Data Analytics

ECON 1008, Semester 1, 2019

Giulio Zanella
University of Adelaide
School of Economics

CHAPTER 10

Estimation:

Describing a single population

Where are we?

In Chapter 9, we studied

- why **statistical inference** is used
- the **central limit theorem**
- the **sampling distributions** of the sample mean
- how to make **probability statements** about sample statistics such as the mean, an exercise in inference

This week we learn more about statistical inference.

Sampling Distribution of the Sample Mean: summary of key results

1. $\mu_{\overline{x}} = \mu_x$

2. $\sigma^2_{\overline{x}} = \dfrac{\sigma^2_x}{n}$

3. *If x is* normal, $\overline{\text{x}}$ is normal. If x is non-normal $\overline{\text{x}}$ is approximately normally distributed for sufficiently large sample size ($n \geq 30$).

Making inferences

Statistical inference is the process by which we acquire information about populations from samples.

There are two procedures for making inferences:

- Estimation (chapter 10, this week)
- Hypothesis testing (chapter 12, next week)

Estimation: examples

- A bank conducts a survey to estimate the number of times customer will actually use ATM machines.

- A survey of eligible voters is conducted to gauge support for the federal government's new carbon emissions reforms.

Concepts of estimation

The objective of estimation is to determine the value of a **population parameter** on the basis of a **sample statistic**. When used for inference, the latter is called **estimator**.

For example,

- the sample mean $\bar{X}$ is an estimator of the population mean, μ (as we have seen in chapter 9)

Concepts of estimation

There are two types of estimators:

