ESTIMATION

1 INTRODUCTION

Estimation is the statistical process of finding an approximate value for a *population parameter*. A population parameter is a characteristic of the distribution of a population such as the *population mean*, the *population variance* or *population proportion*. When no information is available on the parameter under investigation, a sample has to be selected from the population in order to obtain some idea of the value of this parameter. Obviously, we are assuming that a census would be not only impractical but also impossible, given that the population size is of infinite magnitude. The sampling method to be adopted depends on the structure of the population and the sample is to be chosen so as to be as unbiased as possible. It should contain all, if not most of, the characteristics of its parent population. It has to be mentioned that it is very hard to select the perfect sample, as it is impossible to eliminate sampling errors completely.

It is therefore evident that a *sample statistic* will always deviate from its corresponding parameter. A sample statistic is any function of observed data, especially used to estimate a parameter – for example, the sample mean and the sample variance. There are two ways of estimating a population parameter: *point* and *interval estimation*.

2 PROPERTIES OF GOOD ESTIMATORS

A *point estimator* is a single-valued sample statistic which is used to approximate a population parameter. A question of interest is: which statistic should one use to estimate a parameter?

For example, suppose we want to estimate the population mean μ. Should we use the *sample mean* $\bar{x}$, the *median* or the *mode*? The solution is to pick the statistic that tends to produce an estimate closest to the true value. This can be expected to occur if the *estimator* possesses four properties which we will discuss in terms of population and sample means.

2.1 Unbiasedness

An estimator is said to be *unbiased* for the parameter estimated if it is 'centered at the right spot'. Mathematically, the average value of the estimator should be equal to the parameter that it is estimating. In mathematical notation, if the sample statistic T is an unbiased estimator of the population parameter θ, then E[T] = θ. Many estimators are *asymptotically unbiased* in the sense that the biases reduce to practically insignificant values (close to zero) when n becomes sufficiently large. The estimator s^2, the sample variance, is such an example, as will be seen later.

2.2 Consistency

If an estimator *approaches* the parameter it is estimating as the sample size *n* increases, it is then said to be *consistent*. Stated more rigorously, an estimator is said to be consistent if, as n approaches infinity, the probability that it will differ from the parameter is not more than an arbitrary small constant.

For instance, the sample mean is an unbiased estimator of μ, no matter what form the population distribution assumes, while the sample median is unbiased only if the distribution is symmetrical.

In case of large samples, consistency is a desirable property for an estimator to possess. However, in small samples, consistency is of little importance unless the limit of the probability defining consistency is reached even with a relatively small size of the sample.

2.3 Efficiency

The concept of *efficiency* refers to the sampling variability of an estimator. If two competing estimators are both unbiased, the one with smaller variance (for a given sample size) is said to be relatively more *efficient*. Stated in a somewhat different language, estimator *T* is said to be more efficient estimator *U* if the variance of the first is less than that of the second. The smaller the variance of the estimator, the more concentrated is the distribution of the estimator around the parameter being estimated and, therefore, the better this estimator is.

For example, if the population is symmetrically distributed, then both the sample mean and the sample median are consistent and unbiased estimators of μ. Yet the sample mean is better than the sample median as an estimator of μ since it is more efficient.

2.4 Sufficiency

An estimator is said to be *sufficient* if it conveys as much information as is possible about the parameter which is contained in the sample. The significance of sufficiency lies in the fact that, if a sufficient estimator exists, it is absolutely unnecessary to consider any other estimator; a sufficient estimator ensures that all information a sample can furnish with respect to the estimation of a parameter is being utilised.

Many methods have been devised for estimating parameters that may provide estimators satisfying these properties. The two most important methods are the least-squares (*OLS*) and the maximum likelihood (*MLE*).

3 POINT ESTIMATION

In this course, we shall find the point estimators for the population *mean* μ, *variance* σ^2 and *proportion p*. The derivation of these estimators will require a basic knowledge of the properties of *expectation* and variance (read Sections 4.2.1 and 4.2.2 below); the reader might even find these derivations quite mathematically intricate sometimes.

