Statistics for Program Assessment: Has the Program Made a Difference?

Mary R. Anderson-Rowland
Arizona State University

Abstract

As funding becomes scarcer and the demand for accountability increases, creditable assessment and evaluation become more important. For example, funding is generally scarce for programs to establish and to improve activities designed to increase enrollment and retention in engineering. Therefore, almost all funding allocated to these recruitment and retention activities requires an assessment of the program to see if the money and time have been well spent.

This paper describes basic statistical concepts that should be considered when assessing a program or activity. Examples are given to illustrate both good and poor program assessment. Warnings are given for data that may turn out to be useless and suggestions presented on ways to enhance data presentation. What it takes for data to be "significant" will also be discussed, as well as the problem of sample size.

Without the proper planning of assessments and data collection, it may be very difficult to show that the program has made a difference. If a program director does not have a good statistical background, they would be well advised to have an assessment person on their team to help plan assessment strategy, to analyze the data, and to draw conclusions.

Keywords: Evaluation, Assessment, Data Analysis, Statistical Testing

I. Introduction

Many university and college budgets are strained. There is not enough money to go around to comfortably support all of the programs worthy of funding. Terms such as accountability, productivity, responsiveness, efficiency, results, impact, and leveraging are used as tough decisions are made to fund and to continue programs. Engineering schools today are engaged in many activities outside of the classroom. Major issues include recruitment, retention, graduation, and K-12 Outreach Programs. To fund these programs, tough decisions needs to be made by engineering deans on how much money goes to support outreach and retention along with hiring faculty, providing seed money for research, and buying equipment. Many engineering programs seek national funding through a government organization such as the National Science Foundation or the Department of Education. To show that the money and time will be well spent on any particular project, an assessment plan is needed. During the project and at the end of a project, a report is usually required to show that the program was successful, that a change was made, or a result was obtained. User-friendly guidebooks have been developed that describe both formative and summative assessment.[1]

In planning a program, the evaluation process needs to be determined. Information is needed to indicate when changes should be made to the program and to show that the program has made a difference. Anecdotal evidence is nice, but usually funders want analytical proof that a program made a difference. It is very important to think through exactly what type of evaluation will be used before data is collected. A statistician or other evaluator's nightmare is to be called in at the end of a project, be given data, and be told to evaluate the data. If the data collection has not been carefully planned, the data may be useless. Without the proper planning of assessments and data collection, it may be very difficult to show that the program made a difference. If the person responsible for evaluating a program does not have a good statistical background, that person would be well advised to have an assessment person on their team to help plan the assessment strategy, to analyze the data, and to draw conclusions.

One should be warned that many "experts" claim to be a part of the statistical analysis arena. Quality control and continuous improvement are current buzzwords and many of these "experts" have studied statistics and assessment only at a very shallow level and tend to only use the formulas they find in a book. If the data does not fit an example in the book, then this "expert" may conclude there is no known way to handle it. This "expert" rarely checks to see if the data is even appropriate (has met all of the required assumptions) for the test chosen. A reality check should always be carried out. Do the results of the analysis ring true with the data or with those who have worked with the data? Of course, there will always be some data surprises that will show a conclusion that was not evident before the analysis. Faculty members in departments such as Industrial Engineering, who are in the Probability, Reliability, Design of Experiments, Quality Control, and the Statistics arena, are generally well aware of the proper and improper use of statistics in evaluation.

One sage wagged, "There are three kinds of lies: lies, damned lies, and statistics." You may have heard the quote, "You can prove anything you want with statistics." It is true that there has been much misuse of statistics. Most of this misuse is due to ignorance; it is quite easy to misapply statistical methods and have them appear perfectly valid. So, advice number one is to align with a good assessment person who understands statistics. Advice number two is to become familiar with some of the common pitfalls in statistics.

II. Has the Program Made a Difference?

In most cases, a program director wants to show that the program made a difference. For example, more students are interested in entering engineering, the grade point average has increased, or more students were retained and graduated due to a program. However, even these simple examples require a word of caution. When looking at graduation rates, one cannot simply compare the number of graduating students this year with the number of freshmen four years before. In those four years, students have transferred in and because they are working, some of the students will graduate in five, six, or seven years. A set group of students must be selected and those particular students must be followed for several years to ascertain if they have graduated, are still in school, or have left the institution. The graduation rates of a particular cohort can best be reported by the percentage that have graduated at the end of four years along with how many are still in school, then how many have graduated at the end of five years and how many are still in school, etc. In reporting an increase in GPA, the researcher again has to make sure that the population is the same for the "before" and "after" comparisons. Another word of caution, if several programs have impacted the student cohort in question, how do you know which program had the greatest impact? Also, how much of an impact do you have to show in order to say that the program had a "significant" impact on a group of students?