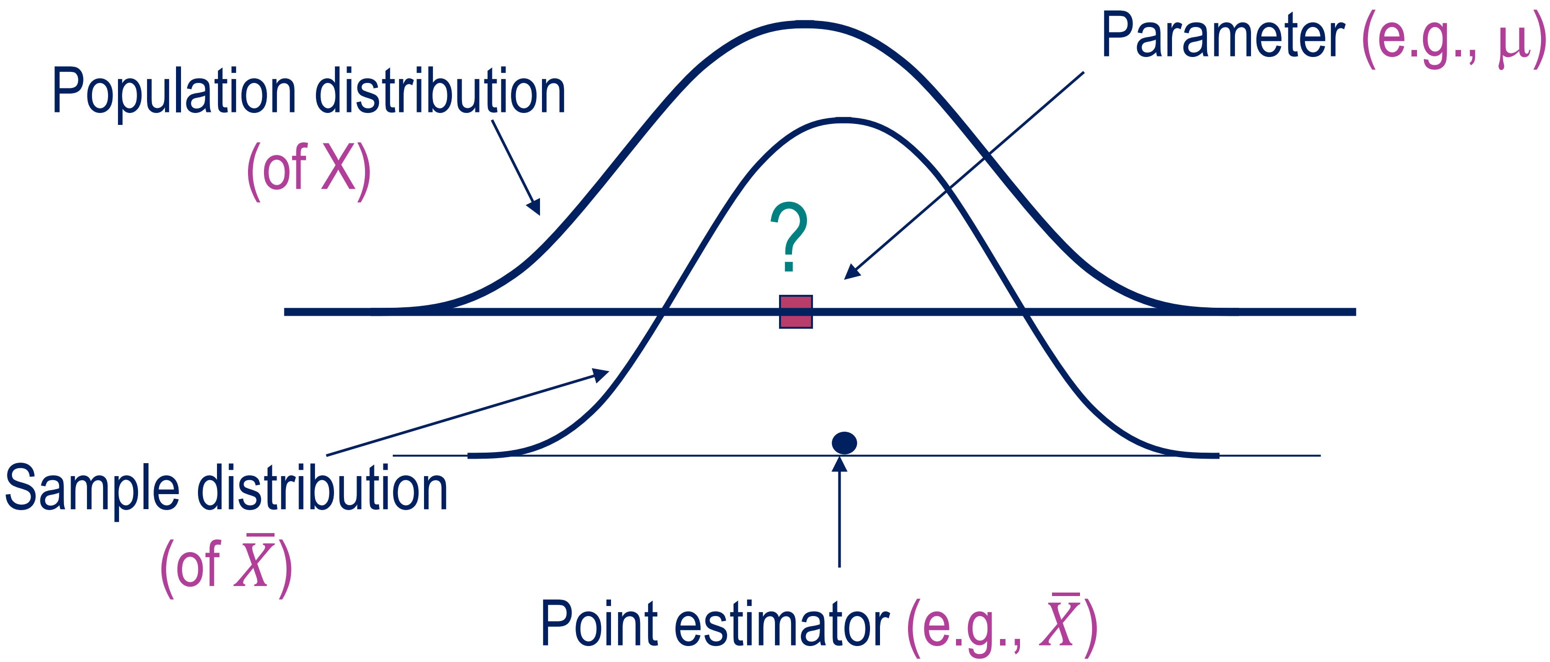
- Point estimator
- Interval estimator

Point Estimator

A *point estimator* estimates the value of an unknown (population) parameter using a single value calculated from the sample data.

For example, the sample mean ($\bar{X}$) calculated from a sample drawn from a population is a point estimator of the population mean (μ).

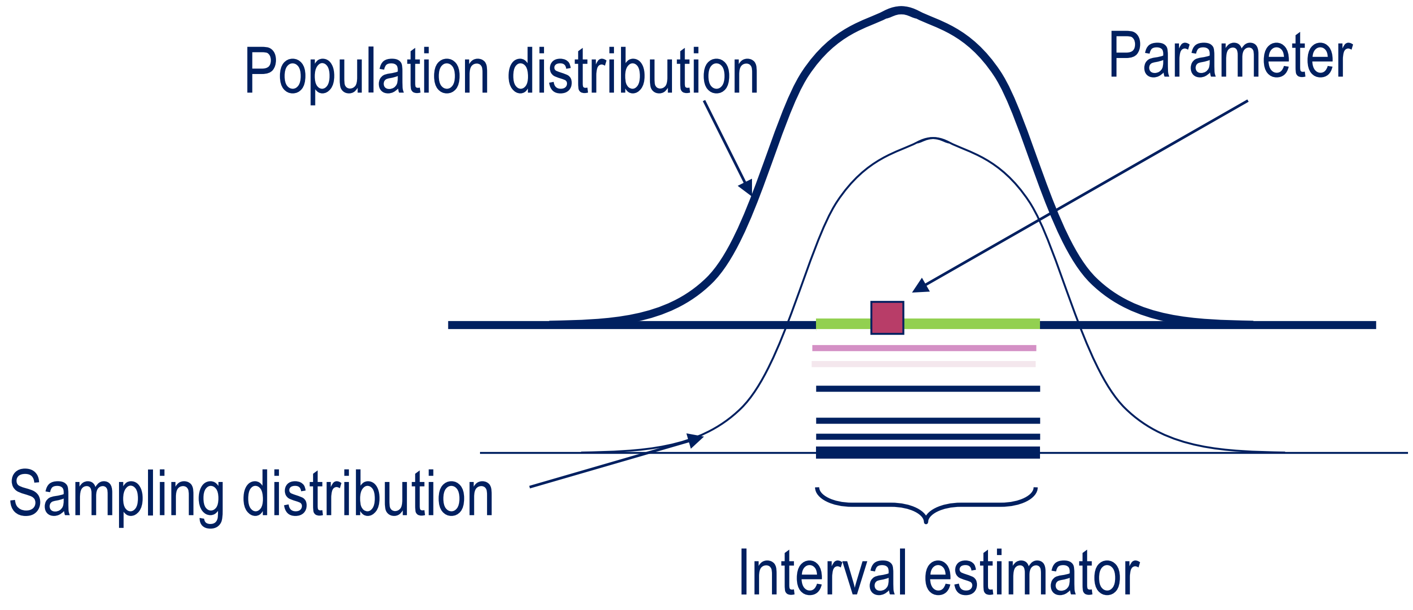
Point Estimator



Interval Estimator

An ***interval estimator*** draws inferences about a population by estimating the value of an unknown population parameter using an interval.

The *interval estimator* is affected by the *sample size*.



Desirable characteristics

A good estimator is

Unbiased: an unbiased estimator is one whose expected value is equal to the parameter it estimates.

