Ch05
Estimation and confidence intervals

This chapter covers the **estimation** of population parameters such as μ and σ^2 while Chapter 6 describes **testing hypotheses** about these parameters. The two procedures are very closely related.

5.1 Point and interval estimation:

- There are basically two ways in which an estimate of a parameter can be presented.
- **A point estimate**, i.e. a single value which is the best estimate of the parameter of interest.
- The point estimate is the one which is most prevalent in everyday usage; for example, the average Briton surfs the internet for 30 minutes per day. Although this is presented as a fact, it is actually an estimate, obtained from a survey of people's use of personal computers.
- Since it is obtained from a sample there must be some doubt about its accuracy: the sample will probably not exactly represent the whole population.
- For this reason **interval estimates** are also used, which give some idea of the likely accuracy of the estimate. If the sample size is small, for example, then it is quite possible that the estimate is not very close to the true value and this would be reflected in a wide interval estimate, for example, that the average Briton spends between 5 and 55 minutes surfing the net per day.

5.1 Point and interval estimation:

- A larger sample, or a better method of estimation, would allow a narrower interval to be derived and thus a more precise estimate of the parameter to be obtained, such as an average surfing time of between 20 and 40 minutes.
- Interval estimates are better for the consumer of the statistics, since they not only show the estimate of the parameter but also give an idea of the confidence which the researcher has in that estimate.
- ***Rules and criteria for finding estimates***
- It is important to distinguish between an **estimator**, a rule and an **estimate**, which is the value derived as a result of applying the rule to the data.
- There are many possible estimators for any parameter, so it is important to be able to distinguish between good and bad estimators. The following examples provide some possible estimators of the population mean:
- (1) the sample mean; (2) the smallest sample observation; (3) the first sample observation.
- A set of criteria is needed for discriminating between good and bad estimators.
- Which of the above three estimators is ‘best’? Two important criteria by which to judge estimators are bias and precision.

5.1 Point and interval estimation:

- *Bias*
- It is impossible to know if a single estimate of a parameter, derived by applying a particular estimator to the sample data, gives a correct estimate of the parameter or not.
- The estimate might be too low or too high and, since the parameter is unknown, it is impossible to check this. What is possible, however, is to say whether an estimator gives the correct answer on average.
- An estimator which gives the correct answer on average is said to be unbiased.
- Another way of expressing this is to say that an unbiased estimator does not systematically mislead the researcher away from the correct value of the parameter.
- It is, however, important to remember, that even using an unbiased estimator does not guarantee that a single use of the estimator will yield a correct estimate of the parameter.
- Bias (or the lack of it) is a theoretical property.
- An estimator is unbiased if its expected value is equal to the parameter being estimated.

5.1 Point and interval estimation:

- Consider trying to estimate the population mean using the three estimators suggested above.
- Taking the sample mean first, we have already learned that its expected value is μ, i.e. $E(\bar{x}) = \mu$
- which immediately shows that the sample mean is an unbiased estimator.
- The second estimator (the smallest observation in the sample) can easily be shown to be biased, using the result derived above.
- Since the smallest sample observation must be less than the sample mean, its expected value must be less than μ.
- Denote the smallest observation by x_s, then

$$E(x_s) < \mu$$

- So this estimator is biased downwards. It underestimates the population mean.
- The size of the bias is simply the difference between the expected value of the estimator and the value of the parameter, so the bias in this case is

$$Bias = E(x_s) - \mu$$

- For the sample mean $\bar{x}$ the bias is obviously zero.

5.1 Point and interval estimation:

- Turning to the third rule (the first sample observation) this can be shown to be another unbiased estimator.
- Choosing the first observation from the sample is equivalent to taking a random sample of size one from the population in the first place.
- Thus the single observation may be considered as the sample mean from a random sample of size one. Since it is a sample mean it is unbiased, as demonstrated earlier.
- *Precision*
- precision is a relative concept, comparing one estimator to another.
- Given two estimators A and B, A is more precise than B if the estimates it yields (from all possible samples) are **less spread out** than those of estimator B. A precise estimator will tend to give similar estimates for all possible samples.

5.1 Point and interval estimation:

- Consider the two unbiased estimators found above: how do they compare on the criteria of precision?
- It turns out that the sample mean is the more precise of the two, and it is not difficult to understand why.
- Taking just a single sample observation means that it is quite likely to be unrepresentative of the population as a whole, and thus leads to a poor estimate of the population mean.
- The sample mean on the other hand is based on all the sample observations and it is unlikely that all of them are unrepresentative of the population.
- The sample mean is therefore a good estimator of the population mean, being more precise than the single observation estimator.
- Just as bias was related to the expected value of the estimator, so precision can be defined in terms of the variance. One estimator is more precise than another if it has a smaller variance.

