THE USE OF STATISTICS IN FINAL YEAR APPLIED RESEARCH PROJECTS

Loi Soh Loi and Wu Yuan, Nanyang Technological University, Singapore

An Applied Research Project (ARP) is a requirement for the first degree in Nanyang Business School, Nanyang Technological University (NTU) in Singapore. A survey was taken of statistical techniques used by business undergraduates in 196 ARP for the academic year 1993/94, 225 ARP for 1994/95, and a random sample of 40 ARP for 1995/96. More than 63% used statistical methods. The use of statistics is becoming substantial. However, the reliability and validity of the research finding and the quality of the statistical techniques used in many projects can be improved. After giving a general picture of the survey and the common statistical techniques used in the ARP, this paper points out the common unaware errors and misuses in these projects. Some recommendations and guidelines as references for future students.

INTRODUCTION

Final year undergraduate students are required to undertake the applied research projects (ARP) as part of the requirement for the first degree in Nanyang Business School, Nanyang Technological University (NTU) in Singapore. The ARP take the form of a field study, a case study, a literature review, a survey or an empirical study of any field that has potentially significant applications in business. Although the ARP are not necessarily the statistical projects, statistics plays an important part of it, especially for the survey type or an empirical study project. Because statistics is widely used by students to establish conclusions, it is crucial that statistical techniques be employed to the data concerned. Any misunderstanding of statistics might lead to the misuse of statistical methods resulting in wrong conclusions. If this happens, it may be disastrous for the user in terms of the validity of his/her analytical reports.

A survey of statistical techniques used in the ARP by our students in Nanyang Business School for the academic year 1992/93 was carried out by Loi and Lian (1993). The study shows that more than 70% of the ARP involved statistical analysis. It also shows an enormous diversity of statistical methods used. However, the reliability and validity of the research findings and the quality of the statistical techniques used in many ARP can be improved. Incidentally, even in good international journals, many of the papers subsequently examined by researchers have doubtful statistical results. For example, Schor and Karten (1966) reviewed 295 papers published in ten medical journals

and concluded that only 28% of the papers were statistically acceptable, 68% were deficient, and 5% were “un-salvageable”.

In the survey done by Loi and Lian (1993) for the academic year 1992/93, the focus was on the presentation of data, especially graphical techniques. In this paper, we devote our attention to the data analysis. Firstly, we proceed to identify the common statistical techniques employed by students. Then, from these common statistical techniques, we point out some of the common misuses and errors. We hope that by pointing out these misuses and errors, future students will avoid making the same mistakes.

COMMON STATISTICAL TECHNIQUES USED IN THE PROJECTS

In our survey, we examined only these projects done by final year business students. There were a total of 196 ARP done by business students in the academic year 1993/94 and 225 ARP done in 1994/95. All ARP in the academic year 1993/94 and 1994/95 were examined, and a random sample of 40 ARP for the academic year 1995/95 was reviewed. Tables 1 and 2 give a summary of our findings.

Table 1: Number of statistical techniques used in the projects

Year	1993/94	1994/95	1995/96
Projects with some statistical techniques	125	145	32
Total	196	225	40
%	63.7	64.4	72.7

Table 1 shows that 63.7% of the ARP in the academic 1993/94, 64.4% of the ARP in 1994/95 and 72.7% of the ARP in 1995/96 used statistical techniques. Among those projects that used some statistical techniques, it can been seen from Table 2, most employed summary table for providing the descriptive statistics and graphics for presentation. The *t*-test, regression analysis, ANOVA and chi-squared test were commonly used for data analysis in the ARP.

ERRORS IN COMMON STATISTICAL TECHNIQUES

Most of the projects used the *t*-test, ANOVA, chi-squared test, or regression analysis for conducting the hypothesis testing and making comparisons. Various unaware errors or weaknesses can be summarised as follows:

Table 2: Percentage of types of statistical techniques used

Year	1993-94	1994-95	1995-96
Sample size	125	145	32
Descriptive statistics			
Summary table	56	65	72
Graphic: Pie chart	27	31	43
Bar chart	29	48	47
Line chart	18	23	38
Inferential statistics			
Simple regression	6	7	6
Multiple regression	11	14	13
ANOVA	10	10	19
Student's *t*-test	22	18	34
Chi-squared test	8	10	28
Pearson's correlation	10	4	6
Factor Analysis	4	5	0
Non-parametric	4	3	0
Others	11	11	13

Note: The totals exceed 100% as multiple selections were allowed.

t-tests

The *t*-test was used widely to compare two groups of measurements. The problems usually relate to the data not complying with the underlying statistical assumption. For example, in some projects, students collected sets of data of ordinal/nominal scale from the response on certain qualitative factors, and the sample size was small. Then the *t*-test was used to compare two groups of data. Students were not aware of the appropriate situation to use the *t*-test.

