6. Estimating the Variance of a Population

We now have an excellent estimator of the mean of a population. The remaining task is to estimate the variance. If $Y \sim \left(\beta, \sigma^2\right)$, the variance is defined as

$$E\left[Y-\beta\right]^2 \qquad (6.1)$$

An expected value is an "average" of sorts, so if we knew β we could estimate the variance by using the sample analog of (6.1),

$$\tilde{\sigma}^2 = \frac{\sum\left(Y_i-\beta\right)^2}{T} \qquad (6.2)$$

Unfortunately we do not know β so we can not use (6.2). We can estimate β using the sample mean, so we might consider using the variance estimator

$$\tilde{\sigma}^2 = \frac{\sum\left(Y_i-b\right)^2}{T} = \frac{\sum\left(Y_i-\overline{Y}\right)^2}{T} \qquad (6.3)$$

The estimator in (6.3) is not a bad one, but it is not unbiased. To make it unbiased we divide by $T-1$ instead of T. That is, the "sample variance" is given by

$$\hat{\sigma}^2 = \frac{\sum\left(Y_i-b\right)^2}{T-1} = \frac{\sum\left(Y_i-\overline{Y}\right)^2}{T-1} \qquad (6.4)$$

We will not prove that $\hat{\sigma}^2$ is unbiased in general, however in the next sections we do derive its properties in the case in which the population is normal.

(6.1) The Chi-square Probability Distribution

Assume the population distribution is normal, $Y \sim N\left(\beta, \sigma^2\right)$. Given a random sample from this population, $Y_1, Y_2, \ldots, Y_T$, the standardized variables

$$Z_i = \frac{Y_i-\beta}{\sigma} \sim N\left(0,1\right)$$

Chi-square random variables arise when standard normal random variables are squared. Thus

$$V = \left[N\left(0,1\right)\right]^2 \sim \chi^2_{(1)}$$

V is said to have a chi-square distribution with one **degree of freedom**. If $Z_1, Z_2, \ldots, Z_m$ are independent $N(0,1)$ random variables, then

$$V = Z_1^2 + Z_2^2 + \cdots + Z_m^2 \sim \chi^2_{(m)} \qquad (6.5)$$

In this case V has a chi-square distribution with m degrees of freedom, indicating the number of independent normal random variables that are squared and summed to form V. In advanced statistics courses it is shown that the mean and variance of V are

$$\begin{aligned} E(V) &= m \\ \operatorname{var}(V) &= 2m \end{aligned} \tag{6.6}$$

In Figure 5 we depict the probability density function of chi-square random variables with various degrees of freedom. Note that since V is formed by *squaring* standard normal random variables, the values v of V are non-negative, $v \geq 0$. The distribution has a long tail to the right, and is said to be **skewed** to the right. As the degrees of freedom parameter m gets larger, the distribution becomes more "bell shaped." The chi-square distribution is important in econometrics as the basis of many test statistics.

