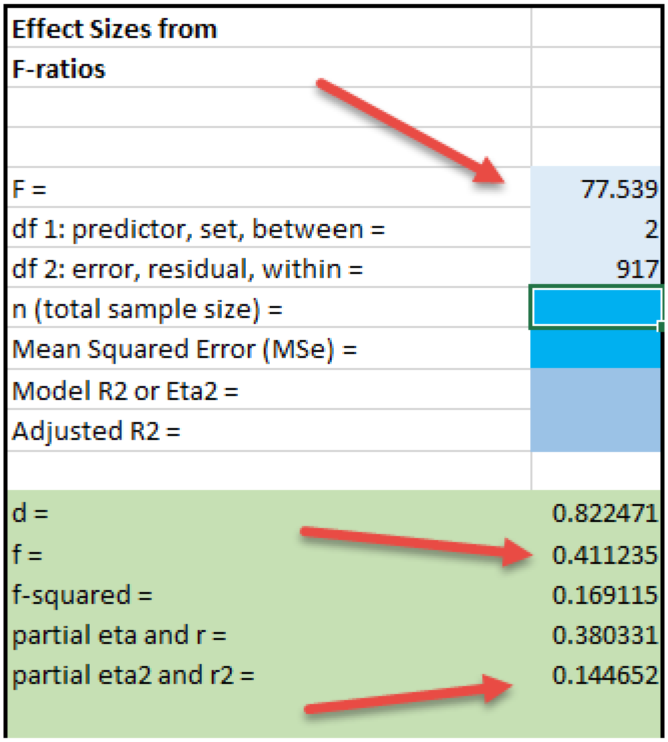
April 2023: Using Excel Effect Size Calculator

To obtain the effect size for an ANOVA, use the F ratio (F = 77.539 above), and the two degrees of freedome (df =2 and df = 917).

According to the calculations above,

f = .41 and
$R^2 = r^2 = .1447$,

and these are the values reported by the effect size calculator as shown below.



Effect Sizes from	
F-ratios	
F =	77.539
df 1: predictor, set, between =	2
df 2: error, residual, within =	917
n (total sample size) =	
Mean Squared Error (MSe) =	
Model R2 or Eta2 =	
Adjusted R2 =	
d =	0.822471
f =	0.411235
f-squared =	0.169115
partial eta and r =	0.380331
partial eta2 and r2 =	0.144652

Question 16

What total sample size is needed for

- f = .41,
- df = 2,
- alpha = .05, and
- power = .80?

Also, how many participants per group are needed for the study of the three groups?

Answer

Total n = 61 which divided among three groups is 61 / 3 = 20 participants per group with one group having 21 participants.

Answer (April 2023 Update)
Using Cohen's Table 2

1. Convert effect size f to closest listed f provided by Cohen. The calculated f = .41 is larger than the largest large effect size f identified by Cohen which is .40, so we can use that.

2. Find the tabled sample size for:

- alpha = .05
- power = .80 (only value provided by Cohen)
- df = 2 (number of groups minus 1, 3 – 1 = 2)
- f = .40 (7. ANOVA, row 3g = 3 groups)

3. Tabled value is 21, and since there are three groups, we must multiply this value by 3, so total sample size is 21*3 = 63.

Example: Math Test Scores and Background Music
We wish to study student performance on mathematics tests under four music conditions:

- unpleasant, disturbing
  - John Coltrane "The Father and the Son and the Holy Ghost"
  - https://www.youtube.com/watch?v=d3XuZ1zu2PI
- calming, soothing
  - Tomaso Albinoni/Remo Giazotto "Adagio in G minor"
  - https://www.youtube.com/watch?v=XMbvcp480Y4
- contemporary (let students pick)
- no music

Prior research in this area is scant or does not provide enough information to allow for calculation of effect sizes so we rely on Cohen's suggestions of effect sizes:

- large f = .40,
- medium f = .25, and
- small f = .10.

We opt for a medium effect since we are unsure that a large effect could be observed by this treatment and since a small effect may be practically meaningless and could also result in a larger sample size than we could afford.

Question 17
What sample size is needed for following? Also, how many participants per group?

