MODULE TWO: MEASURING UNCERTAINTY; AND DRAWING CONCLUSIONS ABOUT POPULATIONS BASED ON SAMPLE DATA

TOPIC 6: CONFIDENCE INTERVALS

+ Learning Objectives

At the completion of this topic, you should be able to:

- calculate estimates and their standard errors
- construct and interpret confidence interval estimates for the mean and the proportion
- determine the sample size necessary to develop a confidence interval for the mean or proportion

+Confidence Intervals

A confidence interval gives a **range** of values

- Takes into consideration variation in sample statistics from sample to sample
- Based on observations from one random sample
- Gives information about closeness to unknown population parameters
- Stated in terms of level of confidence
- Can never be 100% confident

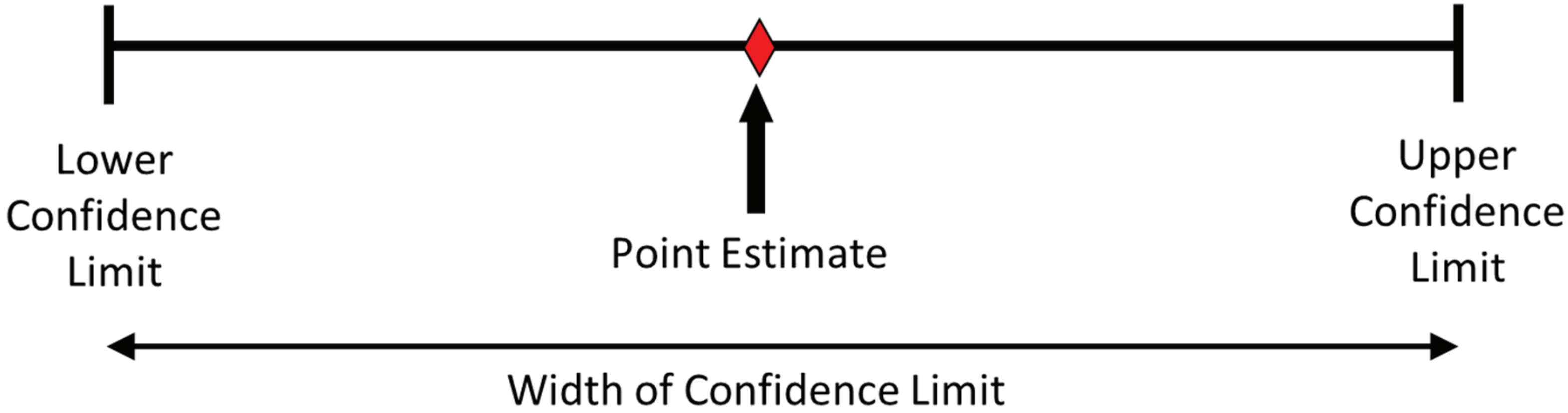
+Point Estimates

We can estimate a Population Parameter…		…with a Sample Statistic (Point Estimate)
Mean	μ	$\overline{X}$
Proportion	π	p

+Point and Interval Estimates

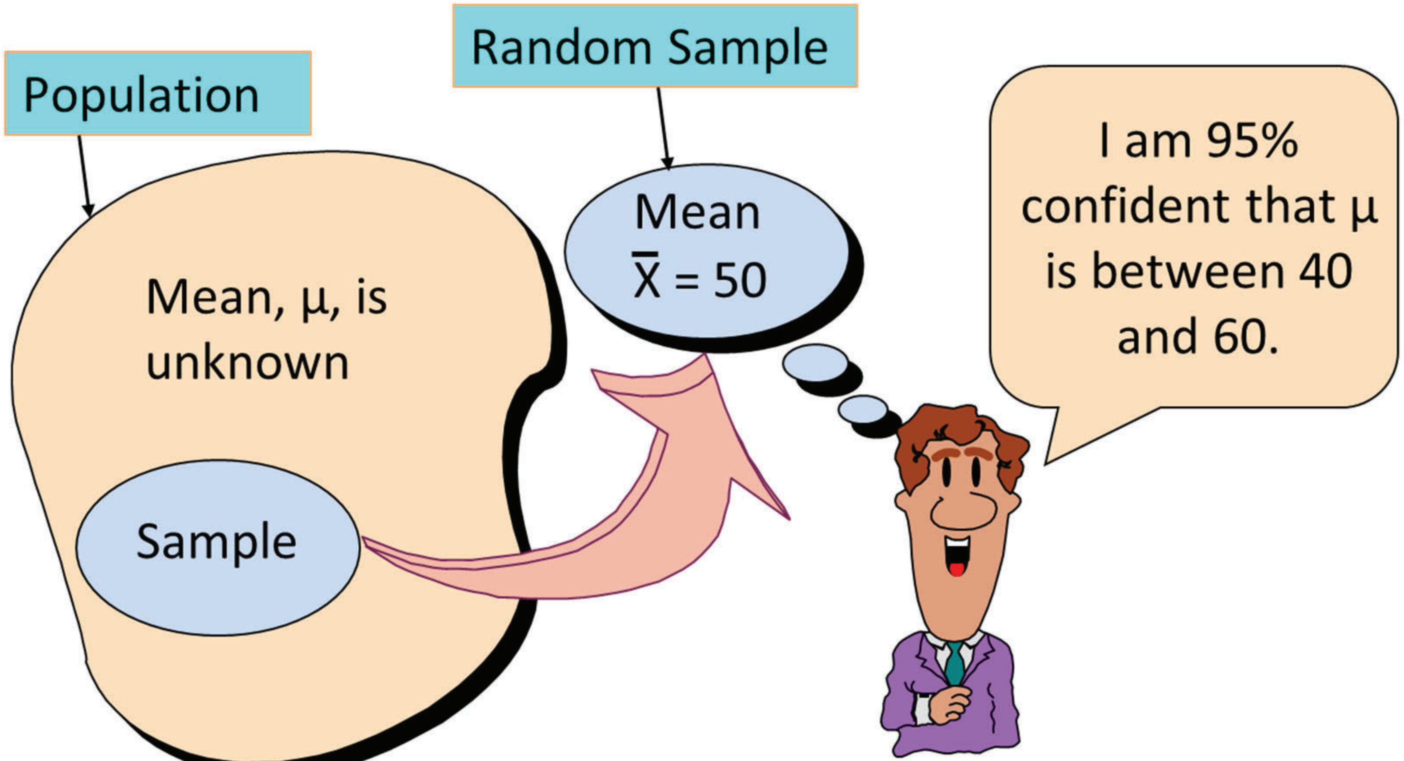
A **point estimate** is the value of a single sample statistic

