Principles of Survey Research
Part 6: Data Analysis

Barbara Kitchenham
Dept. Computer Science
Keele University, Staffs, UK
barbara@cs.keele.ac.uk

Shari Lawrence Pfleeger
RAND Corporation
Arlington VA 22202-5050
shari_pfleeger@rand.org

Abstract

This article is the last of our series of articles on survey research. In it, we discuss how to analyze survey data. We provide examples of correct and incorrect analysis techniques used in software engineering surveys.

Keywords: survey methods, statistical analysis

Introduction

In this article, we assume that you have designed and administered your survey, and now you are ready to analyze the data you have collected. If you have designed your survey properly, you should already have identified the main analysis procedures. Furthermore, if you have undertaken any pre-tests or pilot studies, you should already have tested the analysis procedures.

In this article, we discuss some general issues involved in analyzing survey data. However, we cannot describe in detail how to analyze all types of survey data in a short article, so we concentrate on discussing some of the most common analysis errors and how to avoid them.

As with previous articles in this series, we use three existing software engineering surveys to illustrate common errors:

1. Two related surveys undertaken by Timothy Lethbridge [4] and [5], both aiming to compare what software engineers learned at university with what they needed to know in their current jobs.
2. A survey we ourselves undertook, to investigate what evidence organizations use to assist in making technology adoption decisions.
3. A Finnish survey [9] aimed at investigating project risk and risk management strategies.

Data Validation

Before undertaking any detailed analysis, responses should be vetted for consistency and completeness. It is important to have a policy for handling inconsistent and or incomplete questionnaires. If we find that most respondents answered all questions, we may decide to reject incomplete questionnaires. However, we must investigate the characteristics of rejected questionnaires in the same way that we investigate non-response to ensure that we do not introduce any systematic bias. Alternatively, we may find that most respondents have omitted a few specific questions. In this case, it is more appropriate to remove those questions from the analysis but keep responses to the other questions.

Sometimes we can use all the questionnaires, even when some are incomplete. In this case, we have different sample sizes for each question we analyze, and we must remember to report the actual sample size for each sample statistic. This approach is suitable for analyses such as calculating sample statistics or comparing mean values, but not for correlation or regression studies. Whenever analysis involves two or more questions at the same time, we need an agreed procedure for handling missing values.

For example, suppose we ask respondents their educational background in one question and their opinion about software quality in another. Suppose further that there are inadmissible or incomplete responses for some respondents; for instance, a respondent may leave out an educational background choice (incomplete) or check two categories (inadmissible). We can report measures of central tendency (mean, median, mode) for each of the two questions, but the sample sizes are likely to be different. On the other hand, if we want to investigate the relationship between educational background and opinion on software quality, we must consider the issue of missing values.

In some cases, it is possible to use statistical techniques to "impute" the values of missing data [7]. However, such techniques are usually inappropriate when the amount of missing data is excessive and/or the values are categorical rather than numerical.

It is important to reduce the chance of incomplete questionnaires when we design and test our instruments. A very strong justification for pilot surveys is that misleading questions and/or poor instructions may be detected before the main survey takes place.

The questionnaire related to our technology adoption survey (shown in Appendix 1 in Part 3 of this series) suffered badly in terms of incomplete answers. A review of the instructions to respondents made it clear why this had happened. The instructions said

> "If you are not sure or don't know an answer just leave the line blank; otherwise it is important to answer YES or NO to the first section of every Technique/Technology section."

With these instructions, perhaps it is not surprising that most of the questionnaires had missing values. However, replies were not just incomplete; they were also inconsistent. For example, some respondents left blank question 1 ("Did your company evaluate this technology?") while replying YES to question 2, about the type of evaluation undertaken. Thus, blanks did not just mean "Don't know"; sometimes they also meant YES. Ambiguities of this sort make data analysis extremely difficult and the results dubious.

Partitioning the responses

We often need to partition our responses into more homogeneous sub-groups before analysis. Partitioning is usually done on the basis of demographic information. We may want to compare the responses obtained from different subgroups or simply report the

results for different subgroup separately. In some cases, partitioning can be used to alleviate some initial design errors. Partitioning the responses is related to data validation since it may lead to some replies being omitted from the analysis.

For example, we noted in Part 5 of this series that Lethbridge did not exclude graduates from non-IT related subjects from his population; neither did he exclude people who graduated many years previously. However, he knew a considerable amount about his respondents, because he obtained demographic information from them. In his first paper, he reported that 50% of the respondents had degrees in computer science or software engineering, 30% had degrees in computer engineering or electrical engineering, and 20% had degrees in other disciplines. He also noted that the average time since the first degree was awarded was 11.7 years and 9.6 years since the last degree. Thus, he was in a position to partition the replies and concentrate his analysis on recent IT graduates. However, since he did not partition his data, his results are extremely difficult to interpret.