Unbiased Estimators

Figure 10.1 Sampling distribution of $\overline{X}$

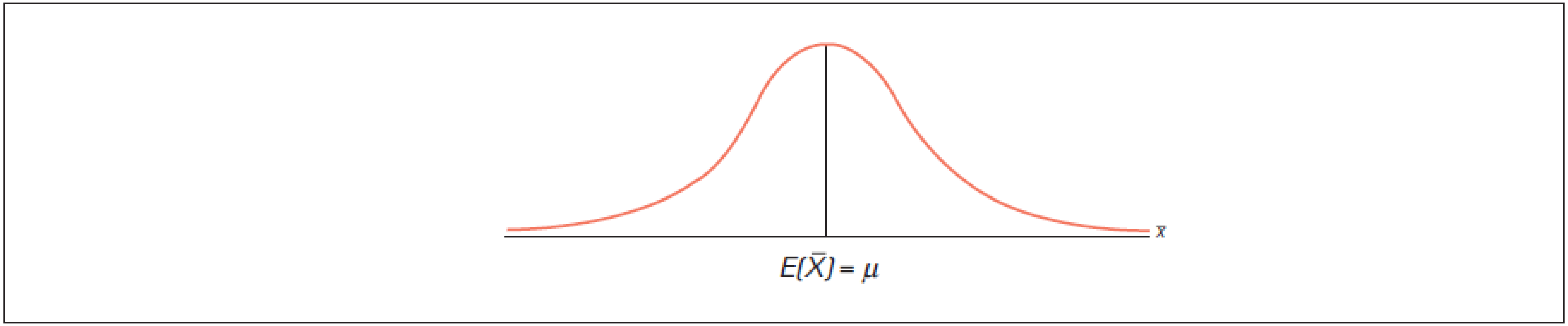
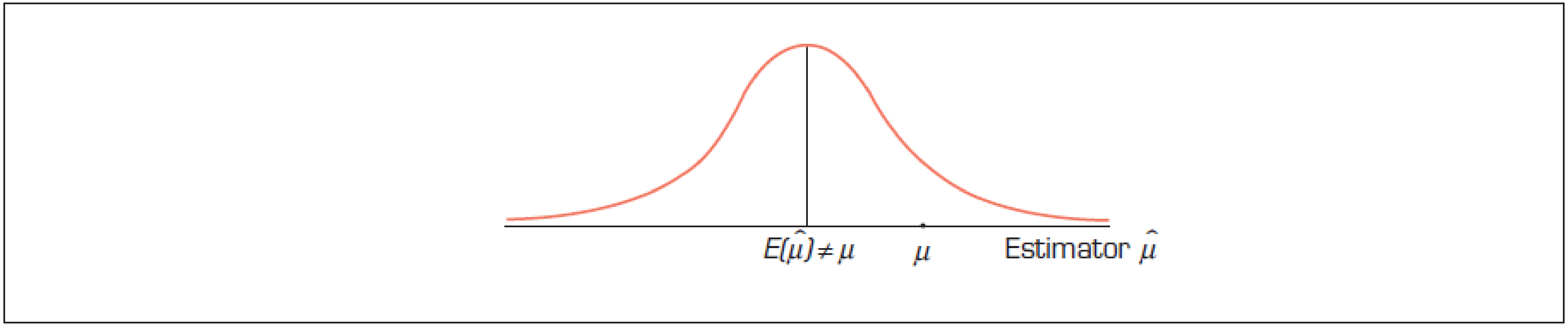


Figure 10.2 Sampling distribution of a biased estimator $\hat{\mu}$ of μ



Unbiased Estimators…

Examples of unbiased estimators

- The sample mean ($\bar{X}$) is an *unbiased* estimator of the population mean (μ), since: $E(\bar{X}) = \mu$
- The sample median (M) is an *unbiased* estimator of the population mean (μ), since: $E(M) = \mu$
- The sample variance (s^2) is an *unbiased* estimator of the population variance (σ^2), since: $E(s^2) = \sigma^2$

Desirable characteristics

A good estimator is

Unbiased: an unbiased estimator is one whose expected value is equal to the parameter it estimates.

Consistent: an unbiased estimator is said to be consistent if the difference between the estimator and the parameter grows smaller as the sample size increases.