A **confidence interval** provides a range of values constructed around the point estimate



Example 1: Average sales is between 45K to 50k
Example 2: Proportion of customers who will pay by cash is between 15–20%

+Estimation Process

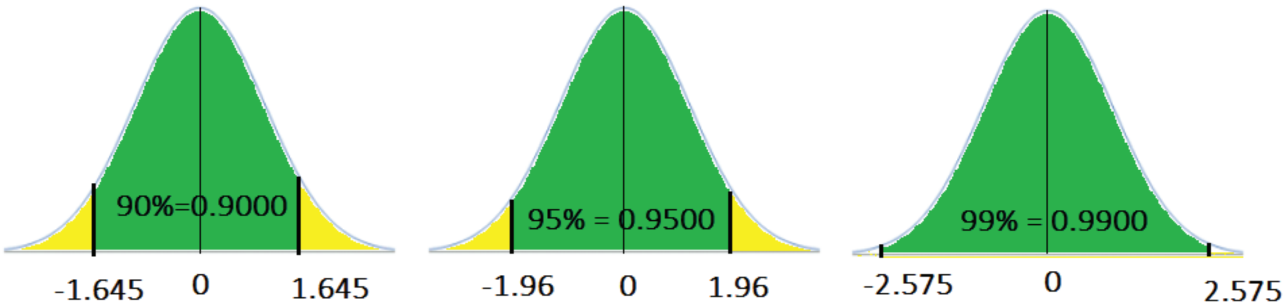


Sampling Error/Margin of error = Difference between sample statistic and parameter

Numerical data: Sample error = $|\mu - \bar{x}|$ or $|\sigma - s|$

Categorical data: Sample error = $|\pi - p|$

+Level of Cor



Common confidence levels = 90%, 95% or 99% (α is therefore 0.1, 0.05 or 0.01)

- Also written (1 - α) = **0.90, 0.95 or 0.99**

A relative frequency interpretation

- In the long run, 90%, 95% or 99% of all the confidence intervals that can be constructed (in repeated samples) will contain the unknown true parameter

A specific interval either will contain or will not contain the true parameter

Confidence Level	***Z value***
80%	1.28
90%	**1.645**
95%	**1.96**
98%	2.33
99%	**2.576**
99.8%	3.08
99.9%	3.27