- f = .25,
- df = 4 -1 = 3,
- alpha = .05, and
- power = .90?

Answer

Total n = 231 and per group sample size is 231 / 4 = 57.75 ~ 58 participants per group.

Answer (April 2023 Update)

Using Cohen's Table 2

1. We have to use power = .80 since table does not provide .90 power. All else remains same as above.

2. Find the tabled sample size for:

- alpha = .05
- power = .80 (only value provided by Cohen)
- df = 3 (number of groups minus 1, 4 – 1 = 3)
- f = .25 (7. ANOVA, row 4g = 4 groups)

3. Tabled value is 45, and since there are four groups, we must multiply this value by 4, so total sample size is 45*4 = 180.

Example: 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. The observed effect size f for this study is f = 0.212.

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

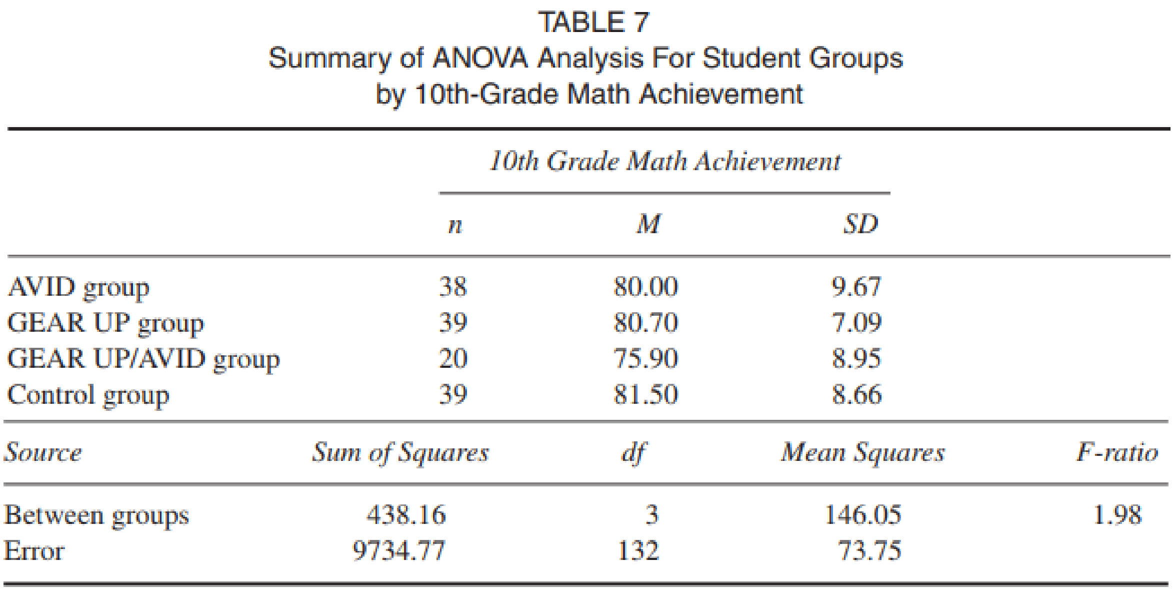
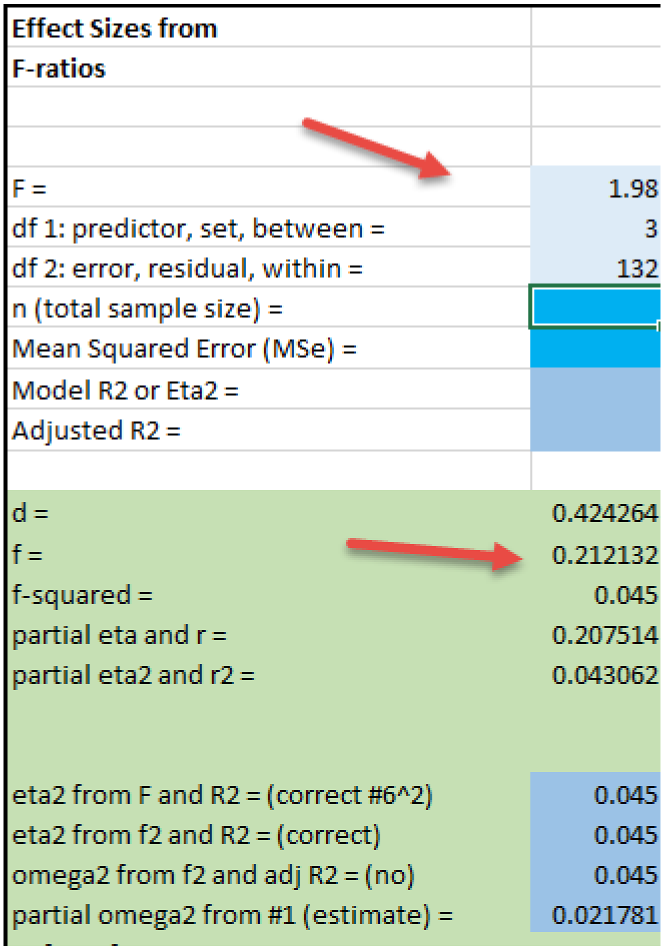