To show that a program has made a difference, it is important to know what the situation was before the program began. It is important to have a baseline, a picture of the landscape before a program was begun to make changes. For example, a college diversity committee is interested in instituting programs to make the college a friendlier place, one built on respect for diversity. To be able to claim in a few years that the college climate has improved, it is important to know what the college climate was before changes were attempted. In this case, since the college may have over 200 faculty, 332 administrators, staff, and academic professionals, and over 6,600 students, it is not feasible to query each person, keep track of them by name or number, and then to reevaluate the climate in a couple of years with the same exact people. In this case, a survey, with a decent return (perhaps over 30%) can paint a general picture of the college climate. Focus groups can also be used to gain specific information. Questions can be asked in focus groups in a reevaluation to determine if changes have occurred. In this case, conclusions about the general climate will be made based on samples, which will vary from the "before" to the "after."

Suppose that one wishes to measure the effect of a program on a smaller group of participants. Consider 35 participants who completed pre-program surveys and 28 participants who completed post-program surveys. Only twenty of the participants took both surveys: furthermore, the surveys were not labeled in a way to pair the responses for those twenty participants. An incorrect program assessment would be to treat the pre-program and post-program groups as independent, use the mean and standard deviation from each group, and calculate the significance using a t-test. However, if a correct assessment is to be made, the pre-data must be identified and marked by the individual participants so that the difference in the pre and post test score for each individual in the program can be calculated. If a student took the pre-program evaluation and is not available to take the post-program evaluation, that pre-program score cannot be used to determine if the program had an effect. Similarly, the post-program score of a student cannot be used if the student did not take the pre-program evaluation. It is very important then that each pre-program score is coded to the individual who produced the score. If this is not done, then an evaluation of the effect of the program cannot be done later with unidentified individual pre-program scores.

As an example of a correct analysis of data, consider the scores on a gender equity survey given to teachers at the beginning of a program. The total number of correct answers is recorded for each teacher. At the end of the program, the same survey is given to the teachers. The difference between the number correct at the end of the course and the number correct at the beginning of the class is recorded. If there is evidence that the scores come from an approximately normal distribution (a plot of the data forms an approximate triangle or bell-shaped curve) and if we have some 20 or more individuals in our class, then a paired t-test can be used to analyze the data.[2] For an example of an application of this test, see Reference 3. The paired t-test automatically takes the sample size into account. However, in general, the paired t-test will not be very accurate with a sample size much smaller than 20 to 30. Consider the application of the paired t-test on a sample of just six participants.

Example: Number of Correct Answers on the Same Gender Equity Test Given at the Beginning and the End of the Gender Equity Program

Participant	Pre-Test Score	Post-Test Score	d_i = difference
1	10	13	-3
2	10	12	-2
3	12	12	0
4	11	11	0
5	9	8	1
6	12	15	-3

The null hypothesis here is that there is no difference in the pre-test and the post-test scores, or that the gender equity training in the program made no difference in the participants. The post-test score is subtracted from the pre-test score for each participant and the difference d_i is recorded. The paired t-test uses the average of the d_i's (in this case, -1.16667) and the standard deviation of the d_i's to calculate the t statistic = -1.659. In this case, the p-value of the t-test is 0.15798. (Statgraphics, a computer statistical package, was used to calculate the p-value and all other such reported values in this paper.) The p-value here tells us that there is a 15.8% chance that the test statistic could have a value at least this extreme and the null hypothesis still be true. In this case, we would generally say that we cannot reject the null hypothesis that there is no difference between the pre-test and the post-test scores. This leads us to the question, how do we know whether the program made a "significant" change in the scores and what is a p-value?

III. Is The Program Effect Significant? (What is a p-value?)

When testing to see if a program made a difference, often the null hypothesis is that the program made no difference. If we accept the null hypothesis, we are saying that, due to natural variation, there may be some change in the measured outcomes, but the changes are not large enough to claim that the program made a difference. How large do the changes have to be to claim that the program made a difference? Often statisticians use a p-value of .05 to mark the difference between a program that made a difference and a program that did not make a difference. If the p-value is .05, we are saying that if we conclude that the program made a difference, we will on average be wrong only 5% of the time. That is, due to natural variation, the program could have made no difference and 5% of the time the data would show that there was a difference. So if a p-value is .05 or smaller, we say that the program made a "significant" difference. If the p-value is .01 or smaller, we say that the program made a "very significant" difference and if the p-value is .001 or smaller, we say that the program made a "very highly significant" difference.