+Confidence Intervals for Different Samples

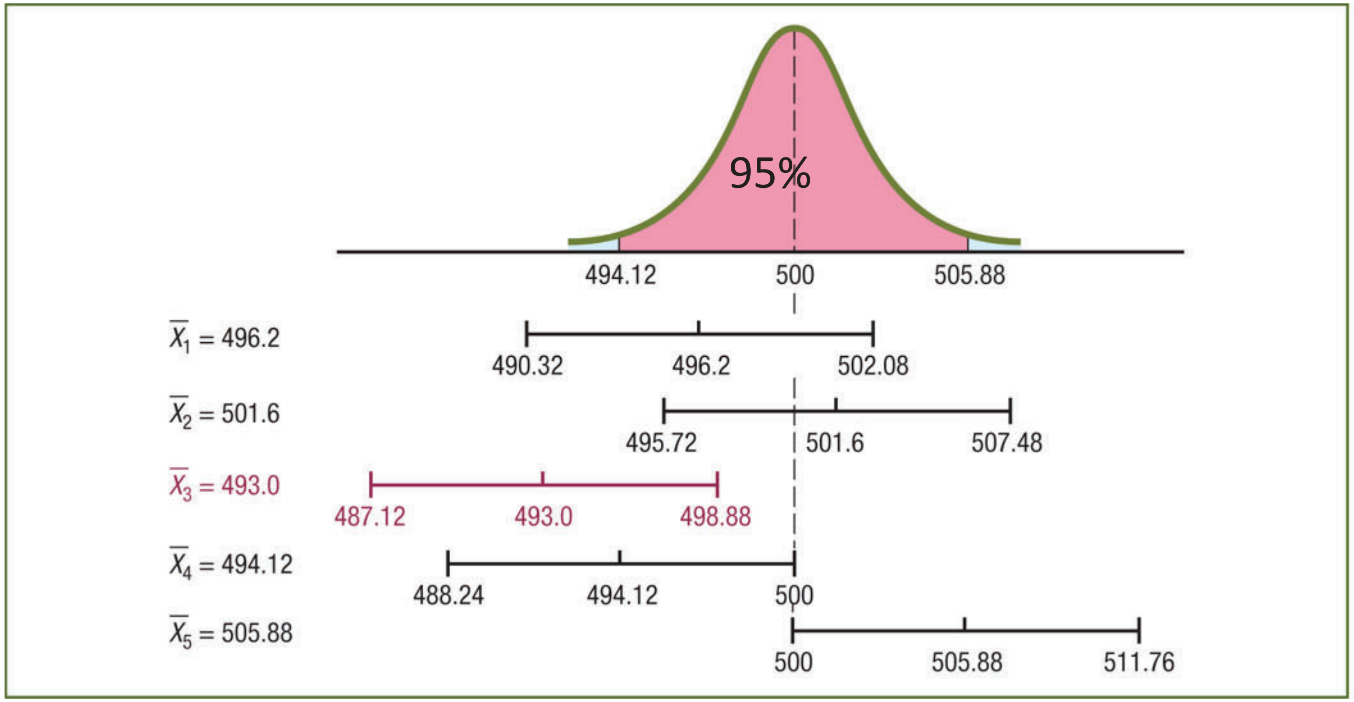


Figure 8.1
Confidence interval estimates for five different samples of $n = 25$ taken from a population where $\mu = 500$ and $\sigma = 15$

