Chapter 2

Methodology

In this chapter we describe the methods for selecting the subjects from the target population and data collection from the subjects, and the statistical methods used to analyse the data.

2.1 Study Design and Sampling Technique

A cross-sectional study design is used for this study. The target population comprised police officers in the southernmost provinces (Pattani, Narathiwat and Yala) of Thailand. Samples were police officers who worked in the three Southern Provinces of Thailand from 12 police stations in 2002. The sample size required for specified precision d with $1-\alpha$ certainty for a continuous outcome with standard deviation σ is (McNeil, 1996:267)

$$n = z_{\alpha/2}^2 \frac{\sigma^2}{d^2}$$

The outcome factor is a standardized score between 0 and 1 based on factor analysis of items in a questionnaire, and might be expected to have standard deviation close to 0.5. A precision of 5% is reasonable in studies such as this.
If the certainty is 95%, the sample size is thus

$$n = \frac{(1.96)^2(0.5)^2}{(0.05)^2}$$

$$= \frac{(3.84)(0.25)}{0.0025} = 384.$$

So a sample size of 384 police officers is needed. However in this study 404 subjects were obtained. The sample acts as a representative of the population. In this study, the stratified random sampling method was used to get a sample. The samples were grouped in each police station by station size, as shown in Table 2.1. These data were sourced from police registrations at each police province.

Province	Number of police	Police station	Target population	Sample
Pattani	1,563	Kapo	77	23
		Saiburee	163	55
		Mealan	74	25
		Muang Pattani	206	34
Yala	1,182	Bannangsata	95	35
		Yaha	111	35
		Ja-kwa	49	24
		MuangYala	241	38
Narathiwat	1,662	Su-ngi-padee	104	44
		Ja-nae	70	28
		Ra-ngc	100	42
		Srisakorn	81	21
	4,407	Total	1,371	404

Table 2.1: The sample size for each police station

The variables of interest for this study comprised 13 determinants and several continuous outcomes based on factor analysis.

2.2 Graphical and Statistical Methods

The graphical and statistical methods comprised the following:

1. Histograms and numerical summaries for data from all variables.
2. Factor analysis, two-sample t-tests, one-way analysis of variance and multiple regression analysis of the variables described by box plots and 95% confidence interval of means.

Univariate and Bivariate Summaries

The mean and standard deviation *(S.D.)* are used to summarise the data for a single variable. They are calculated from the formulas

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \text{ and } S.D = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{X})^2}{N}}$$

Correlation analysis attempted to measure the strength of relationships between two variables by means of a single number called a correlation coefficient. The most widely used measure of linear correlation between two variables is called the Pearson

product-moment correlation coefficient or simply the correlation coefficient. The measure of linear relation between two variables X and Y is estimated by the sample correlation coefficient r, defined as

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

P-value

A null hypothesis as statement that a population parameter takes a specified null value, and its plausibility could be assessed graphically by comparing the null value with a confidence interval for the parameter.

Two-Sample t-test

The two-sample t-test takes the form

$$t = \frac{\bar{y}_1 - \bar{y}2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In this formula, s is the pooled sample standard deviation. If s_1 and s_2 denote the standard deviation of the two samples, respectively, it may be shown that the pooled sample standard deviation is given by the formula

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

A p-value is now obtained from the table of two-tailed t distribution with $n_1 + n_2 - 2$ degrees of freedom. This statistical procedure is called the *two- sample t test* (McNeil, 2001, page 172).

One-way analysis of variance

In this thesis we consider methods for the analysis of data in which the outcome is continuous and each determinant is categorical. This leads to a procedure called the (one-way) analysis of variance (anova). The null hypothesis is that the population means of the outcome variable corresponding to the different categories of the determinant are the same, and this hypothesis is tested by computing a statistic called

the *F*-statistic and comparing it with an appropriate distribution to get a p-value. Suppose that there are n_j observations in sample *j*, denoted by y_{ij} for $i= 1,2,\ldots,n_j$. 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 $$\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n_j} y_{ij}, \bar{y} = \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n_j} y_{ij}, n = \sum_{j=1}^{c} n_j$$

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 obtained by subtracting each 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 (assumed the same for each population). The *p-value* is the area in the tail of the *F*-distribution with *c*-1 and *n-c* degrees of freedom (McNeil, 1996:pp 67-73).