3.1 Properties of expectation

The *expectation* of a random variable is just its *arithmetic mean* or *average*. The expectation of X is denoted by $E[X]$ and defined by

$$E[X] = \sum xP[X = x]$$

Given a constant c and random variables X and Y, then

1. $E[c] = c$
2. $E[cX] = cE[X]$
3. $E[X \pm Y] = E[X] \pm E[Y]$

Note 1. $E[XY] \neq E[X] \times E[Y]$ except if X and Y are *independent*.

2. $E\left[\frac{X}{Y}\right] \neq \frac{E[X]}{E[Y]}$

3.2 Properties of variance

The *variance* of a random variable is a measure of its *spread* or *dispersion*. The variance of X is denoted by $\text{var}[X]$ and defined as

$$\text{var}[X] = E[X^2] - \left(E[X]\right)^2$$

Given a constant c and random variables X and Y, then

1. $\text{var}[c] = 0$
2. $\text{var}[cX] = c^2 \text{ var}[X]$
3. $\text{var}[X \pm Y] = E = \text{var}[X] + \text{var}[Y]$ only if X and Y are *independent.*

Note 1. $\text{var}[XY] \neq \text{var}[X] \times \text{var}[Y]$

2. $\text{var}\left[\frac{X}{Y}\right] \neq \frac{\text{var}[X]}{\text{var}[Y]}$

3.3 Assumptions

During the coming derivations, the following assumptions will be made:

1. The population variable is X where $E[X] = \mu$ and $\text{var}[X] = \sigma^2$.
2. Each observation from the set $\{x_1.x_2,\ \ ,x_n\}$ is *independently* and *identically* distributed (i.i.d), that is, for any observation x_i, $E[X_i] = \mu$ and $\text{var}[X_i] = \sigma^2$ just like for the population variable.

Note that these assumptions are extremely important and will have to be remembered during the derivations.

3.4 Estimation of the population mean μ

If we wish to have an idea of the value of the population mean μ, it is natural to select a sample and calculate the sample mean $\bar{x}$. These values should not be differing by much if the sample is unbiased. Can we conclude that the population mean is very close to $\bar{x}$? Can we use $\bar{x}$ as a substitution for μ whenever the latter is unknown? First, we must show that the sample mean is an *unbiased* estimator for the population mean by proving that $E[\ \bar{X}\ ] =\ \ \mu$.

From definition, $\bar{x} = \dfrac{\sum x}{n}$. We therefore find its expectation as follows:

$$E\left[\frac{\sum X}{n}\right] = \frac{1}{n}E\left[\sum X\right] = \frac{1}{n}E[X_1 + X_2 + ... + X_n]$$
$$= \frac{1}{n}\left\{E[X_1] + E[X_2] + ... + E[X_n]\right\}$$
$$= \frac{1}{n}\left\{\mu + \mu + ... + \mu\right\} = \frac{1}{n}(n\mu) = \mu.$$

We can thus conclude that $\bar{x}$ is an *unbiased point estimator* of μ. In plain and simple English, it means that, whenever the population mean μ is unknown, we may select a sample (as unbiased as possible), calculate its sample mean $\bar{x}$ and consider it as a worthy replacement for μ.

3.5 Estimation of the population variance σ^2

The *sample variance* s^2 seems to be an ideal candidate for being the point estimator of the population variance σ^2. It remains to be checked whether s^2 is unbiased for σ^2.

From the chapter on *Descriptive Statistics*, we know that

$$s^2 = \frac{\sum x^2}{n} - \bar{x}^2$$

The expectation of s^2 is derived as follows:

$$E\left[s^2\right] = E\left[\frac{\sum X^2}{n} - \bar{X}^2\right] = E\left[\frac{\sum X^2}{n}\right] - E\left[\bar{X}^2\right]$$

The right-hand side of the above will be split in order to find the expectation of each term.

