Data Analytics

ECON 1008, Semester 1, 2019

Giulio Zanella
University of Adelaide
School of Economics

CHAPTER 10,

Estimation: continues…

Best point estimator for the mean

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

(i) unbiased,
(ii) consistent
(iii) relatively efficient

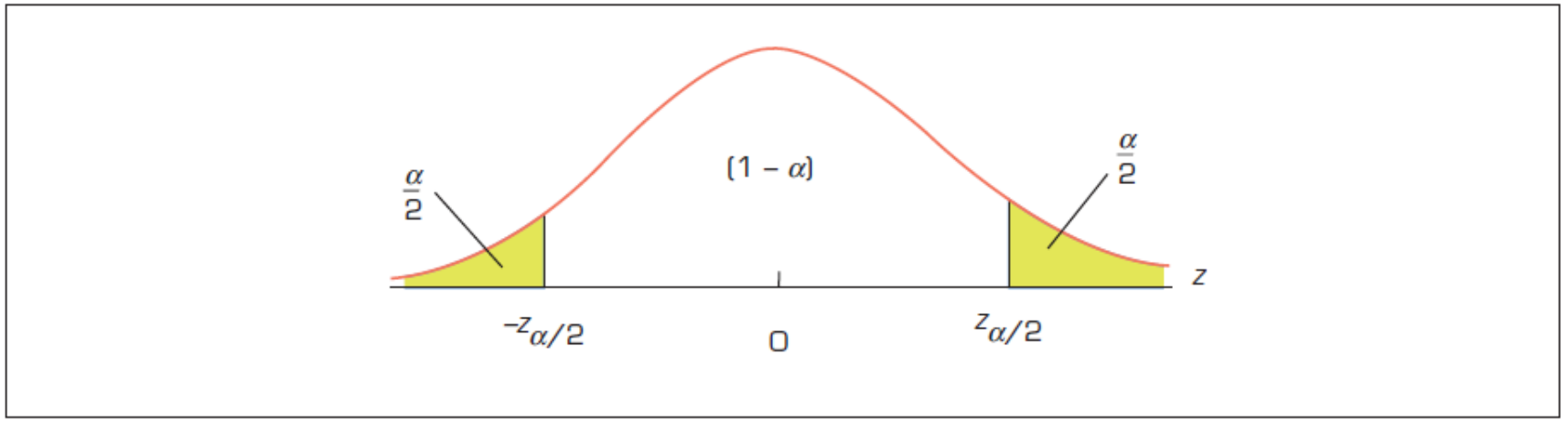
estimator of the **population mean (μ).**

Thus, $\bar{X}$ is the 'best' estimator of μ.

Interval estimator: key ideas

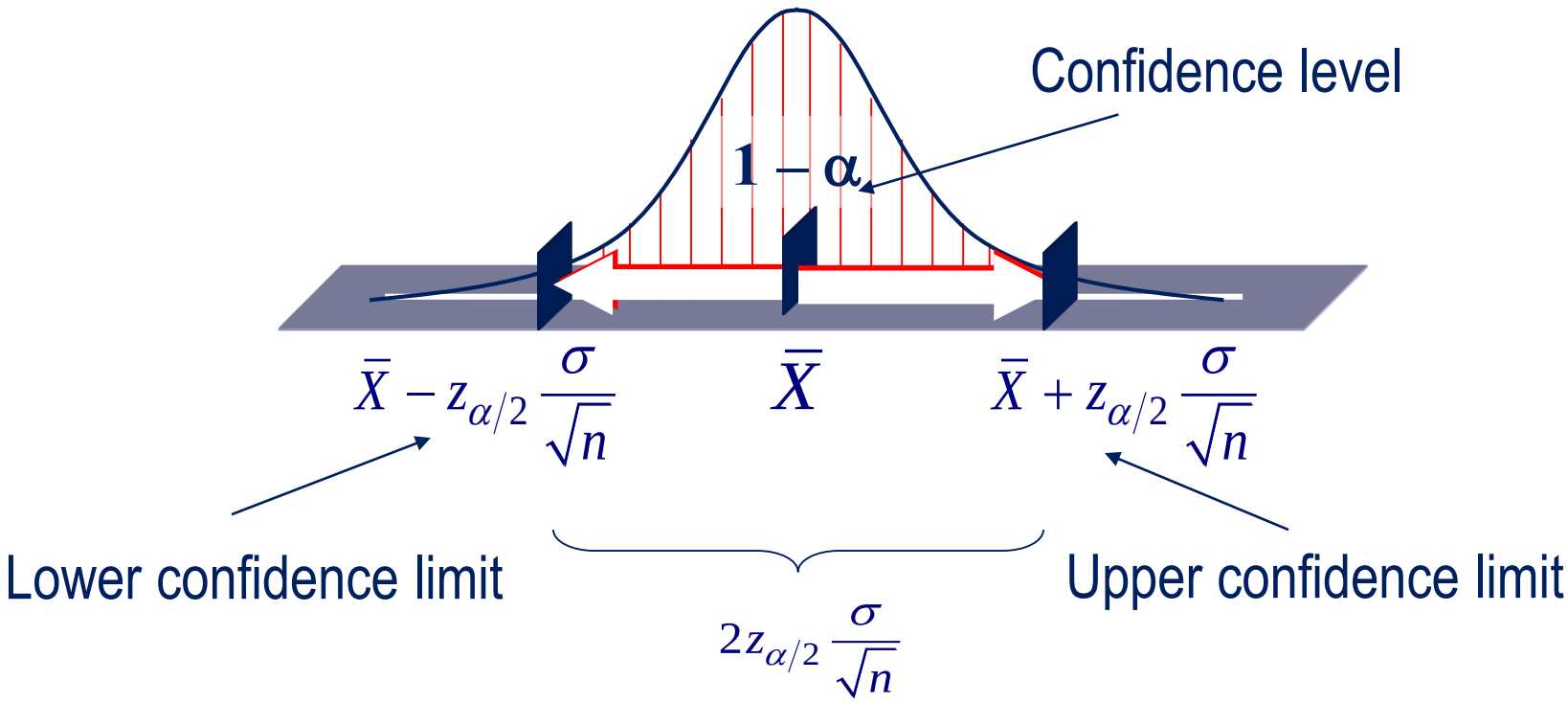
- Construct a confidence interval for μ from a sample, leveraging the CLT which ensures $\bar{X}$ is normally distributed