Correlation Coefficient

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

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

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

Factor Analysis

Since we have multivariate outcomes, factor analysis is used to reduce the dimensionality of these outcomes. Factor analysis is a data reduction technique. It is a group of procedures designed for removing duplicated information from a set of correlated variables and representing the variables with a smaller set of derived variables or factors. There are three procedures involved. The first stage is obtaining the original data matrix. A set of subjects $O_1, O_2, ..., O_n$ are measured with a different number of variables $V_1, V_2, ...V_k$. The second stage involves the creation of a correlation matrix, which is calculated for each combination of two variables: V_1 with V_2, V_1 with V_3, etc, according to the following formula:

If X_i is the data from V_1, and y_i is the data from V_2, then the correlation is given by

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Where s_x and s_y are the sample standard deviations of V_1 and V_2 and n is the number of pairs of observations.

The last stage involves computing the factor loadings. These reveal the extent to which each of the variables contributes to the meaning of each factor. Within any one column of the factor matrix, some of the loadings will be high and some will be low. The variables with the high loading on a factor will be the ones that provide the meaning of the factor

Maximum likelihood factor analysis is a widely used method. This method enables us to carry out test of the goodness of fit of a solution comprising *k* factors. It provides a test of the null hypothesis that *k* common factors are sufficient to describe the data. The algorithms for this method are given as follows.

Suppose we have *P* variables and want to fit *k* factors. Let *R* be the $P \times P$ correlation matrix of the variables, *L* the $P \times k$ matrix of factor loadings, and ψ the vector of length *p* containing the unique variances. Then we need to find values for *L* and ψ that maximise the likelihood function, f (l, ψ).

For the fixed value of ψ, we maximize f (L, ψ) with respect to L. The value of L is then substituted into f (L, ψ). Now f can be reviewed as a function of ψ. A transformation of this function gives

$$m(\psi) = \sum_{m=k+1}^{p} \left[\log \gamma_m + \frac{1}{\gamma_m} - 1 \right],$$

where $\gamma_1 \leq \gamma_2 \ldots \leq \gamma_p$ are the eigenvalues of ψ R^{-1} ψ. We then minimize m (ψ). This gives an estimate of ψ, which is then put into the likelihood $f(L, \psi)$. Then the likelihood is again maximized with respect to L. Then a new value for m (ψ) is computed and so on.

After making the decision on how many factors to extract form the original set of variables we can redefine the factors so that the explained variance is redistributed among the new factors. This technique is used to sharpen the distinction in the meaning of the factors. A redefinition of the factors, with the loading on the various factors either very high or very low, and then eliminating as many mediums sized loading, aids in the interpretation of factors.

Varimax rotation is one of many types of rotation and is regarded as the standard approach. This approach places more emphasis on the simplification of the factors. It tends to avoid a general factor. Using the comprehensibility method to select a number of factors, suppose that three factor are retained. (Prasitrattasin S, 1997:355)

Multiple linear regression

Regression is use to analyse data in which both the determinants and the outcome are continuous variables. When there is just one determinant it can summarise the data in the scatter plot by fitting a straight line. In conventional statistical analysis the line fitted is the least squares line, which minimizes the distances of the points to the line, measured in the vertical direction. If there is more than one determinant, the method generalises to multiple linear regression, in which the regression line extends to the multiple linear relation represented as (McNeil, 1998:185)

$$Y = \beta_0 + \sum \beta_i x_i + \varepsilon$$

where Y is the outcome variable, $\{\beta_0\}$ is a constant, $\{\beta_i\}$ is a set of parameters (i =1 to p), and $\{x_i\}$ is a set of determinants (i =1 to p).

Linear regression analysis resets on three assumptions as follows.

(1) The relation is linear.
(2) The variability of the error (in the outcome variable) is uniform.
(3) These errors are normally distributed.

If these assumptions are not met, a transformation of the data may be appropriate. Linear regression analysis may also be used when one or more of the determinants is categorical. In this case the categorical determinant is broken down into $c-1$ separate binary determinants, where c is the number of categories. The omitted category is taken as the baseline or referent category.

2.3 Computer Programs for Analysis of Data

The following computer programs were used for data analysis and preparation.

Microsoft Excel

This program was mainly used to store the data for this research. Some functions are helpful in finding matrix correlations and plotting graphs.

Microsoft Word

This program was mainly used to write and print the report of this research.

EcStat in Microsoft Excel

EcStat is an add-in to Microsoft Excel 2000. It is a suite of routines for graphing and analysing statistical data using and compatible PC.

EcStat was mainly used in "Comparison" and "Relation" commands. The result of "Comparison" is a one way analysis of variance (anova), and the result of "Relation" is a graph with fitted linear relation shown on the plot.

SPSS PC$^+$ Version 9 was mainly used in "Factor analysis".

MATLAB V.5 (Hanselman & Littlefield, 1997) was mainly used in "Linear regression Analysis". The result of "Linear regression" is a full model and final model used for prediction.