Statistical Inference I: Estimating the Mean and Variance of a Population

1. Introduction

Economists are interested in relationships between economic variables. How much can we expect the sales of Frozen Delight ice cream to rise if we reduce the price by 5%? How much will household food expenditure rise if household income rises by $100 per month? Questions such as these are the focus of Chapters 3 and beyond in *UE/2*.

However some times questions of interest focus on a single economic variable. An airplane seat designer must consider the average hip size of passengers in order to allow adequate room for each person, while still designing the plane to carry the profit maximizing number of passengers. What is the average hip size of U.S. flight passengers? If a seat 18 inches wide is planned, what percent of customers will not be able to fit? Such questions as these must be faced by manufacturers of everything from golf carts to women's jeans. How can we address questions like these? We certainly can not take the measurements of every man, woman and child in the U.S. population. This is a situation when **statistical inference** is used. *Infer* means "to conclude by reasoning from something known or assumed." Statistical inference means that we will draw conclusions about a population based on a sample of data.

2. A Sample of Data

To carry out statistical inference we need data. The data should be obtained from the population in which we are interested. For the airplane seat designer this is essentially the entire U.S. population above the age of two, since small children can fly "free" on the laps of their suffering parents. A separate branch of statistics, called experimental design, is concerned with the question of how to actually collect a representative sample. How would you proceed if your boss asked you to obtain 50 measurements of hip size that is representative of the entire population? This is not such an easy task. Ideally the 50 individuals will be **randomly** chosen from the population, in such a way that there is no pattern of choices. Suppose we focus only the population of adult flyers, since usually there are not that many children on planes, and our experimental design specialist draws the sample in Table 1.

Table 1 Sample hip size data

14.96	14.76	15.97	15.71	17.77
17.34	17.89	17.19	13.53	17.81
16.40	18.36	16.87	17.89	16.90
19.33	17.59	15.26	17.31	19.26
17.69	16.64	13.90	13.71	16.03
17.50	20.23	16.40	17.92	15.86
15.84	16.98	20.40	14.91	16.56
18.69	16.23	15.94	20.00	16.71
18.63	14.21	19.08	19.22	20.23
18.55	20.33	19.40	16.48	15.54

A first step when analyzing a sample of data is to examine it visually. Figure 1 is a histogram of the 50 data points in Table 1.

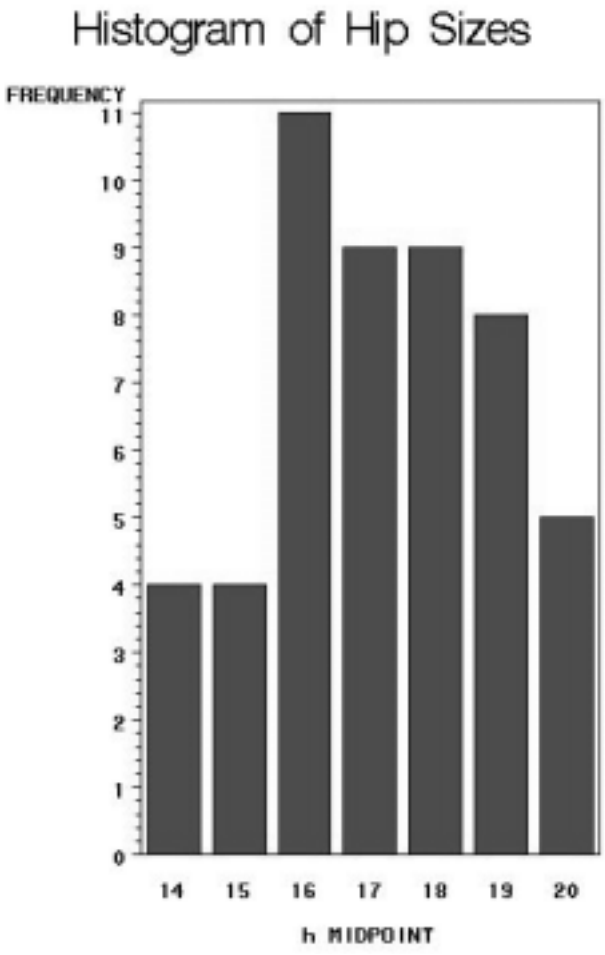

Figure 1. Histogram of hip size data

Based on Figure 1 the "average" hip size in this sample seems to be between 16 and 18 inches. For our profit maximizing designer this casual estimate is not sufficiently precise. In the next section we set up an **econometric model** that will be used as a basis for inference in this problem.