TABLE 7
Summary of ANOVA Analysis For Student Groups
by 10th-Grade Math Achievement

	10th Grade Math Achievement		
	n	*M*	*SD*
AVID group	38	80.00	9.67
GEAR UP group	39	80.70	7.09
GEAR UP/AVID group	20	75.90	8.95
Control group	39	81.50	8.66

Source	*Sum of Squares*	*df*	*Mean Squares*	*F-ratio*
Between groups	438.16	3	146.05	1.98
Error	9734.77	132	73.75	

April 2023: Use the Effect Size Excel Sheet to Calculate Effect Size f



Effect Sizes from	
F-ratios	
F =	1.98
df 1: predictor, set, between =	3
df 2: error, residual, within =	132
n (total sample size) =	
Mean Squared Error (MSe) =	
Model R2 or Eta2 =	
Adjusted R2 =	
d =	0.424264
f =	0.212132
f-squared =	0.045
partial eta and r =	0.207514
partial eta2 and r2 =	0.043062
eta2 from F and R2 = (correct #6^2)	0.045
eta2 from f2 and R2 = (correct)	0.045
omega2 from f2 and adj R2 = (no)	0.045
partial omega2 from #1 (estimate) =	0.021781

Example

Hall, J. M., & Ponton, M. K. (2005). Mathematics self-efficacy of college freshman. *Journal of developmental education*, *28*(3), 26.

Below is a table reported by Hall and Ponton (2005) comparing mathematics self-efficacy between students in calculus 1 and intermediate algebra.

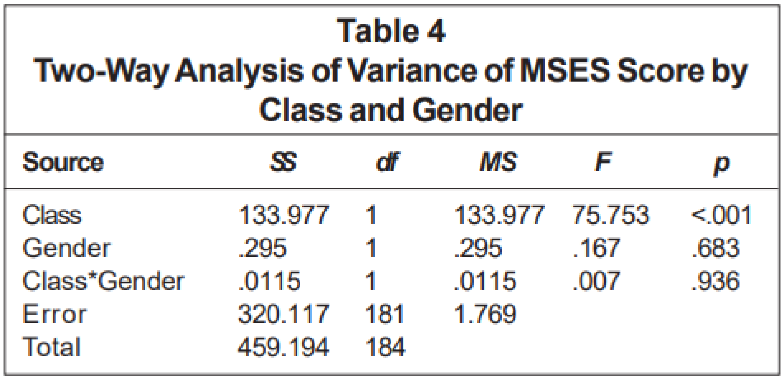


Table 4
Two-Way Analysis of Variance of MSES Score by Class and Gender

Source	*SS*	*df*	*MS*	*F*	*p*
Class	133.977	1	133.977	75.753	<.001
Gender	.295	1	.295	.167	.683
Class*Gender	.0115	1	.0115	.007	.936
Error	320.117	181	1.769		
Total	459.194	184			

