University of Guelph Numeracy Project

Statistical versus Practical Significance

TABLE OF CONTENTS

Statistical versus Practical Significance

What is STATISTICAL SIGNIFICANCE?

- Statistical significance refers to the unlikelihood that mean differences observed in the sample have occurred due to sampling error. Given a large enough sample, despite seemingly insignificant population differences, one might still find statistical significance.

What is PRACTICAL SIGNIFICANCE?

- Practical significance looks at whether the difference is large enough to be of value in a practical sense.

Statistical Significance

Significance Level: Overview

- H_0: Also known as the null hypothesis, it is a statement of "no effect" or "no difference" used in tests of significance. The test of significance is used to obtain substantial evidence against H_0.

- H_a: Also known as the alternative hypothesis, is what we hope to be correct in rejection of the null hypothesis.

- P-Value: A P-value is essentially the probability that the test statistic you compute would acquire a value as extreme or more than what you actually observe. This is after making the assumption that H_0 is true. A common rule is the smaller the P-value, the stronger the evidence is against H_0.

- What does the P-value have to do with significance level?

  We can compare the P-value calculated with a fixed value. This fixed value is called the significance level.

- Significance Level: Is denoted by the Greek letter alpha (α). If the P-value we calculate is as small or smaller than α, we say the data is statistically significant at level α.

- If we chose $\alpha = 0.05$, we are insisting that evidence against H_0 is so strong that it would happen no more than 5% of the time (1 time in 20) when H_0 is true.

- If we chose $\alpha = 0.01$, we are demanding even stronger evidence against H_0, evidence so strong that it would appear only 1% of the time (1 time in 100) if H_0 is true.

What does "statistically significant" mean?

- When we use the term "significant", we are not automatically implying that the data is "important". It is simply a term that is used to assess whether the evidence against the null hypothesis has reached the standard set by α only.

- For example, significance level at 0.05 is often expressed by the statement: "The results were significant at ($P < 0.05$)", where P stands for the P-value.

- The P-value is helpful in providing basic information rather than a statement of significance, because we can then assess significance at any level we choose.

Four Steps Common to All Tests of Significance

- 1) State H_0 and H_a. The test is designed to assess the strength of the evidence against H_0, H_a is the statement that will enable us to reject H_0.

  2) Calculate the value of the test statistic. This statistic usually measures how far the data is from H_0.

  3) Find the P-value for the observed data. Don't forget, this is the probability, assuming H_0 is true, that the test statistic will weigh against H_0.

  4) State a conclusion by choosing a significance level α. If the P-value is less than or equal to α, you conclude that the alternative hypothesis is true. If it is greater than α, you conclude that the data does not provide sufficient evidence to reject the null hypothesis.

Choosing a Level of Significance

- Certain standard levels of significance such as 10%, 5% and 1% are often used. The 5% level ($\alpha = 0.05$) is particularly common. Significance at the 5% level is still a widely accepted criterion for meaningful evidence in research work.

- It is important to note that there is no sharp border between statistically significant and insignificant results. There is no practical distinction between the P-values 0.049 and 0.051. Therefore, it doesn't make sense to treat $\alpha = 0.05$ as a universal rule for what is significant.

Summary

- Beware of placing too much weight on traditional values of α, such as $\alpha = 0.05$.

- Lack of significance does not imply that H_0 is true, make sure to look at the power of the test.

- Significance tests are not always valid. Faulty data collection, outliers in the data and others can invalidate the test.

- Beware of searching for significance. It is often tempting to make significance itself the ultimate goal of your research, especially when investigating a new phenomenon.

Effect Size on Statistical Significance

What is Effect Size?

- To review, when a difference is statistically significant, it does not necessarily mean that it is big, important, or helpful. It simply means you can be confident that there is a difference.

- Effect size is a measure of the strength of the relationship between two variables. In a research setting, it is not only helpful to know whether results have a statistically significant effect, but also the magnitude of any observed effects.

- To know if an observed difference is not only statistically significant but also important or meaningful, you will need to calculate its effect size.

- When H_0 ("no effect" or "no difference") is rejected at the usual levels ($\alpha = 0.05$ or $\alpha = 0.01$), there is good evidence that an effect is present. But, that effect may be extremely minor and trivial.

- When a large sample size is accessible, even tiny deviations from the null hypothesis will be significant.

- If the study is based on a very large sample size, relationships found to be statistically significant may not have much practical significance. Almost any null hypothesis can be rejected if the sample size is large enough.

  This has implications on practical significance, as statistically significant results may be practically applied despite having an extremely small effect size.

- How do we remedy the problem of attaching too much importance to statistical significance?

  This can be done by providing a confidence interval for the parameter for which you are interested in. A confidence interval can actually estimate the size of an effect.

- To review, a confidence interval is an interval that is calculated from repeated sampling. If identical samples are repeatedly drawn from a population and each given a confidence interval, theoretically "C" of these intervals should contain the population mean. "C" is often expressed as a percentage (most commonly 95%).