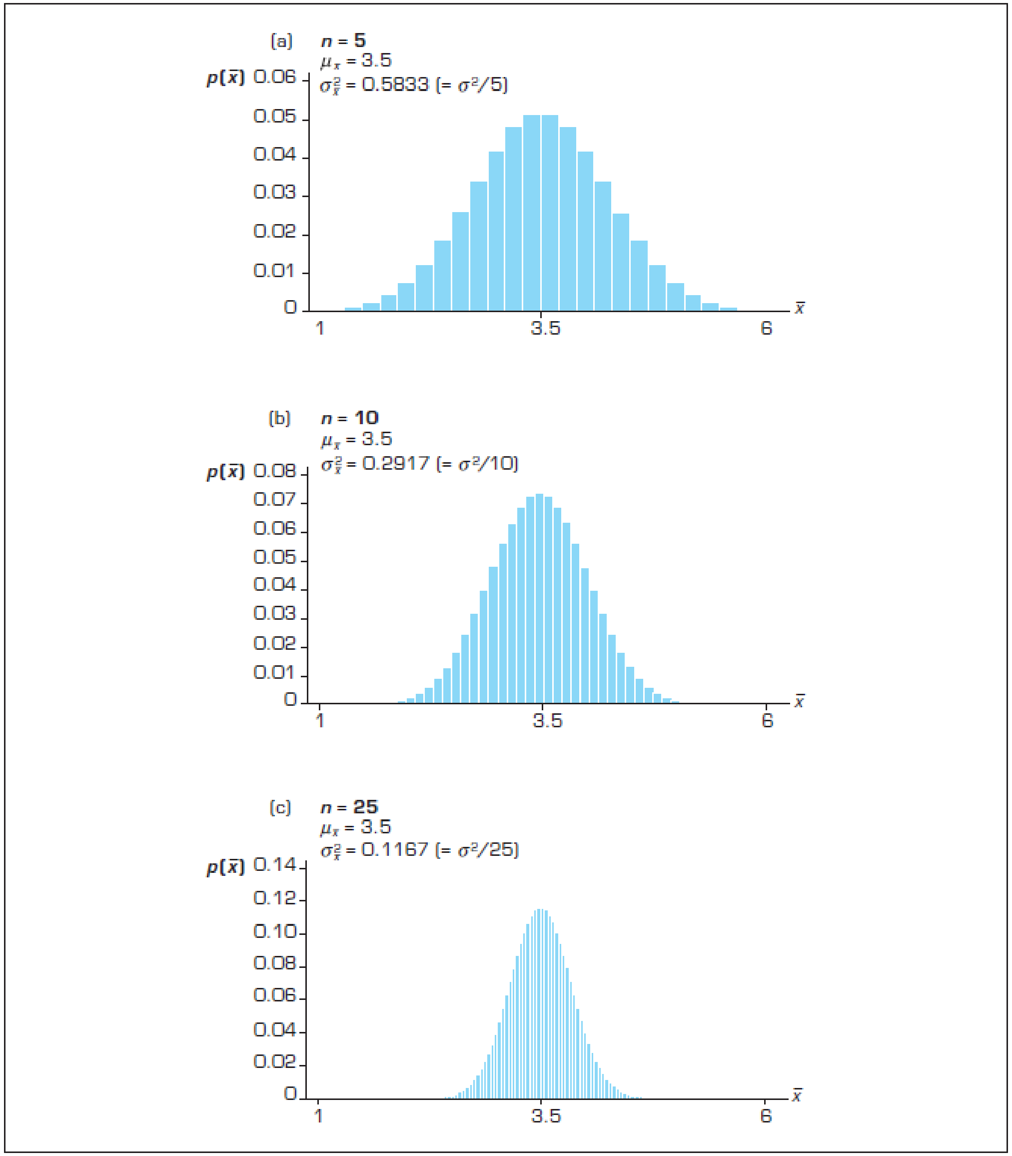
Consistent estimators…

The sample mean ($\bar{X}$) is a consistent estimator of μ because

(i) it is an unbiased estimator of μ and,

(ii) V($\bar{X}$) is σ^2/n and therefore as n grows larger, the variance of $\bar{X}$ grows smaller.