Note that these classifications are quite arbitrary. If we test data and the p-value is .06, we may conclude that the program was not significant; it did not make a difference. Is that, however, the conclusion we wish to make? We may be willing to proceed assuming the program made a difference when we have only a 6% chance of being wrong. This is to say, we are 94% confident that the program made a difference and if we were to run the program again, the probability is 0.94 that the program would show positive results. Since we are not sure what type of risk others may take, it is very useful to report the p-value with the analysis so that someone else can understand what the chances are that the program had no effect. A p-value gives more useful information than just saying that the test statistic is not significant.

An example of poor reporting of program assessment comes from a proposal in which the authors had evaluated their program and were asking for additional funding. At the beginning of their proposal, they stated that they wished to show that the results of their program were "significant." That is, they stated up front that a p-value of 0.05 or less was hoped for, on which they would then conclude that their program made a difference. The authors then did the analysis and unfortunately, the p-value was over 0.05 (say p=0.08). Their conclusion was that the results of their program were not "significant," but that it was still a good program, since anyone could see that the post-program numbers were higher than the pre-program numbers, and that the program should be funded anyway. The authors were misguided into thinking that if a p-value is over 0.05, then that number reports that the program was not effective. If the authors had not emphasized the "importance" of a p-value under 0.05, had just calculated the p-value, and then reported it with the statement that they were 92% confident that the program was effective, they would have had a good proposal. The funders would then need to decide how sure they needed to be that the program was making a difference before they continued the funding.

IV. Is The Sample Size Large Enough?

For the data in our example, the p-value was 0.15798. We might be inclined to say that the program was effective; depending on what magnitude response increase we really wanted to see. We note that the sample is very small, n=6. It is difficult to make a decision based on only six pairs of data. The small sample size is taken into account by the statistical test and thus reflected in the p=0.16. If we had more evidence, say six more data points that replicated the first six, what would the p-value be? It is p=0.032! With six additional points with the same magnitudes of differences as the first six, the p-value, the probability that by chance the program really made no difference and still these differences were observed, has gone down considerably!

Note, however, that neither six, nor even 12, observations is a very large participant number and is not very close to the 20 to 30 observations that would normally be necessary for the t-test to be accurate, unless we knew that the underlying distribution of the d_i's was normal. In fact, if we plot the d_i's in our example, we will see that the plot is not in a triangle or bell-shape and therefore the t-test result may not be very accurate. What can we do when we only have six or 12 observations? Some people report that in such cases there are no appropriate statistical tests. We may not be able to run a test that depends on a normal distribution or a large (>30) sample, but there are other tests, generally called non-parametric tests, that make no assumption on the underlying distribution or on the sample size. We can use one of these tests to conclude if our program was effective. Since non-parametric tests make no assumptions on the data, the p-value will usually be larger than the p-value associated with the t-test, especially if the data is not normal. Most statistical software packages give the results of appropriate non-parametric tests when the sample size is small. For more information on non-parametric tests, see Reference 4 or 5. If we use the non-parametric sign test on our sample of six, the p-value is 0.617 and if we use a non-parametric signed rank test, that uses more information from the data, the p-value is 0.399.

If we use the sign test and the signed rank test on our sample of 12, the p-values are 0.289 and 0.166, respectively.

The conclusion to be drawn is that as much data should be gathered as can easily be gathered (at least 30 data points, if time and expense allow). A larger sample will give more information and a more accurate prediction as to whether the program or project has really been effective. A friendly industrial engineer consultant can also help determine the sample size after which gathering more data will give very little additional information. In many of the programs that need evaluating, the participant size may not be large; therefore as many of the participants as possible should be included. Although complete information may be obtained with one group through statistical testing, one can say with 1-p confidence that if the program were applied to another comparable group, the outcome would be favorable.

V. Which Program Made What Difference?

When more than one program impacts a group of students, it may be very difficult to determine which program had what effect. For example, suppose that we are interested in the one-year retention rate for entering freshmen. We make available several new programs, including expanded tutoring, the addition of a concept building class for some of the freshmen classes, and additional academic and social activities. By the next fall, we know that this class had a higher retention rate than classes that came before. However, how can we determine which programs had a positive impact and made a difference? In this case, we have changed several variables at the same time. The effects of the several variables and their interactions can be measured if we have a proper design of our experiment. Without a proper design, that is, waiting until the end of the year and then looking at the data to see if one can decipher any conclusions, may mean a lot of useless data. The data would be useless, for example, if we had no record of the programs in which each student had participated. This is the statistician's nightmare: to be given data and then asked to draw conclusions. It may be impossible to do so.