+Normal Curve for 95% and 99% Level of Confidence

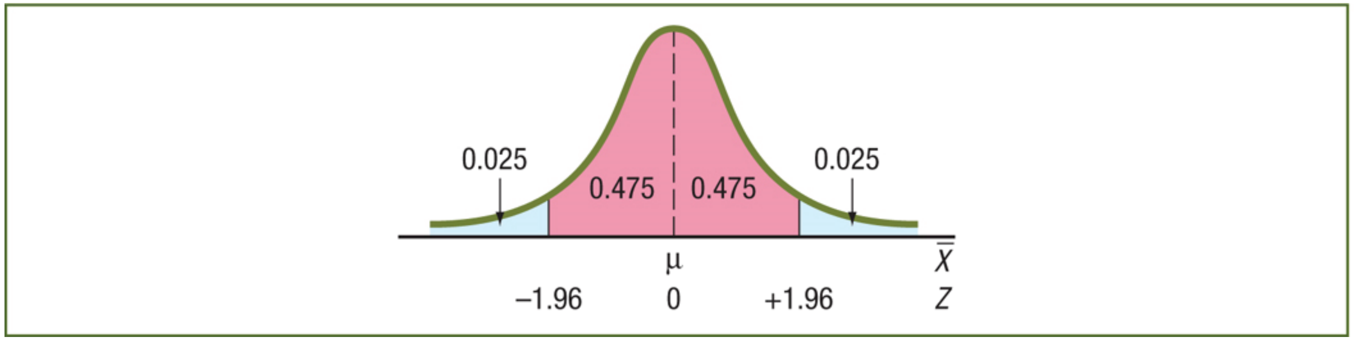


Figure 8.2
Normal curve for determining the Z value needed for 95% confidence

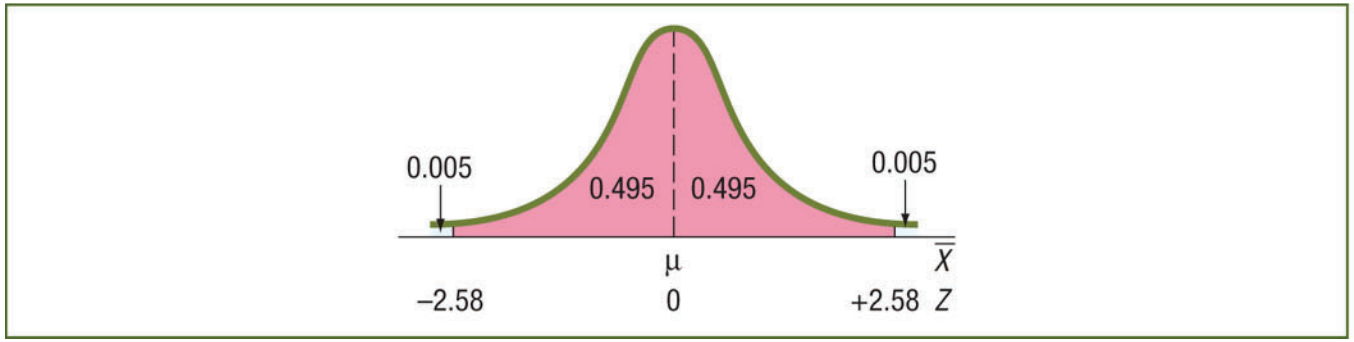
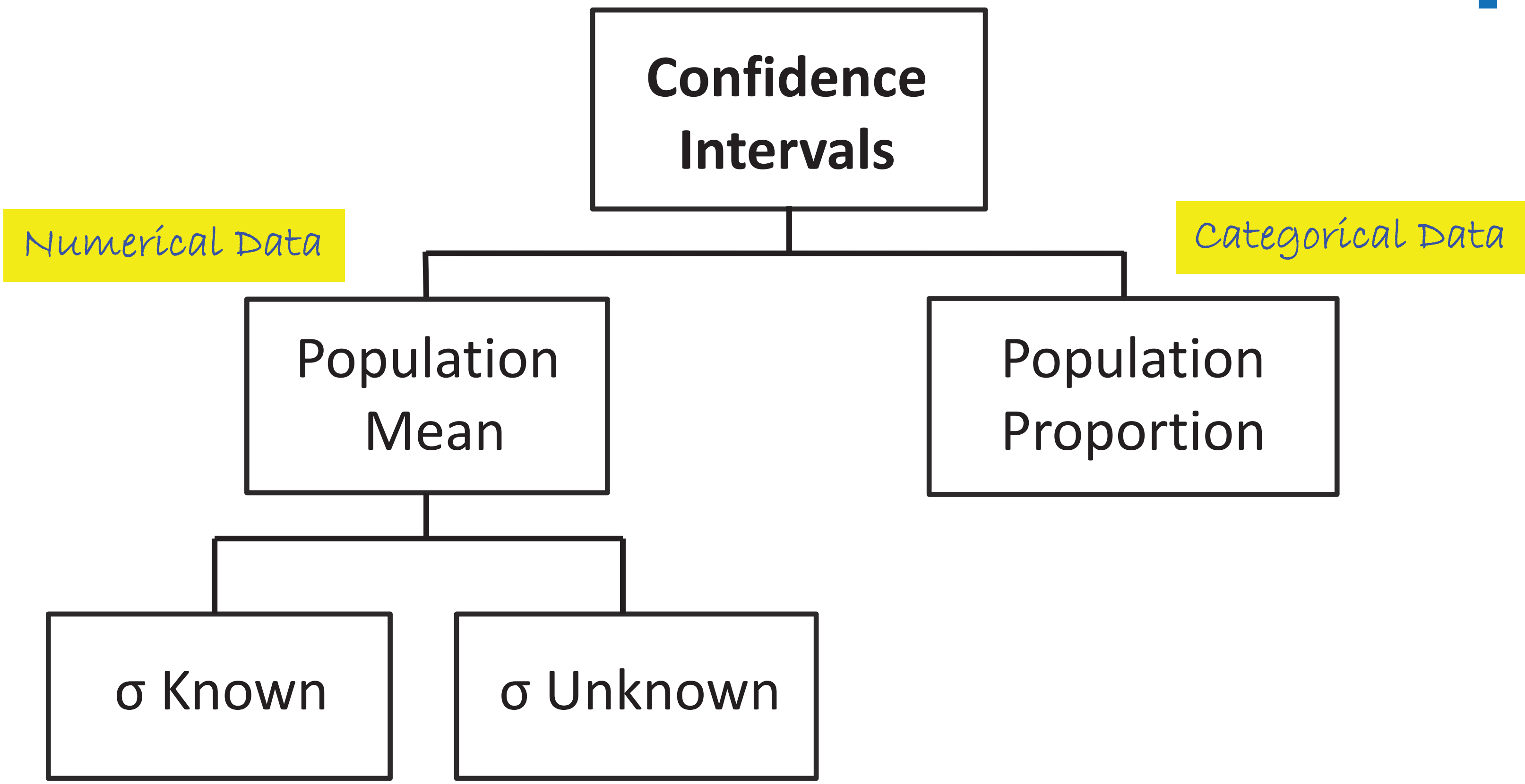


Figure 8.3
Normal curve for determining the Z value needed for 99% confidence

+Confidence Intervals



+Confidence Interval

The general formula for all confidence intervals is:

Point Estimate ± $\underbrace{\text{(Critical Value)*(Standard Error)}}_{\text{Margin of Error/Sampling Error}}$

Represents confidence for which the interval will contain the unknown population parameter

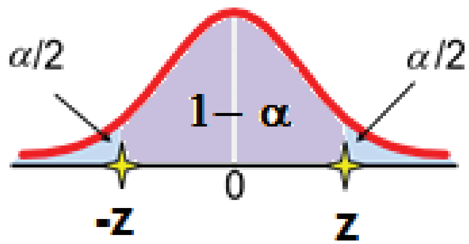
Interval estimate for population mean: $\bar{x}$ ± Margin of error

