Chapter 2

Methodology

This chapter describes the methods used in the study. These methods include following components.

(1) the study design
(2) the conceptual framework
(3) the methods and programs for the analysis of the data

2.1 Study Design

As explained in Chapter 1, we are interested in describing the development of elite sporting performance during the period of the modern Olympics, and we want to see if there are differences in different sports and between men and women. In particular we want to know if men and women have reached the limits of their speed and strength dictated by their human bodies, or whether humans will continue to improve their performances.

To answer these questions, we examine all the winning results of Olympic sports champions in specified events over the period 1928-2000. We look at the four main sports in which men and women have competed unaided by technology. These are swimming, running, jumping and throwing. The jumping events include the high jump, the long jump, and the triple jump, but we exclude the pole vault because technological improvements in the material (from bamboo to metal-enhanced fibre glass poles) have allowed pole-vaulters to improve their performances quite dramatically. We have also included running events up to 1500 meters, but not longer, because long-distance running is really a different sport to sprinting.

There is no question of selecting a sample from a population, because the population of Olympic winners is finite, and can be accommodated quite easily in a relatively small database. But since conventional statistical methods rely on the concept of selecting a random sample from a population, we will imagine that there is a population of conceptual Olympic winners, and that the data constitute a random sample from this population. This is a device that is used quite commonly in such studies. For example, case-control studies involving rare diseases in epidemiology frequently study all the cases that exist.

Thus the scope for our study includes 299 men's and women's swimming results from 26 events, 188 men's and women's running results from 13 events, 84 jumping results from 6 events, and 115 throwing results from 7 events.

The outcome variables are time (for swimming and running) and distance (for jumping and throwing). The determinants are year, sex, type of event and distance (for swimming and running).

2.2 Conceptual Framework

Figure 2.1 shows the directions of the relations between the variables for each of the four sports.

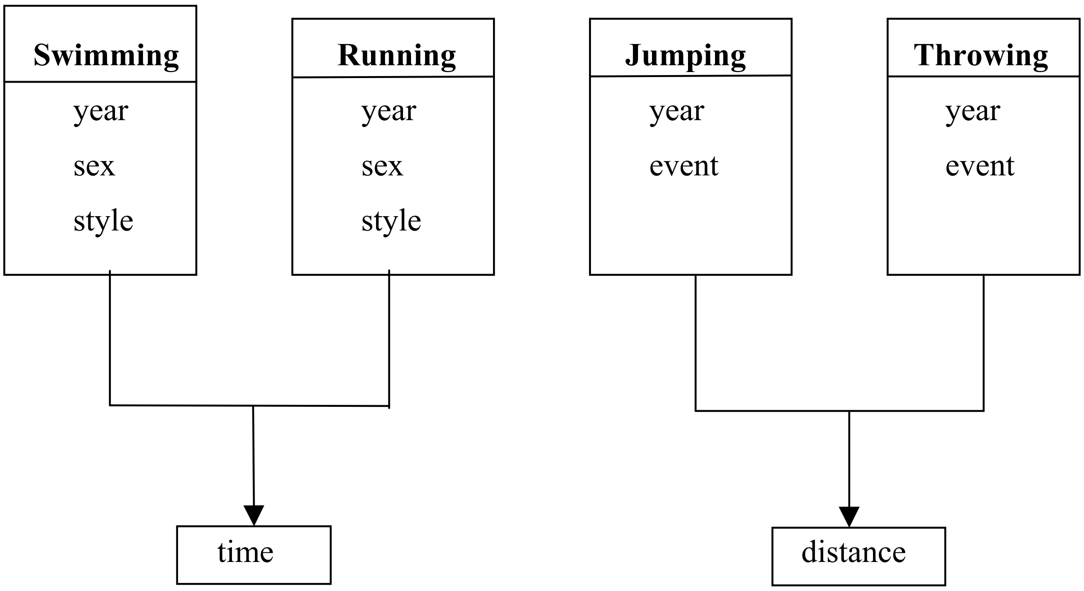

Figure 2.1: Conceptual framework

In each case there are 17 Olympic years from 1928 to 2000. In swimming there are five styles (freestyle, backstroke, butterfly, breaststroke and medley). In running there are two style (sprint and hurdles). In jumping there are three styles (high jump, long jump and triple jump), but to balance the sample sizes these are classified into six events combining style and gender. Similarly, there are seven events in the throwing, corresponding to the men's and women's shot put, discus and javelin throws, and the men's hammer throw.

2.3 Statistical Methods

Numerical and graphical summaries of data and comparisons using one-way and two-way analysis of variance were used in the preliminary analysis.

In further analysis, simple statistical models were used, including multiple linear regression to develop the model of Olympic performances.

One-way analysis of variance

In this thesis we are considering 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: page 67).

Two-way analysis of variance

Considering the analysis of data in which the outcome is continuous and the determinant is categorical, and there is also a categorical covariate, this leads to a modification of anova called two-way analysis of variance. Again suppose there are n_j observations in sample j, denoted by y_{ij} for $i = 1, 2, \ldots, n_j$. The F-statistic is now defined as

$$F = \frac{(S_2 - S_{12})/(c-1)}{S_{12}/(n-c-r+1)}$$

where $$S_2 = \sum_{j=1}^{c}\sum_{i=1}^{r}\left(y_{ij} - \bar{y}\right)^2, S_{12} = \sum_{j=1}^{c}\sum_{i=1}^{r}\left(y_{ij} - \bar{y}_j + \bar{y}\right)^2$$

and $$\bar{y}_i = \frac{1}{c}\sum_{j=1}^{c} y_{ij}, \bar{y}_j = \frac{1}{r}\sum_{i=1}^{r} y_{ij}, \bar{y} = \frac{1}{rc}\sum_{j=1}^{c}\sum_{i=1}^{r} y_{ij}$$

S_2 is the sum of squares of the data after adjusting for row effects, S_{12} is the sum of squares after adjusting for both row effects and column effects. The p-value is the area in the tail of the F-distribution with $c-1$ and $n-r-c+1$ degrees of freedom (McNeil, 1996: page 73).

Regression analysis

Regression used to analyse data in which both the determinant and the outcome are continuous variables. In the simplest case, it can summarise the data in a scatter plot by fitting a straight line. In conventional statistical analysis the line fitted is the least squares line, which minimises the distances of the points to the line, measured in the vertical direction. More generally, the multiple linear regression model is given by

$$Y = \beta_0 + \Sigma\beta_i x_i + \varepsilon,$$

where Y is the outcome variable, β_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).

The model is again fitted to data using least squares, which minimises the sum of squares of the residuals.

Linear regression analysis rests on three assumptions as follows.

(1) The association is linear.

(2) The variability of the error (in the outcome variable) is uniform.

(3) These errors are normally distributed.

If any of these assumptions is not met, a transformation of the data may be appropriate.

Linear regression may also be used to fit data in which some of the determinants could be categorical. In this case, each categorical determinant is replaced by a set of binary indicator variables, with one such variable for each level of the determinant except one, which is taken as the referent or baseline level.

Programs for analysis of data

MATLAB (For regression analysis)

Microsoft Access (for data storage and data management)

Microsoft Excel (Using EcStat add-in for preliminary data analysis)