

# 7

## Point Estimation

In this note we consider the problem of estimating an unknown population parameter  $\theta$  using the available data. Our approach, called *point estimation*, consists of choosing a number which supposedly represents our “best guess” about  $\theta$ .

### 7.1 INFERENCE ABOUT THE POPULATION MEAN

Usually, we draw a sample in order to make an inference about the underlying population. In order to understand what can be learned from a sample, consider the following situation.

Suppose that a population is known to have mean  $\mu$  and variance  $\sigma^2$ . Given a sample  $Z_1, \dots, Z_n$  from this population, it seems plausible to try to estimate  $\mu$  using the sample mean

$$\bar{Z} = \frac{1}{n}(Z_1 + \dots + Z_n) = \frac{1}{n} \sum_{i=1}^n Z_i.$$

The sample mean  $\bar{Z}$  is an example of *estimator*, namely a special type of sample statistic that is employed in order to estimate a population parameter. A particular value of an estimator, corresponding to a given sample, is called an *estimate*. Under repeated sampling, an estimator may be regarded as a random variable whose sampling distribution depends on three elements: (i) the precise form of the estimator, (ii) the probability distribution of the underlying population, and (iii) the way the data have been gathered.

How close will the estimator  $\bar{Z}$  be to the population mean  $\mu$ ? Because  $\bar{Z}$  is a random variable, let us try to derive its probability distribution, or *sampling distribution*, or at least its mean and variance. Notice that the sampling distribution is indeed necessary if we want to compute the degree of concentration of  $\bar{Z}$  about the population target  $\mu$  as measured by

$$\Pr(|\bar{Z} - \mu| \leq \epsilon),$$

for some  $\epsilon > 0$ . In order to do this we have two possibilities:

1. We may draw a number of samples from the given population. For each sample we may compute the sample mean  $\bar{Z}$ . We can then tabulate the frequency distribution or plot the histogram of the values of  $\bar{Z}$  thus obtained. If the number of samples is high, this method, called the *Monte Carlo method*, gives a good approximation to the sampling distribution of  $\bar{Z}$ . Further, the

average and the mean squared deviation of the values of  $\bar{Z}$  give a good approximation to the sampling mean and variance of  $\bar{Z}$ .

2. We can try to use the tools developed in the previous chapters to work out mathematically what the sampling distribution of  $\bar{Z}$  is, or at least its mean and variance. This method has two advantages. First, we may be able to obtain exact results and not just approximations as in the Monte Carlo case. Second, because of the analytical nature of the results, it is easier to carry out experiments of the “if, then” type by modifying the parameters of the problem.

### 7.1.1 SAMPLING MEAN AND VARIANCE OF $\bar{Z}$

We shall derive the sampling mean and the sampling variance of  $\bar{Z}$  under the assumptions that the data  $Z_1, \dots, Z_n$  are a simple random sample from a population whose variability is described by a probability distribution with mean  $\mu$  and finite variance  $\sigma^2$ . Thus, the observations in the sample are independently distributed and follow a common distribution with mean  $\mu$  and variance  $\sigma^2$ . We say, in this case, that the data  $Z_1, \dots, Z_n$  are a *random sample from a distribution with mean  $\mu$  and variance  $\sigma^2$* .

Because  $\bar{Z}$  is a linear sample statistic, that is, is a linear combination of  $Z_1, \dots, Z_n$ , its sampling mean is

$$E(\bar{Z}) = E\left(\frac{1}{n} \sum_{i=1}^n Z_i\right) = \frac{1}{n} \sum_{i=1}^n E(Z_i).$$

Because each observation has mean equal to  $\mu$ , we get

$$E(\bar{Z}) = \frac{1}{n} \sum_{i=1}^n \mu = \mu.$$

Since  $\bar{Z}$  is on average equal to the target parameter  $\mu$ ,  $\bar{Z}$  is said to be an *unbiased* estimator of the population mean  $\mu$ .

By (5.4), the sampling variance of  $\bar{Z}$  is

$$\begin{aligned} \text{Var}(\bar{Z}) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n Z_i\right) \\ &= \frac{1}{n^2} \left[ \sum_{i=1}^n \text{Var}(Z_i) + 2 \sum_{i=1}^n \sum_{j=i+1}^n \text{Cov}(Z_i, Z_j) \right]. \end{aligned}$$

The hypothesis of random sampling implies that the observations are mutually independent and therefore uncorrelated. Hence all covariance terms disappear from the above expression and we obtain

$$\text{Var}(\bar{Z}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Z_i).$$

The hypothesis of random sampling also implies that each observation has variance equal to  $\sigma^2$ . Hence, the sampling variance of  $\bar{Z}$  is

$$\text{Var}(\bar{Z}) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}. \quad (7.1)$$

The (positive) square root of  $\text{Var}(\bar{Z})$  is called the *standard error* of the sample mean, written  $\text{SE}(\bar{Z})$ . Under random sampling,  $\text{SE}(\bar{Z}) = \sigma/\sqrt{n}$ . If  $n > 1$ , then the sampling variance of  $\bar{Z}$  is smaller than the variance  $\text{Var}(Z_i)$  of each individual observation. Thus, the sample mean displays much less sampling variability than the individual sample elements. This occurs because averaging “washes out”, at least partly, some of the extreme values that may appear in a sample.