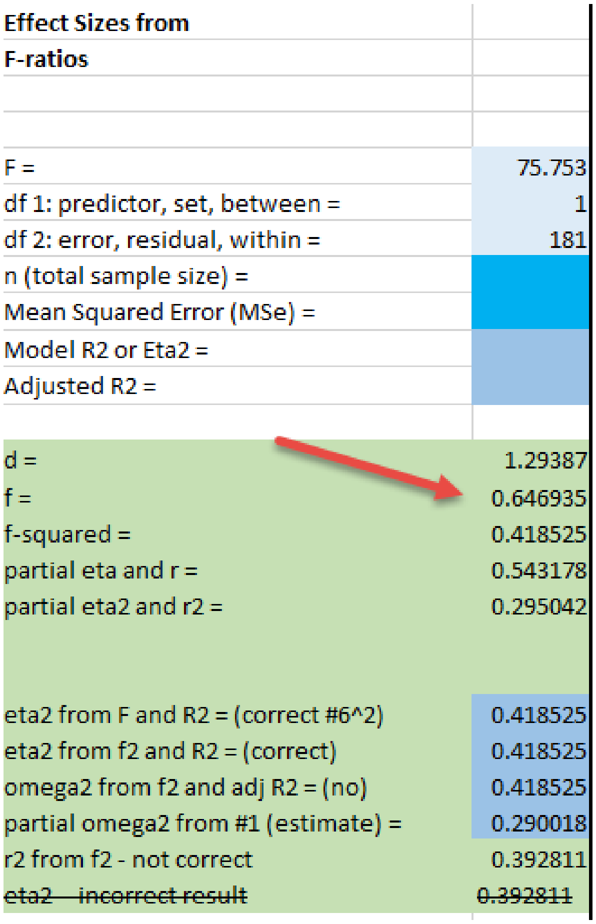
Using the F-ratio, what effect size f is obtained for self-efficacy differences between these two groups?

f = 0.646

Also note that d = 1.293.

Previously when we calculated d from the groups M, SD, and n, we obtained a d = 1.32.

Why the difference in d here?



Effect Sizes from	
F-ratios	
F =	75.753
df 1: predictor, set, between =	1
df 2: error, residual, within =	181
n (total sample size) =	
Mean Squared Error (MSe) =	
Model R2 or Eta2 =	
Adjusted R2 =	
d =	1.29387
f =	0.646935
f-squared =	0.418525
partial eta and r =	0.543178
partial eta2 and r2 =	0.295042
eta2 from F and R2 = (correct #6^2)	0.418525
eta2 from f2 and R2 = (correct)	0.418525
omega2 from f2 and adj R2 = (no)	0.418525
partial omega2 from #1 (estimate) =	0.290018
r2 from f2 - not correct	0.392811
~~eta2 - incorrect result~~	~~0.392811~~

Question 18
Suppose we wish to replicate this study but want a sample size that would give us an 80% chance of detecting a difference, if one exists, at the .05 level of significance. What sized sample is needed, and how many per group?

Answer

A total sample of 247 is needed, which means 247/4 = 61.75 or ~62 per group.

Answer (April 2023 Update)
Using Cohen's Table 2

1. We have to use power = .80 since table does not provide .90 power. All else remains same as above.

2. Find the tabled sample size for:

- alpha = .05
- power = .80 (only value provided by Cohen)
- df = 3 (number of groups minus 1, 4 – 1 = 3)
- f = .25 (7. ANOVA, row 4g = 4 groups)

3. Tabled value is 45, and since there are four groups, we must multiply this value by 4, so total sample size is 45*4 = 180.

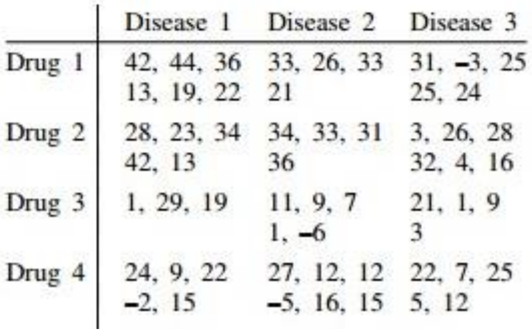
6.2 Multi-way Analysis of Variance, Effect Size f, η^2 and Multiple R^2

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: 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"