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



- For α = 0.05 (i.e., 5%), $P(\bar{X} - 1.96\dfrac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\dfrac{\sigma}{\sqrt{n}}) = 0.95$
- Interval $[\bar{X} - 1.96\dfrac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\dfrac{\sigma}{\sqrt{n}}]$ contains μ 95% of times

Interval estimator: key ideas



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

NOTICE: we are assuming that the population variance is know!

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 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

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, "with 95% confidence" somewhere between

LCL = 25.62 hours and UCL = 28.76 hours

Interpreting the results

- 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.

Factors that determine the width of a confidence interval for μ

Half the width of a confidence interval for μ,

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

The width of the confidence interval ($2B$) is affected by three factors:

1. Level of confidence, 1-α. (B is directly proportional to α, via $z_{\alpha/2}$)
2. Population standard deviation, σ. (B is directly proportional to σ)
3. Sample size, n. (B is inversely proportional to $\sqrt{n}$)

Factors that determine the width of a confidence interval for μ...

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

1. Level of confidence, 1-α. (*B* is directly proportional to α, via $z_{\alpha/2}$)

Confidence level, (1-α)	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)

As the level of confidence (1-α) increases from 0.90 to 0.99, $z_{\alpha/2}$ also increases (from 1.645 to 2.575). Therefore, in general, when the level of confidence increases, the width also increases.

Factors that determine the width of a confidence interval for μ...

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

2. Population standard deviation, σ.

   B is directly proportional to σ and *w*hen σ increases, the width (2B) also increases. For example, if σ is doubled, then the width would also be doubled or if σ is halved, then the width would be halved as well.

3. Sample size, *n*.

   B is inversely proportional to $\sqrt{n}$. When *n* increases, the width (2B) decreases. For example, if n increases by 4 times, the width would be halved.

The width of the confidence interval

A wide interval provides little information.

For example, suppose we estimate with 95% confidence that an accountant's average starting salary is between $15 000 and $100 000.

Contrast this with: a 95% confidence interval estimate of starting salaries between $42 000 and $45 000.

The second estimate is much narrower, providing accounting students more precise information about starting salaries.