Let us look at the <u>*first term*</u>:

$$E\left[\frac{\sum X^2}{n}\right] = \frac{1}{n}E\left[\sum X^2\right] = \frac{1}{n}E\left[X_1^2 + X_2^2 + \ldots + X_n^2\right]$$
$$= \frac{1}{n}\left\{E\left[X_1^2\right] + E\left[X_2^2\right] + \ldots + E\left[X_n^2\right]\right\}$$

Using the definition of variance (in terms of expectation, Section 4.3.2), we have

$\text{var}\left[X_i\right] = E\left[X_i^2\right] - \left(E\left[X_i\right]^2\right)$ so that
$E\left[X_i^2\right] = \text{var}\left[X_i\right] + \left(E\left[X_i\right]^2\right)$ for any *i*.

Thus, $E\left[X_1^2\right] = E\left[X_2^2\right] = E\left[X_3^2\right] = \ldots = E\left[X_n^2\right] = \sigma^2 + \mu^2$

and $E\left[\frac{\sum X^2}{n}\right] = \frac{1}{n}E\left[\sum X^2\right] = \frac{1}{n}n(\sigma^2 + \mu^2) = \sigma^2 + \mu^2$ **Equation (I)**

The <u>*second term*</u> can be simplified similarly, that is, using the definition of variance:

$$E\left[\bar{X}^2\right] = \text{var}\left[\bar{X}\right] + \left(E\left[\bar{X}\right]\right)^2 = \text{var}\left[\bar{X}\right] + \mu^2$$

The term $\operatorname{var}\left[\overline{X}\right]$ can further be simplified as follows:

$$\operatorname{var}\left[\overline{X}\right] = \operatorname{var}\left[\frac{\sum X}{n}\right] = \frac{1}{n^2}\operatorname{var}\left[X_1 + X_2 + X_3 + \ldots X_n\right]$$

$$= \frac{1}{n^2}\left\{\operatorname{var}\left[X_1\right] + \operatorname{var}\left[X_2\right] + \operatorname{var}\left[X_3\right] + \ldots + \operatorname{var}\left[X_n\right]\right\}$$

since it was already assumed that the observations are i.i.d.

Thus, $\operatorname{var}\left[\overline{X}\right] = \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}$.

so that $E\left[\overline{X}^2\right] = \operatorname{var}\left[\overline{X}\right] + \mu^2 = \frac{\sigma^2}{n} + \mu^2$ **Equation (II)**

Combining Equations (I) and (II), we have

$$E\left[s^2\right] = \left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2 - \frac{\sigma^2}{n} = \left(\frac{n-1}{n}\right)\sigma^2$$

It is clear that s^2 is a *biased* estimator of σ^2 since its expectation is not equal to σ^2. However, it is not very far from being unbiased given that (n – 1) and n would approximately equal if n were very large – this is what *asymptotic unbiasedness* is all about!

More interesting would then be to determine the statistic which is the unbiased estimator of σ^2. Using one of the laws of expectation (Section 4.3.1), we have, starting from

$$E\left[s^2\right] = \left(\frac{n-1}{n}\right)\sigma^2$$

that, multiplying both sides by $\left(\frac{n}{n-1}\right)$,

$$\left(\frac{n}{n-1}\right)E\left[s^2\right] = \sigma^2, \text{ implying that } E\left[\left(\frac{n}{n-1}\right)s^2\right] = \sigma^2.$$

Hence, $\left(\frac{n}{n-1}\right)s^2$ is an unbiased estimator of σ^2.

3.6 Estimation of the population proportion p

We often want to know the proportion of individuals in a population which satisfies a certain characteristic. For example, it would be interesting to know the *percentage* of left-handed people in Mauritius or the proportion of books in a library which contain more than 500 pages. As usual, it will be assumed that the population is infinite so that information may only be obtained by selecting a sample. The population proportion is denoted by p.