Notice that the variance of  $\bar{Z}$  tends to vanish as the sample size  $n$  increases. By Chebyshev inequality

$$\Pr(|\bar{Z} - \mu| \geq \epsilon) \leq \frac{\text{Var}(\bar{Z})}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2}.$$

Because  $\sigma^2/n \rightarrow 0$  as  $n \rightarrow \infty$ , it follows that  $\Pr(|\bar{Z} - \mu| \geq \epsilon) \rightarrow 0$  as  $n \rightarrow \infty$ , no matter how small  $\epsilon$  is. Thus, as the sample size grows large, the sampling distribution of  $\bar{Z}$  becomes more and more concentrated about the population parameter  $\mu$ . This result is known as the *Law of Large Numbers* and, because of this property,  $\bar{Z}$  is said to be a *consistent* estimator of the population mean  $\mu$ .

The fact that the sampling variance of  $\bar{Z}$  is equal to  $\sigma^2/n$  when the data are a random sample from a distribution with variance equal to  $\sigma^2$  is of considerable theoretical interest, but is only of practical usefulness if  $\sigma^2$  is known. When  $\sigma^2$  is not known, one may consider estimating the sampling variance of  $\bar{Z}$  by  $\widehat{\text{Var}}(Z_i)/n$ , where  $\widehat{\text{Var}}(Z_i)$  is some estimate of the population variance, such as the mean squared deviation

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})^2,$$

or the sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2.$$

If  $s^2$  is used, then an estimate of  $\text{Var}(\bar{Z})$  is

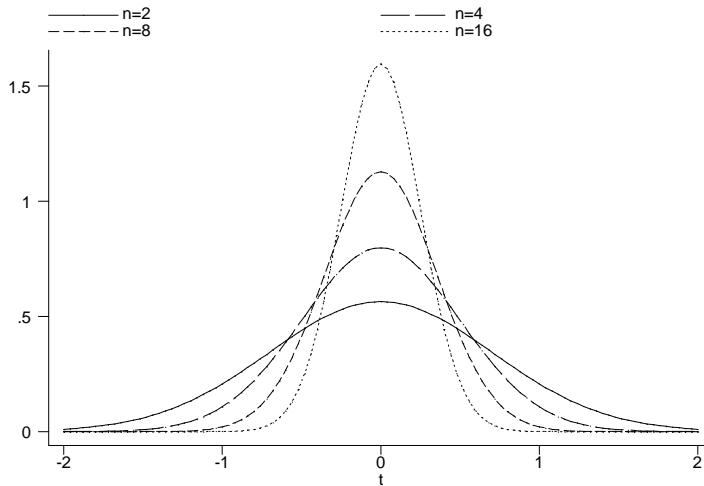
$$\widehat{\text{Var}}(\bar{Z}) = \frac{s^2}{n} = \frac{1}{n(n-1)} \sum_{i=1}^n (Z_i - \bar{Z})^2.$$

**Example 7.1** The Chebyshev bound on  $\Pr(|\bar{Z} - \mu| \geq \epsilon)$  cannot be computed unless  $\sigma^2$  is known. If we estimate  $\text{Var}(\bar{Z})$  by  $\widehat{\text{Var}}(\bar{Z}) = s^2/n$ , then an estimate of the bound is

$$\frac{\widehat{\text{Var}}(\bar{Z})}{\epsilon^2} = \frac{s^2/n}{\epsilon^2}.$$

The quality of this empirical version of the Chebyshev bound also depends on how good  $s^2$  is as an estimator of  $\sigma^2$ .  $\square$

**Figure 17** Sampling distribution of the sample mean for samples of size  $n = 2, 4, 8, 16$  under random sampling from a  $\mathcal{N}(0, 1)$  distribution.



### 7.1.2 SAMPLING DISTRIBUTION OF $\bar{Z}$

In general, obtaining the sampling distribution of  $\bar{Z}$  is not easy except in one important special case.

Thus suppose that the data  $Z_1, \dots, Z_n$  are a simple random sample from a population whose variability is described by a normal (Gaussian) distribution with mean  $\mu$  and finite variance  $\sigma^2$ . Thus, the observations in the sample are independently distributed, each with the same  $\mathcal{N}(\mu, \sigma^2)$  distribution, and we say that the data  $Z_1, \dots, Z_n$  are a *random sample from a  $\mathcal{N}(\mu, \sigma^2)$  distribution*.

