Chapter 2

Methodology

The methodology for this study comprises the following components:

(a) study design and sampling technique,
(b) variables,
(c) data collection and management,
(d) graphical and statistical methods.

Study Design and Sampling Technique

A cross-sectional study design is used for this study. The sample was taken randomly from the target population without regard to the outcome or determinants. The target population comprised graduate students from Prince of Songkla University in Thailand. In this study, the subjects who had graduated during the academic year 1993 – 1997 were selected from this population. The reason for this selection is that they comprise the most complete data set. The sample contains three faculties in Pattani Campus, including Education, Humanities and Social Sciences, and Science and Technology, and seven faculties in Hat Yai Campus, including Engineering, Science, Management Science, Natural Resources, Nursing, Environmental Management, and Agro-Industry.

The sample size needed to obtain a specified accuracy can be calculated by using the following formula (McNeil, et al, 1998 : 119)

$$n = \left(1.96\frac{\sigma}{\delta}\right)^2$$

where σ is the population standard deviation of the outcomes, and δ is half of the width of the 95% confidence interval for the mean of the population (assumed normally distributed). In this study, the value of σ is unknown. However, in practice some value of σ can be assumed based on previous research, and evidence suggests

that the standard deviation of grade point averages of graduate students is close to 0.27 (Choochom and Sukharom, 1988 : 45). Taking δ to be 0.03 (or 3%), we obtain

$$n = \left(1.96\frac{0.27}{0.03}\right)^2$$

$$= 311.17$$

$$\cong 312.$$

So a sample size of 312 is needed. The sample should be representative of the population. In this study, the stratified random sampling method is used to get a sample. Thus, the following two general principles are used to obtain the sample for the study;

1) the sample should be unbiased, that is, it should fairly represent the population from which it is drawn;

2) the sample must be sufficiently large to ensure that a reasonably accurate inference can be made.

The samples are grouped by faculty, so that the items in each group are more nearly alike than are the items in the population as a whole. A sample is drawn from each group by simple random sampling. The samples are combined to form all groups, as shown in Table 2.1. Ensuring that these sample sizes are similar improves the effrecency of the comparison of results between faculties.

Table 2.1: The sample size for each faculty

Faculty	Sample
Engineering	22
Science	30
Pharmaceutical	18
Nursing	26
Management science	39
Environmental management	38
Natural resources	35
Agro-industry	31
Education	36
Humanities & social science	30
Science & technology	7
Total	312

Variables

The variables of interested for this study comprise 15 determinants and the outcome. The details of all variables is as follows:

Table 2.2: Determinant and outcome variables for the study

Variables	Definition
Determinant	
Gender	male and female
Age	age at enrolled in Master's degree
Marital status	marital status at enrolled in Master's degree
Domicile	province which they lived before enrolled in Master's degree
Occupation	occupation at enrolled in Master's degree
Work experience	duration of working in their before enrolling in Master's degree

Table 2.2: Determinant and outcome variables for the study (ctd.)

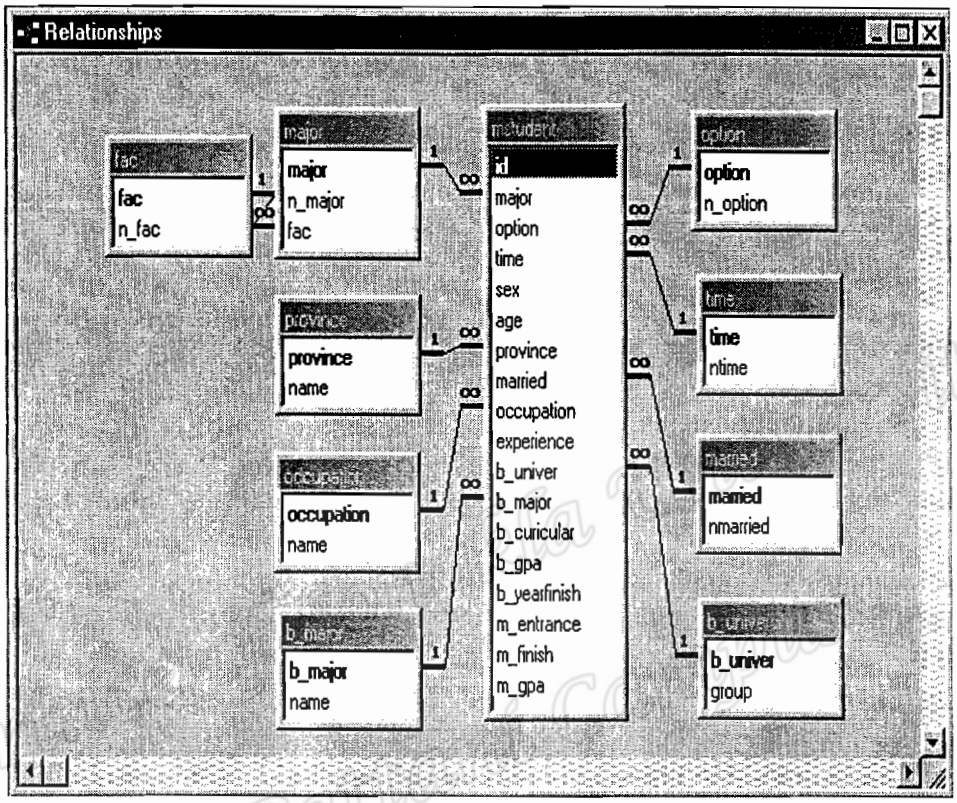
Variables	Definition
Faculty	faculty which finished in Master's degree
Study plan	plan A (requiring a thesis) and plan B (not requiring a thesis)
Type of study	full time and part time
BA University	the university which studied in bachelor's degree level
BA major	the major which studied in bachelor's degree level
BA program	two-year and four-year bachelor's program
BA GPA	grade point average (GPA) at the bachelor's degree
Duration BA-MA	duration of period from finishing bachelor's degree until enrolling in the Master's degree
Total years of study	the total number of years of study in Master's degree
Outcome	
Academic achievement	the grade point average (GPA) of the graduate students