Regression

Regression was widely used to deal with the problems of prediction and estimation. Multiple regression was the most popular techniques because it is realistic to use more than one independent variables relating to the dependent variable. However, the following are some examples of common misuse of regression analysis.

- Did not study the nature of the data and justify if the model is appropriate and realistic for the data. For example, a straight line was fitted where the data show curvature in simple regression. The dependent variable Y is not normally distributed with the same variance for each value of the dependent variable X in multiple regression.
- Predicting the *Y* variable for values of the *X* variable outside the range of the original data set. For example, used data given in 1990-1995 to predict year 2000.
- Interpret the linear correlation as a measure of causal relationship. For example, used regression model in finding the effects of working overtime on the stress level experienced by employees and concluded that "increased hours of overtime led to higher stress levels".
- Not using dummy variables for nominal variables in the multiple regression model. For example, to estimate the value of residential properties by the location of the property, the type of land title, the number of bedrooms and the land area. Dummy variables should be used for the quantitative variables like the location of the property and the type of land title.
- Misinterpreting the results from the printout, say the estimated coefficients and beta coefficients.

Chi-squared test

Generally, students showed a clear understanding of the usage of this test, but some did not know how to state their hypotheses or how to state them clearly. Consequently, they may have made wrong decisions. Another minor mistake is that the expected frequencies or the observations were too few, but yet conclusions were made.

ANOVA

The common error in ANOVA is forgetting to check the underlying assumptions when using it. All the sampled projects that used the ANOVA failed to state and check these assumptions before proceeding. Another error is used when data is in nominal or ordinal scales.

RECOMMENDATIONS

In this section, we attempt to give some guidelines and recommendations as references for future students to refer before embarking on using statistics for the ARP.

Population and Sample

In our studies, we found that a large proportion of students neither stated the target population for their study nor described how their samples were selected randomly. Rarely can we collect data on all the subjects of interest in a particular study. Samples provide a practical and efficient means to collect data. The sample serves as a model of the population. However, for us to extend our findings to the population, the model must be an accurate representation of the population; that is, the sample should be a random sample.

Also, there were many students who gave little thought to sample size, choosing the most convenient number (say, 20, 50, 100, etc.) for the size of the study. Sample size is the most potent method of achieving estimates that are sufficiently precise and reliable for scientific inquiry. Small sample size may contribute to a conservative bias (Type II error) in the application of a statistical test. Increasing sample size obviously has a cost so choice of a sample size cannot be considered in a vacuum. A trade-off in cost, total error, and other design choices must be considered.

Analysis

The basic principle to be adhered to with respect to the statistical aspects of the ARP is that the methods should be described in sufficient detail to be fully understood, and so that anyone else with access to the raw data could, if desired, reproduce the results. The use of unusual forms of analysis should be justified, preferably with a reference, but all analyses should be clearly described. One should describe clearly and exactly what was done. It may be necessary to demonstrate the validity of the assumptions for some analyses, say t-test, ANOVA and regression analysis.

Statistical Methods

Here are a few points to note when using any statistical technique. They can be applied to all methods and students should clearly state these points in their research. They are :

- Assumptions must be clearly stated if there are any.

- Both the null and alternate hypotheses should be stated and explained if possible.
- The level of significance must be specified
- The critical value should be included in the report. This value separates the rejection and non-rejection regions. The purpose of doing this is to help the reader understand why the author arrived at the particular conclusion for the statistical test.

The nature of data determines the different types of tests. In general, for interval/ratio data, parametric tests should be used. For ordinal/nominal data, non-parametric tests should be used.

CONCLUSIONS

After reviewing the entire set of ARP done for academic years 1993/94 and 1994/95 and the sample of 40 ARP for 1995/96, we found that more than 63% of the projects used statistical techniques. A large proportion of these projects had more or less statistical error or misuses. Some of them had errors in which the students applied inappropriate statistical techniques to the data collected. Some did not present their findings and conclusions in an easily understandable form for the reader. The incorrect analysis of data is probably the worst misuse of statistical methods. The mishandling of statistical analysis is as bad as the misuse of any laboratory technique. It is of no value using good statistical techniques to analyze “poor” data (seriously biased), nor is analyzing “good” data with inadequate or invalid statistical techniques.

There needs to be a greater appreciation of the importance of correct statistical thinking and an improvement in the standard of the ARP so that the errors discussed can be eradicated. We hope our findings and discussion will help future undergraduates who are embarking on their ARP to improve the quality of their ARP.

REFERENCES

Schor, S. and Karten, I. (1966). Statistical Evaluation of Medical Journal Manuscripts. *J. Am. Med. Ass., 195,* 1123-8.

Loi, S. L. and Lian, R. (1993). Statistical Techniques Used in Final Year Project. *Working Paper Series No. 33-93*, School of Accountancy and Business, Nanyang Technological University, Singapore.