Chapter 2

Methodology

In this chapter, we describe the methods used in the study. The following topics are covered.

1. Computer applications
2. Study design
3. Path diagram and variables
4. Data collection, management and data analysis
5. Methods for statistical analysis

2.1 Computer implications

The following computer applications were used for data analysis.

Microsoft Excel was used to manage the data for this research.

WebStat, a suit of web-database software engineering tools written in HTML and VB script for graphing and analyzing statistical data stored in a web-based database. It was mainly used to perform preliminary data analysis, using analysis of variance and logistic regression modeling.

2.2 Study design

This study used cross-sectional data obtained from birth certificates in four southern Thai provinces, aggregated by month of birth, age of mother, and district of registration of birth from January 2002 to December 2005, together with female

resident population counts obtained from the 2000 Thai Population and Housing Census (National Statistical Office, 2002). The provinces were selected to include two from each of the east-west locations, Muslim majority (>50% Muslim) and Muslim minority (<50% Muslim). Trang (17% Muslim) and Satun (68% Muslim) on the west coast of the peninsula and Songkhla (23% Muslim) and Pattani (82% Muslim) on the east coast were selected. Birth registration data for 2001 were excluded from the study because for the first seven months of 2001 the age of the father was recorded instead of that of the mother.

A total of 45 districts were sampled and classified into eight regions according to east-west location and four categories of percent Muslim (below 20%, 20-50%, 51-80%, more than 80%) as shown in Figure 2.1.

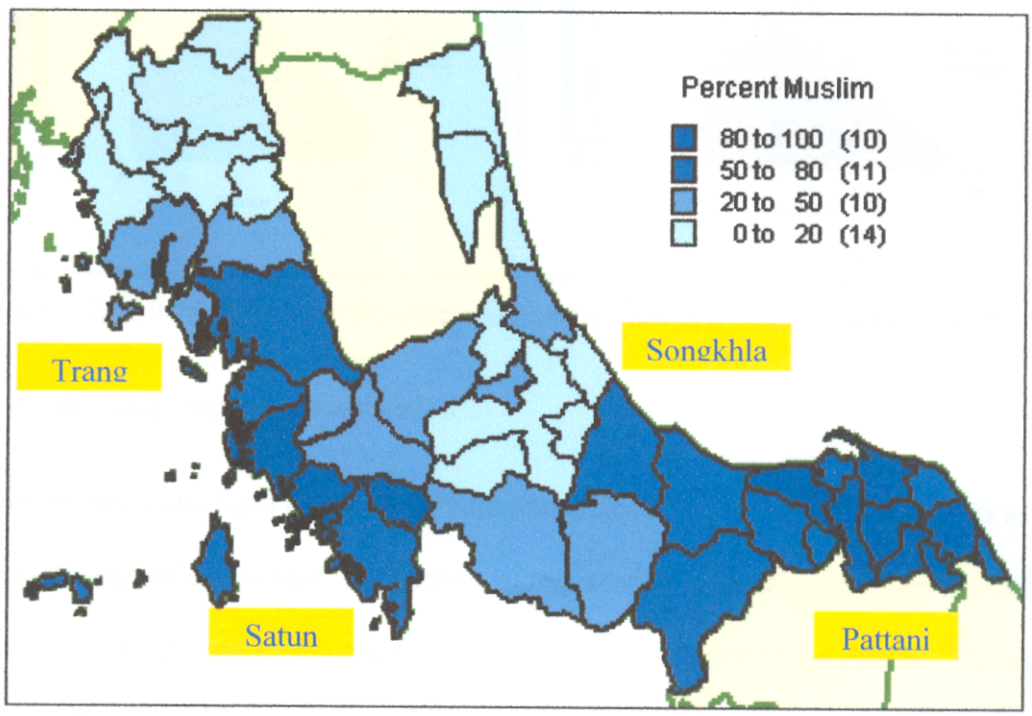


Figure 2.1: Classification of districts in four southern provinces by percent Muslim

The age distributions for the Muslim and non-Muslim populations in these provinces (based on the 2000 population census) are quite different (Figure 2.2), with the young Muslim populations larger than those of the non-Muslim populations, particularly in Pattani.