Data Collection and Management

The data were collected from two divisions of Prince of Songkla University. First, the grade point average of the graduate students was taken from the registration units at Pattani Campus and Hat Yai Campus. Other information was obtained from the application form from the Graduate School in both campuses.

The data were entered into a file called *psu.mdb* using Microsoft Access. There are ten tables in the database *psu.mdb*, as shown the relationship diagram in Figure 2.1.

Figure 2.1: Relationships between tables in database



Graphical and statistical methods

The graphical methods are presented and obtained by using Matlab program version 5 (D Hanselman and B Littlefeeld, 1997) and Asp (McNeil, 1998), as follows.

1. Histograms and statistics of raw data for all variables.
2. Two-sample *t*-test and one-way analysis of variance of the variables described by box plots and 95% confidence intervals of means.

The statistical methods used to analyze the data described as follow:

1. Two-sample t-test

The null hypothesis of two population means are equal can be expressed as

$$H_0 : \mu_1 = \mu_2$$

where μ_1 and μ_2 represent the population means. If samples of size n_1 and n_2 are taken from the two population means, giving sample means $\bar{y}_1$ and $\bar{y}_2$, a t statistic may be used to test this null hypothesis. The two-sample t statistic is

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where s is the pooled sample standard deviation, given by

$$s = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}}$$

and n_1 and n_2 are the size of the sample and s_1 and s_2 the standard deviations of the two samples. There are two assumptions for the two-sample t test, as follows:

1) the two populations from which the samples are drawn have the same spread,

2) the two populations are normally distributed.

Box plots can be graphed to check that the population standard deviations could be the same in the two populations, and normal scores plots are used to assess the normality assumption.

2. One-way Analysis of Variance

One-way analysis of variance is the method used for the analysis of data in which the outcome is continuous and the determinant is categorical. There are two assumptions for this method, namely (1) the standard deviations of the populations are equal, and (2) the populations are normally distributed (McNeil, 1996 : 75).

As in the case of one-way analysis of variance, we can assess the statistical assumptions by first looking at the box plots of the adjusted data. If the spreads of the box plots are similar, the first assumption is reasonable. Turning to the second assumption, it may be checked by graphing the residuals against normal scores and seeing if these data could be linear, so the second assumption is reasonable. The null hypothesis states that the samples have arisen from the same population, so the null hypothesis may be stated as

$$H_0 : \mu_1 = \mu_2 = ... = \mu_c$$

where c is the number of groups and μ_j is the population mean corresponding to group j. This null hypothesis can be tested by computing a statistic called the *F-statistic* and comparing it with an appropriate distribution to get a p-value. Suppose that there n_j observations in sample j, denote by y_{ij} for $i = 1, 2, ..., n_j$. The F-statistic is defined as

$$F = \frac{(S_0 - S_1)/(c-1)}{S_1/(n-c)}$$

where

$$S_0 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2 \, , \; S_1 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$$

and

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} \, , \; \bar{y}_j = \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n_j} y_{ij} \, , \quad n = \sum_{j=1}^{c} n_j$$

where S_0 is the sum of squares of the data after subtracting their overall mean, while S_1 is the sum of squares of the residuals obtain by subtracting each the sample mean. If the population means are the same the numerator and the denominator in the F-statistic are independent estimates of the square of the population standard deviation and the p-value is the area in the tail of the F-distribution with $c-1$ and $n-c$ degrees of freedom. However the F-statistic is based on an assumption that the populations have a common standard deviation.

