What Constitutes Significant Results When Working with Large Data Sets?

Robert L. Andrews, Virginia Commonwealth University
Department of Management, Richmond, VA. 23284-4000
804-828-7101, rlandrew@vcu.edu

Jonathan A. Andrews

ABSTRACT

This paper discusses issues arising from determining if significant results exist when working with large data sets. A procedure is proposed for determining practically significant results. The paper also discusses the value of failing to find significant results.

INTRODUCTION AND OVERVIEW

The electronic collection of data can provide an enormous amount of data. A *New York Times* article [4] states, "In field after field, computing and the Web are creating new realms of data to explore — sensor signals, surveillance tapes, social network chatter, public records and more. And the digital data surge only promises to accelerate, rising fivefold by 2012, according to a projection by IDC, a research firm."

The analysis of data sets with such large sample sizes changes from the analysis of data that consist of a relatively small sample size selected from a process or large population. When analyzing data from a sample that is relatively small, the customary procedure is to test for statistical significance using a significance level α. However, one knows that statistically significant results can be obtained for very little departure from the null hypothesis if the sample size is very large. This departure from the null is referred to as the effect size, but this is a topic that is generally not given much attention in typical statistics courses.

The ubiquitous source of information on the web, Wikipedia [6], defines statistically significant as follows: "In statistics, a result is called statistically significant if it is unlikely to have occurred by chance. 'A statistically significant difference' simply means there is statistical evidence that there is a difference; it does not mean the difference is necessarily large, important, or significant in the common meaning of the word." This definition infers that there is a difference in the meaning of significant between the world of statistics and common every day language. To define this difference in the understanding of significance, the term practically significant is often used to show a contrast to statistically significant. Another web source, Wiki.answers [5], defines practically significant as follows: "There is statistical significance, an arbitrary limit whereby an observed difference could reasonably be assumed to be due to some factor other than pure chance. Then there is practical significance, an arbitrary limit whereby an observed difference is of some practical use in the real world." Unfortunately some may think that this infers that statistics is not of practical use in the real world. Statistical significance is clearly important, but to be of value in the context of a particular problem one needs to also make sure that there is practical significance.

As has been pointed out, it is possible to have statistical significance when the result does not support practical significance. At the same time one can observe a result that would be practically significant if it was truly representative of the phenomenon being studied. However, the result may not be statistically significant, meaning that the magnitude of the difference from the null hypothesis was too likely to have

happened by chance to say that the data support a statistically significant result. Of course, this is not very likely with a very large data set.

LARGE DATA SET ISSUES

Many organizations are able to obtain large sets of data through automated recording of transactional data for processes that are a part of their regular operations. Similarly, computers are able to automatically record all types of data from experiments. The availability of computers has greatly reduced the cost of obtaining data. Hence, when setting up data collection procedures, the general rule now becomes to make sure that data are recorded for anything that might potentially be of future use. With limited amounts of data the challenge is to obtain good data that are representative of the phenomenon being studied and then to use these data to make inferences about the characteristics of the phenomenon. Other challenges exist with large data sets. There is so much information recorded in large data sets that finding information of real value can sometimes become like finding the proverbial needle in the haystack. The availability of data and the computational tools to analyze them have been the catalyst for new areas/disciplines of analysis taking the names of Exploratory Data Analysis, Data Mining, Business Intelligence, and most recently Analytics or Business Analytics. Davenport and Harris [2, page 7] defined analytics as "the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decision and actions." This definition emphasizes that the analysis is taking place for a situation where decisions are being made and actions are being carried out based on these decisions.

Statistical significance can be determined with no real knowledge of the phenomenon. However, determining practical significance requires knowledge of the phenomenon and the anticipated consequence of associated decisions. To be effective for decision making, analysis needs to determine both statistical and practical significance.

The goal of most hypothesis testing is to obtain a significant result. When one does not find a significant result the language that is often used is "fail to reject the null." This indicates that the investigation failed to find a significant result and may as a result be judged as a failure and of no real value. However, if the sample size is large then the result has provided some clear information about the phenomenon. The lack of significance for a relationship means that this relationship need not be considered in decision making.

Murphy and Myors [3, page 35] address some of the above issues about determining significance. They state, "most dictionary definitions of 'significant' include synonyms such as 'important' or 'weighty.' However, these tests do not directly assess the size or importance of treatment effects. Tests of the traditional null hypothesis are more likely to tell about the sensitivity of the study than about the phenomenon being studied. With large samples, statistical tests of the traditional null hypothesis become so sensitive that they can detect the slightest difference between a sample result and the specific value that characterized the null hypothesis, even if the difference is negligibly small. With small samples, on the other hand, it is difficult to establish that anything has a statistically significant effect." In an attempt to rectify the cited difficulty with testing the traditional null hypothesis, they propose testing a minimum-effect hypothesis to detect if the data support that there is some effect beyond what would be considered as being of no real importance. They also state that, "The main advantage of the traditional null hypothesis is that it is simple and objective. If researchers reject the hypothesis that treatments have no effect, they are left with the alternative that they have some effect. On the other hand, testing minimum-effect hypotheses requires value judgments, and requires that some consensus be reached in a particular field of inquiry."

Following the arguments presented about determining significance and large data sets, it should be clear that testing for statistical significance alone is not an adequate analysis procedure. The analyst needs to

use knowledge of the phenomenon in conjunction with statistical knowledge and analysis results to determine if the data indicate the existence of significant results. A significant effect is one that would be practically significant to the degree that knowledge of an effect with this magnitude would be of value in decision making.

