Chapter 2

Methodology

This chapter describes the research methodology used for this study. Data and variables, data management and statistical methods are explained.

2.1 Data and Variables

Data of students enrolled in four-year undergraduate degrees at PSU, Pattani between 1999 and 2007 were obtained from the database of the Registration Office at PSU, Pattani. Students who had studied in the faculties of Education, Humanities and Social Sciences, and Science and Technology, or the College of Islamic Studies were included in the analysis. There are 16, 17, 9 and 5 majors in faculty of Education, Humanities and Social Sciences, Science and Technology and College of Islamic Studies, respectively. The total number of students was 13,232. Of these, 480 transferred their major.

Variables obtained included faculty of study, year of admission, duration of study, gender and religion. We combined faculty, religion and gender into a new variable called faculty-religion-gender containing 14 distinct combinations (2 combinations, non-Muslim males and non-Muslim females admitted to the College of Islamic studies, were omitted due to zero counts). Year of admission is defined as the year of the student enrolled in their degree program at PSU, Pattani. Duration of study is the number of years a student had been studying at the time of this analysis for students who had not transferred or the number of years of study up until the year of transfer

for students who transferred to another major. The outcome is a binary variable denoting whether or not the student transferred to another major during their degree. If a student transferred more than once then the data (duration of study) related to each new major degree was included in the analysis as a separate record. A path diagram is used to summarize the variables considered in the study is shown in Figure 2.1. The distributions of variables are shown in Chapter 3.

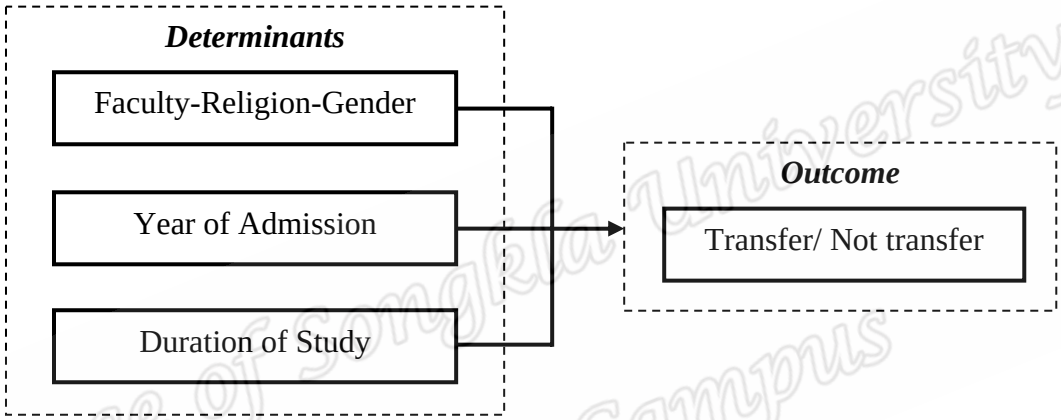

Figure 2.1: Path diagram

2.2 Data Management

Microsoft Office

Excel was used to manage the data for this study. It was used for data cleaning, including data coding and fixing errors.

R Program

R was used to analyze data for the preliminary results and to fit the logistic regression model for the pattern of transfer students.

Data Management

The data were loaded into a Microsoft Excel spreadsheet file for data cleaning and recoding. The faculty-religion-gender, the first digit of coding represents the faculty, 1=Education, 2=Humanities and Social Studies, 3=Science and Technology, 4=Islamic Studies, and second digit represents the religion/gender which are 1=Muslim/Male, 2= non-Muslim/Male, 3=Muslim/Female, 4= non-Muslim/Female). There were 14 categories of combined variables are coding as 11:Edu.Muslim.Male, 12:Edu.non-Muslim.Male, 13:Edu.Muslim.Female, 14:Edu.non-Muslim.Female, 21:Hum.Muslim.Male, 22:Hum.non-Muslim.Male, 23:Hum.Muslim.Female, 24:Hum.non-Muslim.Female, 31:Sci.Muslim.Male, 32:Sci.non-Muslim.Male, 33:Sci.Muslim.Female, 34:Sci.non-Muslim.Female, 41:Isl.Muslim.Male, 43:Isl.Muslim.Female. Duration of study was defined as '2-3 years' and '4 or more years'. The data were transferred into *R* for analyzing, graphing and fitting logistic regression models.

Data used were structured as a multi-way contingency table of counts with 224 possible combinations of factors (combinations with zero counts were excluded) representing the three aforementioned determinants.

2.3 Statistical Methods

Several statistical methods were used in this thesis. Results are described as percentages, and Pearson's chi-squared test with odds ratios were used to assess the associations between the determinants and outcome. Logistic regression was used to obtain independent adjusted effects of the three determinants on the probability of transferring to another major.

Descriptive statistics

Chi-squared test and odds ratio

Pearson's chi-squared test and 95% confidence intervals for odds ratios was used to assess the strength of the associations between the determinants and the outcome of this study. The formulas based on contingency tables (McNeil, 1998) have the following general form.

A. 2 x 2 tables

X is a binary determinant and *Y* is a binary outcome. The data may be presented as a 2-by-2 table with four cells. The cells in this table contain the numbers of observations corresponding to each combination of determinant and outcome (McNeil, 1998). The odds ratio describes the strength of the association between *X* and *Y*; a 2-by-2 table is constructed as follows.