5.1 Point and interval estimation:

- Recall that the probability distribution of the sample mean is

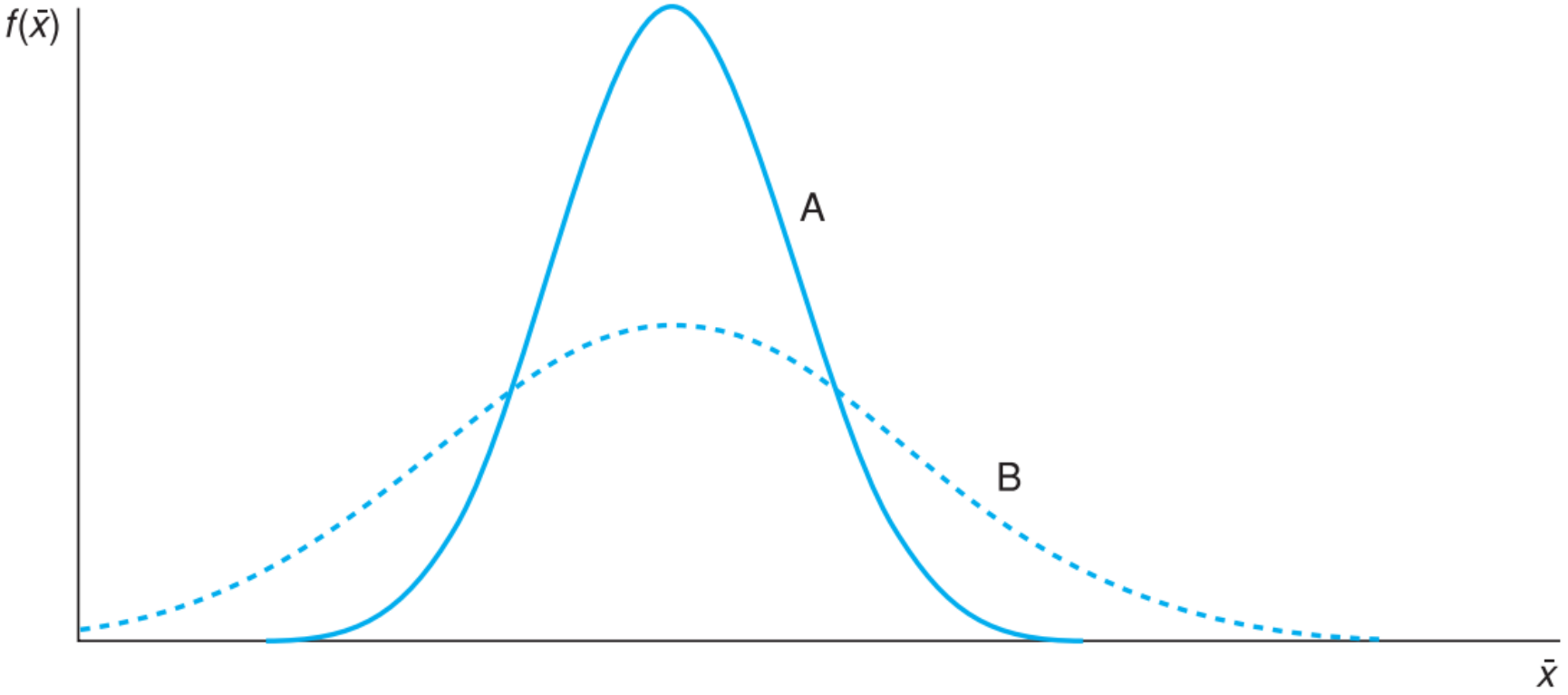
$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- in large samples, so the variance of the sample mean is

$$V(\bar{x}) = \frac{\sigma^2}{n}$$

- As the sample size n becomes larger, the variance of the sample mean becomes smaller, so the estimator becomes more precise.
- For this reason large samples give better estimates than small samples, and so the sample mean is a better estimator than taking just one observation from the sample.
- The two estimators can be compared in a diagram (see next Figure) which draws the probability distributions of the two estimators.

5.1 Point and interval estimation:



Note: Curve *A* shows the distribution of sample means, which is the more precise estimator. *B* shows the distribution of estimates using a single observation.

5.1 Point and interval estimation:

- A related concept is that of efficiency.
- The efficiency of one unbiased estimator, relative to another, is given by the ratio of their sampling variances. Thus the efficiency of the first observation estimator, relative to the sample mean, is given by

