

19. Effect Size

- A. Proportions
- B. Difference between Means
- C. Variance Explained
- D. Exercises

Researchers often seek to learn more than whether the variable under investigation has an effect and/or the direction of the effect. This is particularly true for research that has practical applications. For example, an investigation of the efficacy of a pain-relief drug would seek to determine the extent of the relief and not merely whether there was any relief. Similarly, a study of a test-preparation course's efficacy would seek to determine how much the course raises students' test scores. Finally, a study of the relationship between exercise and blood pressure would seek to determine how much blood pressure decreases for a given amount of exercise. In all of these examples, a significance test would not be sufficient since it would only provide the researcher with information about the existence and direction of the effect. It would not provide any information about the size of the effect.

Before we proceed with a discussion of how to measure effect size, it is important to consider that for some research it is the presence or absence of an effect rather than its size that is important. A controversial example is provided by Bem (2011) who investigated precognition. Bem found statistically significant evidence that subjects' responses are affected by future events. That is, he rejected the null hypothesis that there is no effect. The important question is not the size of the effect but, rather, whether it exists at all. It would be truly remarkable if future events affect present responses even a little. It is important to note that subsequent research (Ritchie, Wiseman, & French, 2012) has failed to replicate Bem's results and the likelihood that the precognition effects he described are real is very low.

Proportions

by David M. Lane

Prerequisites

- none

Learning Objectives

1. Compute absolute risk reduction
2. Compute relative risk reduction
3. Compute number needed to treat

Often the interpretation of a proportion is self-evident. For example, the obesity rate for white non-Hispanic adults living in the United States was estimated by a study conducted between 2006 and 2008 to be 24%. This value of 24% is easily interpretable and indicates the magnitude of the obesity problem in this population.

Often the question of interest involves the comparison of two outcomes. For example, consider the analysis of proportions in the case study “Mediterranean Diet and Health.” In this study, one group of people followed the diet recommended by the American Heart Association (AHA), whereas a second group followed the “Mediterranean Diet.” One interesting comparison is between the proportions of people who were healthy throughout the study as a function of diet. It turned out that 0.79 of the people who followed the AHA diet and 0.90 of those who followed the Mediterranean diet were healthy. How is the effect size of diet best measured?

We will take the perspective that we are assessing the benefits of switching from the AHA diet to the Mediterranean diet. One way to assess the benefits is to compute the difference between the proportion who were not healthy on the AHA diet (0.21) with the proportion who were not healthy on the Mediterranean diet (0.10). Therefore, the difference in proportions is:

$$0.21 - 0.10 = 0.11.$$

This measure of the benefit is called the *Absolute Risk Reduction* (ARR).

To define ARR more formally, let C be the proportion of people in the control group with the ailment of interest and T be the proportion in the treatment group. ARR can then be defined as:

$$ARR = C - T$$

Alternatively, one could measure the difference in terms of percentages. For our example, the proportion of non-healthy people on the Mediterranean diet (0.10) is 52% lower than the proportion of non-healthy people on the AHA diet (0.21). This value is computed as follows:

$$(0.21 - 0.10) / 0.21 \times 100 = 52\%$$

This measure of the benefit is called the Relative Risk Reduction (RRR). The general formula for RRR is:

$$RRR = (C - T) / C \times 100$$

where C and T are defined as before.

A third commonly used measure is the “odds ratio.” For our example, the odds of being healthy on the Mediterranean diet are 90:10 = 9:1; the odds on the AHA diet are 79:21 = 3.76:1. The ratio of these two odds is $9/3.76 = 2.39$. Therefore, the odds of being healthy on the Mediterranean diet is 2.39 times the odds of being healthy on the AHA diet. Note that the odds ratio is the ratio of the odds and not the ratio of the probabilities.

A fourth measure is the number of people who need to be treated in order to prevent one person from having the ailment of interest. In our example, being treated means changing from the AHA diet to the Mediterranean diet. The number who need to be treated can be defined as

$$N = 1/ARR$$

For our example,

$$N = 1/0.11 = 9$$

Therefore, one person who would otherwise not be healthy would be expected to stay healthy for every nine people changing from the AHA diet to the Mediterranean diet.

The obvious question is which of these measures is the best one. Although each measure has its proper uses, the RRR measure can exaggerate the importance of an effect, especially when the absolute risks are low. For example, if a drug

reduced the risk of a certain disease from 1 in 1,000,000 to 1 in 2,000,000, the RRR is 50%. However, since the ARR is only 0.0000005, the practical reduction in risk is minimal.