Estimating the population mean μ when the population variance σ^2 is unknown

Recall that when σ is known, the statistic

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$ is (standard) normally distributed

(approximately, thanks to the CLT)

Estimating μ when σ^2 is unknown…

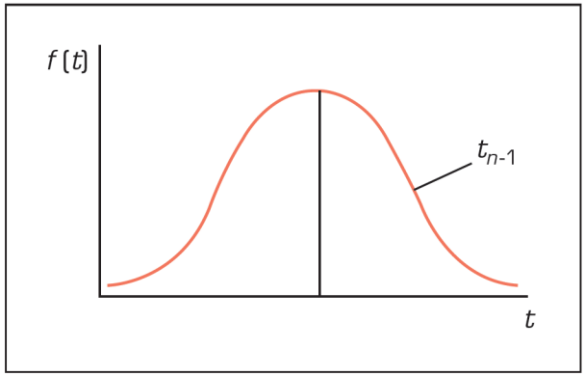
When σ is unknown, the best we can do is to use its point estimator s (sample s.d.). BUT the *z*-statistic is no longer standard normally distributed, but becomes a ***t*-statistic**

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

If the sample is drawn from a normal population,
the t-statistic is **Student t distributed**
with (*n*-1) degrees of freedom.

The *t*-distribution

Figure 10.5 Student *t* distribution



The t distribution is bell-shaped and symmetrical around zero, like the normal distribution

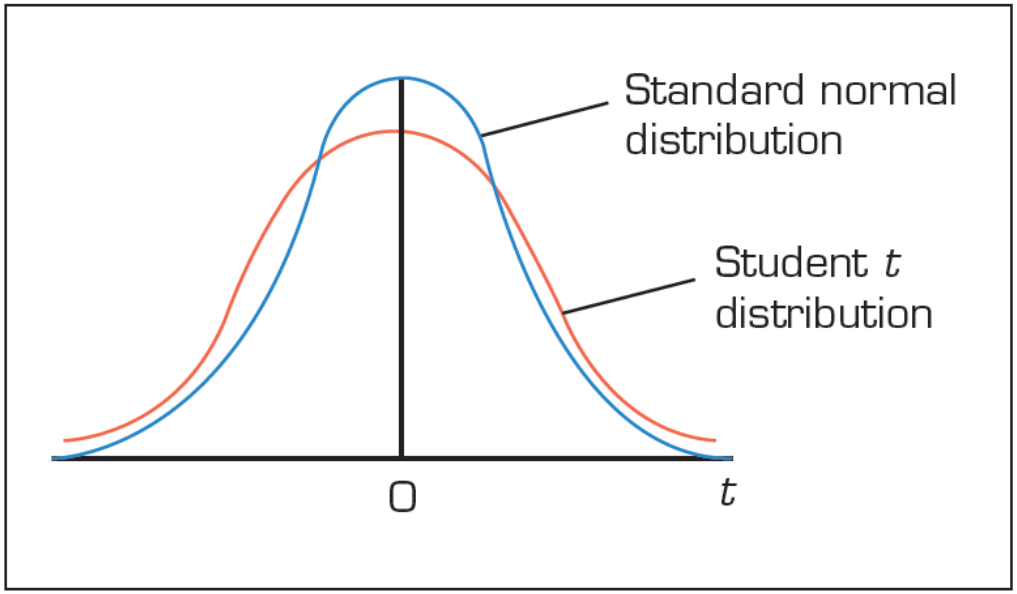
The 'degrees of freedom', *n*-1, a function of the sample size, determines how spread the distribution is (compared to normal distribution).

The *t*-distribution

E(t) = 0 and var (t) = $\frac{(n-1)}{(n-3)}$ > 1. When n is large, var(t) $\rightarrow$ 1.

Therefore, for large values of *n*, the t-distribution is similar to a standard normal distribution.

Figure 10.6 Student *t* and standard normal distribution



Probability Calculations for the *t* Distribution

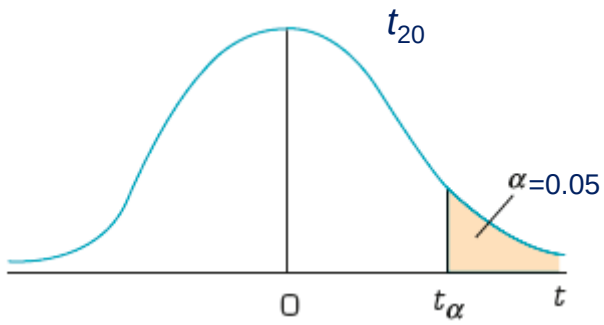
Like the z table, the t table provides critical values for various probabilities of interest but slightly in a different way.

The form of the (tail) probabilities that appear in the table:

$$P(t > t_{\alpha,d.f.}) = \alpha$$

For a given degree of freedom, and for a predetermined right-hand tail probability α, the entry in the table is the corresponding t_{α}.