Figure 9.3 Sampling distributions of $\bar{X}$ when n = 5, 10 and 25



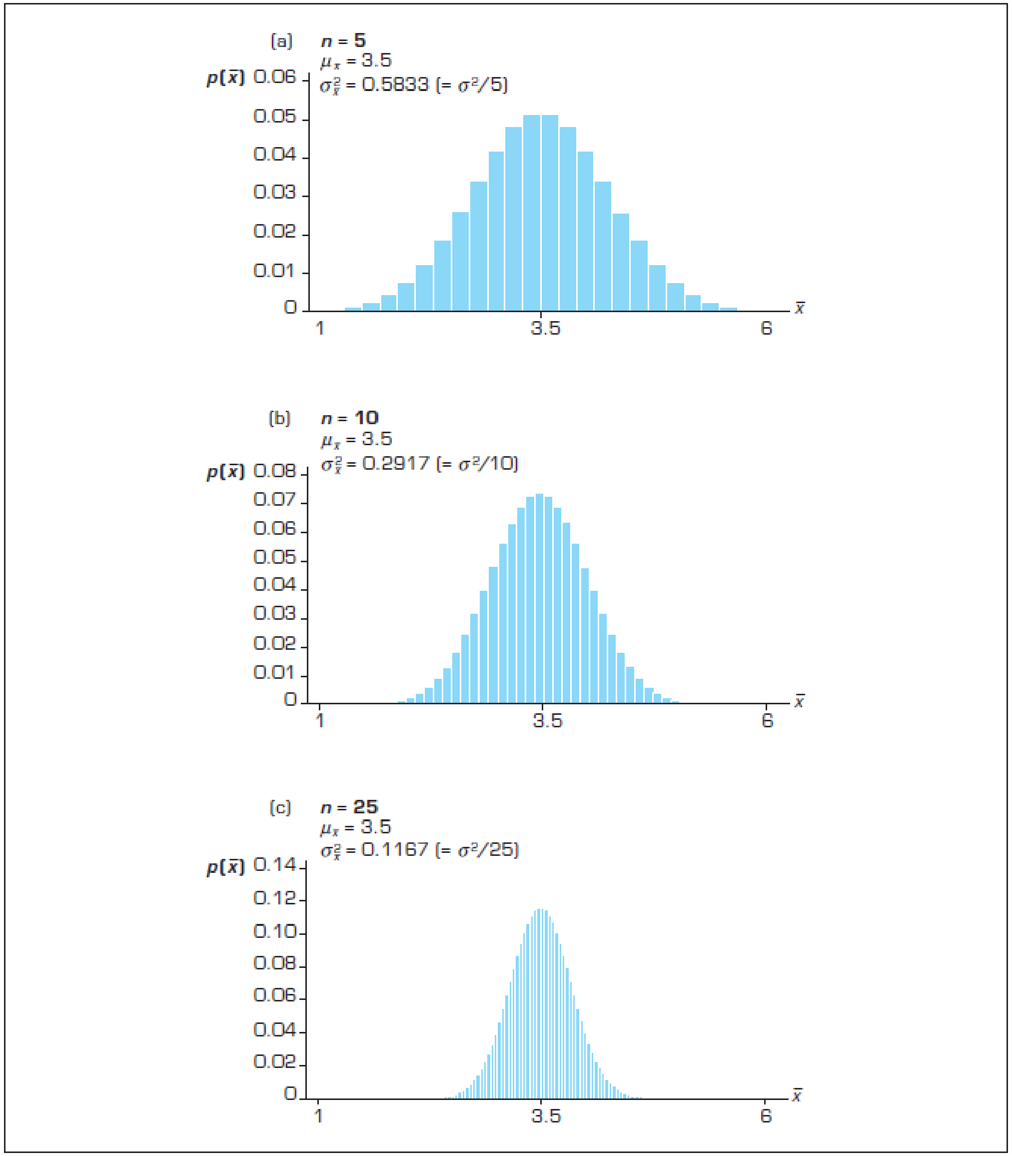
Consistent estimators…

The sample median (M) is a consistent estimator of μ as

(i) it is an unbiased estimator of μ and,

(ii) V(M) is $1.57\sigma^2/n$ and therefore as n grows larger, the variance of M grows smaller.

Figure 9.3 Sampling distributions of $\overline{X}$ when n = 5, 10 and 25



Desirable characteristics

A good estimator is

Unbiased: an unbiased estimator is one whose expected value is equal to the parameter it estimates.

Consistent: an unbiased estimator is said to be consistent if the difference between the estimator and the parameter grows smaller as the sample size increases.

Efficient: if there are two unbiased estimators available, the one with a smaller variance is said to be *relatively efficient*.