Interval estimate for population proportion: p ± Margin of error

+8.1 Confidence Interval Estimation for the Mean (σ Known)

Assumptions:

- Population standard deviation σ is known
- Population is normally distributed
- If population is not normal, use Central Limit Theorem



Confidence interval estimate

$$\overline{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

Where: $\overline{X}$ is the point estimate

Z is the normal distribution critical value for a probability of $\alpha/2$ in each tail

$\sigma/\sqrt{n}$ is the standard error

+Confidence Interval (σ Known)

Example:

A sample of 11 pizza shops from a large normal population has a mean pizza cooking time of 2.2 minutes. We know from past testing that the population standard deviation is 0.35 minutes

Determine a 95% confidence interval for the true mean pizza cooking time of the population

Variable: pizza cooking time
Sample mean = xbar = 2.2
population standard deviation = σ = 0.35
Sample size = n = 11
Data Distribution : Normal
Sampling distribution of sample mean : Normal

+Confidence Interval (σ Known)

$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

$$= 2.2 \pm 1.96 \frac{0.35}{\sqrt{11}}$$

$$= 2.2 \pm 0.2068$$

$$= 1.9932 \leq \mu \leq 2.4068$$

Confidence Level	*Z value*
80%	1.28
90%	**1.645**
$\boxed{\textbf{95\%}}$	$\boxed{\textbf{1.96}}$
98%	2.33
99%	**2.576**
99.8%	3.08
99.9%	3.27

We are 95% confident that the true mean pizza cooking time is somewhere between 1.9932 and 2.4068 minutes

Note: Although the true mean may or may not be in this interval, 95% of intervals formed in this manner (in repeated samples) will contain the true mean

+8.2 Confidence Interval Estimation for the Mean (σ Unknown)

If the population standard deviation σ is unknown, we can substitute the sample standard deviation, S

This introduces extra uncertainty, since S is variable from sample to sample

So we use the Student t distribution instead of the normal distribution

- the t value depends on degrees of freedom denoted by sample size minus 1; i.e. ($df = n - 1$)
- df : number of observations that are free to vary after sample mean has been calculated

+The Concept of Degrees of Freedom

The Degrees of Freedom are the number of observations that are free to vary after a sample mean has been calculated

Example:

The mean of 3 numbers is calculated as 8. If $X_1 = 7$ and $X_2 = 8$, then X_3 must be 9. In other words, X_3 is not free to vary.

For this example, the first two values can be any numbers, but the third is not free to vary for a given Mean

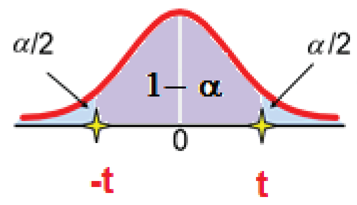
Therefore, if n = 3, the degrees of freedom will be n – 1 or 2

$$\bar{X} = \frac{X_1 + X_2 + X_3}{3} = 8$$

+8.2 Confidence Interval Estimation for the Mean (σ Unknown)

Confidence Interval Estimate (σ Unknown)

$$\bar{X} \pm \underbrace{\boxed{t_{n-1}\frac{S}{\sqrt{n}}}}_{\text{Margin of Error/Sampling Error}}$$



(where t is the critical value of the t distribution with n -1 degrees of freedom and an area of α/2 in each tail)

+Student’s *t* Distribution

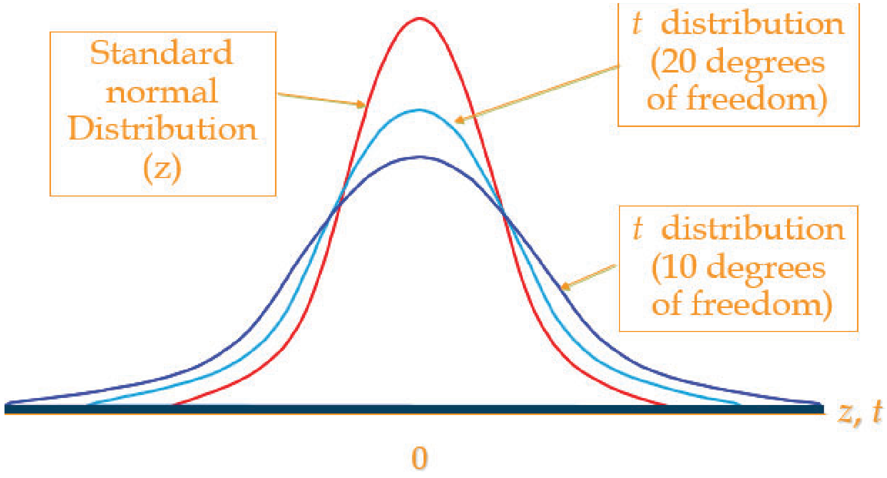
At the beginning of the twentieth century a statistician for Guinness Breweries in Ireland, William S. Gosset, wanted to make inferences about the mean when σ was unknown