Difference Between Two Means

by David M. Lane

Prerequisites

- Chapter 3: Measures of Variability
- Chapter 12: Differences between Two Means (Independent Groups)
- Chapter 16: Chapter Log Transformations

Learning Objectives

1. State how the inherent meaningfulness of the scales affects the type of measure that should be used
2. Compute g
3. Compute d
4. State the effect of the variability of subjects on the size of standardized measures

When the units of a measurement scale are meaningful in their own right, then the difference between means is a good and easily interpretable measure of effect size. For example, a study conducted by Holbrook, Crowther, Lotter, Cheng and King in 2000 investigated the effectiveness of benzodiazepine for the treatment of insomnia. These researchers found that, compared to a placebo, this drug increased total sleep duration by a mean of 61.8 minutes. This difference in means shows clearly the degree to which benzodiazepine is effective. (It is important to note that the drug was found to sometimes have adverse side effects.)

When the dependent variable is measured on a ratio scale, it is often informative to consider the proportional difference between means in addition to the absolute difference. For example, if in the Holbrook et al. study the mean total sleep time for the placebo group were 120 minutes, then the 61.8-minute increase would represent a 51% increase in sleep time. On the other hand, if the mean sleep time for the placebo were 420 minutes, then the 61.8-minute increase would represent a 15% increase in sleep time.

It is interesting to note that if a log transformation is applied to the dependent variable, then equal percent changes on the original scale will result in equal absolute changes on the log scale. For example, suppose the mean sleep time

increased 10% from 400 minutes to 440 in one condition and 10% from 300 to 330 minutes in a second condition. If we take the log base 10 of these values, we find that $\text{Log}(440) - \text{Log}(400) = 2.643 - 2.602 = 0.041$ and, similarly, $\text{Log}(330) - \text{Log}(300) = 2.518 - 2.477 = 0.041$.

Many times the dependent variable is measured on a scale that is not inherently meaningful. For example, in the “Animal Research” case study, attitudes toward animal research were measured on a 7-point scale. The mean rating of women on whether animal research is wrong was 1.47 scale units higher than the mean rating of men. However, it is not clear whether this 1.47-unit difference should be considered a large effect or a small effect, since it is not clear exactly what this difference means.

When the scale of a dependent variable is not inherently meaningful, it is common to consider the difference between means in standardized units. That is, effect size is measured in terms of the number of standard deviations the means differ by. Two commonly used measures are Hedges' g and Cohen's d . Both of these measures consist of the difference between means divided by the standard deviation. They differ only in that Hedges' g uses the version of the standard deviation formula in which you divide by $N-1$, whereas Cohen's d uses the version in which you divide by N . The two formulas are given below.

$$g = \frac{M_1 - M_2}{\sqrt{MSE}}$$

$$d = g \sqrt{\frac{N}{N - 2}}$$

where M_1 is the mean of the first group, M_2 is the mean of the second group, MSE is the mean square error, and N is the total number of observations.

Standardized measures such as Cohen's d and Hedges' g have the advantage that they are scale free. That is, since the dependent variable is standardized, the original units are replaced by standardized units and are interpretable even if the original scale units do not have clear meaning. Consider the Animal Research case study in which attitudes were measured on a 7-point scale. On a rating of whether animal research is wrong, the mean for women was 5.353, the mean for men was

3.882, and MSE was 2.864. Hedges' g can be calculated to be 0.87. It is more meaningful to say that the means were 0.87 standard deviations apart than 1.47 scale units apart since the scale units are not well defined.

It is natural to ask what constitutes a large effect. Although there is no objective answer to this question, the guidelines suggested by Cohen (1988) stating that an effect size of 0.2 is a small effect, an effect size of 0.5 is a medium effect, and an effect size of 0.8 is a large effect have been widely adopted. Based on these guidelines, the effect size of 0.87 is a large effect.

It should be noted, however, that these guidelines are somewhat arbitrary and have not been universally accepted. For example, Lenth (2001) argued that other important factors are ignored if Cohen's definition of effect size is used to choose a sample size to achieve a given level of power.

Interpretational Issues

It is important to realize that the importance of an effect depends on the context. For example, a small effect can make a big difference if only extreme observations are of interest. Consider a situation in which a test is used to select students for a highly selective program. Assume that there are two types of students (red and blue) and that the mean for the red students is 52, the mean for the blue students is 50, both distributions are normal, and the standard deviation for each distribution is 10. The difference in means is therefore only 0.2 standard deviations and would generally be considered to be a small difference. Now assume that only students who scored 70 or higher would be selected for the program. Would there be a big difference between the proportion of blue and red students who would be able to be accepted into the program? It turns out that the proportion of red students who would qualify is 0.036 and the proportion of blue students is 0.023. Although this difference is small in absolute terms, the ratio of red to blue students who qualify is 1.6:1. This means that if 100 students were to be accepted and if equal numbers of randomly-selected red and blue students applied, 62% would be red and 38% would be blue. In most contexts this would be considered an important difference.