Efficient estimators…

For example, consider the two estimators of μ, the sample mean ($\bar{X}$) and the median, *M*. We have

- $E(\bar{X}) = \mu$; $V(\bar{X}) = \sigma^2/n$, and
- $E(M) = \mu$; $V(M) = 1.57(\sigma^2/n)$ $(> \sigma^2/n)$.

Both the sample mean ($\bar{X}$) and the sample median (*M*) are unbiased estimators of the population mean (μ). However, *M* has a greater variance than $\bar{X}$. So, the sample mean ($\bar{X}$) is *relatively efficient* when compared to the sample median (*M*).

'Best' estimator

The sample mean $(\bar{X})$ is

(i) unbiased,

(ii) consistent

(iii) relatively efficient

estimator of the population mean (μ).

Thus, the sample mean $(\bar{X})$ is the ‘best’ estimator of the population mean μ.

Interval estimator

How is an interval estimator produced from a sampling distribution? Recall from chapter 9:

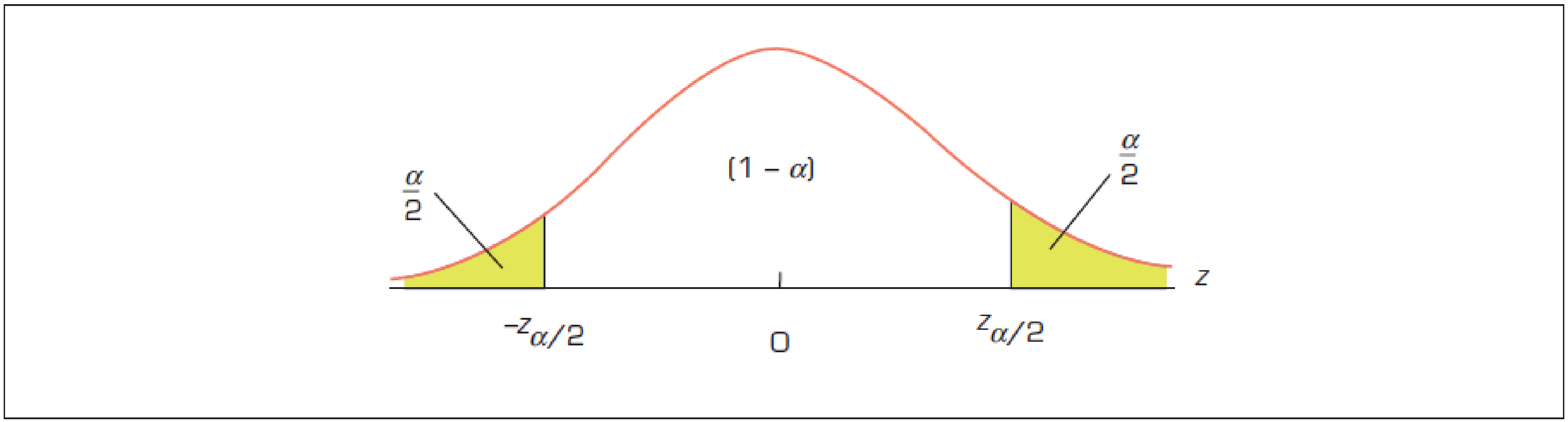
- To estimate μ, a sample of size *n* is drawn from the population, and its mean $(\bar{X})$ is calculated

- Apply the Central Limit Theorem: $\bar{X}$ is normally distributed and so we can construct the standardised statistic, which is the sampling distribution of the standardised sample mean:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim Normal(0,1)$$

Estimating μ when σ² is known…

Figure 10.4 Sampling distribution of $Z = \dfrac{X - \mu}{\sigma/\sqrt{n}}$



Estimating μ when σ^2 is known…

For example, for α = 0.05, the symmetry of the normal distribution with the sampling distribution of the sample mean leads to:

$$P(-1.96 \leq Z \leq 1.96) = 0.95 \quad \text{or} \quad P\left(-1.96 \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

$-z_{0.025}$ $z_{0.025}$

This can be written as

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

which becomes

$$\boxed{P\left(\mu-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95}$$

Estimating μ when σ² is known…

For example, for α = 0.05, the symmetry of the normal distribution with the sampling distribution of the sample mean leads to:

$$P(-1.96 \le Z \le 1.96) = 0.95 \quad \text{or} \quad P(-1.96 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le 1.96) = 0.95$$