		Y = 1	*Y* = 0
X	1	*a*	*b*
	0	*c*	*d*

$$n = a + b + c + d$$

In this notation, *a* and *b* are the numbers of outcomes coded as 1 and 0, respectively, in the first group, while *c* and *d* are the numbers of respective outcomes in the second group. Thus the odds ratios may be expressed in terms of these counts as:

$$OR = \frac{a \times d}{b \times d} \tag{2.1}$$

An asymptotic formula for the standard error of the logarithm of the odds ratio is given by:

$$SE(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \quad (2.2)$$

A 95% confidence interval for the odds ratio is given by:

$$95\%\ \text{CI} = \text{OR} \times \exp(\pm 1.96\ SE\ [\ln OR]) \quad (2.3)$$

For a 2-by-2 table, Pearson's chi-squared statistic of independence is defined as

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} \quad (2.4)$$

The p-value associated with the test of the null hypothesis of independence is the probability that a chi-squared distribution with 1 degree of freedom exceeds this statistic. If the p-value is less than a pre-defined significant level, then the null hypothesis is rejected.

B. r x c tables

Both the determinant (*X*) and outcome (*Y*) are categorical. Such data may be summarised by a table with several rows and several columns, *r* is the number of categories defining the determinants and *c* the number of categories in the outcome. In our the outcome is binary (0, 1) and determinants are nominal (1,…, *r*), an $r \times c$ table is constructed as follows.

		Y	
		1	0
X		a_{11} a_{21} : a_{r1}	a_{12} a_{22} : a_{r2}

An estimate of the odds ratio is given by

$$OR_{ij} = \frac{a_{ij} \times d_{ij}}{b_{ij} \times c_{ij}}, \tag{2.5}$$

where $b_{ij} = \sum_{j=1}^{2} a_{ij} - a_{ij}$, $c_{ij} = \sum_{i=1}^{r} a_{ij} - a_{ij}$, $d_{ij} = n - a_{ij} - b_{ij} - c_{ij}$, $n = \sum_{i=1}^{r}\sum_{j=1}^{c} a_{ij}$

The standard error of the logarithm of the odds ratio is given by

$$SE(\ln OR_{ij}) = \sqrt{\frac{1}{a_{ij}} + \frac{1}{b_{ij}} + \frac{1}{c_{ij}} + \frac{1}{d_{ij}}} \tag{2.6}$$

A 95% confidence interval for the odds ratio is given by:

$$95\%\ \text{CI} = \text{OR} \times \exp(\pm 1.96\, SE\,[\ln OR]) \tag{2.7}$$

Pearson's chi-squared statistic for independence is defined as

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(a_{ij} - \hat{a}_{ij})^2}{\hat{a}_{ij}} \tag{2.8}$$

where $\hat{a}_{ij}$ are the expected frequencies.

The p-value is the probability that a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom exceeds this statistic. If the p-value is small (say less than 0.05) the null hypothesis is rejected (McNeil, 1998).

Logistic regression

Logistic regression (Hosmer and Lemeshow, 2000) is a statistical method widely used for modeling the association between a binary outcome and categorical determinants, although the determinants can also be continuous. Contingency tables of counts for

the categorical determinant were explored. In logistic regression, the outcome is a logit, which is the natural log of the odds. For a single determinant *x*, the logistic model is given by:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta\, x\,, \qquad (2.9)$$

where p is the proportion of outcomes, α is a constant and β is the coefficient estimated by the modeling process, corresponding to the determinant, *x*. The outcome variable Y takes values 0 and 1. Equation (2.9) can be inverted to give an expression for the probability of the event as

$$\text{Prob}[Y=1] = \frac{1}{1+\exp(-\alpha-\beta\, x)} \qquad (2.10)$$

Equation (2.10) ensures that the values of p are always between 0 and 1. If there were 3 determinants in the model (x_1, x_2, x_3), Equation 2.10 may be written as

$$\text{Prob}[Y=1] = \frac{1}{1+\exp(-\alpha-\beta_1 x_1-\beta_2 x_2-\beta_3 x_3)} \qquad (2.11)$$

Logistic regression provides an appropriate statistical method for modeling student transfers. In the model, the outcome is the binary event denoting transfer in the demographic group indexed by faculty-religion-gender, year of admission and duration of study.

A graph of confidence intervals of population proportions is appropriate for comparing the difference of two or more groups. The proportions of positive outcome and their corresponding standard errors may be estimated by fitting a logistic regression model, and again it is appropriate to use weighted sum contrasts to obtain

the standard errors underlying the confidence intervals for comparing these proportions (Tongkumchum and McNeil, 2009).

$$SE = \sqrt{\frac{p(1-p)}{n}} \tag{2.12}$$

95% confidence interval for proportion is thus given by

$$p - 1.96 \times SE,\ p + 1.96 \times SE \tag{2.13}$$

Goodness-of-fit of model

The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected count from model. Pearson's residual is defined as

$$z = \frac{p - \hat{p}}{\sqrt{\hat{p}(1-\hat{p})/n}}, \tag{2.14}$$

p is the proportion of transferring major students observed in the cell, $\hat{p}$ is the corresponding probability given by the logistic regression model and n is the total number of cases in the cell. The goodness-of-fit of the model can be assessed visually by plotting these z-values against corresponding normal scores. An adequate fit could be obtained if the point of observed plot is close to a straight-line with unit slope.

A p-value for the goodness of fit is obtained by subtracting the deviance associated with the saturated model from the deviance obtained from model and comparing this difference with a chi-squared distribution having degrees of freedom equal to $k - m$, where k is the number of cells and m is the number of parameters in the model.