Data Coding

It is sometimes necessary to convert nominal and ordinal scale data from category names to numerical scores prior to the data's being input into electronic data files. This translation is not intended to permit nominal and ordinal scale data to be analyzed as if they were simple numerical values. Rather, it is done because many statistical packages cannot handle categories represented by character strings. In many cases, codes are put into the questionnaire along with category names, so coding is done during questionnaire design rather than during data analysis.

A more difficult coding problem arises for open questions. In this case, response categories need to be constructed after the questionnaires have been returned. It requires human expertise to identify whether two different answers are equivalent or not. In such cases, it is wise to ask several different people to code replies and compare the results, so that bias is not introduced by the categorization.

Standard Data Analysis

This discussion of data analysis assumes we have undertaken probability sampling. If we do not have a probability sample, we can calculate various statistics associated with the data we have collected, but we cannot estimate population statistics.

The specific data analysis you need depends on the survey design and the scale type of replies (nominal ordinal, interval, ratio, etc.). The most common population statistics for numerical values are:

Population Total: $X_T = \sum_{i=1}^{N} X_i$

Population Mean: $\hat{X} = (\sum_{i=1}^{N} X_i) / N$

Population Variance: $\sigma_x^2 = \sum (X_i - \hat{X})^2 / N$

where N is the population size.

For dichotomous (Yes/No or 0/1) variables, the most common population statistics are:

Proportion: $P_Y = (\sum_{i=1}^{N} Y_i) / N$

Variance of a proportion: $\sigma_Y^2 = P_Y (1 - P_Y)$

where Y_i is a dichotomous variable taking the value 1 or 0 and N is the population size.

It we have *a random sample* of size n taken from a population of size N, we can estimate the population statistics from our sample as follows:

Total: $\hat{X}_T = \left(\frac{N \sum_{i=1}^{n} x_i}{n} \right)$

Mean: $\bar{x} = \left(\sum_{i=1}^{n} x_i \right) / n$

Variance: $\hat{\sigma}_x^2 = \left(\frac{N-1}{N} \right) \left(s_x^2 \right)$

where: $s_x^2 = \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right) / (n-1)$

The standard error of the estimate of the total is: $N \left(\frac{s_x}{n} \right) \sqrt{\frac{N-n}{N}}$

The standard error of the estimate of the mean is: $\left(\frac{s_x}{n} \right) \sqrt{\frac{N-n}{N}}$

For a proportion $\hat{P}_Y = \frac{\sum_{i=1}^{n} y_i}{n}$ and the standard error of the estimate of the proportion is: $\sqrt{\frac{N-n}{N}} \sqrt{\frac{\hat{P}_y (1 - \hat{P}_y)}{n-1}}$

These may not be the formulas you might have expected from reading a basic book on statistics and data analysis. The standard errors include the term $\sqrt{(N-n)/N}$ which is referred to as the *finite population correction* (fpc) (see, for example, [6]). The fpc can be re-written as $\sqrt{1-(n/N)}$, from which we can see that as N tends to infinity, the fpc approaches 1 and the standard errors formulas approach the usual formulas. If $n = N$, the standard error terms are zero because the mean, total and proportion values are known and therefore not subject to error.

If you are using a statistical package to analyze your data, you need to check whether it allows population estimates of finite population statistics to be calculated correctly. For example, Levy and Lemeshow [6] give examples of the commands available in the STATA statistical package for analyzing survey data.

Note. The equation we gave for determining sample size (in Part 5 of this series) ignored the fpc, and if used as-is will therefore result in an over-estimate of the required sample size. However, it

is better to have too many in a sample than too few.

Analyzing Ordinal and Nominal Data

Analyzing numerical data is relatively straightforward. However, there are additional problems if your data is ordinal or nominal.

Ordinal Data

A large number of surveys ask people to respond to questions on an ordinal scale, such a five-point agreement scale. For example, respondents are asked to specify the extent to which they agree with a particular statement. They are offered the choice of: strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree. The Finnish survey and Lethbridge's survey both requested answers of this sort. It is common practice to convert the ordinal scale to its numerical equivalent (e.g. the numbers 1 to 5) and to analyze the data as if they were simple numerical data. There are occasions when this approach is reasonable, but it violates the mathematical rules for analyzing ordinal data. Using a conversion from ordinal to numerical entails a risk that subsequent analysis will give misleading results.