Because Guinness employees were not permitted to publish research work under their own names, Gosset adopted the pseudonym 'Student'. The distribution that he developed is known as Student's t distribution

- If the random variable X is normally distributed, then the following statistic has a t distribution with n - 1 degrees of freedom:

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

+Properties of the *t* Di



The t distribution appears very similar to the standardised normal distribution - both distributions are bell shaped

However, the t distribution has more area in the tails and less in the centre than the standardised normal distribution

The degrees of freedom (n – 1) are directly related to the sample size n

As the sample size increases, S becomes a better estimate of σ and the t distribution gradually approaches the standardised normal distribution until the two are virtually identical

+Properties of the *t* Distribution (cont)

With a sample size of about 120 or more, there is little difference between the *t* values and Z values.

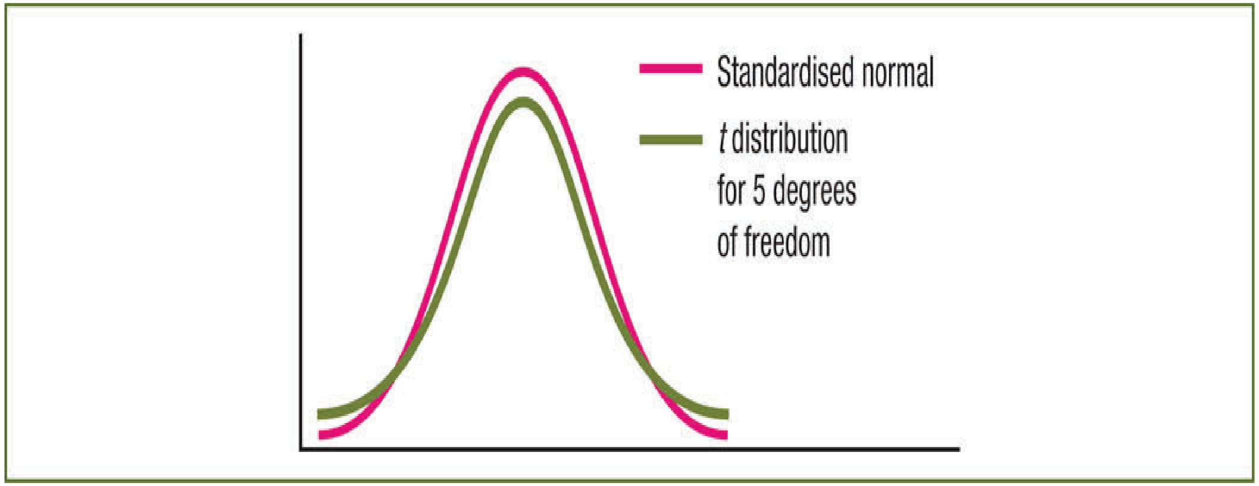


Figure 8.4
Standardised normal distribution and *t* distribution for 5 degrees of freedom

TABLE E.3
Critical Values of *t*
(*Continued*)

Degrees of Freedom	Upper-Tail Areas					
	0.25	0.10	0.05	0.025	0.01	0.005
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259
110	0.6767	1.2893	1.6588	1.9818	2.3607	2.6213
120	0.6765	1.2886	1.6577	1.9799	2.3578	2.6174
∞	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758

+*t* Distribution

Figure 8.5
t distribution with 99 degrees of freedom

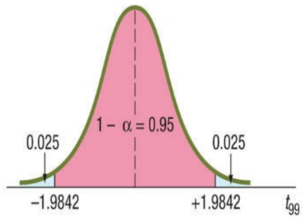


TABLE E.3
Critical Values of *t*

For a particular number of degrees of freedom, entry represents the critical value of *t* corresponding to a specified upper-tail area (α).

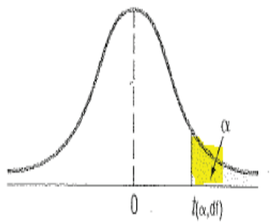


TABLE E.3
Critical Values of *t*
(*Continued*)

Degrees of Freedom	Upper-Tail Areas 0.25	0.10	0.05	0.025	0.01	0.005
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259
110	0.6767	1.2893	1.6588	1.9818	2.3607	2.6213
120	0.6765	1.2886	1.6577	1.9799	2.3578	2.6174
∞	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758

Table 8.1
Determining the critical value from the *t* table for an area of 0.025 in each tail with 99 degrees of freedom (extracted from Table E.3 in Appendix E of this book)

Degrees of freedom	Upper-tail areas .25	.10	.05	.025	.01	.005
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
96	0.6771	1.2904	1.6609	1.9850	2.3658	2.6280
97	0.6770	1.2903	1.6607	1.9847	2.3654	2.6275
98	0.6770	1.2902	1.6606	1.9845	2.3650	2.6269
99	0.6770	1.2902	1.6604	1.9842	2.3646	2.6264
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259

+Confidence Interval (σ Unknown)

