Lecture 5
Confidence Intervals

Confidence interval for μ, the population mean

○ □known

○ □unknown (using the t-distribution)

Confidence interval for π, the population proportion

In the Last Lecture

We looked at the distribution of sample means and sample proportions in repeated sampling. We found that:

- sample means follow a normal distribution when the population is normally distributed
- provided the sample size, n, is 'large enough', sample means are from an approximately normal distribution, with $mean = \mu \; and \; standard\ error = {}^{\sigma}/_{\sqrt{n}}$
- the standard deviation of sample means is called the $standard\ error\ of\ \bar{y}\ or\ se_{\bar{y}}\ or\ \sigma_{\bar{y}}$
- provided n□$and\ n(1$ – □are **both** at least 5, sample proportions follow an approximately normal distribution with $mean = \pi \; and \; standard\ error = \sqrt{\frac{\pi(1-\pi)}{n}}$

Sample Statistics $\xrightarrow{\text{estimate}}$ Population Parameters

Mean	$\bar{y}$	→	μ
Median	$\tilde{y}$	→	$\tilde{\mu}$
Std. dev	s	→	σ
Std. dev	s^2	→	σ^2
se (mean)			${}^{\sigma}/_{\sqrt{n}}$
Proportion	p	→	π
se (p)			$\sqrt{\frac{\pi(1-\pi)}{n}}$

Standardising: z-scores

Individuals scores (y)

$$z = \frac{y-\mu}{\sigma}$$

Sample means $(\bar{y})$

$$z = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$$

Sample proportions (p)

$$z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Review Quiz 1

The sign on an elevator in Sydney says that the maximum weight it should carry is 2500kg. 30 males get into the lift. Assuming the average weight of males in Australia is 85kg with a standard deviation of 10kg, find the probability that the lift will be overloaded.

Solution to Review Quiz 1

Review Quiz 2

In a large class of statistics students, the lecturer asks each student to toss two coins **30** times and calculate the proportion of times his or her coins turned up two heads. The students then report their results and the lecturer provides a histogram of these results:

a. What shape would you expect this histogram to be?

b. Where do you expect the histogram to be centred?

c. How much variability would you expect among these proportions (ie. what is the standard deviation)?

d. Suppose each student had tossed the coins 10 times? Would you still be able to answer the questions above? Explain why or why not.

✓ Solution to Review Quiz 2

Review Quiz 3

Let's look again at the IQ data. IQ scores are normally distributed, with mean μ =100 and standard deviation σ = 15.

Consider taking many samples of size 25 from this population.

a. What is the shape of the distribution of the sample means?

b. What is the mean of this distribution?

c. What is the standard error of this distribution?

d. What proportion of sample means lies within 1 se of μ?

That is, between $100 \pm 1\times\frac{15}{\sqrt{25}}$ ie. between 97 and 103?

e. What proportion of sample means lies within 2 se of μ?

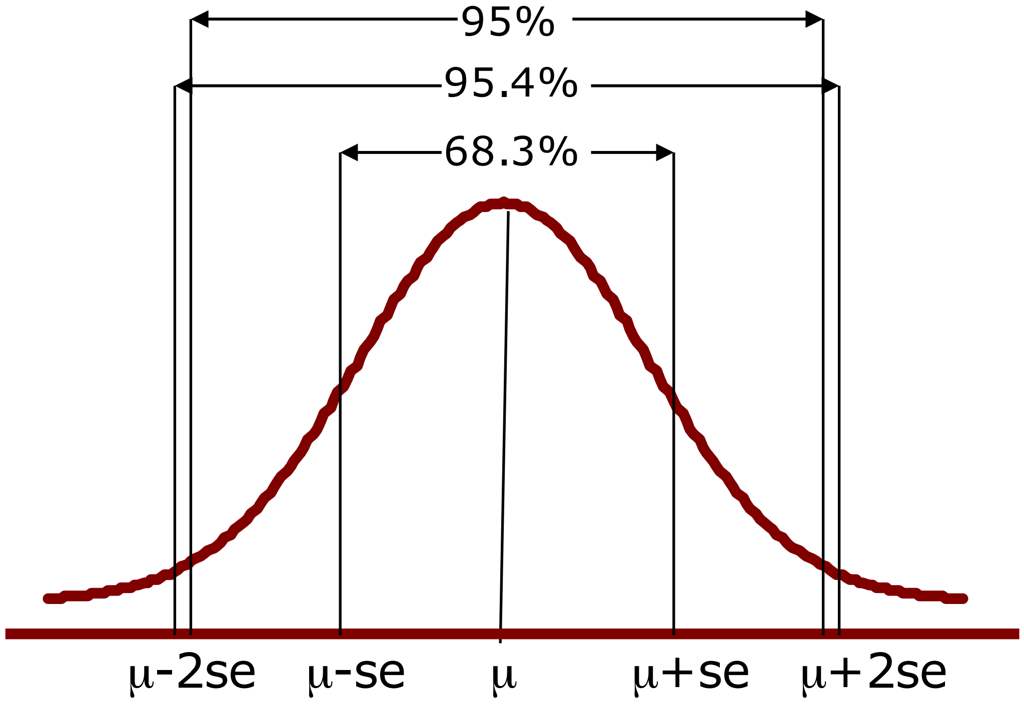
That is, between $100 \pm 2\times\frac{15}{\sqrt{25}}$ ie. between 94 and 106?

f. What proportion of sample means lies within 1.96 se of μ?

That is, between $100 \pm 1.96\times\frac{15}{\sqrt{25}}$ ie. 94.12 to 105.88?