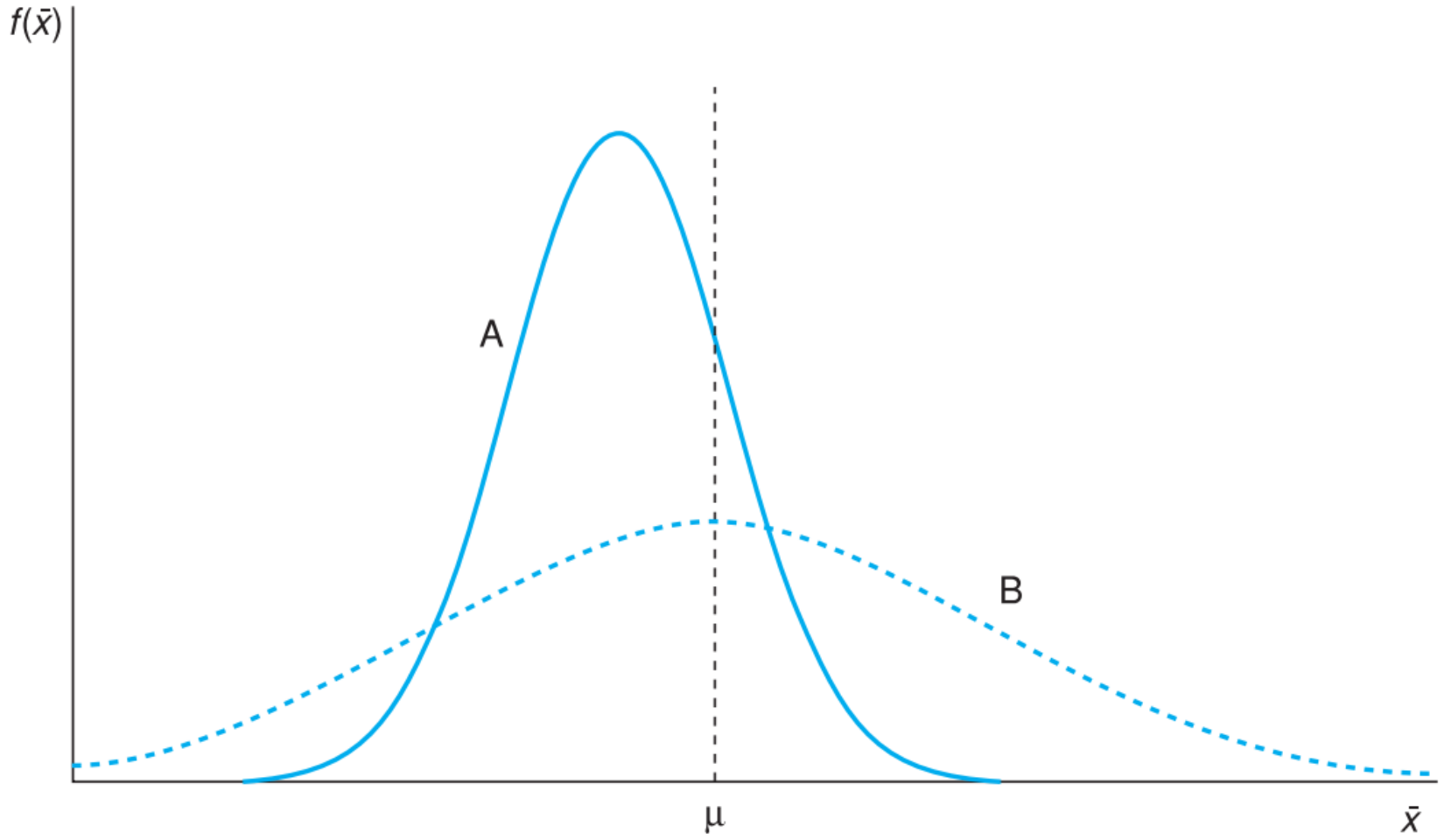
$$Efficiency = \frac{var(\bar{x})}{var\ (x_1)} = \frac{\frac{\sigma^2}{n}}{\sigma^2} = \frac{1}{n}$$

- Thus the efficiency is determined by the relative sample sizes in this case.
- Other things being equal, a more efficient estimator is to be preferred.
- Similarly, the variance of the median can be shown to be (for a Normal distribution) $\frac{\pi}{2} \times \frac{\sigma^2}{n}$. The efficiency of the median is therefore $\frac{2}{\pi} \approx 64\%$.

5.1 Point and interval estimation:

- *The trade-off between bias and precision: the Bill Gates effect*
- It should be noted that just because an estimator is biased does not necessarily mean that it is imprecise.
- Sometimes there is a trade-off between an unbiased, but imprecise, estimator and a biased, but precise, one.
- Figure below illustrates this. Although estimator A is biased it will nearly always yield an estimate which is fairly close to the true value; even though the estimate is expected to be wrong, it is not likely to be far wrong.
- Estimator B, although unbiased, can give estimates which are far away from the true value, so that A might be the preferred estimator.
- As an example of this, suppose we are trying to estimate the average wealth of the US population.
- Consider the following two estimators:
- (1) use the mean wealth of a random sample of Americans;
- (2) use the mean wealth of a random sample of Americans but, if Bill Gates is in the sample, omit him from the calculation.