When the effect size is measured in standard deviation units as it is for Hedges' g and Cohen's d , it is important to recognize that the variability in the subjects has a large influence on the effect size measure. Therefore, if two experiments both compared the same treatment to a control but the subjects were much more homogeneous in Experiment 1 than in Experiment 2, then a

standardized effect size measure would be much larger in the former experiment than in the latter. Consider two hypothetical experiments on the effect of an exercise program on blood pressure. Assume that the mean effect on systolic blood pressure of the program is 10mmHg and that, due to differences in the subject populations sampled in the two experiments, the standard deviation was 20 in Experiment 1 and 30 in Experiment 2. Under these conditions, the standardized measure of effect size would be 0.50 in Experiment 1 and 0.33 in Experiment 2. This standardized difference in effect size occurs even though the effectiveness of the treatment is exactly the same in the two experiments.

Proportion of Variance Explained

by David M. Lane

Prerequisites

- Chapter 15: One-Factor ANOVA (Between Subjects)
- Chapter 14: Partitioning Sums of Squares
- Chapter 14: Multiple Regression

Learning Objectives

1. State the difference in bias between η^2 and ω^2
2. Compute η^2
3. Compute ω^2
4. Distinguish between ω^2 and partial ω^2
5. State the bias in R^2 and what can be done to reduce it

Effect sizes are often measured in terms of the proportion of variance explained by a variable. In this section, we discuss this way to measure effect size in both ANOVA designs and in correlational studies.

ANOVA Designs

Responses of subjects will vary in just about every experiment. Consider, for example, the “*Smiles and Leniency*” case study. A histogram of the dependent variable “leniency” is shown in Figure 1. It is clear that the leniency scores vary considerably. There are many reasons why the scores differ. One, of course, is that subjects were assigned to four different smile conditions and the condition they were in may have affected their leniency score. In addition, it is likely that some subjects are generally more lenient than others, thus contributing to the differences among scores. There are many other possible sources of differences in leniency ratings including, perhaps, that some subjects were in better moods than other subjects and/or that some subjects reacted more negatively than others to the looks or mannerisms of the stimulus person. You can imagine that there are innumerable other reasons why the scores of the subjects could differ.

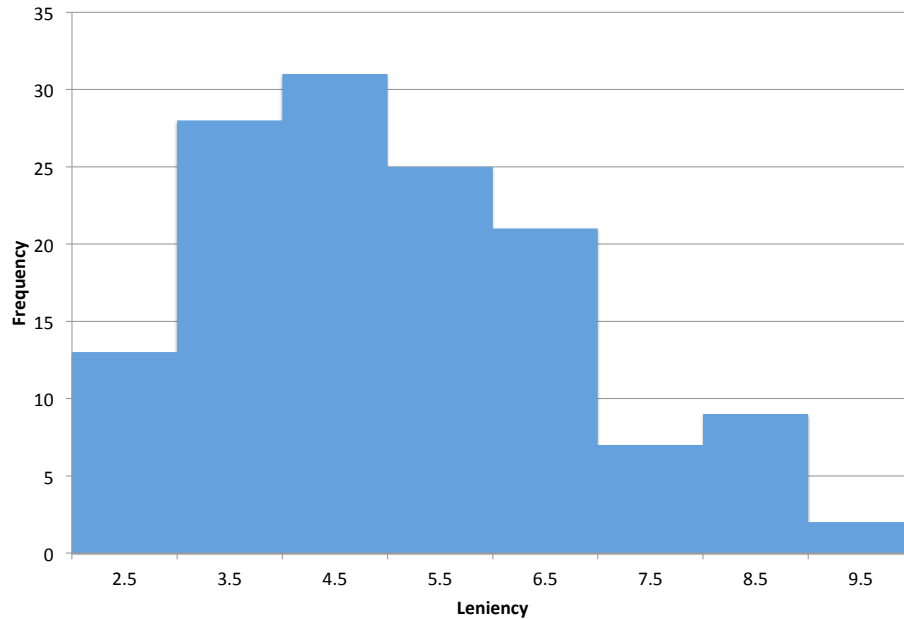


Figure 1. Distribution of leniency scores.