Solution to Review Quiz 3

Where were the Sample Means?



68.3% of the sample means lie between μ–se and μ+se

95.4% of the sample means lie between μ–2se and μ+2se

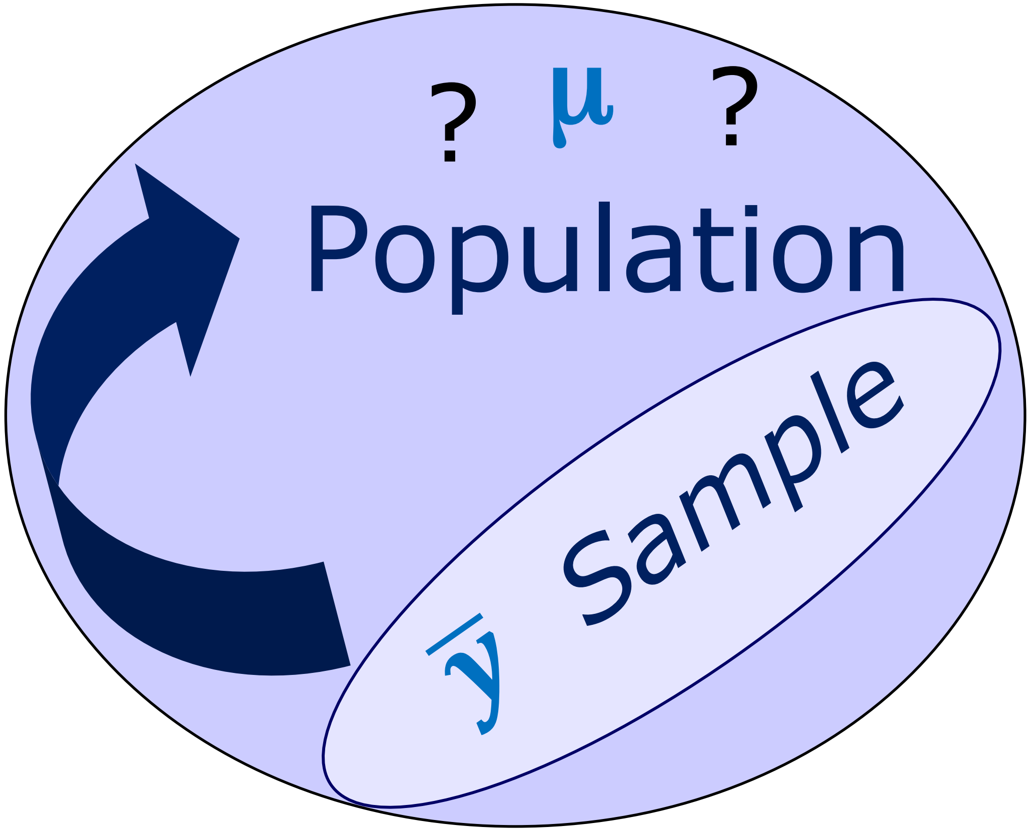
95% of the sample means lie between μ–1.96se and μ+1.96se

Confidence interval for a population mean μ: $\sigma\ known$

Answering Research Questions

Research Questions

We use **sample statistics** to estimate **Population parameters**



Estimating μ: a Point Estimate

- If we don’t know the population mean μ, we can take a sample from the target population, and **estimate** μ from the sample mean $\bar{y}$.

- $\bar{y}$ is called a ***point estimate*** of μ, that is, a single value estimate.

- How ‘good’ is this estimate? - this depends on the standard error.

Estimating μ: an Interval Estimate

- Often it is more informative to give an **interval estimate** for μ, rather than a point estimate.

- An interval estimate gives us a **range of believable values** for μ:

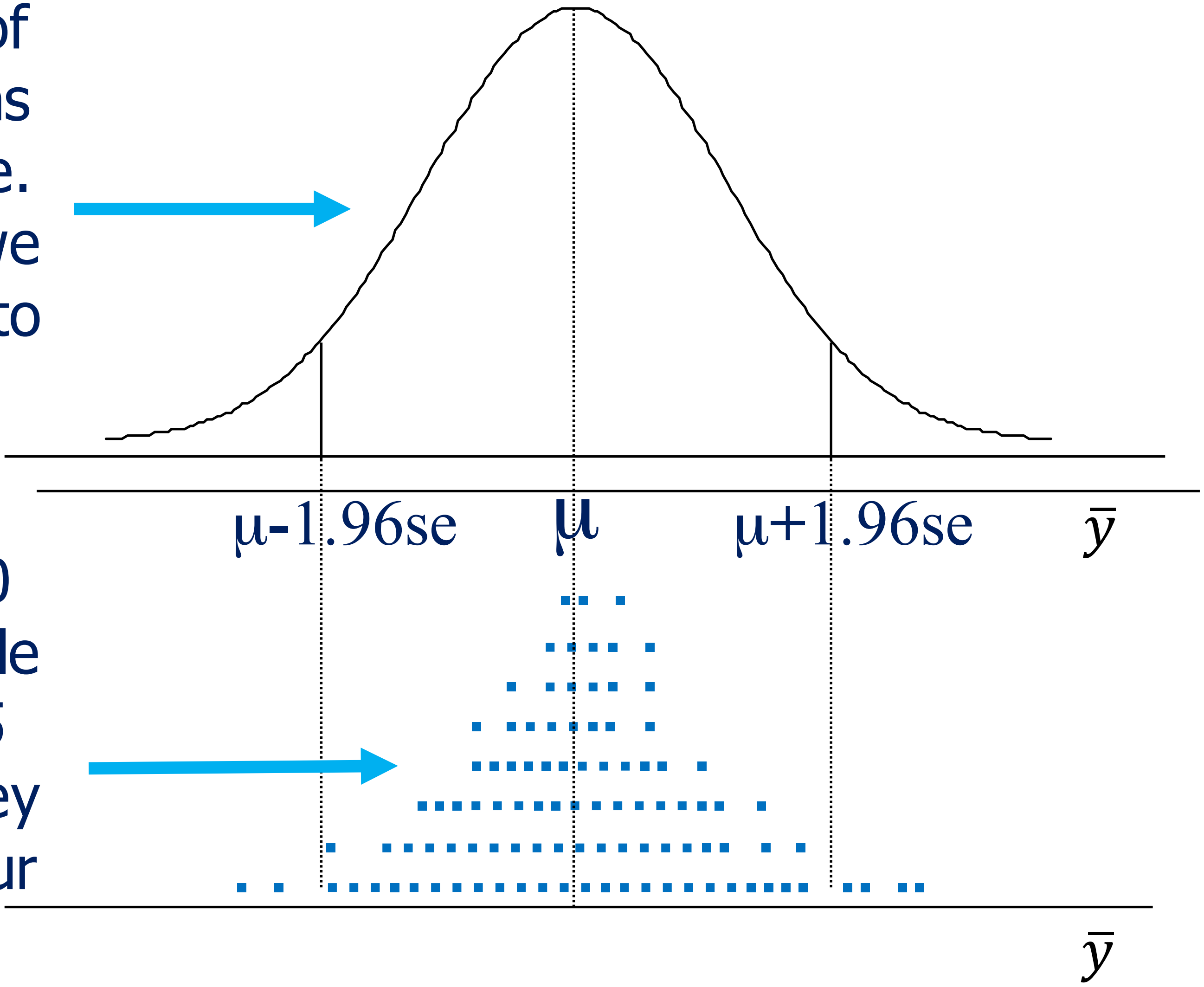
- We call the interval estimate/the range of believable values a **confidence interval for μ.**

- **How** do we find this interval, **why** do we have 'confidence' in it and **how much** confidence do we have in it?

100 Samples: Each of Size n = 25

Distribution of sample means (theoretical ie. this is what we expect them to look like)

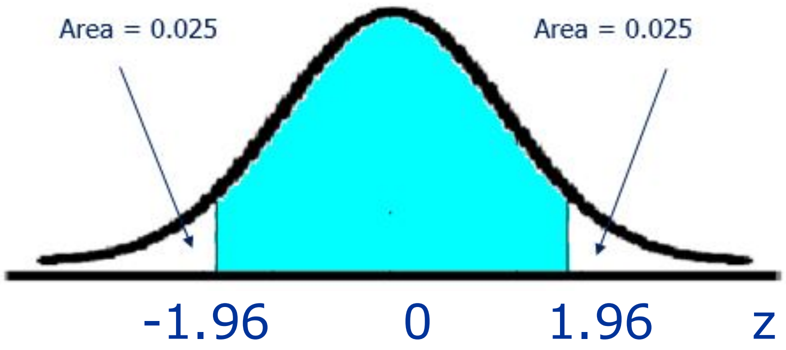
Dot plot of 100 simulated sample means, n = 25 (this is what they looked like in our simulation)



Generalising this Result

- 94 of the 100 simulated sample means lie within 1.96 standard errors of μ.

- Note that 1.96 is the z-score that cuts off an area of 0.025 in *each* tail of the standard normal distribution.



- Approximately 95% of sample means *should* fall within 1.96 standard errors of μ, according to our knowledge of the distribution of sample means.

95% Confidence Interval for μ

- When we take a sample from some population and we know that the sample mean is from a normal distribution, we know the probability that $\bar{y}$ is within 1.96 standard errors of μ is 0.95 or 95%.

- So, if we perform any study an infinite number of times, 95% of the intervals will contain the true population mean μ.

- That is, with 95% confidence, we can say that the interval: $\left(\bar{y} - 1.96 \times {}^{\sigma}/_{\sqrt{n}}, \bar{y} + 1.96 \times {}^{\sigma}/_{\sqrt{n}}\right)$ contains μ.

- This is a **95% confidence interval (CI) for** μ.