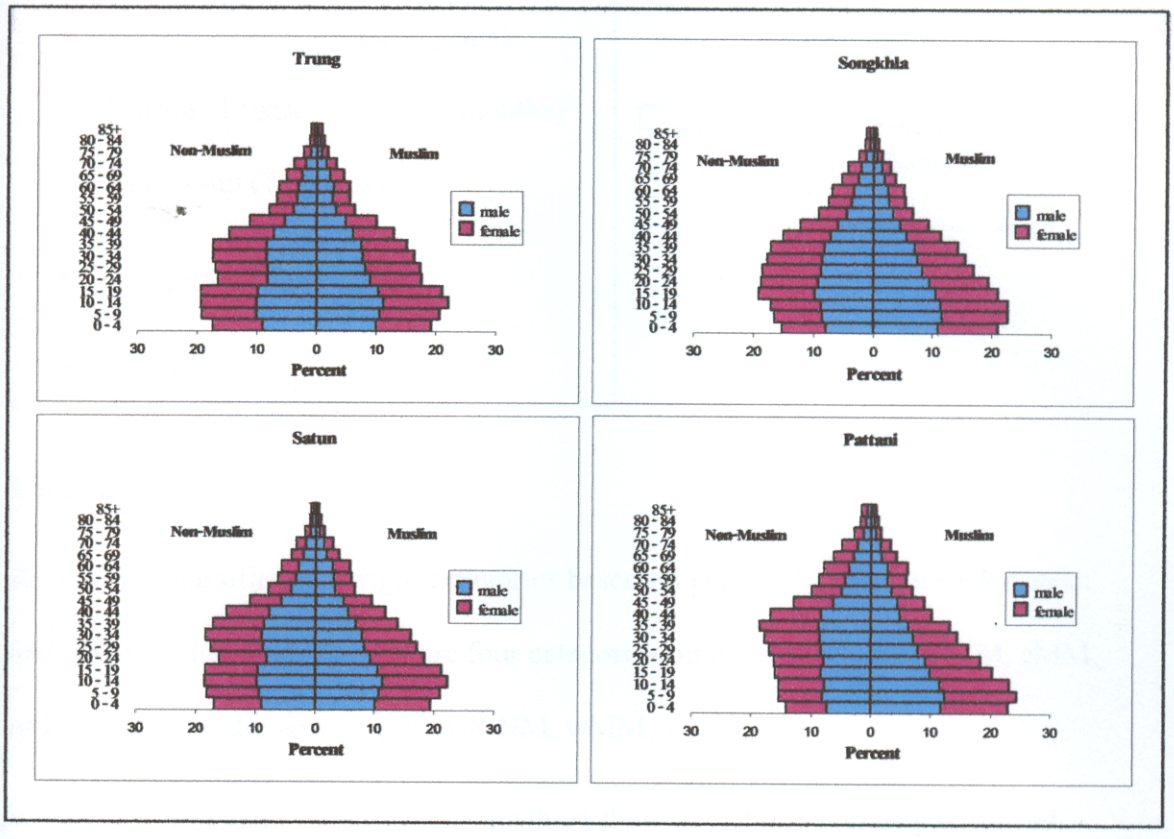
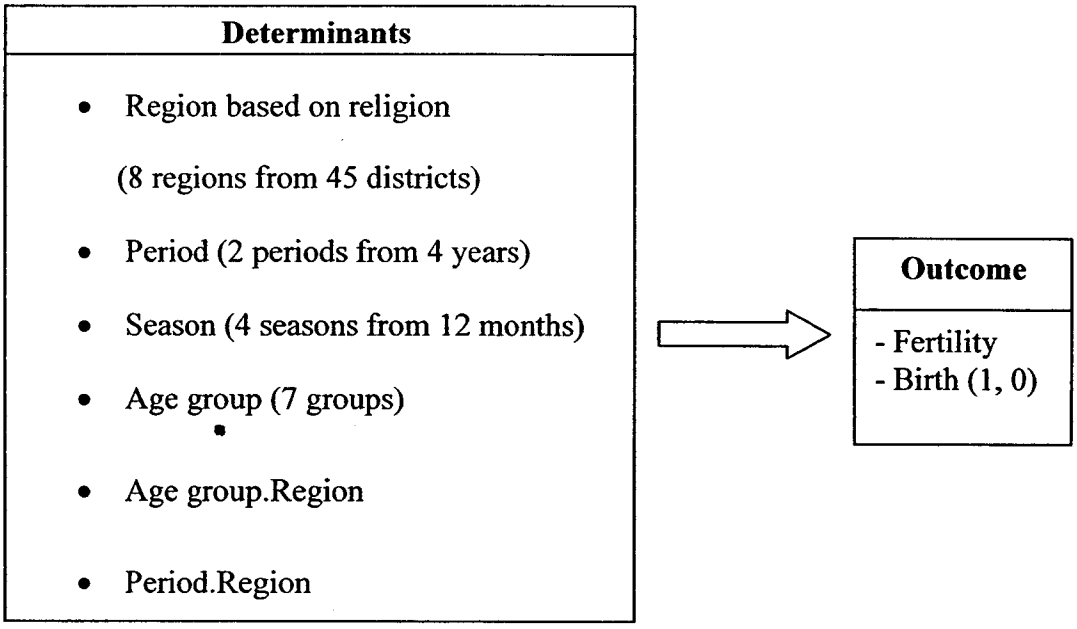


Figure 2.2: Age pyramids for Muslim and non-Muslim populations in four provinces

2.3 Path diagram and variables

The birth certificate form contains the date and district of registration of the birth as well as the age of the mother. The district of registration is the district where the baby is born, which may not be the same as the district of residence of the mother. The age of the mother in years is classified as 20 or less, 21-25, 26-30, 31-35, 36-40, 41-45, and 46 or more.