One way to measure the effect of conditions is to determine the proportion of the variance among subjects' scores that is attributable to conditions. In this example, the variance of scores is 2.794. The question is how this variance compares with what the variance would have been if every subject had been in the same treatment condition. We estimate this by computing the variance within each of the treatment conditions and taking the mean of these variances. For this example, the mean of the variances is 2.649. Since the mean variance within the smile conditions is not that much less than the variance ignoring conditions, it is clear that “Smile Condition” is not responsible for a high percentage of the variance of the scores. The most convenient way to compute the proportion explained is in terms of the sum of squares “conditions” and the sum of squares total. The computations for these sums of squares are shown in the chapter on ANOVA. For the present data, the sum of squares for “Smile Condition” is 27.544 and the sum of squares total is 377.189. Therefore, the proportion explained by “Smile Condition” is:

$$27.544/377.189 = 0.073.$$

Thus, 0.073 or 7.3% of the variance is explained by “Smile Condition.”

An alternative way to look at the variance explained is as the proportion reduction in error. The sum of squares total (377.189) represents the variation when “Smile Condition” is ignored and the sum of squares error ($377.189 - 27.544 =$

349.654) is the variation left over when “Smile Condition” is accounted for. The difference between 377.189 and 349.654 is 27.535. The reduction in error of 27.535 represents a proportional reduction of $27.535/377.189 = 0.073$, the same value as computed in terms of proportion of variance explained.

This measure of effect size, whether computed in terms of variance explained or in terms of percent reduction in error, is called η^2 where η is the Greek letter eta. Unfortunately, η^2 tends to overestimate the variance explained and is therefore a biased estimate of the proportion of variance explained. As such, it is not recommended (despite the fact that it is reported by a leading statistics package).

An alternative measure, ω^2 (omega squared), is unbiased and can be computed from

$$\omega^2 = \frac{SSQ_{condition} - (k - 1)MSE}{SSQ_{total} + MSE}$$

where MSE is the mean square error and k is the number of conditions. For this example, $k = 4$ and $\omega^2 = 0.052$.

It is important to be aware that both the variability of the population sampled and the specific levels of the independent variable are important determinants of the proportion of variance explained. Consider two possible designs of an experiment investigating the effect of alcohol consumption on driving ability. As can be seen in Table 1, Design 1 has a smaller range of doses and a more diverse population than Design 2. What are the implications for the proportion of variance explained by Dose? Variation due to Dose would be greater in Design 2 than Design 1 since alcohol is manipulated more strongly than in Design 1. However, the variance in the population should be greater in Design 1 since it includes a more diverse set of drivers. Since with Design 1 the variance due to Dose would be smaller and the total variance would be larger, the proportion of variance explained by Dose would be much less using Design 1 than using Design 2. Thus, the proportion of variance explained is not a general characteristic of the independent variable. Instead, it is dependent on the specific levels of the independent variable used in the experiment and the variability of the population sampled.

Table 1. Design Parameters

Design	Doses	Population
1	0.00 0.30 0.60	All Drivers between 16 and 80 Years
2	0.00 0.50 1.00	Experienced Drivers between 25 and 30 Years

Factorial Designs

In one-factor designs, the sum of squares total is the sum of squares condition plus the sum of squares error. The proportion of variance explained is defined relative to sum of squares total. In an A x B design, there are three sources of variation (A, B, A x B) in addition to error. The proportion of variance explained for a variable (A, for example) could be defined relative to the sum of squares total ($SSQ_A + SSQ_B + SSQ_{A \times B} + SSQ_{error}$) or relative to $SSQ_A + SSQ_{error}$.

To illustrate with an example, consider a hypothetical experiment on the effects of age (6 and 12 years) and of methods for teaching reading (experimental and control conditions). The means are shown in Table 2. The standard deviation of each of the four cells (Age x Treatment combinations) is 5. (Naturally, for real data, the standard deviations would not be exactly equal and the means would not be whole numbers.) Finally, there were 10 subjects per cell resulting in a total of 40 subjects.

Table 2. Condition Means

	Treatment	
Age	Experimental	Control
6	40	42
12	50	56

The sources of variation, degrees of freedom, and sums of squares from the analysis of variance summary table as well as four measures of effect size are shown in Table 3. Note that the sum of squares for age is very large relative to the

other two effects. This is what would be expected since the difference in reading ability between 6- and 12-year-olds is very large relative to the effect of condition.