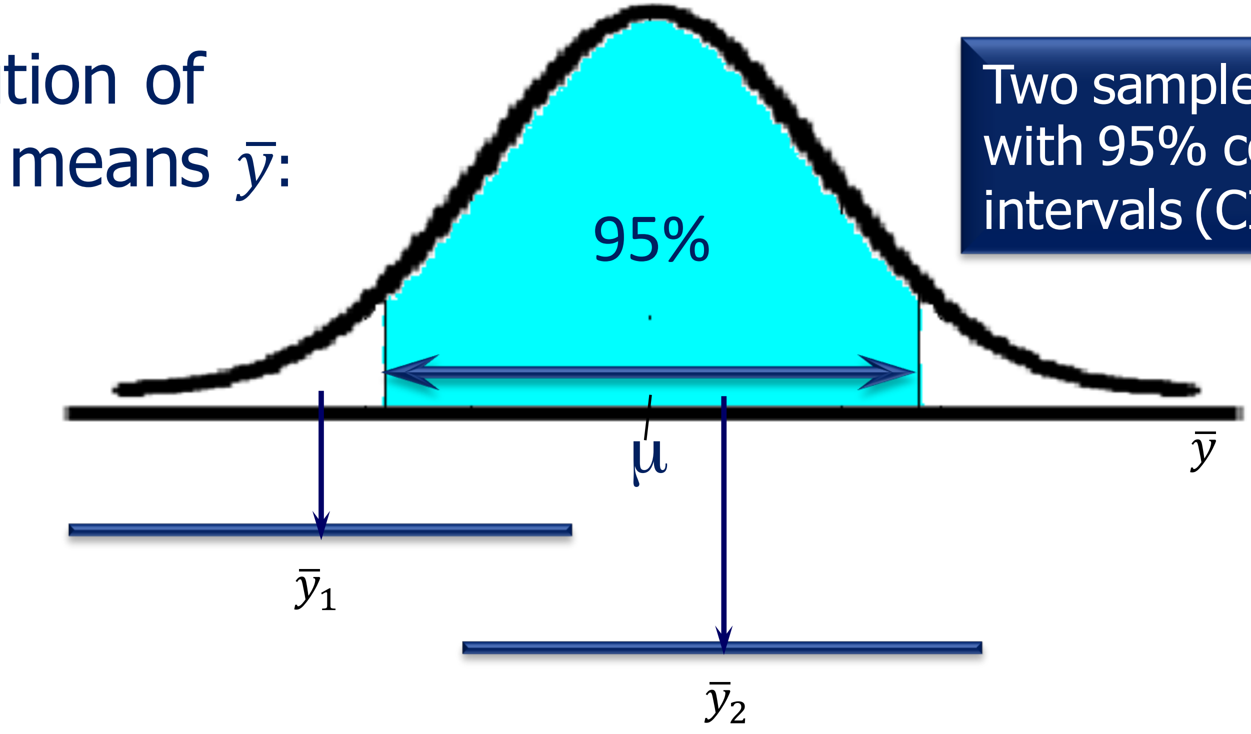
Confidence Intervals and Sampling Error

- In calculating the confidence interval, we use the fact that, if we take many random samples from a population where sample means follow a normal distribution, **in the long run 95% of the sample means will lie within 1.96 standard errors of μ.**

- Similarly, if we take many random samples of the same size and calculate a 95% confidence interval from each sample mean**, in the long run 95% of all these confidence intervals should contain μ.**

Basis for a Confidence Interval

Distribution of sample means $\bar{y}$:



Two sample means, with 95% confidence intervals (CIs) for μ.

Using the mean of sample 1, the CI **does not** contain μ

Using the mean of sample 2, the CI **does** contain μ

In the long run, 95% of CIs should contain μ.

Example: Country School Students

Let's consider IQ scores again. Suppose our target population is students who attend country schools in NSW. We would like to calculate a 95% confidence interval to estimate the mean IQ score in this population.

In the general population, IQ scores follow a normal distribution with standard deviation, **$\sigma = 15$**. For this problem, let's assume that this also applies to our target population, country school students in NSW. Then sample means will also be normally distributed, regardless of sample size. Say we have taken a random sample of 36 students from country schools in NSW and found that the sample mean **$\bar{y} = 103$**.

95% confidence interval for μ:

$$\left(\bar{y} - 1.96 \times {}^{\sigma}/_{\sqrt{n}}, \bar{y} + 1.96 \times {}^{\sigma}/_{\sqrt{n}}\right)$$

Example continued

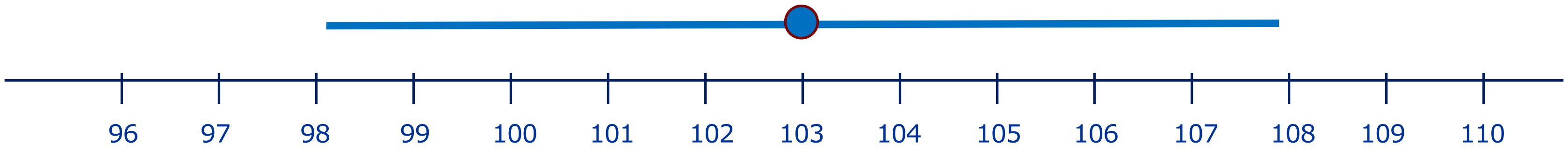
So using this sample with **n = 36 and $\bar{y}$ =103** and assuming that the sample is drawn from a normal distribution (this was given although n is large enough for the CLT to apply here anyway) and also assuming that the variation in IQ scores is the same for country school students as in the general population (ie. σ is known to be 15), we can calculate **the 95% CI for** μ:

$$95\%\ CI\ for\ \mu = \bar{y} \pm 1.96 \times {}^{\sigma}/_{\sqrt{n}}$$

$$= 103 \pm 1.96 \times {}^{15}/_{\sqrt{36}}$$

$$= 103 \pm 1.96 \times 2.5$$

$$= (98.1, 107.9)$$



Interpreting a Confidence Interval

- We *interpret* the confidence interval in terms of the mean of the **target population**, μ.
- The interpretation of a 95% CI for μ is that we have **95% confidence that the *population* mean μ is included in the interval**.
- Therefore, for the country school students example, **we are 95% confident that the *true* mean IQ of <u>all</u> students attending country schools in NSW lies between 98.1 and 107.9**.
- NOTE: We could be 'unlucky': our sample could be unusual, with its interval not containing μ. Then the true mean is *not* within the range of the 95% CI. There is a 5% chance of this happening.

Quiz 4

Estimate the average weekly cost of renting a two-bedroom apartment in Newtown:

The following weekly rentals ($) were recorded from 20 randomly selected 2 bedroom apartments advertised to rent on *domain.com.au* in 2013. You may assume the standard deviation for all weekly rental prices in Newtown is $85 as claimed by a local realtor.