	Disease 1	Disease 2	Disease 3
Drug 1	42, 44, 36 13, 19, 22	33, 26, 33 21	31, −3, 25 25, 24
Drug 2	28, 23, 34 42, 13	34, 33, 31 36	3, 26, 28 32, 4, 16
Drug 3	1, 29, 19	11, 9, 7 1, −6	21, 1, 9 3
Drug 4	24, 9, 22 −2, 15	27, 12, 12 −5, 16, 15	22, 7, 25 5, 12

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

```
. anova systolic drug disease drug#disease

                           Number of obs =         58     R-squared     =  0.4560
                           Root MSE      =    10.5096     Adj R-squared =  0.3259

                  Source | Partial SS         df         MS           F     Prob>F
              -----------+----------------------------------------------------
                   Model |  4259.3385         11    387.21259       3.51   0.0013
                         |
                    drug |  2997.4719          3    999.15729       9.05   0.0001
                 disease |  415.87305          2    207.93652       1.88   0.1637
            drug#disease |  707.26626          6    117.87771       1.07   0.3958
                         |
                Residual |  5080.8167         46    110.45254
              -----------+----------------------------------------------------
                   Total |  9340.1552         57    163.86237
```

Drug f

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

Disease f

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

Drug by Disease Interaction

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

Question 19

If we desired to replicate this study, and have enough power to detect the possible interaction between drug and disease, what sized sample would be need?

- f = .373,
- df = (a-1)(b-1) = (3)(2) = 6,
- alpha = .05, and
- power = .80?

Answer

A sample of about 105 would be needed approximately equally divided among all the variable level combinations.

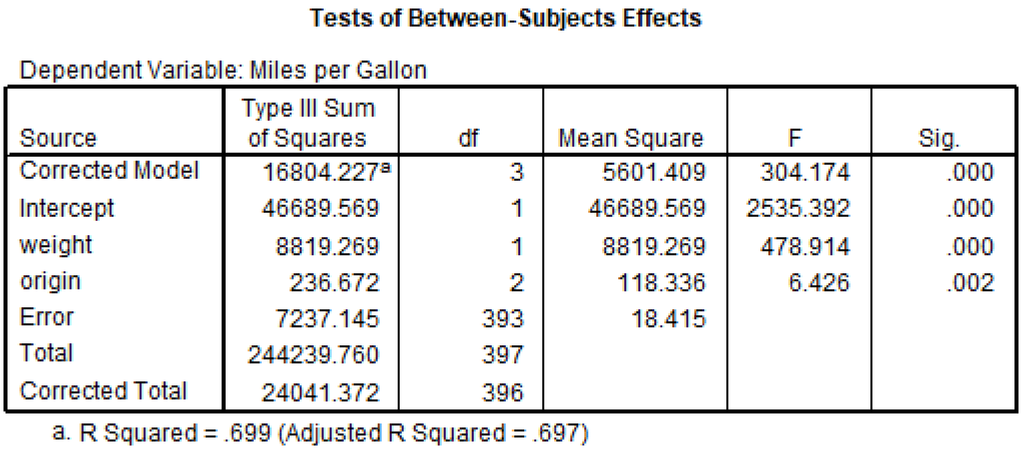
6.4 Analysis of Covariance from Existing Studies

The calculation 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: Car MPG by Origin Controlling for Car Weight

Using the SPSS data file cars.sav (example data file that comes with SPSS), 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.

ANCOVA results and effect size f values are reported below.



Tests of Between-Subjects Effects

Dependent Variable: Miles per Gallon

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	16804.227[a]	3	5601.409	304.174	.000
Intercept	46689.569	1	46689.569	2535.392	.000
weight	8819.269	1	8819.269	478.914	.000
origin	236.672	2	118.336	6.426	.002
Error	7237.145	393	18.415		
Total	244239.760	397			
Corrected Total	24041.372	396			

[a] R Squared = .699 (Adjusted R Squared = .697)

Origin f

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

Weight f

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

Question 20

The cars.sav were collected during the 1970s and 1980s. Suppose we wish to replicate this analysis, but with current cars. We don't have the resources to obtain a sample of about 400 cars like included in the original data file, so, given the effect sizes observed above for Origin and for Weight, what is the minimum sample size we could use to have a power of .80 and Type 1 error rate of .05 to test either Origins or Weight? We must have a sample size to ensure both Origin and Weight will have a minimum power level of .80.

