Chapter 7: Statistical Significance, Effect Size, and Confidence Intervals

I. Overview

a. The key to all inferential statistics is the test for statistical significance
    i. This tells us whether we should conclude that the results we found in our sample(s) was due to something besides random sampling error.

b. Traditionally in statistics, we test a hypothesis to determine whether a result is statistically significant.
    i. The null hypothesis is that the results in our sample were just due to random sampling error.

c. Because tests of statistical significance do not actually indicate how strong our effect was, and are strongly influenced by sample size, researchers should also measure the effect size of their results.
    i. Effect size provides information about *practical significance* of the results.

d. Another indicator of the importance of the statistical result is the confidence interval.

II. Statistical significance

a. Purpose: To determine whether a result generated with sample data is likely to hold true in the population(s) from which the sample(s) were selected.
    i. What should we *infer* from the sample to the population? Hence, inferential statistics.

b. Logic: Whenever a sample is selected from the population, the sample is likely to differ somewhat from the population simply due to random sampling error.
    i. Results that are due solely to random sampling error are not considered meaningful, or significant, because a different sample would have produced a different result.
    ii. So to consider a result meaningful, or significant, the researcher must rule out random sampling error as the cause of the result.
    iii. To make this decision, use probability to determine how likely it is that our statistic was obtained due simply to random chance.
        1. If the probability of obtaining my result *by chance* is too large (usually greater than .05, or 5%, likelihood), I cannot rule out chance as the cause of my results, and therefore will conclude my results are *not* statistically significant.

c. How it works: Compare the statistic to the standard error.
    i. The standard error tells us how much of a difference we should expect to see between the sample and the population just due to random sampling error.
    ii. If the statistic (e.g., difference between a sample mean and a population mean, size of a correlation coefficient, etc.) is quite a bit larger than the standard error, we can conclude that the statistic was probably *not* due to random sampling error, and is therefore significant.

d. An example: Suppose that I have a sample of men and a sample of women and I want to compare their ratings of a movie. On a 10-point scale, I find that the men liked the movie two points more, on average, than the women. Is this a significant difference? Does it indicate that the population of men will like this movie more, on average, than the population of women? Or is the difference in my samples just due to random sampling error?
    i. The difference between my sample means is 2 points. The standard error of the difference between the means is 1 point. So the difference between my sample means is about twice as large as the difference I would expect to get just due to random sampling error. Therefore, it may be significant.

e. Determining whether a result is statistically significant depends on the size of the sample(s) and the type of statistics (e.g., *t* value, *z* score, *F* value, etc.).
    i. Remember that larger samples produce smaller standard errors, and smaller standard errors produce larger *t*, *F*, and *z* values, making them more likely to be statistically significant.

III. Hypothesis testing

a. Logic: To help researchers decide whether a result is statistically significant, they can create a test, called a hypothesis test.

b. The null hypothesis is that the statistic of interest is not different from zero in the population.
    i. E.g., men and women in the population do not differ in how much they liked the movie. Any difference in the sample means was due to random sampling error.
        1. H_o: Sample mean for men = sample mean for women.
        2. H_A: Sample mean for men ≠ sample mean for women.
    ii. When we conclude that a result is statistically significant, we *reject* the null hypothesis.
    iii. When we reject the null hypothesis, we are concluding that the results from our sample were not due to random sampling error, and represent a meaningful effect in the population(s).
        1. But it is always possible that our sample results were due to random sampling error, not matter how unlikely.
        2. So it is always possible that when we reject our null hypothesis, we are making an error.
            a. This is called a Type I error.
        3. It is also possible to retain the null hypothesis and conclude that the the results are *not* statistically significant (i.e., were due to chance) when, in fact, the results were not due to chance.
            a. This is called a Type II error.
    iv. *p* value: The probability of an event occurring by chance (i.e., random sampling) alone.
        1. In inferential statistics, chance refers to random sampling error.
            a. Inferential statistics means reaching a conclusion about population(s) based on sample data.
        2. The normal distribution, *t* distributions, *F* distributions, and other distributions can be used to calculate the exact probability of a statistic occurring by chance (i.e., the *p* value).

3. If the *p* value is less than a certain percent (in the social sciences, usually less than 5%), the result is considered statistically significant.
    a. The 5% cutoff is called the alpha level (α)

IV. Limitations of Hypothesis testing

a. It is designed to answer "yes-or-no" questions. Specifically, is the statistic (e.g., differences between the means, correlation coefficient) different from zero?
    i. Many questions are more complex than that.
        1. How large is the effect?
        2. What is the range of likely values for the population parameter?

b. It is strongly influenced by sample size.
    i. Large samples almost always produce statistically significant results, even for small effects.
        1. Because larger samples produce smaller standard errors, and smaller standard errors produce larger *z*, *t*, and *F* values.

V. Effect size

a. Although statistical significance is important, it is influenced by sample size. Large samples will make even small effects statistically significant.

b. Therefore, researchers often want a measure of the effect size that is independent of sample size.

c. Two common measures of effect size are Cohen's *d* (commonly used for *t* tests) and the percentage of variance explained in the dependent variable
    i. Eta-squared and R^2 are two of the most common measures of explained variance.

VI. Confidence intervals

a. A confidence interval is an interval that is built around a sample statistic that is likely to contain the population parameter.
    i. Researchers generally look at 95% confidence intervals, which means the interval that will contain the population parameter 95% of the time.
        1. Remember that when we only have sample data, we do not know the exact value of the actual population parameter.
        2. But if we know that sample statistic, the standard error, and how confident we want to be with our estimate, we can build a confidence interval.
    ii. Confident intervals can be built around any sample statistic.

b. Example: Building a confidence interval around a sample mean.
    i. Formula: CI_{95} = Sample mean $\pm$ (t_c)(standard error)
    ii. Suppose that we have a random sample of 25 cats, and the average weight of these cats is 10 pounds. The standard error is 1 pound. I want to build a 95% confidence interval.
        1. $CI_{95} = 10 \pm (t_c)(1)$
            a. To find t_c I look in Appendix B, 24 degrees of freedom, alpha level of .05 2-tailed.

b. This value is 2.064. I plug this into the formula.

2. $CI_{95} = 10 \pm (2.064)(1)$
    a. $CI_{95} = 10 \pm 2.064$
    b. $CI_{95} = 7.936, 12.064$
3. I am 95% confident that the population mean for cats is between 7.936 and 12.064 pounds.

c. The smaller the confidence interval, the more confident the researcher is that the sample statistic is close to the value of the population parameter. Larger confidence intervals indicate greater amounts of error in the sample estimate of the population parameter.

VII. Summary

a. Researchers need a way to judge whether a statistical result generated from sample data is meaningful, and not just due to random sampling error.
    i. Hence, the term statistically significant.

b. One way to make this determination is to calculate the probability of obtaining a sample statistic by chance, or random sampling error.
    i. Most agree that if this probability is less than 5%, we can conclude the result was not due to chance, and is therefore significant.

c. The test of statistical significance is the basis for all inferential statistics.

d. Tests of statistical significance are useful, but there is important information that such tests do not provide, and they are heavily influenced by sample size.

e. In addition to testing for statistical significance, researchers often want some measure of the effect size.
    i. Several effect size measures are used, including percentage of variance explained, confidence intervals, and Cohen's *d*.