Path diagram



Variables

Region was classified into eight categories based on percent Muslim in each district and east-west location. There were four categories on the east (eNM, eMNM, eMM, eM) and four on the west (wNM, wMNM, wMM, wM) where

- *NM* (less than 20%) is a region where the non-Muslim majority is located.
- *MNM* (20-49%) is a mixed region with a larger non-Muslim population.
- *MM* (50-79%) is a mixed region with a larger Muslim population.
- *M* (80% or more) is a region where the Muslim majority is located.

Period was classified into 2-year periods, 2002-2003 was a period 1 and 2004-2005 was period 2.

Season was classified into four seasons; January-March, April-June, July-September and October-December.

Age group included seven groups; 15-19, 20-24, 25-29, 30-34, 35-39, 40-44 and 45-49.

Age group.Region was combined from seven age groups and eight regions into 56 groups; 15-19.west<20, 15-19.east<20,..., 45-49.west80+ and 45-49.east80+.

Period.Region was combined from two periods and eight regions into 16 groups; 2002-2003.west<20, 2002-2003.east<20, ..., 2004-2005.west80+ and 2004-2005.east80+.

2.4 Data collection, management and data analysis

Data collection

Data used in the study is comprised of two data sets, which are number of babies and number of female. The number of babies was taken from the birth certificate at The Registrar Administration Center 9, Pattani from January 2001 to December 2005. The female population aged 15-49 was obtained from the 2000 Thai Population and Housing Census, Statistics National Office.

Data management

The data were originally as a report from the Registrar Administration, recorded in a Microsoft Excel spreadsheet file, imported to SQL Server and analyzed using WebStat.

Data analysis

The data records comprised the number of live births B_{ijt} classified by registration district i, age-group j and month t, and P_{jt} the estimated populations of women in the age-group. We estimated this population using the formula

$$P_{jt} = p_t N_j + (1-p_t) N_{j-1}, \qquad (2.1)$$

where N_j is the female population recorded at the 2000 population census in age group j (where $j = 1$ for 0-4, $j = 2$ for 5-9, etc.), which is one year less than the reported age group of the mother, and p_t is the proportion of the 5-year period elapsed from January 2001 until month t. This calculation synchronizes the population age groups obtained from the census with those used in the birth registration form.

2.5 Statistical methods

Total fertility rate

In the preliminary analysis we computed the total fertility rate (*TFR*) for each year and region i using the standard demographic formula (Pollard et al, 1974) for 5-year age groups, namely

$$TFR_i = \sum_{j=4}^{10} \left(5 \sum_t B_{ijt} \Big/ \tfrac{1}{12} \sum_t P_{ijt}\right), \qquad (2.2)$$

where t is summed over all months within the specified year. Note that the contribution from each of the seven age groups ($j = 4, 5 \ldots 0$) are 5 (the width of the age group in years) times the total number of births in the 12-month period of interest divided by the average number of women in the corresponding age group.

Logistic Regression

In further analysis logistic regression was used to model the effects of age group, region and period on the fertility in each 3-month quarter of a year. This fertility is the probability that a woman in a specified age group registers a birth in a specified region and quarter. Logistic regression analysis is used to investigate the association

between determinants and binary outcome. In the simplest case, when there is a single continuously varying determinant x, the model takes the form,

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta x, \tag{2.3}$$

where p is the probability that the outcome is in the specified category. Equation (2.3) can be inverted to give an expression for the probability of the event as

$$p = \frac{1}{1+\exp(-\alpha - \beta x)}. \tag{2.4}$$

The functional form of Equation (2.4) ensures that its values are always between 0 and 1, as they should be given that they are probabilities.

This model is easily extended to handle multiple determinants. For m continuous or binary determinants ($x_1, x_2 \ldots x_m$), it may be written as

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \sum_{j=1}^{m} \beta_j x_j . \tag{2.5}$$

Nominal determinants are handled by separating them into their binary components, giving k–1 such components for a determinant with k categories. Asymptotic results based on statistical theory provide estimates based on maximum likelihood fitting of the model, together with confidence intervals and p-values for testing relevant null hypotheses (Kleinbaum and Klein, 2002).

Goodness-of-fit for a model

For each cell corresponding to a combination of nominal determinants, the Pearson residual is defined as

$$z = \frac{p - \hat{p}}{\sqrt{\hat{p}(1-\hat{p})/n}}, \tag{2.6}$$

where p is the proportion of outcomes observed in the cell, $\hat{p}$ is the corresponding probability given by the 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. The fit is adequate if the points in this plot are 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 model deviance and comparing this difference and comparing this difference and comparing this difference R_g with a chi-squared distribution with degrees of freedom equal to $n_g - m$, where n_g is the number of cells and m is the number of parameters in the model.