Weekly Rental Price - Newtown									
650	500	600	680	560	450	730	650	750	610
585	580	540	530	500	480	585	650	500	650

Calculate a 95% confidence interval to estimate the mean weekly cost of renting a two-bedroom apartment in Newtown in 2013.

Draw and interpret the interval.

Solution to Quiz 4

Checking Feasibility

For these confidence intervals to give valid estimates, we should ensure that the following conditions apply:

- **The measurements recorded MUST be *independent* of each other.**
  - In the country students data we need to make sure, for example, that not all the children are selected from one area or that a number of students are not from the same family, etc.
- **The sample mean is from a normal distribution.**
  - In the case of the country school students it was given that the IQ scores were normally distributed.
  - In the rental apartments example we checked for normality using a histogram.

Confidence interval for a population mean μ: σ *unknown*

○ an introduction to the t-distribution

What about a Confidence Interval for μ when σ is Unknown?

- σ is the standard deviation of the population. We use σ to calculate the standard error for sample means: ${}^{\sigma}/_{\sqrt{n}}$

- If the population standard deviation, σ, is unknown, we can use the **sample standard deviation**, s, as an estimate of σ.

- So when we calculate a confidence interval for the population mean, μ, when σ is unknown, we use the estimated standard error of sample means,

$$est.se_{\bar{y}} = {}^{s}/_{\sqrt{n}}$$

The Students' *t*-distribution

- **There is a catch:**
- Because we're only estimating σ and because bigger samples give better estimates of population parameters, the standard normal distribution isn't the best model for sample means.
- When σ is unknown, we use a distribution which takes sample size into account.
  - There is a distribution which does just that. It is the **student's t-distribution**.

The larger the sample the better our estimate s will be of σ.
→ If we don't know σ , we should take the sample size into account

Confidence Interval for μ when σ Unknown

To calculate a 95% confidence interval for μ when σ is not known, we replace the critical z-value of 1.96 with a critical t-value:

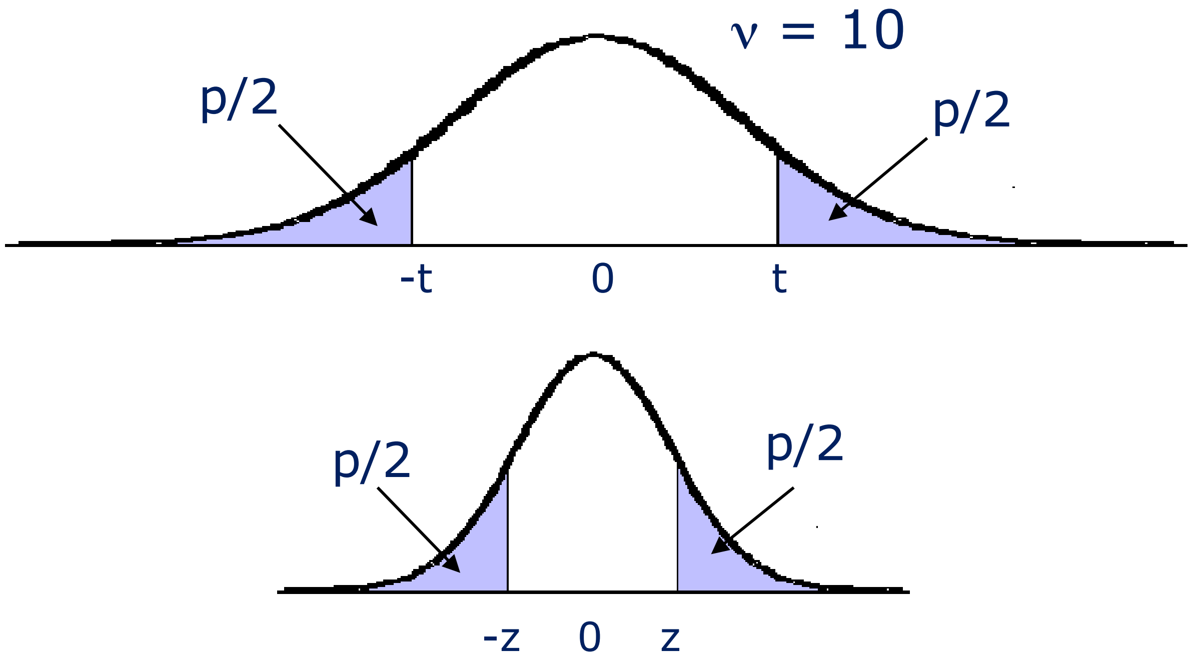
$$\left(\overline{y} - t_{crit} \times {}^{s}/_{\sqrt{n}},\ \overline{y} + t_{crit} \times {}^{s}/_{\sqrt{n}}\right)$$

This **t-value** also cuts off an area of 0.025 in each tail of its distribution but, unlike the z-value, **it is different for different sample sizes**.

Important note: We may only use the t-distribution if we are reasonably certain that the **variable of interest is normally distributed**. This can be checked by plotting the data.

This is particularly true of small samples. Note, however, that **if our sample size is large then the t-distribution is fairly robust against departures from normality**.