Answer

Origin

- f = .18
- df = 3 -1 = 2
- alpha = .05
- power = .80
- n = 301

Weight

- f = 1.10
- df = 1
- alpha = .05
- power = .80
- n = 9

The relation between Weight and MPG is so strong that a sample of only 9 cars is needed to detect that relationship. However, Origin requires a sample of 301 once Weight is considered, so a sample of only 9 would not provide adequate power to detect the Origin effect, the minimum sample needed is 301.

6.5 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 = t^2. 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.

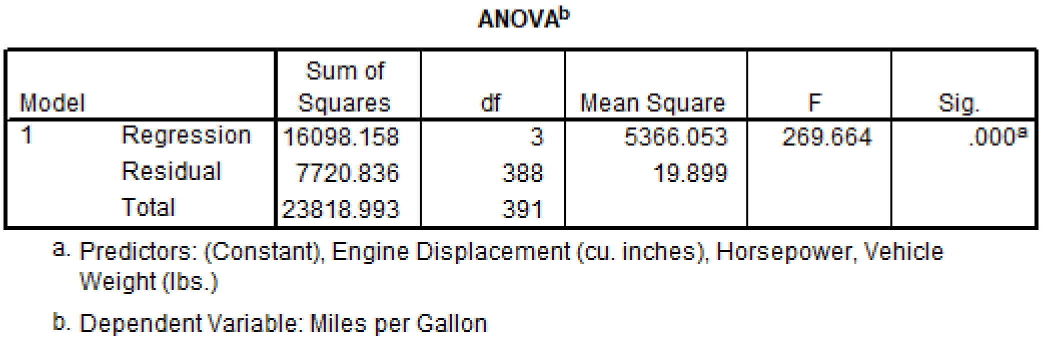
Given this equivalence, formulas posted above continue to work for regression, and the t-test can be converted to f by squaring t and using formula F6.

$$\text{f} = \sqrt{\frac{F(df_{predictors})}{df_{error}}}$$

$$\text{f} = \sqrt{\frac{t^2}{df_{error}}}$$

Example: 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 and ES f 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.

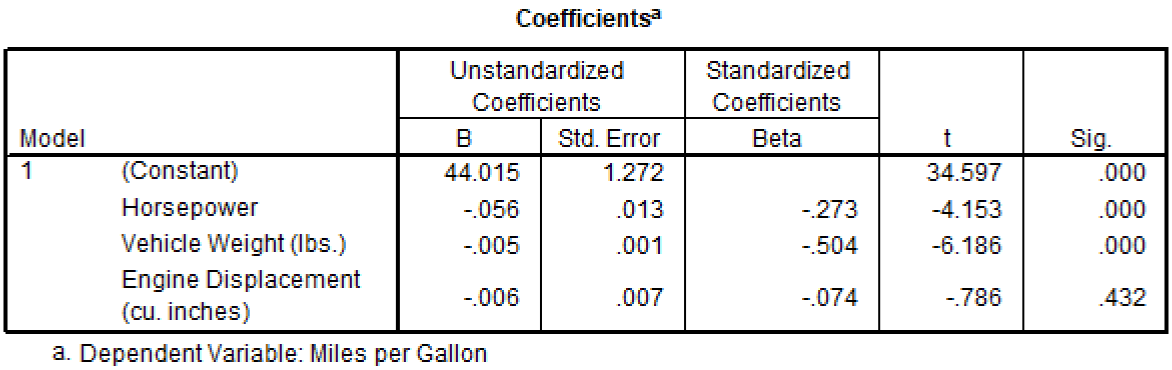


ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16098.158	3	5366.053	269.664	.000[a]
	Residual	7720.836	388	19.899		
	Total	23818.993	391			

a. Predictors: (Constant), Engine Displacement (cu. inches), Horsepower, Vehicle Weight (lbs.)