Figure 1 represents a survey with three questions, each on a 5-point ordinal scale (labeled SP1, SP2, SP3, SP4, SP5), where we have 100 respondents. Figure 1 shows the number of responses assigned to each scale point; for example, for question 1, respondents chose SP1 10 times, SP2 20 times, SP3 40 times, SP4 20 times and SP5 10 times. If we convert the scale points to their numerical equivalents (1,…,5) and determine the mean value, we find the mean is 3 for all three questions. However, we cannot conclude that all the responses are equivalent. In the case of question 1, we have a symmetric single-peaked distribution. This may be regarded as an approximately Normal distribution with a mean value of 3. In the case of question 2, we have a bimodal distribution. For bimodal distributions, the data are not Normal. Furthermore, there is no central tendency, so the mean is not a useful statistic. In the case of the third question, we have an equal number of responses in each category, typical of a uniform distribution. A uniform distribution has no central tendency, so again the concept of a mean is not useful.

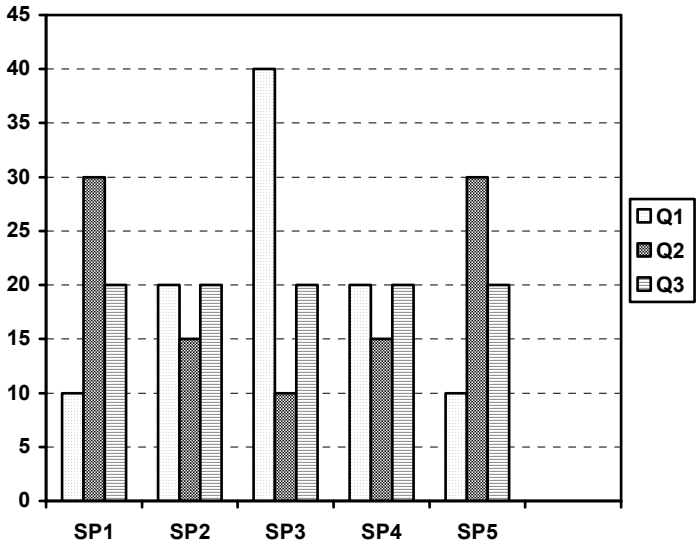

Figure 1 Responses to three five-point ordinal scale questions

In general, if our data are single peaked and approximately Normal, our risks of misanalysis are low if we convert to numerical values. However, we should also consider whether such a conversion is necessary. There are three approaches that can be used if we want to avoid scale violations:

1. We can use the properties of the multinomial distribution to estimate the proportion of the population in each category and then determine the standard error of the estimate. For example, Moses uses a Bayesian probability model of the multinomial distribution to assess the consistency of subjective ratings of ordinal scale cohesion measures [8].
2. We may be able to convert an ordinal scale to a dichotomous variable. For example, if we are interested in comparing whether the proportion who agree or strongly agree is greater in one group than another, we can re-code our responses into a dichotomous variable (for example, we can code "strongly agree" or "agree" as 1 and all other responses as 0) and use the properties of the binomial distribution. This technique is also useful if we want to assess the impact of other variables on an ordinal scale variable. If we can convert to a dichotomous scale, we can use logistic regression.
3. We can use Spearman's rank correlation or Kendall's tau [11] to measure association among ordinal scale variables.

There are two occasions where there is no real alternative to scale violations:

1. If we want to assess the reliability of our survey instrument using Cronbach's alpha statistic [1].
2. If we want to add together ordinal scale measures of related variables to give overall scores for a concept.

However, in both cases, if we do not have an approximately Normal response, the results of analyzing the data may be misleading.

We believe it is important to understand the scale type of our data and analyze it appropriately. Thus, we do not agree with Lethbridge's request for respondents to interpolate between his scale points as they saw fit (i.e. to give a reply of 3.4 if they wanted to).

Nominal Data

The most common form of analysis applied to nominal data is to determine the proportion of responses in each category. Thus, unless there are only two categories, there is no choice except to use the properties of the multinomial (or possibly the hypergeometric) distribution to determine standard errors for the proportions. However, it is still possible to use multi-way tables and chi-squared tests to measure associations among nominal scale variables (see [11], Section 9.1).

Questionnaire Size and Multiple Tests

In an earlier article, we pointed out that respondents do not like to answer excessively long questionnaires. It must also be noted that statisticians don't like analyzing excessively long questionnaires either. Unlike respondents, the problem is not one of simple fatigue; it is one of methodology. It is important to realize that the more tests we perform, the more likely we are to find spurious results. For example, if we have an alpha level of 0.05, we have a 5% chance of falsely detecting a significant difference in a data set. Thus, if we perform 50 tests, the binomial distribution indicates that we can expect 2.5 ±4.7 spurious statistically significant results. This problem is especially pervasive when researchers dig for significance, administering test after test until a significant

difference is found. Digging is particularly common among graduate students seeking something important for their masters or doctoral theses; no one likes to report a result claiming no difference between two techniques.