3. Correlation Coefficient

The correlation coefficient is a measure of the linear or straight-line, relationship between variables and level of relation. The model of correlation coefficient is defined as (McNeil, et al, 1998 : 181)

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

It shown that r ranges from a minimum of -1 to a maximum value of 1. A correlation coefficient equal to 0 indicates no linear relationship between the two variables.

4. Multiple Regression Analysis

Multiple regression is a method of analyzing the collective and separate contributions of two or more independent variables, x_i, to the variation of a dependent variable y. This method for measuring the effects of several factors concurrently. Multiple regression uses a model of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

where y is outcome and α is a constant and β_1, β_2, ... β_k are regression coefficients associated with the independent variables x_1, x_2, ..., x_k. Here we predict from the x's to y using the α and βs. There are three assumptions underlying the statistical methods described in the preceding section, as follows: (Mcneil, 1998 : 185)

1) The relation between x and y in the population is *linear*;

2) The standard deviation of the distribution of errors is *constant*, that is, it dose not depend on x;

3) The distribution of the errors is *normal.*

These assumptions can be assessed by plotting the residuals against the fitted values to assess the first two assumptions and by plotting the residuals against normal scores to assess the normality assumption.

5. Partial Correlation Coefficient

Partial correlation is a new set of correlations which involve the variables. After chosen the first predictor variable placed in the model as the one, which is most correlated with y, that is, the variable x_1 which r_{1y} is the largest of all the variables. The model $y = \alpha + \beta_1 x_1$ is fitted. As $r_{2y.1}$, meaning the correlation of variable x_2 and y, and read as the partial correlation of variable x_2 and y after both have been adjusted for variable x_1. This partial correlation can be calculated by (N.R. Draper and H. Smith, 1981 : 266)

$$r_{2y.1} = \frac{r_{2y} - r_{12} r_{1y}}{\sqrt{(1 - r_{1y}^2)(1 - r_{12}^2)}}.$$

Partial correlation is not limited to three variables. So-called higher-order partial correlations can be calculated. The order of partial *r*'s is determined by the number of variables being partialed. $r_{2y.1}$ is first-order partial, while $r_{3y.12}$ is a second-order. The reasoning and procedure outlined above apply to higher-order partial *r*'s, which can be calculated by using successive partialing. For example,

$$r_{3y.12} = \frac{r_{3y.1} - r_{32.1} r_{2y.1}}{\sqrt{(1 - r_{2y.1}^2)(1 - r_{32.1}^2)}}$$

Note that the formula uses first-order partial. For third-order partials the formulas and calculation are cumbersome, but the pattern is the same. (F. Kerlinger and E. Pedhazur, 1973 : 89-90)

6. Stepwise Regression Methods

Stepwise regression methods used for eliminating redundant predictors. Three methods are available (stepwise regression, forward selection, and backward elimination). The procedure for this method is as follows: (S. Dowdy and S. Wearden, 1983 : 421)

Step 1. Compute all simple correlation between y and the possible independent variables *x*, that is, r_{y1}, $r_{y2,...}$, r_{yk}. Select the *x* with the largest simple correlation coefficient, say, x_j. Compute the regression $\hat{y} = a + b_j x_j$ and test $H_0 : \beta_j = 0$. If this test is not significant, stop; there is no appropriate model using these *x*'s. If it is significant, proceed to Step 2.

Step 2. Compute partial correlation coefficients with the effect of *j* removed. Select the *x* with the largest partial correlation coefficient, say, x_i. Compute the regression $\hat{y} = a + b_i x_i + b_j x_j$. Test this model for significance, if it is not significant, use $\hat{y} = a + b_j x_j$. If it is significant, proceed to Step 3.

Step 3. Test x_i and x_j as if each were the last to enter the equation. If both are significant, go to Step 4, but if x_i is not significant, leave it out and use $\hat{y} = a + b_j x_j$ as

the model. If x_i is significant but not x_j, remove x_j, put it back in the pool of possible independent variables, and next go to Step 4.

Step 4. Compute the partial correlation coefficients between y and each x that is not in the equation with the effects of the included x's removed. Select the x with the largest partial correlation coefficient. Add this x to the equation, and test the enlarged equation for significance, if not, remove this last x and use the previous equation. If it is significant, proceed to Step 5.

Step 5. Check all of the x's as if they were the last to enter the equation. If any are not significant, remove them and return them to the pool of possible independent variables. Compute the equation for the reduced set of x's. Repeat Step 4 with this reduced set of included x's. If all are significant, repeat Step 4 with this full set of included x's.

Step 6. Repeat Steps 4 and 5 until the x with the largest partial correlation coefficient does not make a significant contribution or until all x's are included.