In this case, since  $\bar{Z}$  is a linear combination of normally distributed random variables, its sampling distribution is itself normal with mean equal to mean  $\mu$  and variance equal to  $\sigma^2/n$ , that is,

$$\bar{Z} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

(Figure 17) or, equivalently,

$$\frac{\sqrt{n}(\bar{Z} - \mu)}{\sigma} \sim \mathcal{N}(0, 1). \quad (7.2)$$

It then follows that

$$\begin{aligned}\Pr(|\bar{Z} - \mu| \leq \epsilon) &= \Pr\left(\frac{\sqrt{n}|\bar{Z} - \mu|}{\sigma} \leq \frac{\epsilon\sqrt{n}}{\sigma}\right) \\ &= \Pr\left(-\frac{\epsilon\sqrt{n}}{\sigma} \leq X \leq \frac{\epsilon\sqrt{n}}{\sigma}\right) \\ &= 2\Pr\left(0 \leq X \leq \frac{\epsilon\sqrt{n}}{\sigma}\right),\end{aligned}$$

where  $X \sim \mathcal{N}(0,1)$ . This probability can easily be computed using the normal probability tables.

### 7.1.3 THE CENTRAL LIMIT THEOREM

It is a remarkable result from the theory of probability that, under simple random sampling from a population with finite variance, (7.2) also holds approximately, even if the population is not Gaussian, provided that the sample size  $n$  is large enough (say, at least  $n \geq 30$ ). This is known as the *Central Limit Theorem (CLT)*.

Thus, if the observations are randomly drawn from the same population, one that has mean equal to  $\mu$  and finite variance equal to  $\sigma^2$  but is not necessarily Gaussian (it may not even be symmetric, nor unimodal, nor continuous), then (7.2) holds approximately provided that the sample size  $n$  is large enough. This means that the sampling distribution of  $\bar{Z}$  is well approximated by a Gaussian distribution with mean equal to  $\mu$  and variance equal to  $\sigma^2/n$ . Hence, probabilities such as  $\Pr(|\bar{Z} - \mu| \leq \epsilon)$  may be approximated using the normal probability tables for, by the CLT,

$$\Pr(|\bar{Z} - \mu| \leq \epsilon) \approx 2\Pr\left(0 \leq X \leq \frac{\epsilon\sqrt{n}}{\sigma}\right),$$

where  $X$  is a standard normal random variable. Further, the quality of this approximations increases with the sample size.

### 7.1.4 NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

We now study an important example of application of the CLT. Besides being of interest by itself, this example provides an illustration of the power and usefulness of the normal approximation implied by the CLT.

Consider a random variable  $Z$  that can only take two values, 0 (“failure”) and 1 (“success”). The distribution function of  $Z$  is completely specified by the probability of success

$$\Pr(Z = 1) = \pi, \quad 0 < \pi < 1. \quad (7.3)$$

Clearly,  $\Pr(Z = 0) = 1 - \Pr(Z = 1) = 1 - \pi$ . Such a random variable is called a *Bernoulli random variable* and its distribution is simply a binomial with index 1 and parameter  $\pi$ .

Using (7.3), let us first compute the mean and the variance of  $Z$ . Since  $Z$  can only take the values 0 and 1 we have

$$E(Z) = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi,$$

Thus, the mean of  $Z$  coincides with the probability of success. Further

$$\begin{aligned}\text{Var}(Z) &= E(Z^2) - [E(Z)]^2 \\ &= 1 \cdot \pi + 0 \cdot (1 - \pi) - \pi^2 \\ &= \pi - \pi^2 = \pi(1 - \pi).\end{aligned}$$

Now suppose that we draw a random sample of size  $n = 10$  from the population represented by the random variable  $Z$ . A possible sample consists of a sequence of 0 and 1, for example:

$$\begin{array}{cccccccccc} Z_1 & Z_2 & Z_3 & Z_4 & Z_5 & Z_6 & Z_7 & Z_8 & Z_9 & Z_{10} \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \end{array}$$

Because sampling is at random, each of the observations  $Z_i$  may be viewed as a replica of the basic random variable  $Z$ .

The *sample proportion* of successes is given by

$$P = \frac{\text{no. of successes}}{n} = \frac{1}{n} \sum_{i=1}^n Z_i = \bar{Z}.$$

Thus, the sample proportion  $P$  is just the sample average of observations that can only take the values 0 and 1. It then follows that the sampling mean and variance of  $P$  are

$$E(P) = E(Z) = \pi, \quad \text{Var}(P) = \frac{\text{Var}(Z)}{n} = \frac{\pi(1 - \pi)}{n}. \quad (7.4)$$

It is clear, on the other hand, that the sampling distribution of  $P$  cannot be normal, because the sample observations are definitely not normal. The exact distribution of  $P$  can in principle be obtained from the fact that

$$S = \text{“no. of successes in } n \text{ Bernoulli trials”} = nP,$$

where  $S$  has a binomial distribution with index  $n$  and parameter  $\pi$ . If  $p = s/n$ , then