In a controlled experiment, we place some students in only one of each of the new programs. Other students are placed in two of the programs, and if there are three variables (new programs) being evaluated, then some students are placed in all three programs. In this way, the effects and interactions of the programs can be determined at the end of the measurement time. Obviously this type of data collection takes up front planning. There may be other students in the programs, whose program effect will not be used in the data analysis. A trained statistician can easily help with the set-up and analysis of such an experiment.

Why not just vary one variable at a time, so there is no confusion on which new program had an effect? If only one program is being evaluated, then this is the way to go. However, if one wants to know the effect of more than one program on retention, higher grades, or creating interest in majoring in engineering, then changing only one variable at a time may take a long time and will not give the full program effects if there is any interaction between the programs. For example, it may be true that students who take advantage of individual tutoring do much better if they also participate in a concept building class.

A designed experiment is a test or series of tests in which purposeful changes are made to the input variables of a process (new programs in our case) so that we may observe and identify corresponding changes in the output response (such as retention).[6] For example, if we were testing the effect of two retention programs, A and B, we could calculate the effect of program A by itself, program B by itself, and of programs A and B together, if we have the proper design and data collection. This means that we would have data from students who were not involved in either program, some that were in program A only, some in program B only, and some students in both of the retention programs, A and B. Since it is usually very difficult to control the placing of students in certain programs, the evaluation could be done retrospectively if good records were kept on which students attended which program.

VI. Conclusions

While qualitative data, including anecdotes, can be useful in evaluating a program, funders and/or decision makers on the future of a program usually want quantitative evidence that the program made a difference. If the program did not make a difference, then other efforts need to be tried. Both formative and summative evaluation needs to be planned before a program commences so that the proper data can be obtained. It is especially important to have baseline data so that the results due to a program can be compared with the situation before the program took effect. Before and after data need to be identified carefully so the comparison of the program effect with the pre-program situation can be made on the same person. Usually some of the faculty in an Industrial Engineering Department can assist in planning the evaluation to make sure that the correct data is collected, as well as to assist in the analysis of the data and the making of conclusions.

P-values are useful in describing the importance of the effect of a program and in giving a more accurate description of the probability that the program did have an effect. To say that there is a 90% ($p=0.10$) chance that a program is effective gives more information than simply saying that the results of the program were not significant ($p>.05$). Statistical analysis can be done on small samples: nonparametric statistics are designed to deal with small samples from unknown distributions. With small samples, care needs to be given that the sample distribution meets the requirements of the statistical test used. However, if larger samples are possible, within the constraints of time and money, then larger samples should be used since they can give a more accurate picture of the whole population. When several programs are introduced at the same time, a careful evaluation needs to be made in order to calculate the effect of any one program or the interactions of the programs. Again, a statistician can help plan to make this type of evaluation.

Quantitative evaluation with statistical testing is a powerful way to assess a program. If the program evaluator is not trained as a statistician, assistance should be sought at the beginning of a program or project from someone who is a statistician to make sure that the proper data will be collected and that it will be analyzed correctly. Also, remember, statisticians love to do this type of work!

References

1. "User-friendly Handbook for Project Evaluation: Science, Mathematics, Engineering and Technology Education," National Science Foundation, Directorate for Education and Human Resources and Division of Research, Evaluation and Communication, NSF 93-152 (Reprinted 6/97).
2. Walpole, Ronald E., and Myers, Raymond H., Probability and Statistics for Engineers and Scientists, Third Edition, Macmillan Publishing Company, New York, 1985.
3. Secola, Patricia M.; Smiley, Bettie A.; Anderson-Rowland, Mary R.; and Baker, Dale R., "Evaluating the effectiveness of Gender Equity Training in Engineering Summer Workshops with Pre-College Teachers and Counselors," 2001 Proceedings of American Society for Engineering Education Annual Conference, Albuquerque, New Mexico, June 2001, CD Rom, Session 1692, 14 pages.
4. Daniel, Wayne W., Applied Nonparametric Statistics, Second Edition, Duxbury, 1989.
5. Conover, W. J.; O'Sullivan, Mary; and Wiley, Brad, Editors, Practical Nonparametric Statistics, Third Edition, John Wiley & Sons, 1998.
6. Montgomery, Douglas C., Introduction to Statistical Quality Control, Third Edition, John Wiley & Sons, Inc., New York, 1996.

Biography

MARY R. ANDERSON-ROWLAND
Mary R. Anderson-Rowland is the Associate Dean of Student Affairs in the CEAS at ASU. She received her PhD in statistics from the U. of Iowa. Her awards include the YWCA Tribute to Women 2001 Award (Scientist/ Researcher) and ASEE Fellow in 2001. A frequent speaker on the career opportunities in engineering, especially for women and minority students, she is a faculty member in Industrial Engineering and does statistical consulting.