However, there are alternatives to applying a plethora of tests. One method for dealing with the results of multiple tests on the same data set is to adjust the significance level for individual tests to achieve a required overall level of significance, as described by Rosenberger [8] or Keppel [2]. For example, if you perform ten independent tests and require an overall significance level of 0.05, the *Bonferroni adjustment* requires a significance level of 0.005 for each individual test. Rosenberger describes other, less severe approaches, but each still requires much higher levels of significance for individual tests than the customary 0.05 in order to achieve an overall significance level of 0.05.

An alternative to adjusting significance levels is to report the number of results that are likely to have occurred by chance given the number of tests performed. What should be avoided is reporting only positive results with no indication of the number of tests performed. For example, Ropponen and Lyytinen [9] reported 38 significant correlations but did not report how many correlation coefficients they tested.

Final thoughts

Throughout this series of articles, we have discussed the lessons we learned in designing and administering a survey. We used our own work, plus the reported techniques in two other surveys, as examples of the dos and don'ts in survey research. In many cases, we have criticized the approaches taken by the researchers; we hope you realize that our criticism was of general and commonly-observed mistakes, not of the individual researchers and their abilities. We recognize that sometimes, as with aspects of our own survey, some errors occur because of issues beyond our control or simply beyond our ken. Thus, we end this series as we began it, by making more visible several ways to address the most critical issues that must be improved if the reported surveys were replicated today.

Consider first our own survey of technology adoption. The survey published in *Applied Software Development* was somewhat premature. The specific goals of the survey are not clear, and neither is the target population. We believe the best approach would have been to form a focus group to discuss the specific goals and research questions we should address, and to consider whether a self-assessment questionnaire was the right approach.

Lethbridge's survey [4] and [5] would have been better focused if it had been organized by a university or a company. A university could have surveyed its own graduates, giving it a clear target population to sample. A company can survey its new hires in the context of its own hiring and training policies. The survey instrument would also have benefited from reducing the number of questions.

Ropponen and Lyytinen's survey was generally good methodologically. Some methodological improvements might be to perform a proper pilot study to assess reliability independently of the survey proper, and to address the problem of multiple tests. Another potential problem with the Finnish study is that principal component analysis can be used either to test a hypothesis or to generate hypotheses. However, it cannot do both things at once. Thus, the risk factors identified using principal component analysis represent hypotheses/theories that should have been confirmed by an independent study before investigating risk strategies. The underlying problem in this study (and many others) was trying to do too much in one investigation.

We plan to continue our work in examining existing studies and providing guidelines for improvement. For example, *IEEE Transactions on Software Engineering* has accepted for publication a set of guidelines we have developed on what to keep in mind when performing or evaluating empirical research in software engineering [3]. Our goal is to assist the software engineering community in understanding key issues in empirical software engineering research, and to make us more effective in using such research to further our knowledge about practices and products.

References

[1] L. J. Cronbach, Coefficient alpha and internal structure of tests, *Psychometrika*, 16(2), 1951, pp. 297-334.

[2] G. Keppel. *Design and Analysis: A Researcher's Handbook*, third edition, Prentice Hall, 1991.

[3] Barbara A. Kitchenham, Shari Lawrence Pfleeger, Lesley M. Pickard, Peter W. Jones, David C. Hoaglin, Khaled El Emam, and Jarrett Rosenberg. Preliminary guidelines for empirical research in software engineering, *IEEE Transactions on Software Engineering*. Accepted for publication

[4] Timothy Lethbridge, A survey of the relevance of computer science and software engineering education, *Proceedings of the 11th International Conference on Software Engineering*, IEEE Computer Society Press, 1998.

[5] Timothy Lethbridge, What knowledge is important to a software professional, *IEEE Computer*, May 2000.

[6] P. S. Levy and S. Lemeshow, *Sampling and Populations*, Third Edition, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, 1999.

[7] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley, New York, 1987.

[8] J. Moses, Bayesian probability distributions for assessing measurement of subjective software attributes, *Information and Software Technology*, 42(8), 2000, pp 533-546.

[9] J. Ropponen and K. Lyytinen, Components of software development risk: How to address them. A project manager survey, *IEEE Transactions on Software Engineering* 26(2), February 2000.

[10] W. F. Rosenberger, Dealing with multiplicities in pharmacoepidemiologic studies, *Pharmacoepidemiology and Drug Safety*, 5, 1996, pp. 95-100.

[11] S. Siegel and N. J. Castellan, *Nonparametric Statistics for the Behavioral Sciences*, 2nd Edition, McGraw-Hill Book Company, N.Y., 1998.

Acknowledgements

Our thanks to Alberto Sampaio for his helpful comments on earlier drafts of all the articles in this series.