3. An Econometric Model

The data in Table 1 were obtained by sampling. Sampling from a population is an experiment. The outcome variable of interest in this experiment is an individual's hip size. Before the experiment is performed we do not know what the values will be, thus the hip size of a randomly chosen person is a **random variable**. Let us denote this random variable as *Y*. When choosing 50 individuals we actually have $T = 50$ random variables, $Y_1, Y_2, \ldots, Y_T$, where each Y_i represents the hip size of another person. The data values in Table 1 are **values** of the random variable which we denote as $y_1, y_2, \ldots, y_T$, where $T = 50$ is the sample size. We assume that the population has a center, which we describe by the expected value of the random variable *Y*,

$$E[Y] = \beta \tag{3.1}$$

We use the Greek letter β to denote the mean of the random variable *Y*, and also the mean of the population we are studying. Thus if we knew β we would have the answer to the question "What is the average hip size of adults in the U.S.?" To indicate its importance to us in describing the population we call β a population **parameter**, or just a parameter. Our objective is to use the sample of data in Table 1 to make inferences, or judgments about the unknown population parameter β.

The other random variable characteristic of interest is its variability, which we measure by its variance,

$$\text{var}(Y) = E[Y - E(Y)]^2 = E[Y - \beta]^2 = \sigma^2 \tag{3.2}$$

The variance σ^2 is also an unknown population parameter. As described in Chapter 2 of *UE/2* the variance of a random variable measures the dispersion within a population as the average squared distance between the values of the random variable and their mean β. In the context of the hip data, the variance tells us how much hip sizes can vary from one person to the next.

To economize on space we will denote the mean and variance of a random variable as

$$Y \sim (\beta, \sigma^2) \tag{3.3}$$

where "~" means "is distributed as." The first element in parentheses is the population mean and the second is the population variance. So far we have not said what kind of probability distribution we think *Y* has.

The econometric model is not complete. If our sample is drawn randomly, we can assume that $Y_1, Y_2, \ldots, Y_T$ are statistically independent. Thus knowing the hip size of any one individual tells us nothing about the hip size of another randomly drawn individual. Furthermore, we assume that each of the observations we collect is from the population of interest, so each random variable $Y_i \sim (\beta, \sigma^2)$. If $Y_1, Y_2, \ldots, Y_T$ are statistically independent with identical probability distributions then the Y_i constitute a **random sample**, in the statistical sense.

It is sometimes reasonable to assume that population values are *normally* distributed. In this case we will represent them as

$$Y \sim N(\beta, \sigma^2) \tag{3.4}$$

4. Estimating the Mean of a Population

How shall we estimate the population mean β given our sample of data values in Table 1, which we represent generally as $y_1, y_2, \ldots, y_T$. The population mean is $E[Y] = \beta$. The "expected value" of a random variable is its "average" value and also the "center" of its probability density function. Thus finding the center of the sample data seems to be a reasonable way to estimate the center of the population data, β.

One way to find the center of a sample is to use the *principle of least squares*. Under this principle the center of the sample values $y_1, y_2, \ldots, y_T$ is the value of β that minimizes

$$S = \sum_{i=1}^{T} (y_i - \beta)^2 \tag{4.1}$$

where *S* is the sum of squared deviations of the data values from β.

The motivation for this approach can be deduced from the following example. Suppose you are going shopping at a number of shops along a certain street. Your plan is to shop at one store and return to your car to deposit your purchases. Then you visit a second store and return again to your car, and so on. After visiting each shop you return to your car. Where would you park to minimize the total amount of walking between your car and the shops you visit? You want to minimize the *distance* traveled. Think of the street along which you shop as a number line. The Euclidean distance between a shop located at y_t and your car at point β is

$$d_i = \sqrt{(y_i - \beta)^2} \tag{4.2}$$

The squared distance, which is mathematically more convenient to work with, is

$$d_i^2 = (y_i - \beta)^2 \tag{4.3}$$

To minimize the total squared distance between your parking spot β and all the shops located at $y_1, y_2, \ldots, y_T$ you would minimize

$$S = \sum_{i=1}^{T} d_i^2 = \sum_{i=1}^{T} (y_i - \beta)^2 \tag{4.3}$$

as given in (4.1). Thus the least squares principle is really the least *squared distance* principle.

Since the values of y_i are known, and in this case given in Table 1, the sum of squares function S is a function of the unknown parameter β. Multiplying out (4.4) we have

$$S = \sum_{i=1}^{T} y_i^2 - 2\beta \sum_{i=1}^{T} y_i + T\beta^2 \ = a_0 - 2a_1\beta + a_2\beta^2 \tag{4.4}$$

For the data in Table 1 we have

$$a_0 = \sum y_i^2 = 14880.2, \quad a_1 = \sum y_i = 857.9094207, \quad a_2 = T = 50 \tag{4.5}$$

The plot of the sum of squares parabola is shown in Figure 2. The minimizing value appears to be a bit larger than 17 on the figure. Now we will determine the minimizing value exactly.