5.1 Point and interval estimation:



5.1 Point and interval estimation:

- Bill Gates is the Chairman of Microsoft and one of the world's richest men. Because of this, he is a dollar billionaire. His presence in a sample of, say, 30 observations would swamp the sample and give a highly misleading result.
- Assuming Bill Gates has $50bn and the others each have $200000 of wealth, the average wealth would be estimated at about $1.6bn, which is surely wrong.
- The first rule could therefore give us a wildly incorrect answer, although the rule is unbiased.
- The second rule is clearly biased but does rule out the possibility of such an unlucky sample. We can work out the approximate bias. It is the difference between the average wealth of all Americans and the average wealth of all Americans except Bill Gates. If the true average of all 250 million Americans is $200 000, then total wealth is $50 000bn. Subtracting Bill's $50bn leaves $49 950bn shared among the rest, giving $199 800 each, a difference of 0.1%.
- This is what we would expect the bias to be. It might seem worthwhile therefore to accept this degree of bias in order to improve the precision of the estimate. Furthermore, if we did use the biased rule, we could always adjust the sample mean upwards by 0.1% to get an approximately unbiased estimate.

5.2 Estimation with large samples:

- For the type of problem encountered in this chapter the method of estimation differs according to the size of the sample.
- ‘Large’ samples, by which is meant sample sizes of 25 or more, are dealt with first, using the Normal distribution.
- Small samples are considered in a later section, where the t distribution is used instead of the Normal. The differences are relatively minor in practical terms and there is a close theoretical relationship between the t and Normal distributions.
- With large samples there are three types of estimation problem we will consider.
- (1) The estimation of a mean from a sample of data.
- (2) The estimation of a proportion on the basis of sample evidence. This would consider a problem such as estimating the proportion of the population intending to buy an iPhone, based on a sample of individuals. Each person in the sample would simply indicate whether they have bought, or intend to buy, an iPhone. The principles of estimation are the same as in the first case but the formulae used for calculation are slightly different.
- (3) The estimation of the difference of two means (or proportions), for example a problem such as estimating the difference between men and women’s expenditure on clothes. Once again, the principles are the same, the formulae different.

5.2 Estimation with large samples:

- ***Estimating a mean***

- To demonstrate the principles and practice of estimating the population mean, we shall take the example of estimating the average wealth of the UK population, the full data for which were given in Chapter 1.

- Suppose that we did not have this information but were required to estimate the average wealth from a sample of data. In particular, let us suppose that the sample size is $n = 100$, the sample mean is $\bar{x} = 130$ (in £000) and the sample variance is $s^2 = 50000$.

- Obviously, this sample has got fairly close to the true values (see Chapter 1) but we could not know that from the sample alone. What can we infer about the population mean μ from the sample data alone?

- For the point estimate of μ the sample mean is a good candidate since it is unbiased, and it is more precise than other sample statistics such as the median.

- The point estimate of μ is simply £130 000, therefore. The point estimate does not give an idea of the uncertainty associated with the estimate. We are not absolutely sure that the mean is £130 000 (in fact, it isn't – it is £146 984).

- The interval estimate gives some idea of the uncertainty. It is centred on the sample mean, but gives a range of values to express the uncertainty.

5.2 Estimation with large samples:

- To obtain the interval estimate we first require the probability distribution of $\bar{x}$,

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

- From this, it was calculated that there is a 95% probability of the sample mean lying within 1.96 standard errors of μ, i.e.

$$\Pr\left(\mu - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \bar{x} \leq \mu + 1.96\sqrt{\frac{\sigma^2}{n}}\right)$$

- We can manipulate each of the inequalities within the brackets to make μ the subject of the expression