In general, when we select individuals, they either satisfy or do not satisfy the characteristic under investigation. If it ever happens that an individual falls in both categories simultaneously (for example, someone *ambidextrous*), then that individual is automatically discarded for the sake of calculations. It is thus quite natural to use the *binomial distribution* because each individual will either be labelled as '*success*' or '*failure*', depending on whether it satisfies the characteristic or not. If we want to have an idea of the value of p, we select a sample of size n and count the number, x, of individuals satisfying the required characteristic. It is obvious that a natural point estimate for p, usually denoted by p_s, would be $\frac{x}{n}$.

For the reader who is unfamiliar with the binomial distribution (discrete), it is sufficient to know that if X is a binomial variable with parameters n and p, then $E[X] = np$ and $\text{var}[X] = np(1-p)$. These results will certainly help in the following derivation.

The expectation of the sample proportion is obtained as follows:

$$E\left[\frac{X}{n}\right] = \frac{1}{n}E[X] = \frac{1}{n}(np) = p$$

Thus, the sample proportion is an *unbiased* estimator of the population proportion.

We summarise our findings in the following table.

	Parameter	**Sample statistic**	**Unbiased estimator**
Mean	μ	$\bar{x} = \frac{\sum x}{n}$	$\hat{\mu} = \frac{\sum x}{n}$
Variance	σ^2	$s^2 = \frac{\sum x^2}{n} - \bar{x}^2$	$\hat{\sigma}^2 = \frac{n}{n-1}\left(\frac{\sum x^2}{n} - \bar{x}^2\right)$
Proportion	p	$p_s = \frac{x}{n}$	$\hat{p} = \frac{x}{n}$

Fig. 3.6

4 INTERVAL ESTIMATION

This aspect of estimation is an attempt to find the *lower* and *upper bou*ndaries of an interval that may contain the parameter under investigation.

The length of this interval depends on the *confidence level* as specified in the problem. The confidence level is the *degree of certainty* with which we can say that the interval will contain the parameter. It is given in the $100(1-\alpha)\,\%$ form, wherẹ α is known as the *significance level* or the margin of error. More formally, a 95% *confidence interval* would be defined as one that has a probability of 0.95 of containing the parameter.

It has to be mentioned here that 0.95 is *not* the probability that the parameter lies in the interval. There is a subtle difference between these two statements in the sense that it is not the parameter which varies but the boundaries of the interval.

Let us first become familiar with the above theory and notations by means of an example.

Example Imagine that we wish to find an interval estimate for the *population mean* weight of people who are 45 years old in a given population. The first step would be to select a sample of reasonable size, as unbiased as possible, and find the *point estimate* of the population mean, that is, the *sample mean*. This point estimate will be a guideline to the construction of the interval, which also requires the confidence level. If the sample mean is, say, 57.2 kg, then the population mean should not be 'very far' from this figure, taking into consideration the fact that the sample is as unbiased as possible. To maximise the probability of being correct in our interval estimation, it is *logical* to place the sample mean in the middle of the interval. The obvious reason is that, if there did exist some sampling error during the sampling process, then any amount of deviation from the sample mean will still yield a *true* figure for the population mean which is contained in the interval (think about it very carefully). Furthermore, it is clear that the greater the confidence level, the larger will be the interval since the margin of error has to be minimised.

The logical argument given above can also be statistically explained – since we always select relatively large samples in order to obtain maximum information on the population parameter, we can make use of an extremely powerful theorem to support our argument.

4.1 The Central Limit Theorem

If $x_1.x_2, \ , x_n$ are observations of a random sample of size *n* from *any* distribution with mean μ and variance σ^2, then, *for large n*, the distribution of the sample mean $\overline{X}$ is approximately normal such that $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ where $\overline{x} = \dfrac{\sum x}{n}$.

4.2 Construction of an interval

We can thus make use of the *normal distribution theory* to show that the probability of being correct in our estimation is maximised whenever the interval is *symmetric* about the point estimate of the parameter.

The following diagrams show two intervals of the same length but placed at different locations on the *x*-axis of a normal curve. It is obvious that maximum probability is achieved when the interval is placed in the centre of the distribution (*LB* and *UB* stand for the lower boundary and upper boundary of each interval respectively).