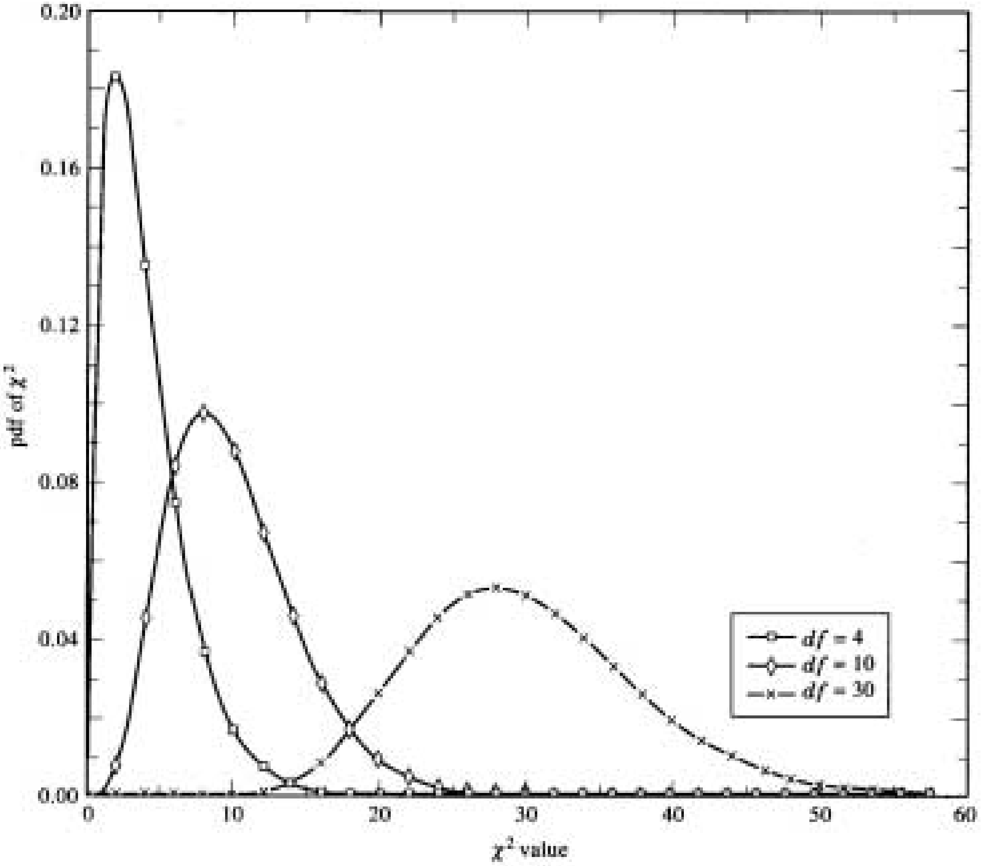

Figure 5 The Chi-square Distribution

(6.2) The Probability Distribution of $\hat{\sigma}^2$ When the Data are Normal

If we assume that $Y \sim N\left(\beta,\sigma^2\right)$, then we know that

$$Z_i = \frac{Y_i - \beta}{\sigma} \sim N\left(0,1\right)$$

Using (6.5) we also know that if the data are a random sample,

$$\left(\frac{Y_1 - \beta}{\sigma}\right)^2 + \left(\frac{Y_2 - \beta}{\sigma}\right)^2 + \cdots + \left(\frac{Y_T - \beta}{\sigma}\right)^2 \sim \chi^2_{(T)} \tag{6.7}$$

since each of the individual terms is the square of a $N(0,1)$ random variable. Unfortunately (6.7) includes the unknown parameter β. If we replace β by the least squares estimator b we obtain

$$V = \left(\frac{Y_1 - b}{\sigma}\right)^2 + \left(\frac{Y_2 - b}{\sigma}\right)^2 + \cdots + \left(\frac{Y_T - b}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum\left(Y_i - b\right)^2 = \frac{(T-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(T-1)} \tag{6.8}$$

The fact that we have used b in place of β causes the degrees of freedom to go down by one, since the terms in the sum (6.8) are no longer independent. Using (6.8) we can say that

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{T-1}V \tag{6.9}$$

where $V \sim \chi^2_{(T-1)}$. Then using (6.6) we can say that

$$E\left(\hat{\sigma}^2\right) \sim \frac{\sigma^2}{T-1}E\left(V\right) = \frac{\sigma^2}{T-1}\left(T-1\right) = \sigma^2 \tag{6.10}$$

which proves that in a normal population the sample variance is an unbiased estimator of the population variance. The sample variance is also unbiased in non-normal populations, but it no longer has a chi-square distribution in small samples.

Using the sample variance we can estimate the variance of the least squares estimator as

$$\hat{\text{var}}\left(b\right) = \hat{\sigma}^2 / T \tag{6.11}$$

The square root of the estimated variance is called the **standard error** of b,

$$se\left(b\right) = \sqrt{\hat{\text{var}}\left(b\right)} = \hat{\sigma}/\sqrt{T} \tag{6.12}$$

(6.3) The Hip Data Revisited

The sample variance for the hip data is

$$\hat{\sigma}^2 = \frac{\sum\left(y_i - b\right)^2}{T-1} = \frac{\sum\left(y_i - 17.158\right)^2}{49} = \frac{160.067}{49} = 3.267 \tag{6.13}$$

This means that the estimated variance of the sample mean is

$$\hat{\text{var}}(b) = \hat{\sigma}^2 / T = \frac{3.267}{50} = .0653 \tag{6.14}$$

How can we summarize what we have learned? Our estimates suggest that the hip size of U.S. adults is normally distributed with mean 17.158 inches with a variance of 3.267, $Y \sim N(17.158, 3.267)$. Based on this information, what is the answer to the question we posed in Section 6.1. If an airplane seat is 18 inches wide, what percentage of customers will not be able to fit? We can recast this question as asking, what is the probability that a randomly drawn person will have hips larger than 18 inches,

$$P(Y > 18) = P\left(\frac{Y-\beta}{\sigma} > \frac{18-\beta}{\sigma}\right) \tag{6.15}$$

We can estimate an answer to this question by replacing the unknown parameters by their estimates,

$$\widehat{P(Y > 18)} = P\left(\frac{Y-b}{\hat{\sigma}} > \frac{18-17.158}{1.807}\right) = P(Z > 0.465) = 0.32 \tag{6.16}$$

Based on our estimates, 32% of the population would not be able to fit into a seat 18 inches wide, which is a pretty hefty share.

How large would a seat have to be to fit 95% of the population?

$$\widehat{P(Y \le Y^*)} = P\left(\frac{Y-b}{\hat{\sigma}} \le \frac{Y^* - 17.158}{1.807}\right) = P\left(Z \le \frac{Y^* - 17.158}{1.807}\right) = 0.95 \tag{6.17}$$

Using your computer software, or the table of normal probabilities, the value of Z such that $P(Z \le z^*) = .95$ is z^* = 1.645. Then

$$\frac{Y^* - 17.158}{1.807} = 1.645 \Rightarrow Y^* = 20.131 \tag{6.18}$$

Thus to accommodate 95% of U.S. adult passengers, we estimate that the seats should be slightly greater than 20 inches wide.