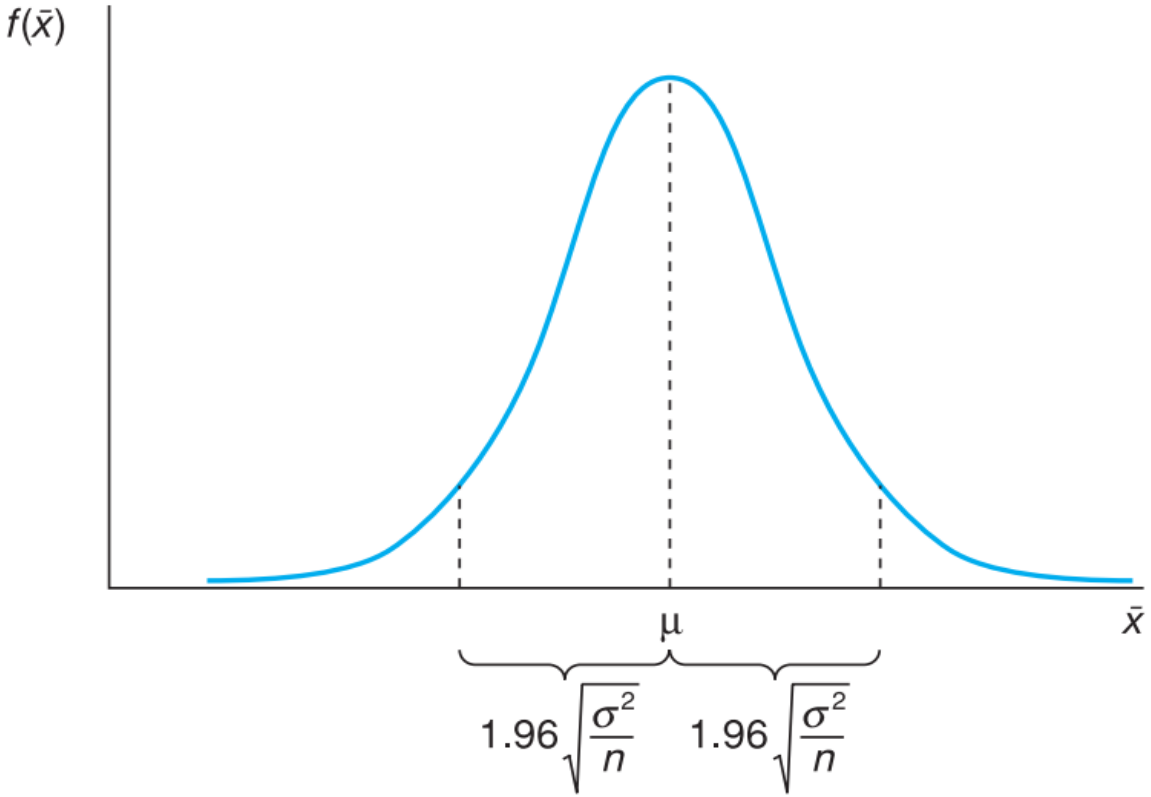
$$\mu - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \bar{x} \Rightarrow \mu \leq \bar{x} + 1.96\sqrt{\frac{\sigma^2}{n}} \text{ and } \bar{x} \leq \mu + 1.96\sqrt{\frac{\sigma^2}{n}} \Rightarrow \bar{x} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu$$

- Combining these two new expressions we obtain

$$\left[\bar{x} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + 1.96\sqrt{\frac{\sigma^2}{n}}\right]$$

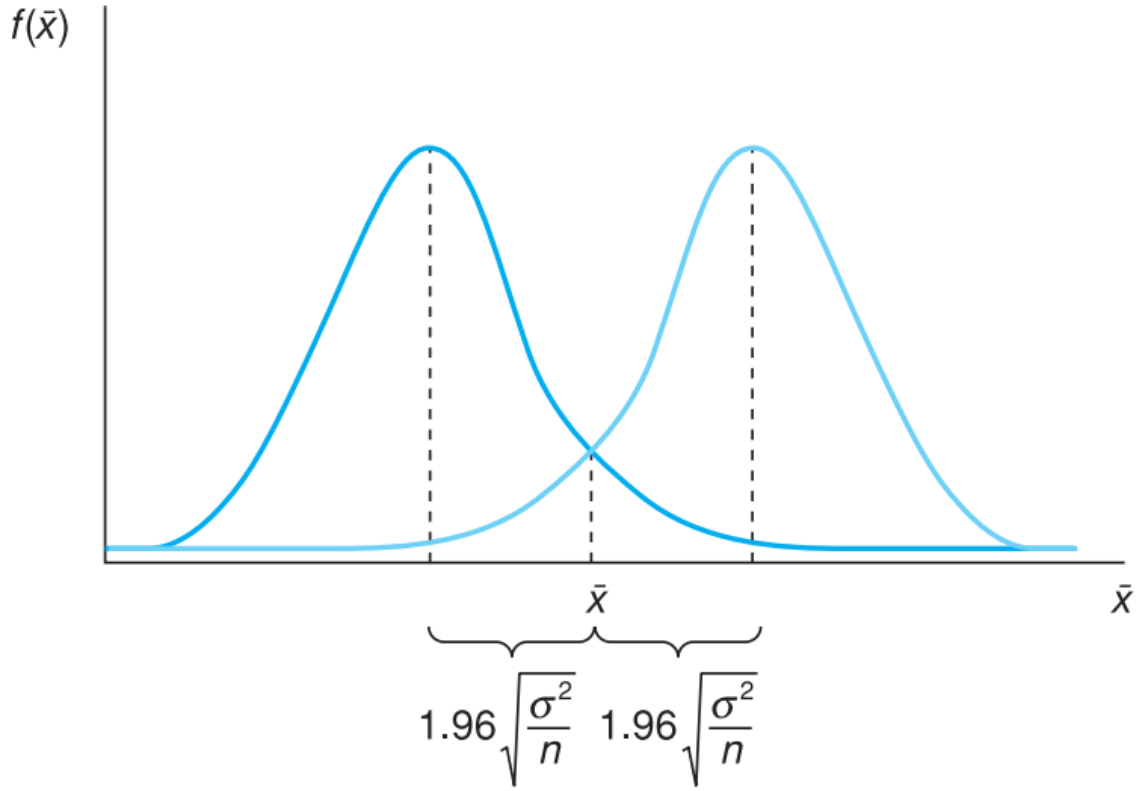
5.2 Estimation with large samples:

- We have transformed the probability interval. Instead of saying $\bar{x}$ lies within 1.96 standard errors of μ, we now say μ lies within 1.96 standard errors of $\bar{x}$.
- Figure below shows μ at the centre of a probability interval for $\bar{x}$. Next figure shows a sample mean $\bar{x}$ at the centre of an interval relating to the possible positions of μ.