A t-distribution vs the Normal (z) Distribution

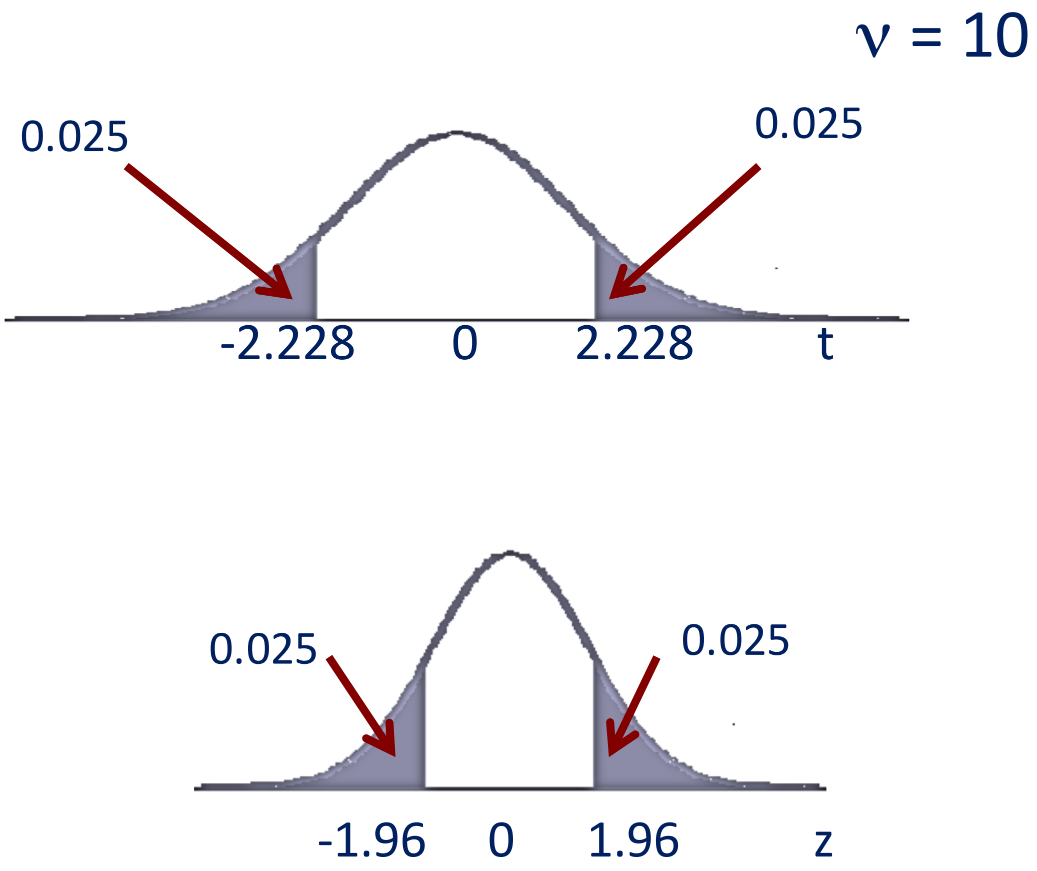


Heavier tailed than the standard normal distribution.

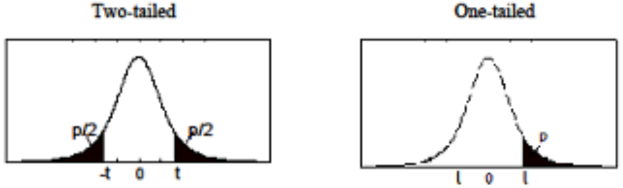
Has an extra parameter, ν, (nu) the degrees of freedom (df) which depends on the sample size. (ν = n – 1)

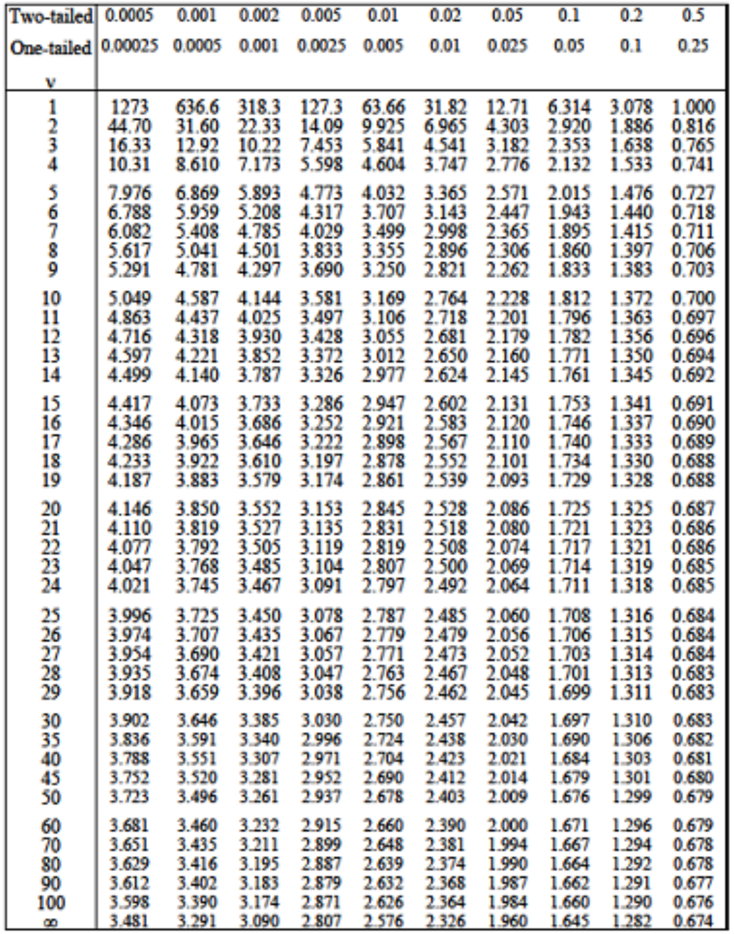
As ν → ∞ the t-distribution approaches the normal distribution.