SUGGESTED METHODS FOR DETECTING PRACTICAL SIGNIFICANCE

Traditional procedures for hypothesis testing provide three different methods for testing for statistical significance using some value of α for the level of statistical significance. These three procedures consist of using critical values for the test statistic, p-values based on the test statistic value or a confidence interval for a parameter of interest. Andrews [1] discussed hypothesis testing for location parameters. He specifically cited situations where one is testing to determine if

1. the phenomenon mean equals some specified value,
2. the phenomenon proportion equals some specified value,
3. the difference in two phenomenon means equals some specified value,
4. two phenomenon proportions are equal, or
5. the phenomenon regression coefficient equals some specified value.

In all of these situations the analyst can create a $100(1-\alpha)\%$ confidence interval for the parameter of interest or difference in two parameters. Based on knowledge of the situation, decision makers can use this confidence interval and conclude whether practical significance exits or not. Practical significance is concluded if knowing that the parameter had any of the values in the range presented by the confidence interval would have an impact on decision making. However, if there was at least some set of values in the interval for the parameter that would not really have any impact on decision making then the conclusion would be that the results were not practically significant. This decision can be reached in one of two ways. One way would be to decide after the analysis if all of the values in the confidence interval have some practical impact. The other would involve decision maker input prior to or apart from knowing the analysis results. The decision maker can establish a range of values of the parameter that would have no effect on decision making and range that would have an effect. There would be a demarcation value or values separating these two regions. In this case one could conclude practical significance only if all of the values in the confidence interval were in the range for practical significance. If any part of the "no effect" range is inside the confidence interval then one would conclude no practical significance.

The other two hypothesis testing methods use a test statistic. When testing for statistical significance, the analyst can use α and the probability distribution for the test statistic when the null is true to establish a critical value or values that define a rejection region that is in either one or two tails of the distribution. If the observed test statistic calculated from the data is in the rejection region then a statistically significant result exists. For practical significance, the demarcation value or values determined by subjective knowledge of the situation as described above will be used rather than a null hypothesized value for the parameter. For situations with a single range for practically significant parameter values corresponding to a one-sided rejection region, the demarcation value for the region would be used as the null hypothesized to calculate the test statistic and the direction of the one-sided test would be in the direction of the practically significant values relative to the demarcation value. Then a one-sided test can be conducted using either the critical value or p-value decision rule. If there is an upper and lower range for practically significant parameter values, then the test should be performed using the demarcation value that is closer to the observed statistic. For the critical value method of testing in this situation, the rejection region would be in the direction of the practically significant values relative to the demarcation value and the tail probability area would be $\alpha/2$. For the p-value method of testing, the tail area for the test statistic would be doubled to obtain a two-sided p-value.

The described procedures effectively test simultaneously for both practical and statistical significance. The procedures use a standard significance level α and knowledge about what would constitute an effect of practical value in decision making.

VALUE OF FAILURE TO CONCLUDE THAT A SIGNIFICANT RESULT EXISTS

As was indicated earlier, failure to reject the null and conclude that a significant result does not exist is often viewed as a failure for the testing procedure and of no value. If the sample size is small then failure to observe a significant result may just indicate that there was not enough power in the testing procedure to detect a significant result. However, if the result was obtained from a large data set then that essentially assures that the procedure would have the necessary power. Hence, in many decision making situations where a large data set is available, failing to find a significant result is clearly of value, if the data set is truly indicative of the phenomenon of interest. Often the reason that the analysis was performed was to confirm beliefs currently held by the decision makers. Without data, the decision makers would still tend to use their beliefs when making decisions about the phenomenon. Suppose that the decision makers believe that a certain controllable variable has a significant impact on an outcome variable. Without data to support that any such impact is really negligible and of no practical significance, decision makers would attempt to manipulate the controllable variable thinking that they are having impact on the outcome. But with the knowledge that this variable was not significant, they would no longer pursue any future attempts to manipulate the variable with the false sense of improvement.

CONCLUSION

Analysis of large data sets presents different issues from smaller data sets. One needs to consider both statistical and practical significance when working with large data sets. Considering both allows one to avoid taking action on something that is statistically significant but has no practical significance, and would be a waste of effort. Conversely it also allows one to avoid taking action on something that is practically significant but is not statistically significant, and thereby actual a random occurrence of no value to act on. We have presented a procedure that simultaneously tests for statistical and practical significance. We have also discussed how failure to detect a significant difference for a large data set provides different information than reaching this conclusion for a small data set.

REFERENCES

[1] Andrews, Robert L., "Suggestions for Hypothesis Testing for Location Parameters," Proceedings, Southeast Region of the Decision Sciences Institute, February, 2004, pp. 177-179

[2] Davenport, Thomas C. and Harris, Jeanne G., Competing on Analytics, Harvard Business School Press, Boston, MA, 2007.

[3] Murphy, Kevin R. and Myors, Brett, Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests, Second Edition, Lawrence Erlbaum Associates, Mahwah, NJ, 2004.

[4] *New York Times*, (5/13/2009) http://www.nytimes.com/2009/08/06/technology/06stats.html?_r=1&emc=eta1

[5] Wiki.answers, (5/13/2009) http://wiki.answers.com/Q/What_is_practical_significance

[6] Wikipedia, (5/13/2009) http://en.wikipedia.org/wiki/Statistical_significance