$-z_{0.025}$ $z_{0.025}$

This can be written as

$$P(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{x}-\mu \le 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

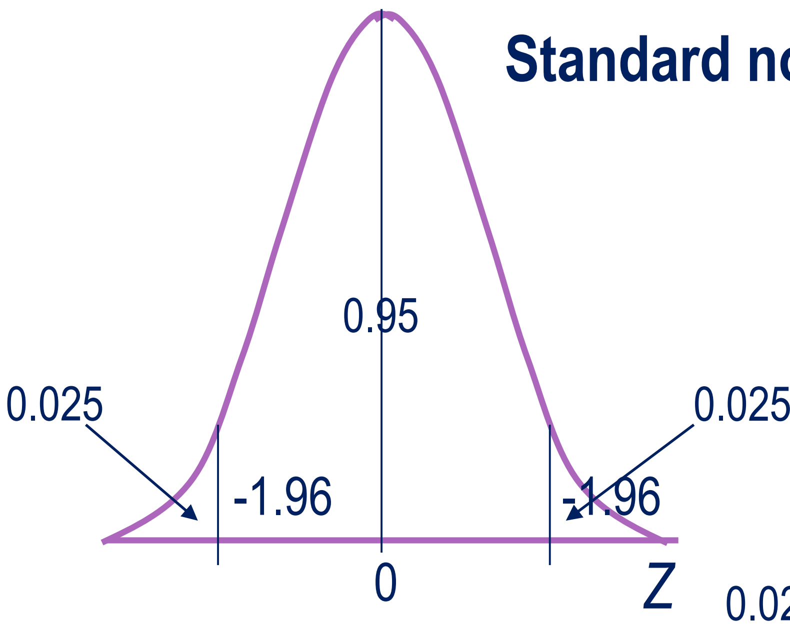
which becomes

$$\boxed{P(\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \bar{x} \le \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95}$$

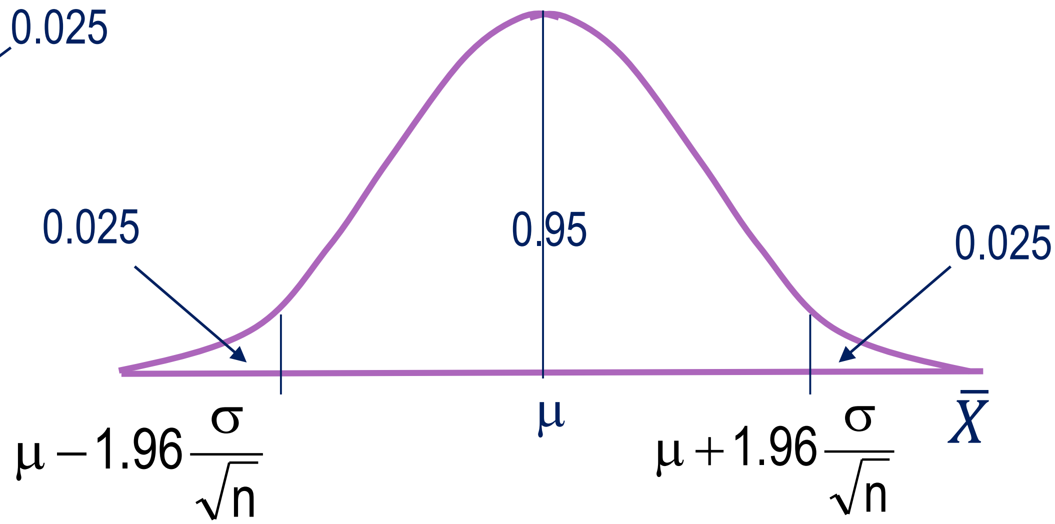
i.e., the probability that the sample mean is in this interval is 95%: this is the 95% confidence interval

Estimating μ when σ^2 is known…

Standard normal distribution Z



Normal distribution of $\bar{X}$



Estimating μ when σ² is known…

In general, $$P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{X} \le \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