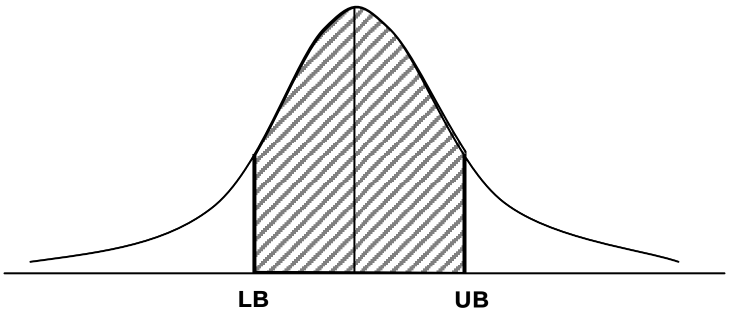

Fig. 4.2.1

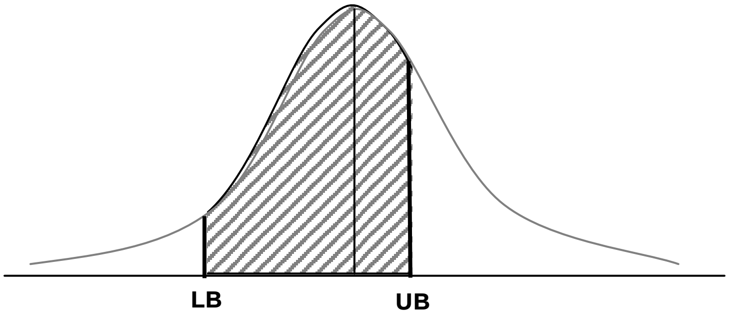

Fig. 4.2.2

It is clear that the shaded area in **Fig. 4.2.1** is larger than that of **Fig. 4.2.2**, hence justifying our choice of centering the interval on the point estimator.

Fig. 4.2.3 below is an overall view of a confidence interval.

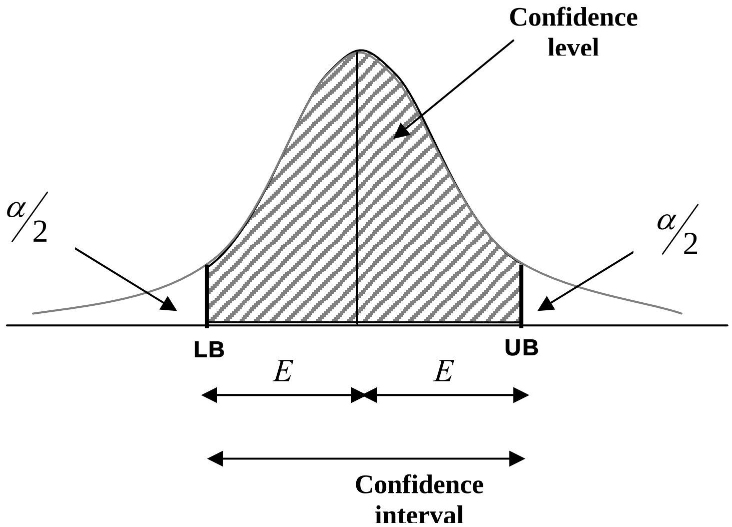

Fig. 4.2.3

The quantity E is just half of the length of the confidence interval. To obtain the respective values of the lower and upper boundaries, it suffices to evaluate E and, in turn, subtract it from, and add it to, the value of the point estimate (since the boundaries are equidistant to the centre of the distribution).

The procedure for interval calculation is as follows:

1. Given the confidence level, we subtract it from 1 and divide by two to obtain the half of the significance level.
2. Use this new value to get its corresponding z-value from the standard normal table – this is the number of standard deviations between any boundary and the centre.
3. Calculate the value of one standard deviation of the estimator.
4. Multiply the standard deviation by the z-value in order to obtain E.
5. The confidence interval will thus be

$$(Point\ Estimate - E,\ Point\ Estimate + E)$$

This procedure will be used for calculating confidence intervals for both population means and proportions. The major differences will just be the point estimates and their standard deviations.