5.2 Estimation with large samples:

- The interval shown in equation is called the 95% confidence interval and this is the interval estimate for μ.
- In our example the value of σ^2 is unknown, but in large ($n > 25$) samples it can be replaced by s^2 from the sample. s^2 is here used as an estimate of σ^2 which is unbiased and sufficiently precise in large ($n > 25$ or so) samples.



5.2 Estimation with large samples:

- By examining the equation one can see that the confidence interval is wider
- ● the smaller the sample size;
- ● the greater the standard deviation of the sample.
- The greater uncertainty which is associated with smaller sample sizes is manifested in a wider confidence interval estimate of the population mean. This occurs because a smaller sample has more chance of being unrepresentative (just because of an unlucky sample).
- Greater variation in the sample data also leads to greater uncertainty about the population mean and a wider confidence interval. Greater sample variation suggests greater variation in the population so, again, a given sample could include observations which are a long way off the mean. Note that in this example there is great variation of wealth in the population and hence in the sample also. This means that a sample of 100 is not very informative (the confidence interval is quite wide). We would need a substantially larger sample to obtain a more precise estimate.

5.2 Estimation with large samples:

- ***Example:*** A sample of 50 school students found that they spent 45 minutes doing homework each evening, with a standard deviation of 15 minutes. Estimate the average time spent on homework by all students.
- The sample data are $\bar{x} = 45$, $s = 15$ and $n = 50$.
- **Exercise 1**
- (a) A sample of 100 is drawn from a population. The sample mean is 25 and the sample standard deviation is 50. Calculate the point and 95% confidence interval estimates for the population mean.
- (b) If the sample size were 64, how would this alter the point and interval estimates?
- **Exercise 2**
- A sample of size 40 is drawn with sample mean 50 and standard deviation 30. Is it likely that the true population mean is 60?

5.3 Estimating the difference between two means

- We have two samples and want to know whether there is a difference between their respective populations.
- One sample might be of men, the other of women, or we could be comparing two different countries, etc. A point estimate of the difference is easy to obtain but once again there is some uncertainty around this figure, because it is based on samples. Hence we measure that uncertainty via a confidence interval. All we require are the appropriate formulae.
- Consider the following example. Sixty students from school 1 scored an average mark of 62% in an exam, with a standard deviation of 18%; 35 students from school 2 scored an average of 70% with standard deviation 12%.
- Estimate the true difference between the two schools in the average mark obtained.
- This is a more complicated problem than those previously treated since it involves two samples rather than one. An estimate has to be found for $\mu_1 - \mu_2$ (the true difference in the mean marks of the schools), in the form of both point and interval estimates.
- The students taking the exams may be thought of as samples of all students in the schools who could potentially take the exams.

5.3 Estimating the difference between two means

- Notice that this is a problem about sample means, not proportions, even though the question deals in percentages. The point is that each observation in the sample (i.e. each student's mark) can take a value between 0 and 100, and one can calculate the standard deviation of the marks.
- For this to be a problem of sample proportions the mark for each student would each have to be of the pass/fail type, so that one could only calculate the proportion who passed.
- It might be thought that the way to approach this problem is to derive one confidence interval for each sample (along the lines set out above), and then to somehow combine them; for example, the degree of overlap of the two confidence intervals could be assessed. This is not the best approach, however. It is sometimes a good strategy, when faced with an unfamiliar problem to solve, to translate it into a more familiar problem and then solve it using known methods.
- This is the procedure which will be followed here. The essential point is to keep in mind the concept of a random variable and its probability distribution.