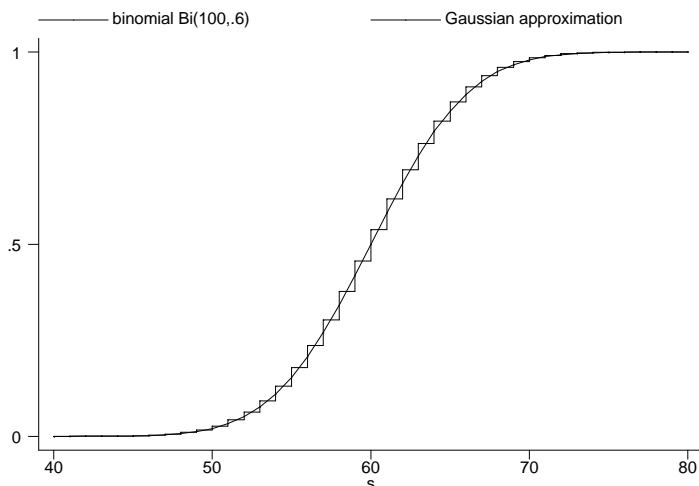
$$\Pr(P = p) = \Pr(S = s) = \binom{n}{s} \pi^s (1 - \pi)^{n-s}$$

and

$$\begin{aligned}\Pr(P \leq p) &= \Pr(S \leq s) = \Pr(S = 0) + \cdots + \Pr(S = s) \\ &= \sum_{j=0}^s \binom{n}{j} \pi^j (1 - \pi)^{n-j},\end{aligned} \quad (7.5)$$

for  $s = 0, 1, \dots, n$ . These probabilities may be difficult to compute if  $n$  is large or the binomial tables are not readily available. In these cases, one can use the CLT to find a simple approximation to (7.5).

**Figure 18** Distribution function of the binomial distribution with  $n = 100$  and  $\pi = .6$  and normal approximation.



If  $n$  is large enough, then the CLT implies that the sampling distribution of  $P$  is well approximated by a Gaussian distribution with mean equal to  $E(Z) = \pi$  and variance equal to  $\text{Var}(Z)/n = \pi(1 - \pi)/n$ . Therefore, if  $n$  is large enough, we get

$$\begin{aligned} \Pr(P \leq p) &= \Pr\left(\frac{P - \pi}{\sqrt{\pi(1 - \pi)/n}} \leq \frac{p - \pi}{\sqrt{\pi(1 - \pi)/n}}\right) \\ &\approx \Pr\left(X \leq \frac{\sqrt{n}(p - \pi)}{\sqrt{\pi(1 - \pi)}}\right), \end{aligned}$$

where  $X \sim \mathcal{N}(0, 1)$ . The probability on the right-hand side is easily computed from the normal probability tables. The normal approximation works particularly well if the probability of success  $\pi$  is not too close to 0 or 1.

Notice that  $S = nP$  is just a linear transformation of the random variable  $P$ . It then follows from (7.4) that

$$E(S) = nE(P) = n\pi,$$

and

$$\text{Var}(S) = n^2 \text{Var}(P) = n^2 \frac{\pi(1 - \pi)}{n} = n\pi(1 - \pi).$$

Finally, provided that  $n$  is large and  $\pi$  not too close to 0 or 1, we have that the binomial distribution of  $S$  is well approximated by a Gaussian with mean equal to  $n\pi$  and variance equal to  $n\pi(1 - \pi)$  (Figure 18). We can therefore approximate the

binomial probability  $\Pr(S \leq s)$  using the fact that

$$\begin{aligned} \Pr(S \leq s) &= \Pr\left(\frac{S - n\pi}{\sqrt{n\pi(1-\pi)}} \leq \frac{s - n\pi}{\sqrt{n\pi(1-\pi)}}\right) \\ &\approx \Pr\left(X \leq \frac{s - n\pi}{\sqrt{n\pi(1-\pi)}}\right), \end{aligned}$$

where  $X \sim \mathcal{N}(0, 1)$ . This probability can also be computed very easily from the normal probability tables.

## 7.2 SELECTING BETWEEN ESTIMATORS

Consider a population which is symmetrically distributed about its mean  $\mu$ . Given data  $Z_1, \dots, Z_n$ , the sample mean  $\bar{Z}$  is one of the possible estimators of  $\mu$ . Although the value of the sample mean depends on the data, the sample observations are combined according to a rather simple rule, namely

$$\bar{Z} = g(Z_1, \dots, Z_n) = n^{-1} \sum_{i=1}^n Z_i,$$

where the function  $g(\cdot)$  is linear. This fact was crucial in order to derive the sampling distribution of  $\bar{Z}$ .

Of course  $\bar{Z}$  is not the only estimate of  $\mu$  that we may consider. Another possibility is to use instead the sample median  $\tilde{Z}$ . If there are  $n$  observations, this is obtained by first ordering the observations as

$$Z_{[1]} \leq Z_{[2]} \leq \dots \leq Z_{[n]},$$

where  $Z_{[i]}$  denotes the  $i$ -th ordered value, and then putting

$$\tilde{Z} = \begin{cases} Z_{[(n+1)/2]}, & \text{if } n \text{ is odd,} \\ (Z_{[n/2]} + Z_{[(n/2)+1]})/2, & \text{if } n \text{ is even.} \end{cases}$$

Thus, the sample median is a nonlinear function of the data and this function is rather complicated to describe. For this reason, the sampling distribution of  $\tilde{Z}$  is not as easy to derive as that of  $\bar{Z}$ . In any case, both  $\bar{Z}$  and  $\tilde{Z}$  share the feature that their sampling distribution becomes more and more concentrated about  $\mu$  as the sample size  $n$  gets larger and larger. Which of the two estimators should we select?