This leads to the relationship

$$\boxed{P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha}$$

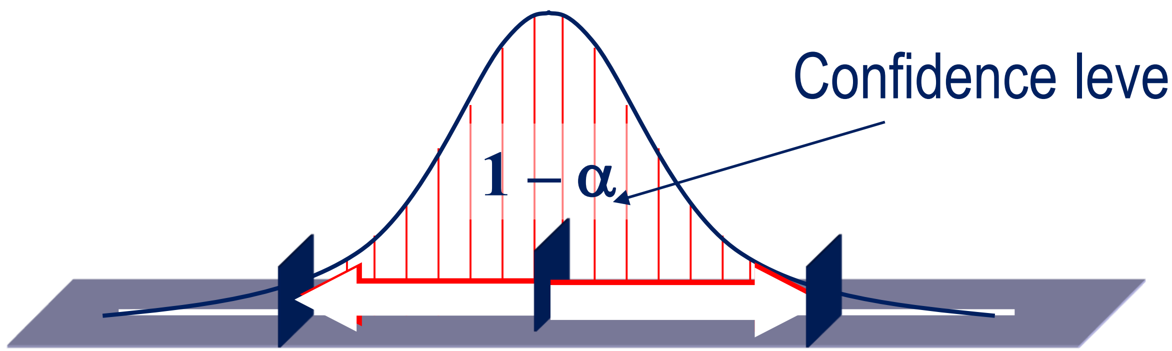
Estimating μ when σ^2 is known…

$(1-\alpha)100\%$ of all the values of $\bar{X}$ obtained in repeated sampling from this distribution, construct an interval

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

that includes (covers) the population mean μ.

Estimating μ when σ² is known…



$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \bar{X} \qquad \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Lower confidence limit

Upper confidence limit

$$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Estimating μ when σ^2 is known…

Three commonly used *confidence levels*, $(1 - \alpha)$

Confidence level, $(1-\alpha)$	α	$\alpha/2$	$z_{\alpha/2}$
0.90	0.10	0.05	1.645 (or 1.64 or 1.65)
0.95	0.05	0.025	1.96
0.99	0.01	0.005	2.575 (or 2.57 or 2.58)

Example 1

(Example 10.1, page 375)

The sponsors of television shows targeted at children wanted to know the amount of time children spend watching television, since the types and number of programs and commercials presented are greatly influenced by this information. As a result, a survey was conducted to estimate the average number of hours Australian children spend watching television per week. From past experience, it is known that the population standard deviation σ is 8.0 hours.

The following are the data gathered from a sample of 100 children. Find the 95% confidence interval estimate of the average number of hours Australian children spend watching television.

Example 1

Amount of time spent watching television each week

39.7	21.5	40.6	15.5	43.9	33.0	21.0	15.8	27.1	23.8	18.3	23.4	20.6
28.4	29.8	41.3	36.8	35.5	27.2	21.0	19.7	22.8	30.0	22.1	30.8	34.7
15.0	23.6	38.9	29.1	28.7	29.3	20.3	36.1	21.6	15.1	43.8	29.0	30.2
26.5	20.5	24.1	29.3	14.7	13.9	37.1	32.5	24.4	22.9	24.5	19.5	29.9
46.4	31.6	20.6	38.0	21.8	23.2	22.0	35.3	17.0	24.4	34.9	24.0	32.9
15.1	23.4	19.5	26.5	42.4	38.6	23.4	37.8	26.5	22.7	27.0	16.4	39.4
38.7	9.5	20.6	21.3	33.5	23.0	35.7	23.4	30.8	27.7	25.2	50.3	31.3
28.9	31.2	15.6	32.8	17.0	11.3	26.9	26.9	21.9				

Example 1: Solution

Calculating manually

Therefore, a 95% confidence interval estimator of μ is

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 27.191 \pm 1.96\frac{8.0}{\sqrt{100}}$$

$$= 27.191 \pm 1.568$$

$$= [27.191 - 1.568, 27.191 + 1.568] = [25.623, 28.759]$$

We therefore estimate that the average number of hours children spend watching television each week lies somewhere between

LCL = 25.62 hours and UCL = 28.76 hours

Interpreting the results

That is, the average time spent watching TV by Australian children is between 25.6 hours and 28.8 hours. This type of estimate is correct 95% of the time.

As a consequence, a network executive may decide (for example) that, since the average child watches at least 25.6 hours of television per week, the number of commercials children see is sufficiently high to satisfy the programs' sponsors.

Incidentally, the media often refer to the 95% confidence interval as '19 out of 20 times', which emphasises the **long-run** aspect of a confidence interval.