5.3 Estimating the difference between two means

- The current problem deals with two samples and therefore there are two random variables to consider, i.e. the two sample means X_1 and X_2. Since the aim is to estimate $\mu_1 - \mu_2$, an obvious candidate for an estimator is the difference between the two sample means, $\bar{x}_1 - \bar{x}_2$. We can think of this as a single random variable (even though two means are involved) and use the methods we have already learned. We therefore need to establish the sampling distribution of $\bar{x}_1 - \bar{x}_2$ as

$$\bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2})$$

- The above equation shows that $\bar{x}_1 - \bar{x}_2$ that is an unbiased estimator of $\mu_1 - \mu_2$.
- The difference between the sample means will therefore be used as the point estimate of $\mu_1 - \mu_2$.
- Thus the point estimate of the true difference between the schools is

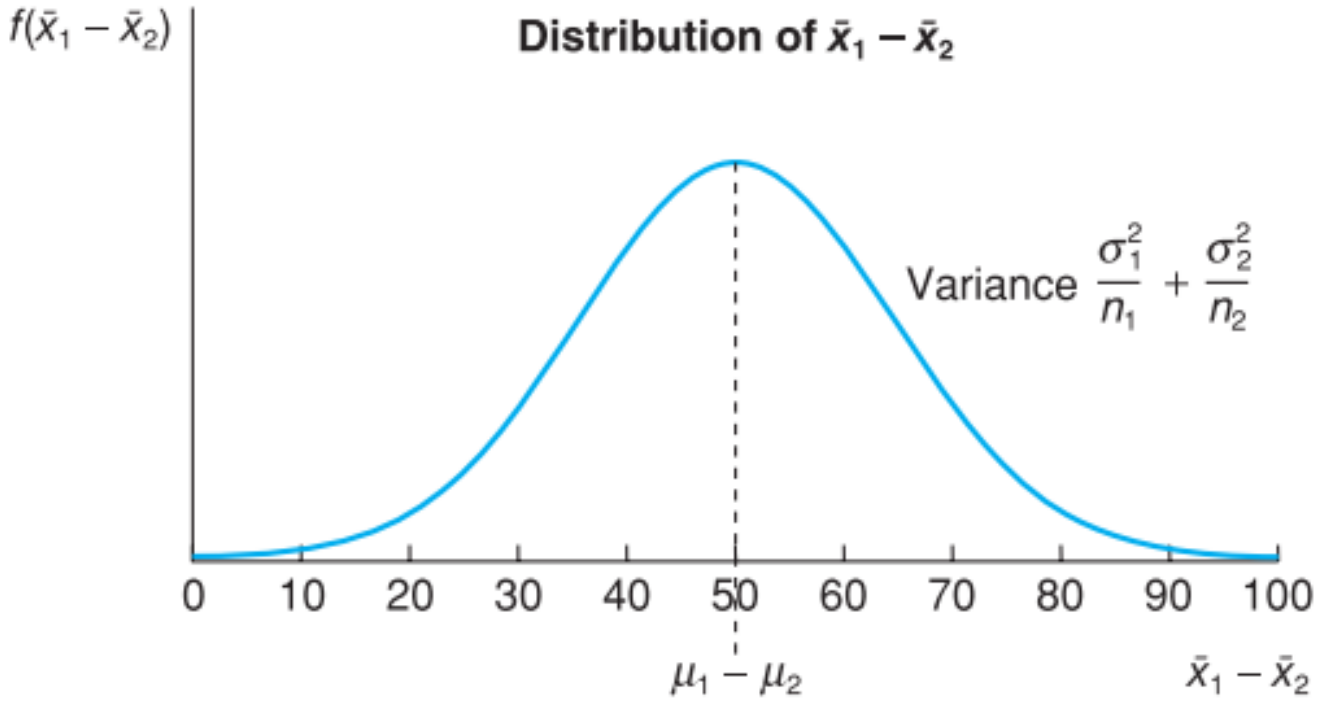
$$\bar{x}_1 - \bar{x}_2 = 62 - 70 = -8\%$$

- The 95% confidence interval estimate is derived in the same manner as before, making use of the standard error of the random variable.

$$\left[(\bar{x}_1 - \bar{x}_2) - 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

5.3 Estimating the difference between two means

- The distribution of $\bar{x}_1 - \bar{x}_2$ is illustrated in the Figure below.



5.3 Estimating the difference between two means

- Example; A survey of holidaymakers found that on average women spent 3 hours per day sunbathing, men spent 2 hours. The sample sizes were 36 in each case and the standard deviations were 1.1 hours and 1.2 hours respectively.
- Estimate the true difference between men and women in sunbathing habits. Use the 99% confidence level.
- **Exercise 1**
- A survey of 50 16-year old girls revealed that 40% had a boyfriend. A survey of 100 16-year old boys revealed 20% with a girlfriend. Estimate the true difference in proportions between the sexes.