More generally, suppose that we are interested in estimating a population parameter  $\theta$ , and that several possible estimators (that is, functions of the data) are available. Which one do we select? The next sections introduce some criteria for selecting among alternative estimators.

### 7.2.1 UNBIASEDNESS

Let  $\hat{\theta}$  be an estimator of a population parameter  $\theta$ . We say that  $\hat{\theta}$  is unbiased for  $\theta$  if, no matter what  $\theta$  is,

$$\mathbf{E}(\hat{\theta}) = \theta,$$

where  $\mathbf{E}(\hat{\theta})$  is the sampling mean of  $\hat{\theta}$ .

**Example 7.2** If the data are a random sample from a population with a finite mean  $\mu$ , then the sample mean is an unbiased estimator of  $\mu$ . If the population distribution is symmetric about  $\mu$ , then the sample median can also be shown to be an unbiased estimator of  $\mu$ .  $\square$

If  $E(\hat{\theta}) \neq \theta$ , then we say that  $\hat{\theta}$  is biased, and the difference

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta,$$

is called the *bias* of  $\hat{\theta}$  as an estimator of  $\theta$ . In particular, if  $\text{Bias}(\hat{\theta}) < 0$ , that is,  $E(\hat{\theta}) < \theta$ , then we say that  $\hat{\theta}$  is a *downward biased* estimator of  $\theta$ , or that it displays a negative bias, or that it tends to underestimate  $\theta$ . We have a corresponding terminology if  $\text{Bias}(\hat{\theta}) > 0$ .

**Example 7.3** It can be shown that, under random sampling from a population with a finite variance  $\sigma^2$ , the mean squared deviation

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (Z_i - \bar{Z})^2$$

is a downward biased estimator of  $\sigma^2$ , for

$$E(\hat{\sigma}^2) = \left(1 - \frac{1}{n}\right) \sigma^2 < \sigma^2 \quad (7.6)$$

and therefore

$$\text{Bias}(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2 = -\frac{\sigma^2}{n} < 0.$$

This implies that using  $\hat{\sigma}^2/n$  as an estimator of  $\sigma^2/n$  tends to underestimate the true sampling variability of the sample mean.

Finding an unbiased estimator of  $\sigma^2$  is, in this case, straightforward. Because (7.6) implies that

$$\frac{n}{n-1} E(\hat{\sigma}^2) = \sigma^2,$$

the sample variance

$$s^2 = \frac{n}{n-1} \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2,$$

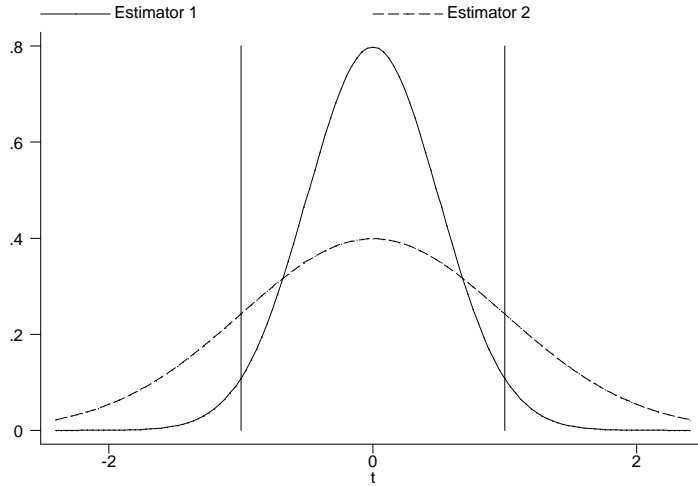
is clearly unbiased for  $\sigma^2$ . This explains why  $s^2$  is often preferred to the mean squared deviation as a measure of dispersion.  $\square$

### 7.2.2 EFFICIENCY

Suppose that there are two unbiased estimators  $\hat{\theta}$  and  $\tilde{\theta}$  of the same population parameter  $\theta$ , that is,

$$E(\hat{\theta}) = E(\tilde{\theta}) = \theta.$$

**Figure 19** Sampling distribution of two unbiased estimators of the same population parameter  $\theta = 0$ .



In this situation, it seems reasonable to select the estimator which, for a given sample size  $n$ , has smaller sampling variance. If

$$\text{Var}(\hat{\theta}) < \text{Var}(\tilde{\theta}), \quad (7.7)$$

then we say that  $\hat{\theta}$  is *efficient relative to*  $\tilde{\theta}$ . As it is clear from Figure 19, the efficiency criterion (7.7) lead to selecting the first estimator ( $\hat{\theta}$ ) because, for any  $\epsilon > 0$ , we have

$$\Pr(|\hat{\theta} - \theta| \leq \epsilon) > \Pr(|\tilde{\theta} - \theta| \leq \epsilon).$$

Instead of using the criterion (7.7), we may equivalently compare  $\hat{\theta}$  and  $\tilde{\theta}$  using the ratio