- It is helpful to examine a confidence interval so that you can determine the MAGNITUDE of the effect.

- Confidence intervals are often underused when it comes to interpreting the results of a significance test.

The Power of a Test

- An important aspect of planning a study is to verify that the test you plan to use has a high probability of detecting an effect of the size you hope to find. This probability is called the Power of the test.

- In reference to your tests, power is a measure of the probability that you will reject the null hypothesis when you should. It is common practice to require a power of .8 or greater. This refers to an 80% or greater chance of discovering a statistically significant difference when one is present.

- Calculating power consists of 3 steps:

  1) State H0, Ha and the significance level α.

  2) Find the mean value (x) that will direct us to reject H0.

  3) Calculate the likelihood of observing these values of x when the alternative is true.

  Calculations of power are beyond the scope of this particular tutorial. It is essential to understand though, that power is an important consideration when planning a study.

- Power calculations are often used to determine the sample size needed prior to a study. To ensure that your sample size is big enough, you will need to conduct a power analysis calculation. This is also beyond the scope of this tutorial. Unfortunately, these calculations are not easy to do by hand, so unless you are a statistician, you will want the help of a software program.

- Overall, power calculations are useful when planning studies. Using a significance test with low power can be detrimental to the study. Low power makes it unlikely that you will find a significant effect even if the alternative hypothesis is favoured.

Practical Significance

Practical Significance

- Practical significance is not the same as statistical significance.

- Recall, a general problem with traditional statistics is that if you take large enough samples, almost any difference or any correlation will be significant.

- Due to this problem, many editors and publishers are requiring authors to include some information surrounding the practical significance of their findings.

- Significant study results vary based on context. Can significant study results ever be translated into recommendations for the general public? We know so far that:

  1) Tests of statistical significance rarely tell us about the importance of a research result.

  2) Effect size tells us about magnitude of difference, which is important, but it is difficult for practice-oriented practitioners to comprehend.

Solutions

- There are several solutions to help translate statistically significant data into results that may be practically applied to real life situations:

  1) Comparison of a Sample to a Meaningful Reference Group

  Some research results are in the form of test scores that can be compared to a table of norms. By comparing the mean score of the sample to a table of norms, researchers and practitioners can evaluate how well the sample performed in relation to the population on whom the test was performed. This can help them draw conclusions about whether the sample is average, superior, or inferior when compared to the population.

2) Confidence Limits

Confidence limits are the lower and upper values of a confidence interval. Confidence limits give researchers the opportunity to estimate population values from sample statistics, and are typically expressed as "margins of errors". Although the public may not understand the statistical meaning of a margin of error, it can be translated into a percentage that is meaningful to the general population.

Confidence limits can also help researchers determine the importance of a result. If the confidence limits are small, researchers have a high degree of assurance that results are true or nearly true. Conversely, if the confidence limits are large, researchers must be more cautious.

3) Know when to detect faulty data

When reading statistically significant study results, keep the following warning signs in mind:

-Observe the sample size used to obtain study results. Remember that if the study is based on a very large sample size, relationships found to be statistically significant may not have much practical significance.

-The media has the tendency to report only "sexy" results; beware of reports where it is evident that many tests were conducted, but where results of only 1 or 2 studies are presented as "significant".

Glossary

Statistical significance:	Refers to the unlikelihood that mean differences observed in the sample have occurred due to sampling error. Given a large enough sample, despite seemingly insignificant population differences, one might still find statistical significance.
Practical significance:	Looks at whether the difference is large enough to be of value in a practical sense.
H_0, null hypothesis:	The statement being tested in a test of significance.
H_a, alternative hypothesis:	The statement we hope or suspect is true instead of H_0.
P-Value:	Is essentially the probability that the test statistic you compute would acquire a value as extreme or more than what you actually observe. This is after making the assumption that H_0 is true. A common rule is the smaller the P-value, the stronger the evidence is against H_0.
Significance Level:	Is denoted by the Greek letter alpha (α). If the P-value calculated is as small or smaller than α, we say the data is statistically significant at level α.
Effect size:	Is a measure of the strength of the relationship between two variables.

Confidence interval:	Is an interval that is calculated from repeated sampling. If identical samples are repeatedly drawn from a population and each given a confidence interval, theoretically "C" of these intervals should contain the population mean. "C" is often expressed as a percentage (most commonly 95%).
Power of a test:	Power is a measure of the probability that you will reject the null hypothesis when you should.

References

Gall, M.D. *Figuring out the Importance of Research Results: Statistical Significance versus Practical Significance*. 2001. Retrieved from www.uoregon.edu/~mgall/statistical_significance_v.htm.

McCabe, George P. & Moore, David S. *Introduction to the Practice of Statistics*: Fourth Edition. 2003. W.H. Freeman and Company: New York, USA.

Utts, Jessica M. *Seeing Through Statistics*: Third Edition. 2005. Thomson Learning Inc: Toronto, Canada.