4.3 Estimation of the population mean μ

When finding an interval estimate for the population mean μ, we should first select a sample and determine the value of the sample mean or point estimate $\bar{x}$. From the *Central Limit Theorem*, the standard deviation for $\bar{x}$ is $\sigma/\sqrt{n}$. *However, if the population standard deviation* σ *is unknown, we have to replace it by* $\hat{\sigma}$*, its unbiased estimate*. We then follow the procedure as given in the previous section. This is illustrated by the example below.

Example

A random sample of 250 adult men undergoing a routine medical inspection had their height (x cm) measured to the nearest centimetre and the following data were obtained: $\sum x = 43205$, $\sum x^2 = 7469107$. Calculate unbiased estimates for the population mean and variance and hence a 99% symmetric confidence interval for the population mean.

Solution

We use the following information:

$$n = 250,\ \sum x = 43205,\ \sum x^2 = 7469107\,.$$

The unbiased estimate for the population mean is

$$\bar{x} = \frac{\sum x}{n} = \frac{43205}{250} = 172.82$$

The unbiased estimate for the population variance is

$$\frac{ns^2}{n-1} = \left(\frac{n}{n-1}\right)\left(\frac{\sum x^2}{n} - \bar{x}^2\right) = \left(\frac{250}{249}\right)\left(\frac{7469107}{250} - (172.82)^2\right) = 9.7144$$

Since the confidence level is 99%, half the significance level is 0.005, which gives us a z-value of 2.576 from the standard normal table.

Thus, $E = \frac{z\hat{\sigma}}{\sqrt{n}} = \frac{(2.576)(\sqrt{9.7144})}{\sqrt{250}} = 0.51$ (2 decimal places, that is, the same degree of accuracy of the sample mean).

A 99% *confidence interval for the population m*ea*n* is therefore

$$172.82 - 0.51 < \mu < 172.82 + 0.51,$$

$$\Rightarrow\ 172.31 < \mu < 173.33\,.$$

4.4 Estimation of the population proportion *p*

As has been proven in Section 3.6, the unbiased estimator of the population proportion *p* is the sample proportion $\frac{x}{n}$. Given that the binomial distribution is used to determine the unbiasedness of the sample proportion, we use the same distribution to find its variance (or standard deviation). We know that if *X* is a binomial variable, $\text{var}[X] = np(1-p)$. Thus, the variance of $\frac{x}{n}$ would be $\text{var}\left[\frac{X}{n}\right] = \frac{1}{n^2}\text{var}[X] = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$.

But then again, we ask ourselves the question: 'How can we use the value of *p* in the formula for the variance of $\frac{x}{n}$ if we are precisely looking for a confidence interval for *p*?' The answer is simple – whenever *p* is unknown, we replace it by its unbiased estimator! Hence, the variance of $\hat{p}$, that is, $\frac{x}{n}$, will be $\frac{\hat{p}(1-\hat{p})}{n}$ or its standard deviation is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Let us now calculate the confidence interval for a population proportion by means of an example.

Example

A survey was carried out to investigate the proportion of people who are left-handed in a population. To that effect, a sample of 1000 people revealed that only 115 of them were left-handed. Calculate 95% confidence limits for the population proportion of left-handed people.

Solution

We use the following information:

$n = 1000, x = 115.$

The sample proportion, $\hat{p}$, is thus $\frac{115}{1000} = 0.115$

Since the confidence level is 95%, half the significance level is 0.025, which gives us a *z*-value of 1.96 from the standard normal table.

Thus, $E = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = (1.96)\sqrt{\frac{(0.115)(0.885)}{1000}} = 0.020$ (3 decimal places, that is, the same degree of accuracy of the sample proportion).

A 99% *confidence interval for the population m*ean is therefore

$$0.115 - 0.020 < p < 0.115 + 0.020,$$

$$\Rightarrow\ 0.095 < p < 0.135.$$