b. Dependent Variable: Miles per Gallon



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	44.015	1.272		34.597	.000
	Horsepower	-.056	.013	-.273	-4.153	.000
	Vehicle Weight (lbs.)	-.005	.001	-.504	-6.186	.000
	Engine Displacement (cu. inches)	-.006	.007	-.074	-.786	.432

a. Dependent Variable: Miles per Gallon

Horsepower

$$\text{f} = \sqrt{\frac{t^2}{df_{error}}} = \sqrt{\frac{-4.153^2}{388}} = .21$$

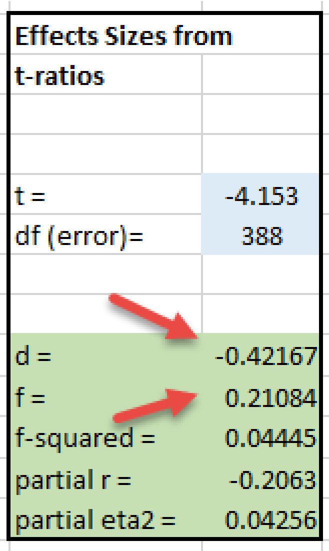
Weight

$$f = \sqrt{\frac{t^2}{df_{error}}} = \sqrt{\frac{-6.186^2}{388}} = .31$$

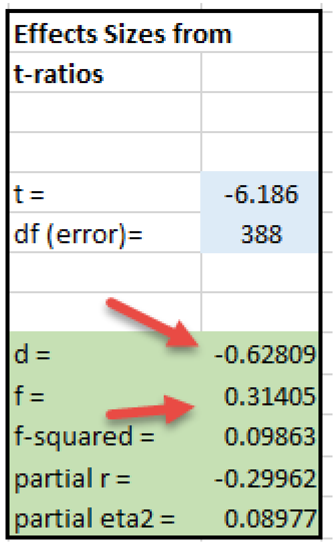
Displacement

$$f = \sqrt{\frac{t^2}{df_{error}}} = \sqrt{\frac{-0.786^2}{388}} = .039$$

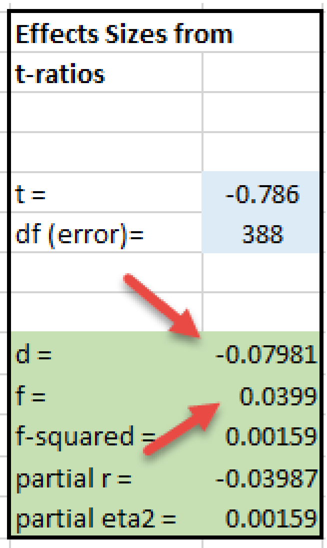
Or, just use the t-ratios with the Excel effect size sheet and use d instead of f. Below are the effect sizes for each of the three.



Effects Sizes from t-ratios	
t =	-4.153
df (error)=	388
d =	-0.42167
f =	0.21084
f-squared =	0.04445
partial r =	-0.2063
partial eta2 =	0.04256



Effects Sizes from t-ratios	
t =	-6.186
df (error)=	388
d =	-0.62809
f =	0.31405
f-squared =	0.09863
partial r =	-0.29962
partial eta2 =	0.08977



Effects Sizes from t-ratios	
t =	-0.786
df (error)=	388
d =	-0.07981
f =	0.0399
f-squared =	0.00159
partial r =	-0.03987
partial eta2 =	0.00159

Question 21

If we wished to replicate this study with current production cars, what sized sample would be needed to detect the smallest effect size observed in the above regression with power = .90 and alpha = .01?

Answer

The smallest effect size is associated with Displacement with f = 0.039. Keep in mind this is a partialed effect size that represents statistical control that takes into account both Weight and Horsepower. The replication study must also include these predictors otherwise the effect size observed for Displacement may be invalid without the presence of these two predictors.

- f = .039
- df = 1
- alpha = .01
- power = .90
- n = 9,786

This result show that a very large sample (n = 9,786) is required to detect the small contribution that Displacement added to the model over the effects of Horsepower and Weight. Given the small effect of Displacement once Horsepower and Weight are controlled, I probably would not replicate this study and instead would focus on just Horsepower and Weight if I could see similar results from a second data file or study.