These values are used in computing interval estimates and performing hypothesis tests when σ is unknown (more later)



$t_{0.05,20} = 1.725$

Degrees of Freedom	$t_{.100}$	$t_{.05}$	$t_{.025}$	$t_{.01}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.92	4.303	6.965	9.925
.	.	.	.	.	.
.	.	.	.	.	.
20	1.325	1.725	2.086	2.528	2.845
.	.	.	.	.	.
.	.	.	.	.	.
200	1.286	1.653	1.972	2.345	2.601
∞	1.282	1.645	1.96	2.326	2.576

Above 200, the standard normal distribution is a good approximation and we switch to the z-table

Example 2

(Example 10.2, page 393)

XM10-02 As you are probably aware, a taxi fare is determined by distance travelled as well as the amount of time taken for the trip. In preparing to apply for a rate increase, the general manager of a fleet of taxis wanted to know the distance customers travel by taxi on an average trip. She organised a survey in which she asked taxi drivers to record the number of kilometres (to the nearest one-tenth) travelled by randomly selected customers. A sample of 41 customers was produced.

The results appear below. The general manager wants to estimate the mean distance travelled with 95% confidence.

Example 2

Distance travelled by taxi (km)

8.2	9.1	11.2	5.0	6.4	9.5	10.1	7.9	8.3	6.8	6.9	7.9	1.1	6.7	11.4	6.9
6.5	8.0	1.5	8.2	7.6	14.1	7.0	10.0	7.1	8.0	8.1	4.4	5.9	2.3	13.3	9.2
2.8	13.0	8.3	10.4	9.0	3.5	9.8	6.5	7.7							

Example 2: Solution…

A 95% confidence interval estimator of μ is

$$[\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}]$$

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 7.70 \pm 2.021\frac{2.93}{\sqrt{41}}$$

$$= 7.70 \pm 0.92$$

We therefore estimate that the mean distance travelled by taxi lies between

LCL = 6.77kms and UCL = 8.62kms

Interpreting the results

We estimate that the mean distance travelled by taxi lies between 6.77 and 8.62 kilometres.

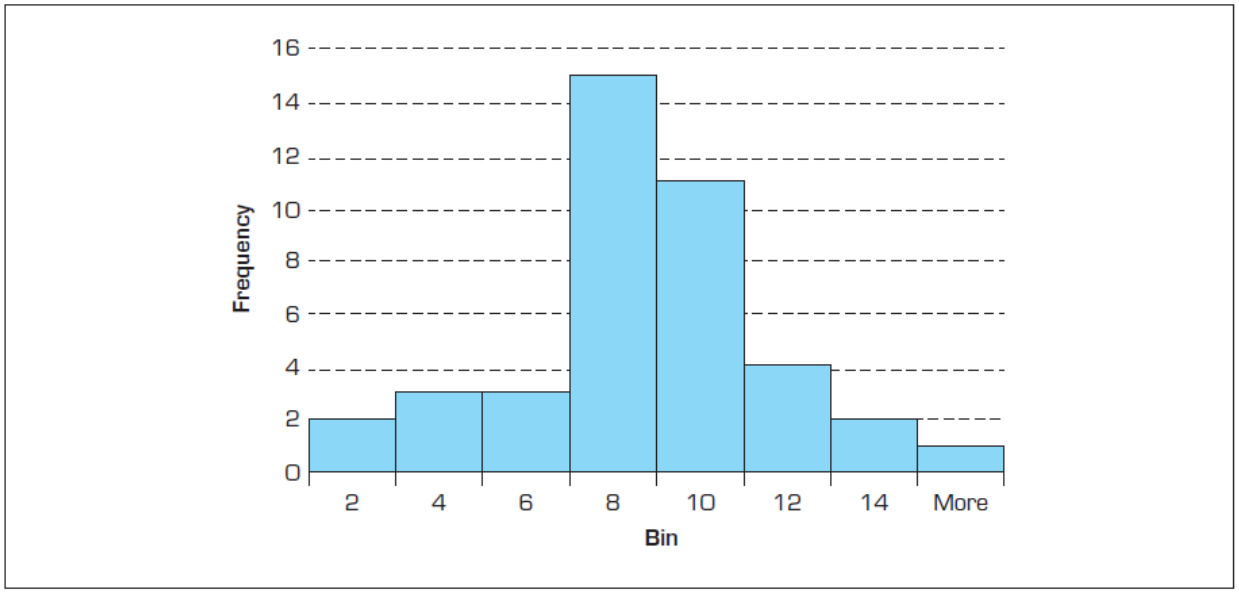
It is worth noting that the accuracy of the interval estimate is dependent upon the validity of the sampling process and the distribution of the distances (they are required to be normal). If the distribution is extremely non-normal, the inference may be invalid.