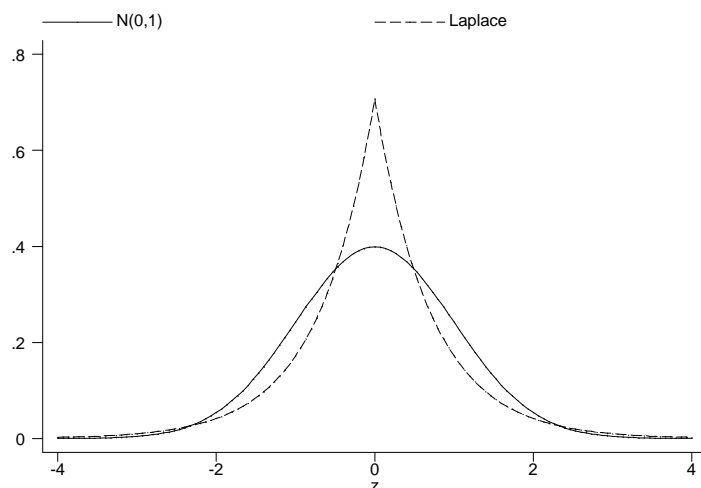
$$\text{Eff}(\hat{\theta}, \tilde{\theta}) = \frac{\text{Var}(\tilde{\theta})}{\text{Var}(\hat{\theta})},$$

called the *relative efficiency of  $\hat{\theta}$  compared with  $\tilde{\theta}$* . Clearly,  $\hat{\theta}$  is efficient relative to  $\tilde{\theta}$  if  $\text{Eff}(\hat{\theta}, \tilde{\theta}) > 1$ .

The variance of an estimator generally depends on the sample size  $n$  and decreases as  $n$  gets larger. Hence,  $\text{Eff}(\hat{\theta}, \tilde{\theta})$  measures the ratio of the sample sizes needed for the two estimators to have the same variance. Thus, suppose that  $\text{Var}(\hat{\theta}) = \text{Var}(\tilde{\theta})$  when  $\hat{\theta}$  is based on a sample of  $n_1$  observations and  $\tilde{\theta}$  is based on a sample of  $n_2$  observations. If  $\text{Eff}(\hat{\theta}, \tilde{\theta}) > 1$ , then we must have that  $n_2 < n_1$ .

Because the sample mean is the “natural” unbiased estimator of the population mean  $\mu$ , it is interesting to ask whether it is efficient relative to other alternative estimators of  $\mu$ . It can be shown that, if the data are a random sample from a  $\mathcal{N}(\mu, \sigma^2)$

**Figure 20** Densities of the standard Gaussian and the Laplace distribution with zero mean and unit variance.



distribution, then the sample mean  $\bar{Z}$  has smaller sampling variance than any other unbiased estimator of  $\mu$ , that is, the sample mean is the most efficient of all unbiased estimators of the mean of a Gaussian population.

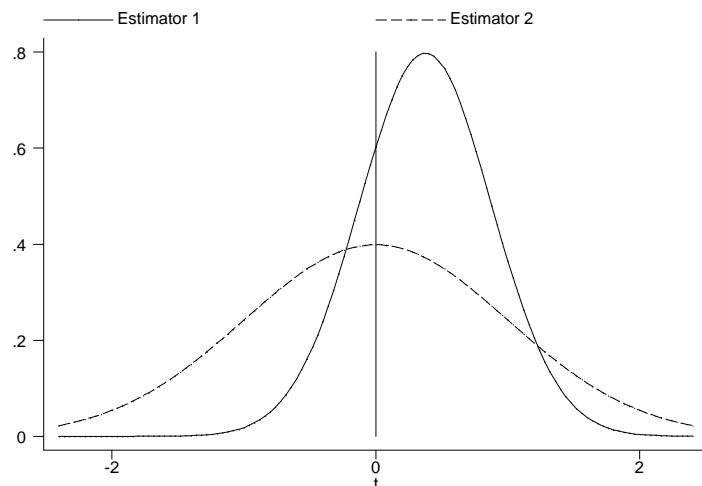
**Example 7.4** Because  $\bar{Z}$  is the most efficient of all unbiased estimators of the population mean  $\mu$  when the data are a random sample from a  $\mathcal{N}(\mu, \sigma^2)$  distribution, it has smaller sampling variance than the sample median  $\tilde{Z}$ , that is,  $\text{Eff}(\bar{Z}, \tilde{Z}) > 1$ . This result is completely reversed if the data are instead a random sample from a *double exponential* or *Laplace distribution*, that is, a continuous distribution with density function of the form

$$f(z) \propto \exp(-|z - \mu|),$$

where  $\propto$  means “proportional to”. The Laplace is also symmetric about its mean  $\mu$ , but unlike the normal distribution has a spike at  $\mu$  and has somewhat fatter tails, which makes the occurrence of outliers more likely (Figure 20). In this case we have  $\text{Var}(\tilde{Z}) > \text{Var}(\bar{Z})$  or, equivalently,  $\text{Eff}(\bar{Z}, \tilde{Z}) < 1$ .  $\square$

If the data do not come from a Gaussian distribution, then the sample mean  $\bar{Z}$  can only be shown to satisfy a much weaker property (known as the *Gauss–Markov theorem*), namely  $\bar{Z}$  has smaller sampling variance than any other estimator  $\hat{\mu}$  of the population mean  $\mu$  that is unbiased and linear, that is, of the form  $\hat{\mu} = \sum_i w_i Z_i$ , where  $w_1, \dots, w_n$  are fixed weights (this excludes the median or any trimmed mean). Because of this property, the sample mean is often said to be *Best Linear Unbiased (BLU)* for the population mean  $\mu$ .