Example:

A random sample of 25 people have a mean age of 50 and sample standard deviation of 8

Calculate a 95% confidence interval for μ

- Degrees of Freedom are 25-1 = 24
- t = 2.0639 (from Table E3 - Critical values of t)

$$\overline{X} \pm t_{n-1} \frac{S}{\sqrt{n}}$$

$$50 \pm 2.0639 \frac{8}{\sqrt{25}}$$

$$50 \pm 3.30224$$

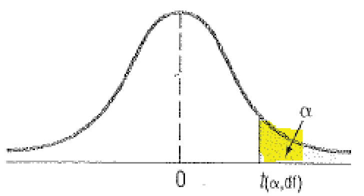
$$46.69776 \leq \mu \leq 53.30224$$

Note: Table E3 is Upper Tail only so need to use α/2

TABLE E.3

Critical Values of t

For a particular number of degrees of freedom, entry represents the critical value of t corresponding to a specified upper-tail area (α).



+MS Excel® Template to Calculate Mean Confidence Interval Estimate

Figure 8.6
Microsoft Excel 2013 worksheet to calculate a confidence interval estimate for the mean sales invoice amount for Dianella Landscaping Supplies

	A	B	
1	**Estimate for the mean sales invoice amount**		
2			
3	**Data**		
4	**Sample standard deviation**	**52.62**	
5	**Sample mean**	**230.27**	
6	**Sample size**	**100**	
7	**Confidence level**	**95%**	
8			
9	Intermediate calculations		
10	Standard error of the mean	5.262	**=B4/SQRT(B6)**
11	Degrees of freedom	99	**=B6 – 1**
12	*t* value	1.984217	**=T.INV.2T(1 – B7,B11)**
13	Interval half width	10.44095	**=B12 * B10**
14			
15	**Confidence interval**		
16	**Interval lower limit**	**219.8291**	**=B5 – B13**
17	**Interval upper limit**	**240.7109**	**=B5 + B13**

+8.3 Confidence Interval Estimation for the Proportion (Categorical Data)

Confidence interval estimate

The Upper and lower confidence limits for the population proportion (π) are calculated with the formula

$$p \pm Z\sqrt{\frac{p(1-p)}{n}}$$

Where: Z is the standard normal value for the level of confidence desired

p is the sample proportion

n is the sample size

$$\text{Point Estimate} \pm \underbrace{\text{(Critical Value)*(Standard Error)}}_{\text{Margin of Error/Sampling Error}}$$

Assumptions: $np \geq 5$ and $n(1-p) \geq 5$

+8.3 Confidence Interval Estimation for the Proportion (cont)

Example:

$np \geq 5$ and $n(1-p) \geq 5$

100*25 =25 and 100*(1-0.25) =75

A random sample of 100 people shows that 25 are left-handed p =25/100=0.25=25%

Calculate a 95% confidence interval for the true proportion of left-handers

Answer:

We are 95% certain that the true proportion of all people who are left handed lies somewhere between 16.51% and 33.49%

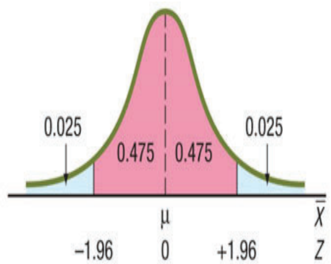
$$p \pm Z\sqrt{\frac{p(1-p)}{n}}$$

$$0.25 \pm 1.96\sqrt{\frac{0.25(1-0.25)}{100}}$$

$$0.25 \pm 1.96\sqrt{\frac{0.25(1-0.25)}{100}}$$

$$0.25 \pm 0.0849$$

$$0.1651 \leq \pi \leq 0.3349$$



+MS Excel® Template to Calculate Proportion Confidence Interval Est.

Figure 8.10
Microsoft Excel 2013 worksheet to form a confidence interval estimate for the proportion of sales invoices that contain errors

	A	B	
1	**Proportion of in-error sales invoices**		
2			
3	**Data**		
4	**Sample size**	**100**	
5	**Number of success**	**10**	
6	**Confidence level**	**95%**	
7			
8	Intermediate calculations		
9	Sample proportion	0.1	**=B5/B4**
10	*Z* value	1.96	**=NORM.S.INV((1 + B6)/2)**
11	Standard error of the proportion	0.03	**=SQRT(B9 * (1 – B9)/B4)**
12	Interval half width	0.0588	**=(B10 * B11)**
13			
14	**Confidence interval**		
15	**Interval lower limit**	**0.0412**	**=B9 – B12**
16	**Interval upper limit**	**0.1588**	**=B9 + B12**

+Determining Sample Size for the Mean

$$\text{Point Estimate} \pm \underbrace{(\text{Critical Value})*(\text{Standard Error})}_{\text{Margin of Error/Sampling Error}}$$

In order to determine the required sample size for the mean, we need to know:

- the desired level of confidence (1 - α), which determines the critical Z value
- an acceptable sampling error, e
- the standard deviation, σ (can be estimated with past data or pilot sample)

