19.3: Difference Between Two Means

Learning Objectives

- State how the inherent meaningfulness of the scales affects the type of measure that should be used
- Compute g
- Compute d
- 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 was 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

$$\log_{10}(440) - \log_{10}(400) = 2.643 - 2.602 = 0.041 \tag{19.3.1}$$

Similarly,

$$\log_{10}(330) - \log_{10}(300) = 2.518 - 2.477 = 0.041 \tag{19.3.2}$$

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.

$$g = \frac{M_1 - M_2}{\sqrt{MSE}} \tag{19.3.3}$$

$$d = g\sqrt{\frac{N}{N-2}} \tag{19.3.4}$$

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 $\text{Experiment } 1$ than in $\text{Experiment } 2$, 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 $10 mm\ Hg$ and that, due to differences in the subject populations sampled in the two experiments, the standard deviation was 20 in $\text{Experiment } 1$ and 30 in $\text{Experiment } 2$. Under these conditions, the standardized measure of effect size would be 0.50 in $\text{Experiment } 1$ and 0.33 in $\text{Experiment } 2$. This standardized difference in effect size occurs even though the effectiveness of the treatment is exactly the same in the two experiments.

Reference

1. Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences (second ed.). Lawrence Erlbaum Associates.
2. Lenth, R. V. (2001) Some Practical Guidelines for Effective Sample Size Determination. The American Statistician, 55, 187-193.