Critical t-value vs Critical z-value for a 95% Confidence Interval



t-table



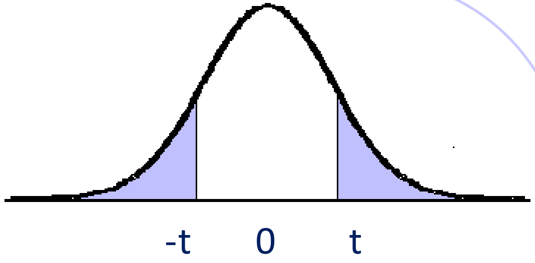


Two-tailed	0.0005	0.001	0.002	0.005	0.01	0.02	0.05	0.1	0.2	0.5
One-tailed	0.00025	0.0005	0.001	0.0025	0.005	0.01	0.025	0.05	0.1	0.25
v										
1	1273	636.6	318.3	127.3	63.66	31.82	12.71	6.314	3.078	1.000
2	44.70	31.60	22.33	14.09	9.925	6.965	4.303	2.920	1.886	0.816
3	16.33	12.92	10.22	7.453	5.841	4.541	3.182	2.353	1.638	0.765
4	10.31	8.610	7.173	5.598	4.604	3.747	2.776	2.132	1.533	0.741
5	7.976	6.869	5.893	4.773	4.032	3.365	2.571	2.015	1.476	0.727
6	6.788	5.959	5.208	4.317	3.707	3.143	2.447	1.943	1.440	0.718
7	6.082	5.408	4.785	4.029	3.499	2.998	2.365	1.895	1.415	0.711
8	5.617	5.041	4.501	3.833	3.355	2.896	2.306	1.860	1.397	0.706
9	5.291	4.781	4.297	3.690	3.250	2.821	2.262	1.833	1.383	0.703
10	5.049	4.587	4.144	3.581	3.169	2.764	2.228	1.812	1.372	0.700
11	4.863	4.437	4.025	3.497	3.106	2.718	2.201	1.796	1.363	0.697
12	4.716	4.318	3.930	3.428	3.055	2.681	2.179	1.782	1.356	0.696
13	4.597	4.221	3.852	3.372	3.012	2.650	2.160	1.771	1.350	0.694
14	4.499	4.140	3.787	3.326	2.977	2.624	2.145	1.761	1.345	0.692
15	4.417	4.073	3.733	3.286	2.947	2.602	2.131	1.753	1.341	0.691
16	4.346	4.015	3.686	3.252	2.921	2.583	2.120	1.746	1.337	0.690
17	4.286	3.965	3.646	3.222	2.898	2.567	2.110	1.740	1.333	0.689
18	4.233	3.922	3.610	3.197	2.878	2.552	2.101	1.734	1.330	0.688
19	4.187	3.883	3.579	3.174	2.861	2.539	2.093	1.729	1.328	0.688
20	4.146	3.850	3.552	3.153	2.845	2.528	2.086	1.725	1.325	0.687
21	4.110	3.819	3.527	3.135	2.831	2.518	2.080	1.721	1.323	0.686
22	4.077	3.792	3.505	3.119	2.819	2.508	2.074	1.717	1.321	0.686
23	4.047	3.768	3.485	3.104	2.807	2.500	2.069	1.714	1.319	0.685
24	4.021	3.745	3.467	3.091	2.797	2.492	2.064	1.711	1.318	0.685
25	3.996	3.725	3.450	3.078	2.787	2.485	2.060	1.708	1.316	0.684
26	3.974	3.707	3.435	3.067	2.779	2.479	2.056	1.706	1.315	0.684
27	3.954	3.690	3.421	3.057	2.771	2.473	2.052	1.703	1.314	0.684
28	3.935	3.674	3.408	3.047	2.763	2.467	2.048	1.701	1.313	0.683
29	3.918	3.659	3.396	3.038	2.756	2.462	2.045	1.699	1.311	0.683
30	3.902	3.646	3.385	3.030	2.750	2.457	2.042	1.697	1.310	0.683
35	3.836	3.591	3.340	2.996	2.724	2.438	2.030	1.690	1.306	0.682
40	3.788	3.551	3.307	2.971	2.704	2.423	2.021	1.684	1.303	0.681
45	3.752	3.520	3.281	2.952	2.690	2.412	2.014	1.679	1.301	0.680
50	3.723	3.496	3.261	2.937	2.678	2.403	2.009	1.676	1.299	0.679
60	3.681	3.460	3.232	2.915	2.660	2.390	2.000	1.671	1.296	0.679
70	3.651	3.435	3.211	2.899	2.648	2.381	1.994	1.667	1.294	0.678
80	3.629	3.416	3.195	2.887	2.639	2.374	1.990	1.664	1.292	0.678
90	3.612	3.402	3.183	2.879	2.632	2.368	1.987	1.662	1.291	0.677
100	3.598	3.390	3.174	2.871	2.626	2.364	1.984	1.660	1.290	0.676
∞	3.481	3.291	3.090	2.807	2.576	2.326	1.960	1.645	1.282	0.674

t-Distribution: Values of |t| corresponding to two-tailed and one-tailed p-values for Student's t- distribution

Finding a Critical t-value for a 95% Confidence Interval

p two-tailed:	0.01	0.02	0.05	0.10	..
p one-tailed:	0.005	0.01	0.025	0.05	..
ν					
1	63.66	31.82	12.71	6.314	..
2	9.925	6.965	4.303	2.920	
3	5.841	4.541	3.182	2.353	..
4	4.604	3.747	2.776	2.132	
5	4.032	3.365	2.571	2.015	..
6	3.707	3.143	2.447	1.943	
7	3.499	2.998	2.365	1.895	..
8	3.355	2.896	2.306	1.860	
9	3.250	2.821	2.262	1.833	..
10	3.169	2.764	**2.228**	1.812	
11	3.106	2.718	2.201	1.796	..
.	...	...	...	...	..