5.4 Estimation with small samples: the t distribution

- So far only large samples (defined as sample sizes in excess of 25) have been dealt with, which means that (by the Central Limit Theorem) the sampling distribution of X follows a Normal distribution, whatever the distribution of the parent population. Remember, from the two theorems of Chapter 3, that:
  - if the population follows a Normal distribution, X is also Normally distributed;
  - and
  - if the population is not Normally distributed, X is approximately Normally distributed in large samples ($n \geq 25$).
- In both cases, confidence intervals can be constructed based on the fact that

$$\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

- so the standard Normal distribution is used to find the values which cut off the extreme 5% of the distribution (z = ±1.96). In practical examples, we had to replace σ by its estimate, s. Thus the confidence interval was based on the fact that $\frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}} \sim N(0,1)$

5.4 Estimation with small samples: the t distribution

- For small sample sizes, above equation is no longer true.
- Instead, the relevant distribution is the t distribution and we have

$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1}$$

- The random variable defined in this equation has a t distribution with $n - 1$ degrees of freedom.
- As the sample size increases, the t distribution approaches the standard Normal, so the latter can be used for large samples.
- The t distribution is in many ways similar to the standard Normal, insofar as it is:
- ● unimodal; ● symmetric;
- ● centred on zero; ● bell-shaped; ● extends from minus infinity to plus infinity

5.4 Estimation with small samples: the t distribution

- The differences are that it is more spread out (has a larger variance) than the standard Normal distribution, and has only one parameter rather than two: the degrees of freedom, denoted by ν.
- In problems involving the estimation of a sample mean the degrees of freedom are given by the sample size minus one, i.e. $\nu = n - 1$.
- To summarise the argument so far, when
  - the sample size is small, and
  - the sample variance is used to estimate the population variance,
- then the t distribution should be used for constructing confidence intervals, not the standard Normal. This results in a slightly wider interval than would be obtained using the standard Normal distribution, which reflects the slightly greater uncertainty involved when s^2 is used as an estimate of σ^2 if the sample size is small.
- Apart from this, the methods are exactly as before and are illustrated by the examples below. We look first at estimating a single mean, then at estimating the difference of two means. The t distribution cannot be used for small sample proportions (explained below) so these cases are not considered.

5.4 Estimation with small samples: the t distribution

- *Estimating a mean*
- A sample of 15 bottles of juice showed an average specific gravity of 1035.6, with standard deviation 2.7. Estimate the true specific gravity of the brew.
- The sample information may be summarised as $\bar{X} = 1035.6$, $s = 2.7$ and $n = 15$
- The sample mean is still an unbiased estimator of μ (this is true regardless of the distribution of the population) and serves as point estimate of μ. The point estimate of μ is therefore 1035.6. Since σ is unknown, the sample size is small and it can be assumed that the specific gravity of all bottles is Normally distributed (numerous small random factors affect the specific gravity) we should use the t distribution.
- The structure of the t distribution table is different from that of the standard Normal table. The first column of the table gives the degrees of freedom. In this
- example we want the row corresponding to $\nu = n - 1 = 14$. The appropriate column of the table is the one headed '0.025' which indicates the area cut off in each tail.

5.4 Estimation with small samples: the *t* distribution

Table 4.1 Percentage points of the *t* distribution (excerpt from Table A3)

	Area (α) in each tail						
ν	0.4	0.25	0.10	0.05	0.025	0.01	0.005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.656
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947

Note: The appropriate *t* value for constructing the confidence interval is found at the intersection of the shaded row and column.

5.4 Estimation with small samples: the *t* distribution

Estimating the difference between two means

As in the case of a single mean the **t-distribution** needs to be used in small samples when the population variances are unknown. Again, both parent populations must be Normally distributed and in addition it must be assumed that the population variances are equal, i.e. $\sigma_1^2 = \sigma_2^2$ (this is required in the mathematical derivation of the t distribution).

This latter assumption was not required in the large-sample case using the Normal distribution.

Consider the following example as an illustration of the method.

$\bar{X}_1 = 175$, $s_1 = 25$ and $n_1 = 20$; $\bar{X}_2 = 158$, $s_2 = 30$ and $n_2 = 15$

We wish to estimate $\mu_1 - \mu_2$. The point estimate of this is $\bar{X}_1 - \bar{X}_2$ which is an unbiased estimate.

This gives $175 - 158 = 17$ as the expected difference between the two sets of authorities.

For the confidence interval, the t distribution has to be used since the sample sizes are small and the population variances unknown. It is assumed that the populations are Normally distributed and that the samples have been independently drawn. We also assume that the population variances are equal, which seems justified since s_1 and s_2 do not differ by much.

5.4 Estimation with small samples: the *t* distribution

$$\left[(\bar{x}_1 - \bar{x}_2) - t_\nu \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} \leq \mu \leq (\bar{x}_1 - \bar{x}_2) + t_\nu \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}\right]$$

Where $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ (the pooled variance)

- **Exercise 1**

A sample of size n = 16 is drawn from a population which is known to be Normally distributed. The sample mean and variance are calculated as 74 and 121. Find the 99% confidence interval estimate for the true mean

- **Exercise 2**

Samples are drawn from two populations to see if they share a common mean. The sample data are:

$\bar{X}_1 = 45$; $\bar{X}_2 = 55$

$s_1 = 18$; $s_2 = 21$

$n_1 = 15$; $n_2 = 20$

Find the 95% confidence interval estimate of the difference between the two population means.