Checking the required conditions

We can plot the histogram of the data set using Excel to check the shape of the distribution.

The variable looks normally distributed or at least not extremely non-normal.

Figure 10.9 Histogram for Example 10.2



Example 3

Suppose that the amount of time teenagers spend on the Internet is normally distributed, with a standard deviation of 1.5 hours. A sample of 100 teenagers is selected at random, and the sample mean is computed as 6.5 hours. Determine the 95% confidence interval estimate of the population mean. Interpret

Example 3

Suppose that the amount of time teenagers spend on the Internet is normally distributed, with a standard deviation of 1.5 hours. A sample of 100 teenagers is selected at random, and the sample mean is computed as 6.5 hours. Determine the 95% confidence interval estimate of the population mean. Interpret

$$[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}]$$

Example 3

Suppose that the amount of time teenagers spend on the Internet is normally distributed, with a standard deviation of 1.5 hours. A sample of 100 teenagers is selected at random, and the sample mean is computed as 6.5 hours. Determine the 95% confidence interval estimate of the population mean. Interpret

$$[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}]$$

$$[6.5 - 1.96(1.5/10) , 6.5 + 1.96(1.5/10)]$$

$$[6.206 , 6.794]$$

Example 3

Suppose that the amount of time teenagers spend on the Internet is normally distributed, with a standard deviation of 1.5 hours. A sample of 100 teenagers is selected at random, and the sample mean is computed as 6.5 hours. Determine the 95% confidence interval estimate of the population mean. Interpret

6.206 to 6.794 hours.

If we repeatedly draw samples of size 100 from the population of teenagers, 95% of the values of x_bar will be such that the population mean amount of time teenagers spend on the internet, the true population parameter, would lie somewhere between 6.206 and 6.794, and 5% will produce intervals that would not include mu.

Example 4

- Find and interpret a 98% confidence interval for the mean number of animals visited by a veterinarian per day. A random sample of 21 veterinarians, found that they had a sample mean of 12.5 animals and a sample variance of 9 animals.

Example 4

- Find and interpret a 98% confidence interval for the mean number of animals visited by a veterinarian per day. A random sample of 21 veterinarians, found that they had a sample mean of 12.5 animals and a sample variance of 9 animals.

$$[\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}]$$

Example 4

- Find and interpret a 98% confidence interval for the mean number of animals visited by a veterinarian per day. A random sample of 21 veterinarians, found that they had a sample mean of 12.5 animals and a sample variance of 9 animals.

$$[\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}]$$

$$[12.5 - 2.53\ (3/4.58)\ ,\ 12.5 + 2.53(3/4.58)]$$

$$[10.84\ ,\ 14.16]$$

Example 4

- Find and interpret a 98% confidence interval for the mean number of animals visited by a veterinarian per day. A random sample of 21 veterinarians, found that they had a sample mean of 12.5 animals and a sample variance of 9 animals.

$$[10.84 , 14.16]$$

- We are 98% confident that the population mean number of animals visited by a veterinarian per day lies between 10.84 animals and 14.16 animals.

Determining the required sample size

- A wide-interval estimator provides little information.
- A narrower interval provides more meaningful information.

Determining the sample size for estimating a population mean

We can control the width of the interval estimate by changing the sample size.

Thus, we determine the interval width first, and derive the required sample size.

The phrase ‘estimate the mean to within B units’ translates to an interval estimate of the form

$$\bar{X} \pm B \quad \text{where} \quad B = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Determining the sample size for estimating a population mean

By rearrangement, the required sample size to estimate the mean can be written as

$$n = \left[\frac{z_{\alpha/2}\sigma}{B} \right]^2$$

Note:

- As n is proportional to $z_{\alpha/2}$, higher confidence level $(1-\alpha)$ requires a larger sample size.
- As n is proportional to σ, larger values of the standard deviation (σ) requires a larger sample size.
- As n is inversely proportional to B, narrower confidence intervals (B) requires a larger sample size.

Example 5

To estimate the amount of lumber that can be harvested in a tract of land, the mean diameter of trees in the tract must be estimated to within 1 cm with 99% confidence. What sample size should be taken? Assume that diameters are normally distributed with standard deviation σ = 6 cm.

Example 5: Solution

The estimate accuracy is ±1 cm. That is, B = 1.

The confidence level 99% leads to α = 0.01, thus

$$z_{\alpha/2} = z_{0.005} = 2.575.$$

The population standard deviation σ is 6.

We compute

$$n = \left[\frac{z_{\alpha/2}\sigma}{B}\right]^2 = \left[\frac{2.575(6)}{1}\right]^2 = 239$$