Table 3. ANOVA Summary Table

Source	df	SSQ	η^2	partial η^2	ω^2	partial ω^2
Age	1	1440	0.567	0.615	0.552	0.586
Condition	1	160	0.063	0.151	0.053	0.119
A x C	1	40	0.016	0.043	0.006	0.015
Error	36	900				
Total	39	2540				

First, we consider the two methods of computing η^2 , labeled η^2 and partial η^2 . The value of η^2 for an effect is simply the sum of squares for this effect divided by the sum of squares total. For example, the η^2 for Age is $1440/2540 = 0.567$. As in a one-factor design, η^2 is the proportion of the total variation explained by a variable. Partial η^2 for Age is SSQ_{Age} divided by $(SSQ_{\text{Age}} + SSQ_{\text{error}})$ which is $1440/2340 = 0.615$.

As you can see, the partial η^2 is larger than η^2 . This is because the denominator is smaller for the partial η^2 . The difference between η^2 and partial η^2 is even larger for the effect of condition. This is because SSQ_{Age} is large and it makes a big difference whether or not it is included in the denominator.

As noted previously, it is better to use ω^2 than η^2 because η^2 has a positive bias. You can see that the values for ω^2 are smaller than for η^2 . The calculations for ω^2 are shown below:

$$\omega^2 = \frac{SSQ_{\text{effect}} - df_{\text{effect}}MS_{\text{error}}}{SSQ_{\text{total}} + MS_{\text{error}}}$$

$$\omega_{\text{partial}}^2 = \frac{SSQ_{\text{effect}} - df_{\text{effect}}MS_{\text{error}}}{SSQ_{\text{effect}} + (N - df_{\text{effect}})MS_{\text{error}}}$$

where N is the total number of observations.

The choice of whether to use ω^2 or the partial ω^2 is subjective; neither one is correct or incorrect. However, it is important to understand the difference and, if you are using computer software, to know which version is being computed. (Beware, at least one software package labels the statistics incorrectly).

Correlational Studies

In the section “*Partitioning the Sums of Squares*” in the *Regression chapter*, we saw that the sum of squares for Y (the criterion variable) can be partitioned into the sum of squares explained and the sum of squares error. The proportion of variance explained in multiple regression is therefore:

$$SSQ_{\text{explained}} / SSQ_{\text{total}}$$

In simple regression, the proportion of variance explained is equal to r^2 ; in multiple regression, it is equal to R^2 .

In general, R^2 is analogous to η^2 and is a biased estimate of the variance explained. The following formula for adjusted R^2 is analogous to ω^2 and is less biased (although not completely unbiased):

$$R_{\text{adjusted}}^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where N is the total number of observations and p is the number of predictor variables.

References

- Bem, D. J. (201). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425.
- Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (second ed.). Lawrence Erlbaum Associates.
- Lenth, R. V. (2001) Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician*, 55, 187-193.
- Ritchie, S. J., Wiseman R., and French, C. C. (2012) Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PLoS ONE* 7.

Statistical Literacy

by David M. Lane

Prerequisites

- Chapter 19:

[This article](#) describes some health effects of drinking coffee. Among the key findings were (a) women who drank four or more cups a day reduced their risk of endometrial cancer by 25% compared with those who drank less than one cup a day and (b) men who drank six or more cups had a 60% lower risk of developing the most deadly form of prostate cancer than those who drank less than one cup a day.

What do you think?

What is the technical term for the measure of risk reduction reported? What measures of risk reduction cannot be determined from the article? What additional information would have been helpful for assessing risk reduction?

This is called the "relative risk reduction." The article does not provide information necessary to compute the absolute risk reduction, the odds ratio, or the number needed to treat. It would have been helpful if the article had reported the proportion of women drinking less than one cup a day who developed endometrial cancer as well as the analogous statistic for men and prostate cancer.

Exercises

Prerequisites

All content in this chapter

1. If the probability of a disease is .34 without treatment and .22 with treatment then what is the
 - (a) absolute risk reduction
 - (b) relative risk reduction
 - (c) Odds ratio
 - (d) Number needed to treat
2. When is it meaningful to compute the proportional difference between means?
3. The mean for an experimental group is 12, the mean for the control group were 8, the MSE from the ANOVA is 16, and N, the number of observations is 20, compute g and d.
4. Two experiments investigated the same variables but one of the experiment had subject who differed greatly from each other whereas the subjects in the other experiment were relatively homogeneous. Which experiment would likely have the larger value of g?
5. Why is ω^2 preferable to η^2 ?
6. What is the difference between η^2 and partial η^2 ?

Questions from Case Studies

Teacher Ratings (TR)

7. (TR) What are the values of d and g?
8. (TR) What are the values of ω^2 and η^2 ?

Smiles and Leniency (SL)

9. (SL) What are the values of ω^2 and η^2 ?

Obesity and Bias (OB)

10. For compute ω^2 and partial ω^2 for the effect of “Weight” in a “Weight x Relatedness” ANOVA.