7. Effect Size f^2

Discussion to be added and expanded. In short, f^2 is just effect size f squared, so little new information is obtained from f^2. All examples shown above with f also work with f^2 and provide the same answer. Cohen introduced f^2 in discussion of power and sample size with regression since f was defined as a measure of group mean variability which is more common in ANOVA models than regression models.

8. Effect Size f and ANCOVA for Prospective, Experimental Studies
(Not covered in EDUR 9131 spring 2023)

(Note that this approach assumes variable targeted for adjustment is uncorrelated with added covariates or predictors, so this approach works well for true experimental studies, but not for studies in which groups are not randomly formed since covariates are likely to be correlated with targeted variable and resultant adjustment is difficult to predict – better to find existing study with same predictors and obtain ES from existing study.)

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 effect, 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:

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

where effect size refers to d or f; neither squared effect sizes such as f^2 or r^2, 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:

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

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

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 R^2 of about .25.

The resulting adjusted d value can be found as follows:

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

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

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

the required sample size for this ANCOVA is n = 372 or about 186 females and 186 males. This sample size estimate agrees (within rounding error) with that obtained from Stata (version 12), a statistical software program that can be used to estimate n for a two-group ANCOVA study.

Stata Sample Size Estimate

```
. sampsi 0 -.27, pre(1) post(1) p(0.85) r01(.5)  sd1(1) sd2(1) m(ancova)

Estimated sample size for two samples with repeated measures
Assumptions:
```

```
                                        alpha  =      0.0500  (two-sided)
                                        power =       0.8500
                                        m1     =      0
                                        m2     =      -.27
                                        sd1    =      1
                                        sd2    =      1
                                        n2/n1 =       1.00
number of follow-up measurements                      =        1
number of baseline measurements               =       1
correlation between baseline & follow-up =       0.500

Method: ANCOVA
        relative efficiency     =       1.333
        adjustment to sd        =       0.866
        adjusted sd1            =       0.866
        adjusted sd2            =       0.866

Estimated required sample sizes:
        n1 = 185
        n2 = 185
```

Question 22

(Problematic example, for adjustment to work as modeled, covariates must be uncorrelated with instructor reputation, otherwise very difficult to predict how the adjustment will affect the original effect size)

Recall the instructor reputation and overall rating of instruction example. Suppose we include three covariates in this study:

- student intrinsic motivation in the course subject,
- student GPA, and
- student rating of course difficulty.

In the original analysis R^2 = .145 for the sole predictor of instructor reputation, which computes to an effect size f = .412.

We believe these three predictors will uniquely account for about 6% of the variance in student ratings, so their R^2 contribution is estimated to be .06.

The adjusted effect size f after adding this predictor is:

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

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

What would be the sample size needed now that these covariates are taken into account?

- adjusted f = .425,
- df = 2,
- alpha = .05, and
- power = .80?

Also, how many participants per group?

Answer

Total n = 57

Per group sample size is 57 / 3 = 19 per group

Question 23

The music experiment in Question 17 required a total sample of 231. Suppose a pretest of mathematics is administered to students and used as a covariate to help equate groups and reduce model error.

Often pretest measures of achievement correlate well with posttest achievement scores, so anticipate a correlation of .40 between pretest and posttest. Converting this correlation to $r^2 = .4x.4 = .16$.

In the original study we expected f = .25; the adjusted effect size f would be

adjusted f$= \frac{f}{\sqrt{(1-r^2)}} = \frac{.25}{\sqrt{(1-.16)}} = \frac{.25}{\sqrt{.84}}$

$$= \frac{.25}{.9165} = 0.273$$

What sample size is needed for

- adjusted f = .273,
- df = 3,
- alpha = .05, and
- power = .90?

Also, how many participants per group? Lastly, how does this sample size compare with the original of 231?

Answer

Total n = 195

Sample size per group is 195 / 4 = 48.75 or 49 per group.

Original study n = 231, but addition of covariate reduced this number to 195, a potential savings in resources and effort.

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)

References

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

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