Statistical Significance, Effect Size, and Practical Significance

Eva Lawrence
Guilford College
October, 2017

Definitions

**Descriptive statistics:** Statistical analyses used to describe characteristics of a sample.

**Inferential statistics:** Statistical analyses used to draw conclusions about a population based on a sample. Inferential statistics provide information to determine if results are statistically significant.

**Effect size:** Strength or magnitude of an effect or relationship.

**Practical significance:** Usefulness or everyday impact of results.
Inferential Statistics

Descriptive statistics describe only the sample whereas one uses inferential statistics to draw conclusions about a population based on a sample. There are many different types of inferential statistics that you will learn about in later chapters, including $t$ tests, chi-square tests, correlational analyses, and ANOVAs.

When you calculate any inferential statistic in SPSS, you will obtain a $p$ value. The $p$ value indicates the probability that your results are due to chance alone, or due to error alone.

Researchers usually choose $p < .05$ as a reasonable amount of error (meaning that there is less than a 5% chance that results are due to error alone).

\[ p = \]

\[ \ldots .0001 \quad .001 \quad .01 \quad .02 \quad .03 \quad .04 \quad .05 \quad .06 \quad .07 \quad .08 \quad .09 \quad .10 \quad .11 \ldots \]

$\text{\textcolor{blue}{$p < .05$}}$

A $p$ value less than .05 indicates statistically significant results.

\textbf{Reject} the null.

There is a chance of a \textbf{Type I error}, and the exact probability of that error is indicated by the $p$ value. For example, $p = .02$ indicates a 2% chance of a Type I error.

$\text{\textcolor{blue}{$p \geq .05$}}$

A $p$ value of .05 or greater indicates results are not statistically significant.

\textbf{Retain} the null.

There is a chance of a \textbf{Type II error}.

\textbf{Note:} Researchers sometimes use more stringent criteria for statistical significance, such as $p < .01$, to further reduce the chance of a Type I error. On rare occasions, researchers might use a less stringent criteria, such as $p < .10$, allowing for a higher probability of a Type I error.
**Reporting and interpreting p values**

Round your p values to two decimal places except in cases where the third decimal place provides important information about your results, such as if rounding would change the interpretation of the results. For example, report $p = .049$ because this indicates statistically significant results whereas rounding to $p = .05$ indicates that results are not statistically significant.

Report the exact p value, except when SPSS reports a p value of .000. This does NOT mean there is no chance of error, the error rate is simply beyond three decimal places. In this situation, report that $p < .001$.

*Examples:*

If SPSS reports $p = .034$
You report: $p = .03$
Interpretation: Results are statistically significant. Reject the null hypothesis.

There is a 3% chance of a Type I error.

If SPSS reports $p = .341$
You report: $p = .34$
Interpretation: Results are not statistically significant. Retain the null.

You have a chance of a Type II error.

If SPSS reports $p = .065$
You report: $p = .07$
Interpretation: Results are not statistically significant. Retain the null.

You have a chance of a Type II error.

If SPSS reports $p = .000$
You report: $p < .001$
Interpretation: Results are statistically significant. Reject the null hypothesis.

There is less than .1% chance of a Type I error.

If SPSS reports $p = .046$
You report: $p = .046$
Interpretation: Results are statistically significant. Reject the null hypothesis.

There is a 4.6% chance of a Type I error.

*What if $p = .05$ or just slightly higher?* In such cases, you might say that the results “approach significance” and report the exact p value. Some researchers will use this language of “approaching significance” with numbers up to $p = .10$, although others do not agree with this practice.
**The theory behind statistical significance**
Statistical significance testing is based on probability theory in which you test your results against a population in which the null hypothesis is true. This population distribution is normally distributed.

For example, a *t* test is an inferential statistic used to compare the means of two groups.

The null hypothesis is that there is no difference between the groups, and for a *t* test, the null hypothesis is that $t = 0$.

![Diagram showing normal distribution and percentages between different SDs](image)

The alternative hypothesis is that your sample mean is different from your population mean. Therefore, you want the *t* to be further away from 0, typically at least 2 *SDs* away from 0 so that your sample falls in the extremes of the population representing the null hypothesis.

Why 2 *SDs* away? Because about 95% of the scores will fall within 2 *SDs* of the mean. If your *t* is greater than or less than 2 *SDs*, there is less than a 5% probability (*p*) that it represents this null hypothesis population distribution.
In statistical significance testing, the areas at the extreme of the population representing the null hypothesis are called the “regions of rejection.”

When you set your criteria for statistical significance at $p < .05$, the regions of rejection represent scores above and below 2 $SD$s from the mean. If your score falls in the region of rejection, you can reject your null hypothesis.

The critical value of $t$ ($t_{crit}$) marks the cutoff scores for the region of rejection. In order for an observed $t$ to be considered statistically significant, it must be the same or stronger than the $t_{crit}$. The $t_{crit}$ depends on sample size - the larger your sample, the smaller the $t_{crit}$ required for the results to be considered statistically significant.

This is a “two-tailed” test because the regions of rejection are at both tails: $+2$ and $-2$ $SD$s from the mean.

Even with a directional alternative hypothesis where you might use a one-tailed test (with only one region of rejection at either the positive or negative extreme), most researchers still run a two-tailed test because it is more stringent.
**Effect Sizes**
Whereas results of inferential statistics will tell you whether or not your results met the criteria for statistical significance, the effect size will give you information on the strength of the effect or relationship you examined. You can have a strong but not statistically significant effect and you can have a weak but statistically significant effect. In other words, results of inferential statistics and effect sizes can vary independently, and it is therefore important that you report both.

Two basic types of effect size are **proportion of variance accounted for** and a **standardized difference score** (most often Cohen's $d$, although there are others).