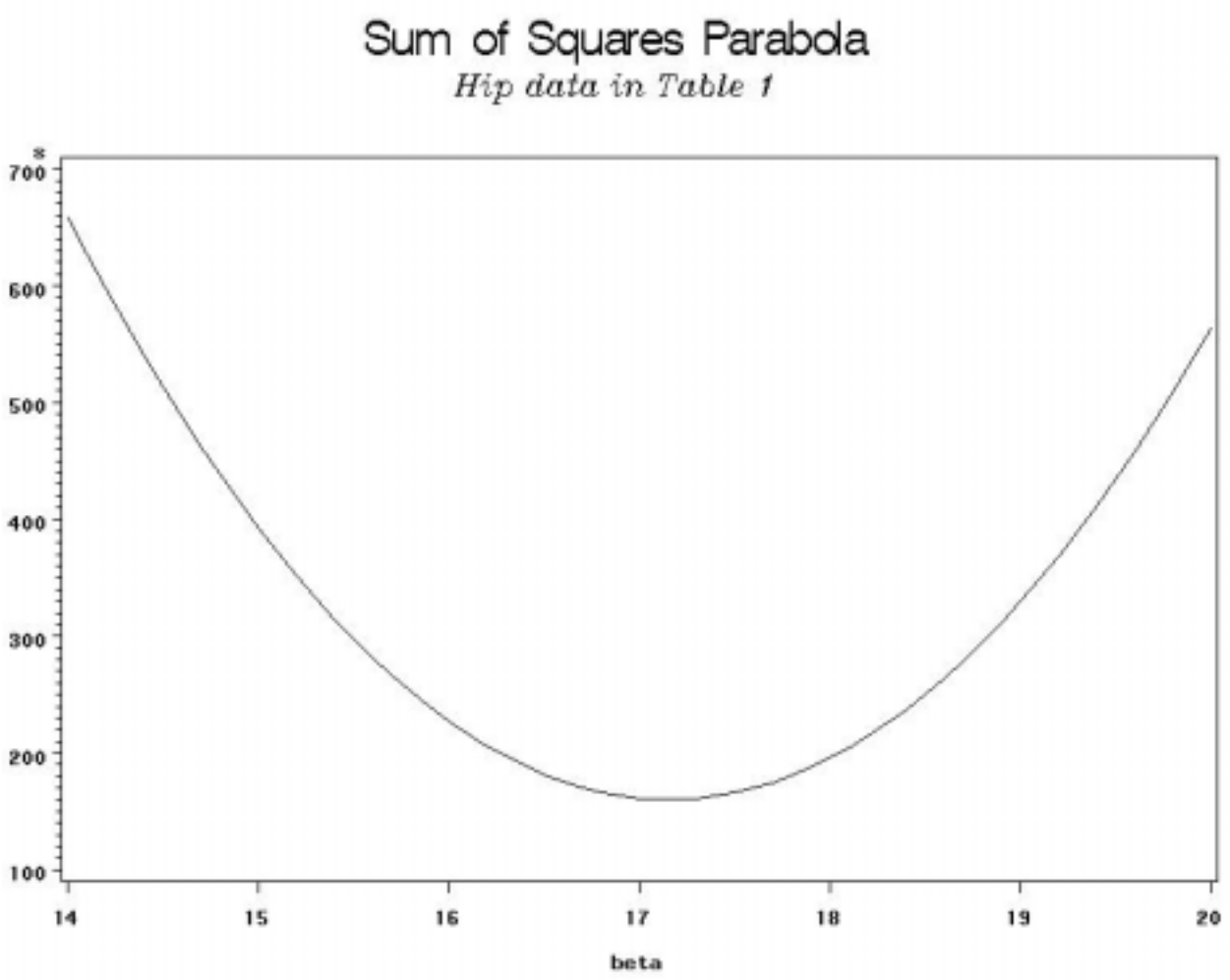

Figure 2 The sum of squares parabola

The value of β that minimizes *S* is the "least squares estimate." From calculus, we know that the minimum of the function occurs where its slope is zero. The function's derivative gives its slope, so by equating the first derivative of *S* to zero and solving, we can obtain the minimizing value exactly. The derivative of *S* is

$$\frac{dS}{d\beta} = -2a_1 + 2a_2\beta \qquad (4.6)$$

Setting the derivative to zero determines the least squares estimate of β which we denote as *b*. Setting (4.7) to zero,

$$-2a_1 + 2a_2 b = 0 \qquad (4.7)$$

Solving for *b* yields the formula for the least squares estimate,

$$b = \frac{a_1}{a_2} = \frac{\sum_{i=1}^{T} y_i}{T} = \bar{y} \tag{4.8}$$

Thus the least squares estimate of the population mean is the sample mean, $\bar{y}$. For the hip data in Table 1

$$b = \frac{\sum_{i=1}^{T} y_i}{T} = \frac{857.9094207}{50} = 17.158188 \simeq 17.158 \tag{4.9}$$

Thus we estimate that the average hip size in the population is 17.158 inches.