**Figure 21** Efficiency comparisons between a biased and an unbiased estimator of the same population parameter  $\theta = 0$ .



### 7.2.3 MEAN SQUARED ERROR

Sometimes it may happen that an estimator  $\tilde{\theta}$  is slightly biased, but has a smaller variance than another unbiased estimator  $\hat{\theta}$  (Figure 21). In this case, we may want to select  $\tilde{\theta}$  because, although it is slightly biased, it is much less spread out than  $\hat{\theta}$ . In fact, in this case we may have

$$\Pr(|\hat{\theta} - \theta| \leq \epsilon) < \Pr(|\tilde{\theta} - \theta| \leq \epsilon).$$

A measure that combines together the bias and the sampling variance of an estimator is its *mean squared error* (MSE)

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= \text{E}[(\hat{\theta} - \theta)^2] \\ &= \text{E}[(\hat{\theta} - \text{E}(\hat{\theta}))^2] + [\text{E}(\hat{\theta}) - \theta]^2 \\ &= \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2, \end{aligned}$$

since  $\text{E}[\hat{\theta} - \text{E}(\hat{\theta})] = 0$ . If  $\hat{\theta}$  is unbiased, then clearly  $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta})$ .

Given two estimators,  $\hat{\theta}$  and  $\tilde{\theta}$ , the *MSE criterion* leads us to select  $\tilde{\theta}$  if, for a given sample size  $n$ ,

$$\text{MSE}(\hat{\theta}) > \text{MSE}(\tilde{\theta}),$$

that is, if

$$\text{Var}(\hat{\theta}) - \text{Var}(\tilde{\theta}) > \text{Bias}(\tilde{\theta})^2 - \text{Bias}(\hat{\theta})^2.$$

If both estimators are unbiased for  $\theta$ , this is just the criterion based on the sampling variance. If  $\hat{\theta}$  is unbiased for  $\theta$  but  $\tilde{\theta}$  is not, then we would still choose  $\tilde{\theta}$  if

$$\text{Var}(\hat{\theta}) - \text{Var}(\tilde{\theta}) > \text{Bias}(\tilde{\theta})^2.$$

The MSE criterion is equivalent to comparing  $\hat{\theta}$  and  $\tilde{\theta}$  on the basis of the ratio

$$\text{Eff}(\hat{\theta}, \tilde{\theta}) = \frac{\text{MSE}(\tilde{\theta})}{\text{MSE}(\hat{\theta})},$$

and selecting  $\tilde{\theta}$  if  $\text{Eff}(\hat{\theta}, \tilde{\theta}) < 1$ .

**Example 7.5** If the data  $Z_1, \dots, Z_n$  are a random sample of size  $n \geq 2$  from a  $\mathcal{N}(\mu, \sigma^2)$  distribution, then the sample variance  $s^2$  is an unbiased estimator of  $\sigma^2$ , and its MSE can be shown to be

$$\text{MSE}(s^2) = \text{Var}(s^2) = \frac{2\sigma^4}{n-1}. \quad (7.8)$$

From Example 7.3, the bias of the mean squared deviation  $\hat{\sigma}^2 = (1 - n^{-1})s^2$  is

$$\text{Bias}(\hat{\sigma}^2) = -\frac{\sigma^2}{n}.$$

It also follows from (7.8) that

$$\text{Var}(\hat{\sigma}^2) = \left(1 - \frac{1}{n}\right)^2 \text{Var}(s^2) = \left(1 - \frac{1}{n}\right) \frac{2\sigma^4}{n}.$$

Notice that, although biased,  $\hat{\sigma}^2$  has smaller variance than  $s^2$ . The MSE of  $\hat{\sigma}^2$  is

$$\text{MSE}(\hat{\sigma}^2) = \left(1 - \frac{1}{n}\right) \frac{2\sigma^4}{n} + \frac{\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}.$$

Because

$$\frac{(2n-1)(n-1)}{n^2} = \left(2 - \frac{1}{n}\right) \left(1 - \frac{1}{n}\right) < 2,$$

we have that  $\text{MSE}(\hat{\sigma}^2) < \text{MSE}(s^2)$ . Hence, the MSE criterion leads to selecting  $\hat{\sigma}^2$  over  $s^2$ .  $\square$

#### 7.2.4 ROBUSTNESS

Loosely speaking, an estimator is *robust* if its value changes little under small changes in the data. This is an important property because it ensures that the value of an estimator cannot be entirely dominated by a few data points. In turns, this ensures some protection against outliers and gross-errors in the data.