We rework the Sampling Error (the Margin of Error) of the Confidence Interval Estimation for the Mean to resolve n

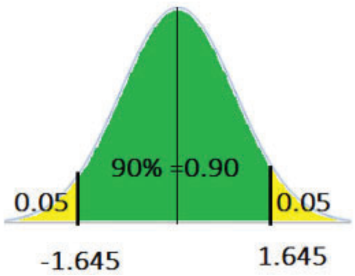
$\boxed{\bar{X} \pm Z \frac{\sigma}{\sqrt{n}}}$ is denoted $\boxed{e = Z \frac{\sigma}{\sqrt{n}}}$ then isolate $\boxed{n = \frac{Z^2 \sigma^2}{e^2}}$

+Determining Sample Size for the Mean (cont)

Example:

What sample size is needed to estimate the mean within ± 5 with 90% confidence? If σ = 45

$$n = \frac{Z^2\,\sigma^2}{e^2} = \frac{1.645^2 * 45^2}{5^2} = 219.19$$



Therefore, the required sample size is **n = 220** (always round up)

+MS Excel® Template to Calculate Sample Size for a Mean

Figure 8.11
Microsoft Excel 2013 worksheet for determining sample size for estimating the mean sales invoice amount for Dianella Landscaping Supplies Pty Ltd

	A	B	
1	**For the mean sales invoice amount**		
2			
3	**Data**		
4	**Population standard deviation**	**25**	
5	**Sampling error**	**5**	
6	**Confidence level**	**95%**	
7			
8	Intermediate calculations		
9	*Z* value	1.9600	**=NORM.S.INV((1 + B6)/2)**
10	Calculated sample size	96.0365	**=((B9 * B4)/B5)^2**
11			
12	**Result**		
13	**Sample size needed**	**97**	**=ROUNDUP(B10,0)**

$$n = \frac{Z^2 \sigma^2}{e^2}$$

Determining Sample Size for the Proportion

Point Estimate ± $\underbrace{\text{(Critical Value)*(Standard Error)}}_{\text{Margin of Error/Sampling Error}}$

In order to determine the required sample size for the Proportion, we need to know:

- the desired level of confidence (1 - α), which determines the critical Z value
- an acceptable margin/sampling error, e
- true proportion of 'successes', π (can be estimated with past data or pilot sample, p, **or** conservatively use π = 0.5)

We rework the Sampling Error (the Margin of Error) of the Confidence Interval Estimation for the Proportion to resolve n

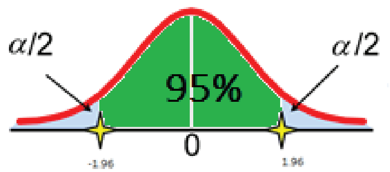
$\pi \pm Z\sqrt{\frac{\pi(1-\pi)}{n}}$ is denoted $e = Z\sqrt{\frac{\pi(1-\pi)}{n}}$ then isolate $n = \frac{Z^2\,\pi(1-\pi)}{e^2}$

+Determining Sample Size for the Proportion (cont)

Example:

What sample size is needed to estimate the true proportion of defective parts in a large population to within ±3%, with 95% confidence? (Assume a pilot sample yielded 12% defects. i.e p = 0.12)

$$n = \frac{Z^2 \pi(1-\pi)}{e^2} = \frac{1.96^2 * 0.12(1-0.12)}{0.03^2} = 450.74$$



Therefore, the required sample size is **n = 451** (always round up)

+

Example: Newspaper election polls

An election poll is to be undertaken to estimate with 95% confidence the proportion of voter's who will vote for the labour party to within 3% accuracy.

Use π= 0.5. (No better estimate of π is available)

$$\boxed{n = \frac{Z^2 \pi (1-\pi)}{e^2}} \qquad n = \frac{1.96^2 * 0.5^2 * 0.5^2}{0.03^2} = 1067.1$$

ALWAYS ROUND UP!!

Thus, **n = 1068** is the minimum sample size needed for our specifications.

+MS Excel® Template to Calculate Sample Size for a Proportion

Figure 8.12
Microsoft Excel 2013 worksheet for determining sample size for estimating the proportion of sales invoices with errors for Dianella Landscaping Supplies Pty Ltd

	A	B	
1	**For the proportion of in-error sales invoices**		
2			
3	**Data**		
4	**Estimate of true proportion**	**0.15**	
5	**Sampling error**	**0.07**	
6	**Confidence level**	**95%**	
7			
8	Intermediate calculations		
9	*Z* value	1.9600	**=NORM.S.INV((1 + B6)/2)**
10	Calculated sample size	99.9563	**=(B9^2 * B4 * (1 – B4))/B5^2**
11			
12	**Result**		
13	**Sample size needed**	**100**	**=ROUNDUP(B10,0)**

+8.6 More on Confidence Interval Estimation and Ethical Issues

- A confidence interval estimate (reflecting sampling error) should always be included when reporting a point estimate
- The level of confidence should always be reported
- The sample size should be disclosed
- An interpretation of the confidence interval estimate should also be provided