**Proportion of Variance Accounted For**

Recall that the standard deviation ($SD$) summarizes the degree to which scores differ from the mean ($M$). Variance is the standard deviation squared ($SD^2$), which is important to know if you are calculating statistics by hand. For our purposes, however, think of variance as providing the same type of information as the standard deviation in that they both tell us how much scores vary from the mean.

When we examine the relationship between variables, or the effect of one variable on another, the proportion of variance accounted for tells us how much variance in one variable is accounted for by the other. There are different statistics for proportion of variance accounted for based on the types of variables you are examining. If you have two variables that are both interval or ratio, it is $r^2$ or the coefficient of determination. If one of your variables is nominal and dichotomous and the other is interval or ratio, it is $r_{pb}^2$ or the squared point-biserial correlation.

An example might help make this clearer:

A professor examining final exam scores might wonder if those who studied more did better on the exam. In proportion of variance accounted for terms, the professor wonders how much of the variance in final exam scores is accounted for by study time.

If there is no relationship between study time and exam score, then 0% of the variance in one variable is accounted for by the other. If there is a perfect relationship, then 100% of the variance is accounted for.

![Diagram showing 0% of variance accounted for by study time vs. 100% of variance accounted for by study time]
These extremes (0% vs. 100%) are both unlikely. We would expect that there will be at least some relationship between studying and scores, but we would not expect that the relationship would be perfect. In addition to study time, exam scores are usually dependent on lots of different factors such as overall performance and attention in the class, test taking skills, stress levels, how well one slept, etc. Measurement error is also a factor in test scores, as it is with all scores.

This example extends to other constructs in that there are multiple factors impacting any score, especially in the social sciences when we are interested in constructs such as love, motivation, personality, etc. Consequently, for $r^2$ and $r_{pb}^2$, Cohen (1988) suggested that about 25% of the variance accounted for is a large effect in the social sciences, and about 9% is moderate, and about 1% is small.
**Standardized Difference Score**
The simple difference between the mean of two groups can provide important information about the magnitude of the difference between the two means in the sample, but we would not be able to easily evaluate the size of the difference beyond the sample. Consequently, it is better to calculate a standardized difference score that takes into account error and that is comparable across samples.

*Cohen’s* $d$ is a standardized difference score that is on standard deviation units. It is a useful and commonly used effect size when comparing the mean of two groups.

For example:

A professor examines the difference in final exam scores between those who studied less than 6 hours and those who studied six hours or more.

If there was absolutely no difference between the groups, Cohen’s $d$ would be 0 and the distributions would completely overlap, as shown below:

![Diagram showing two overlapping normal distributions for exam scores](http://rpsychologist.com/d3/cohend/)

If there are three standard deviations between the means of the two groups, Cohen’s $d$ would be 3.

*Note.* The images on this and the next page are from Magnussen (n.d.) The site is interactive – check it out for yourself: [http://rpsychologist.com/d3/cohend/](http://rpsychologist.com/d3/cohend/)
When interpreting Cohen’s $d$ scores, Cohen (1988) recommended interpreting about .80 as large, about .50 as moderate, and about .20 as small.

<table>
<thead>
<tr>
<th>Proportion of Variance Accounted for ($r^2$ or $r_{pb}^2$)</th>
<th>Cohen’s $d$</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\approx 1%$</td>
<td>$\approx .20$</td>
<td>Small/Weak</td>
</tr>
<tr>
<td>$\approx 9%$</td>
<td>$\approx .50$</td>
<td>Medium/Moderate</td>
</tr>
<tr>
<td>$\approx 25%$</td>
<td>$\approx .80$</td>
<td>Large/Strong</td>
</tr>
</tbody>
</table>
Practical Significance

Practical significance refers to the usefulness of our results or findings from our study. In other words: How do the results affect or apply to daily life? Even if we find statistical significance, the difference we find may not be noticeable or noticeably affect people’s lives. On the other hand, a study that did not find statistical significance or have a large effect size may still yield important, practical results.

A classic example of this is found in the medical literature. Findings from a study comparing heart attacks among those who took aspirin versus a placebo was stopped before completion because preliminary results were so clearly in favor of aspirin’s benefits. The effect size was miniscule, with aspirin accounting for only a tenth of a percentage point in the reduction of heart attacks (Rosenthal, 1990). But the difference in outcomes revealed that those who took aspirin were slightly less likely to have a heart attack and much less likely to die from a heart attack, plus there were no clear health problems resulting from aspirin (Steering Committee of the Physician’s Health Study Research Group, 1989). When we are talking about a life and death situation, even a very small difference can be meaningful.

An example that might hit closer to home is the impact of time spent studying on exam scores. If it turns out that those who studied at least six hours for an exam scored a full letter grade higher than those who studied less than six hours, those results would have practical significance for students regardless if the difference in tests scores was statistically significant. On the other hand, if time spent studying had no meaningful impact on grades, one might decide the difference is not practically important. This assessment of practical significance would be the same if the results were statistically significant or not, and if the effect size was large or small.

There are no clear-cut guidelines to evaluate the practical significance of results. Evaluation of practical significance requires a broader perspective and a consideration of context. Interpretations of practical significance are usually included in the Discussion section of a research report.

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Where to report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive Statistics</td>
<td>Results Section</td>
</tr>
<tr>
<td>Confidence Intervals (optional but recommended)</td>
<td>(and Method Section, as applicable)</td>
</tr>
<tr>
<td>Inferential Statistics</td>
<td></td>
</tr>
<tr>
<td>Effect Sizes</td>
<td></td>
</tr>
<tr>
<td>Practical Significance</td>
<td>Discussion Section</td>
</tr>
</tbody>
</table>