We have already seen that the sample mean is not robust, because changing a fraction of the data equal to  $1/n$  is enough to completely alter its value. We have also seen that one way of robustifying the mean is to use a symmetrically trimmed mean. The sample median is an extreme version of trimmed mean, corresponding to symmetrically trimming about 50 percent of the data on either side, which ensures the highest degree of robustness.

### 7.3 ASYMPTOTIC PROPERTIES

Often the sampling distribution of an estimator, or even simpler properties such as unbiasedness, are difficult to establish in finite samples and one relies on approximations valid for large sample sizes. Properties valid for arbitrarily large sample sizes are called *asymptotic*.

We now introduce a few properties that are often required from a sequence  $\{\hat{\theta}_n\} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \dots, \hat{\theta}_n, \dots)$  of estimators, corresponding to increasing sample sizes.

#### 7.3.1 ASYMPTOTIC UNBIASEDNESS

We say that a sequence  $\{\hat{\theta}_n\}$  of estimators is *asymptotically unbiased* for  $\theta$ , or simply that  $\hat{\theta}_n$  is asymptotically unbiased for  $\theta$ , if  $E(\hat{\theta}_n) \rightarrow \theta$  as  $n \rightarrow \infty$  or, equivalently,  $\text{Bias}(\hat{\theta}_n) \rightarrow 0$  as  $n \rightarrow \infty$ .

**Example 7.6** Under random sampling from a population with finite variance  $\sigma^2$ , the bias of the mean squared deviation  $\hat{\sigma}^2$  is equal to  $-\sigma^2/n$  and goes to zero as  $n \rightarrow \infty$ . Hence, the mean squared deviation  $\hat{\sigma}^2$  is asymptotically unbiased for the population variance  $\sigma^2$ .  $\square$

#### 7.3.2 CONSISTENCY

Consider the probability that  $|\hat{\theta} - \theta| > \epsilon$ , where  $\epsilon > 0$  may be arbitrary small. If

$$\Pr(|\hat{\theta} - \theta| > \epsilon) \rightarrow 0,$$

as  $n \rightarrow \infty$ , no matter how small  $\epsilon$  is, then we say that the sequence  $\{\hat{\theta}_n\}$  is *consistent* for  $\theta$ , or simply that  $\hat{\theta}_n$  is consistent for  $\theta$ , written  $\hat{\theta}_n \xrightarrow{P} \theta$  or  $\text{plim } \hat{\theta}_n = \theta$ .

From Chebyshev inequality

$$\Pr(|\hat{\theta}_n - E(\hat{\theta}_n)| > \epsilon) \leq \frac{\text{Var}(\hat{\theta}_n)}{\epsilon^2},$$

no matter how small  $\epsilon$  is. Hence, sufficient conditions for  $\hat{\theta}_n$  to be consistent for  $\theta$  are:

1.  $E(\hat{\theta}_n) \rightarrow \theta$  as  $n \rightarrow \infty$ , that is,  $\hat{\theta}_n$  is asymptotically unbiased,
2.  $\text{Var}(\hat{\theta}_n) \rightarrow 0$  as  $n \rightarrow \infty$ .

**Example 7.7** Consider the behavior of the sample mean, the sample variance and the sample mean square deviation under random sampling from a population with mean  $\mu$  and finite variance  $\sigma^2$ . We already know that the sample mean is consistent for the population mean  $\mu$ . Because it is unbiased and its sampling variance is equal to  $2\sigma^4/(n-1)$ , the sample variance is consistent for the population variance  $\sigma^2$ . Finally, because it is asymptotically unbiased and its sampling variance is equal to  $2\sigma^2(1-n^{-1})/n$ , the mean squared deviation is also consistent for  $\sigma^2$ .  $\square$

#### 7.3.3 ASYMPTOTIC NORMALITY

We have already seen that if the data  $Z_1, \dots, Z_n$  are a random sample from a distribution that has mean  $\mu$  and variance  $\sigma^2$  but is not necessarily normal, then

the CLT implies that  $\sqrt{n}(\bar{Z} - \mu)/\sigma$  is approximately distributed as  $\mathcal{N}(0, 1)$  in large samples. This is often written as

$$\frac{\sqrt{n}(\bar{Z} - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1).$$

In general, we say that a sequence  $\{\hat{\theta}_n\}$  of estimators is *asymptotically normal* with asymptotic mean  $\theta$ , if there exists a positive number  $V$ , called the *asymptotic variance* of  $\hat{\theta}_n$ , such that

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{V}} \xrightarrow{d} \mathcal{N}(0, 1).$$

This means that, if the sample size  $n$  is large enough, the sampling distribution of  $\hat{\theta}_n$  is well approximated by a Gaussian distribution with mean equal to  $\theta$  and variance equal to  $V/n$ , called the *asymptotic distribution* of  $\hat{\theta}_n$ . When the exact distribution of  $\hat{\theta}_n$  is hard to derive but  $n$  is large, this Gaussian distribution provides a simple and useful approximation.