Suppose $n = 11$.
Then ν (df) $= n - 1$
$= 10$

For 95% confidence:
use the column with
p two-tailed = 0.05 (ie.
p one-tailed = 0.025)

For df = 10:
use the row with $\nu = 10$

This gives $t_{crit} = 2.228$

Example: Calculating a 95% Confidence Interval, σ Unknown

A STAT170 tutorial exercise involved asking students to memorise a list of words for one minute.

In one class, the sample of 41 students was able to recall 12.3 words, on average, with a standard deviation of 2.4 words.

Suppose we want to calculate a 95% confidence interval for μ, the mean number of words recalled in the target population.

We don't know σ but we can use the sample standard deviation s = 2.4 to estimate σ.

Then we can **estimate** the standard error and use the **t distribution** to obtain the confidence interval. *We don't know whether the number of words recalled follows a normal distribution, but the sample size is large so provided the distribution is not highly skewed the t-distribution should be robust against non-normality.*

Example continued

From the sample, $\bar{y}$ = 12.3 and s = 2.4

so the estimated $se_{\bar{y}}$ = ${}^{s}/_{\sqrt{n}}$ = ${}^{2.4}/_{\sqrt{41}}$ = 0.375

The 95% confidence interval for μ is:

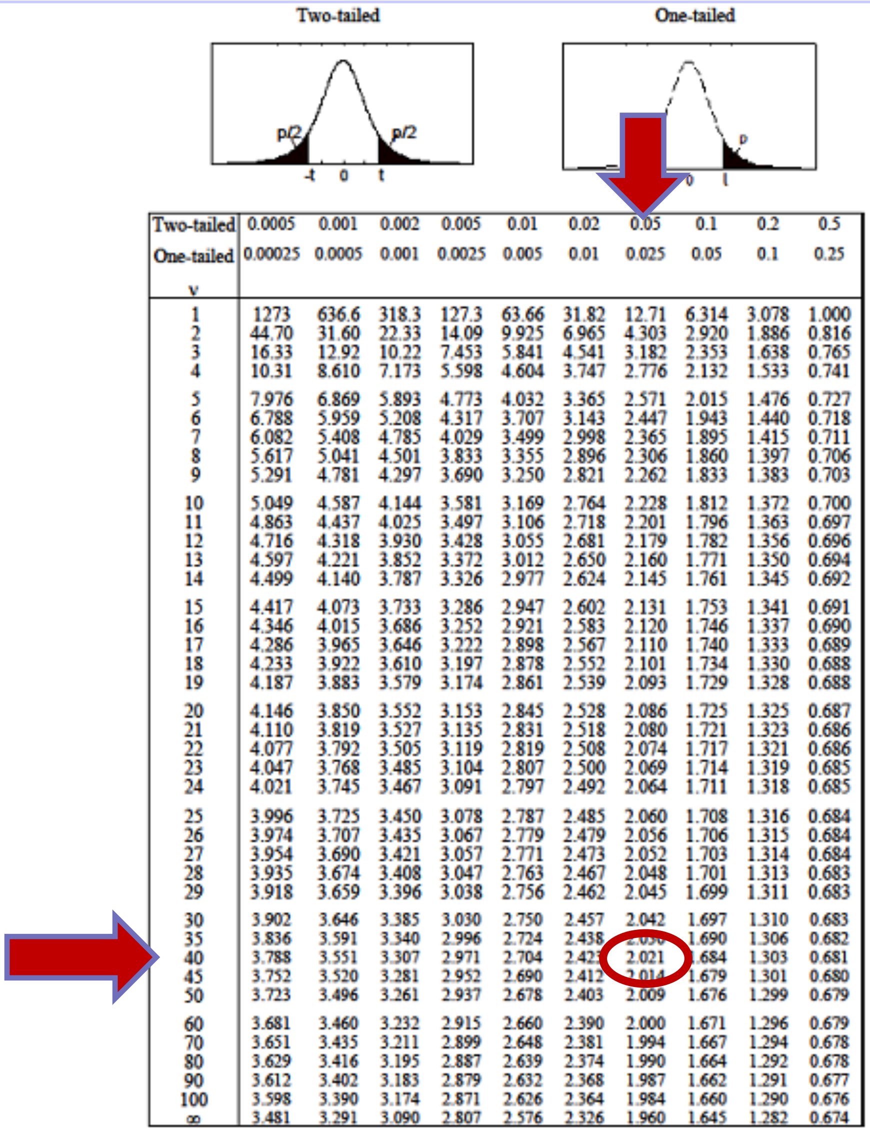
$$95\%\ CI\ for\ \mu = \bar{y} \pm t_{crit} \times {}^{s}/_{\sqrt{n}}$$

$$= 12.3 \pm 2.021 \times {}^{2.4}/_{\sqrt{41}}$$

$$= (11.54, 13.06)$$

We are 95% confident that the average number of words recalled for all STAT170 students (ie. the population mean) was between 11.5 and 13.

Finding a Critical t-value for a Confidence Interval


Two-tailed
p/2
p/2
-t
0
t
One-tailed
p
0
t

Two-tailed	0.0005	0.001	0.002	0.005	0.01	0.02	0.05	0.1	0.2	0.5
One-tailed	0.00025	0.0005	0.001	0.0025	0.005	0.01	0.025	0.05	0.1	0.25
v										
1	1273	636.6	318.3	127.3	63.66	31.82	12.71	6.314	3.078	1.000
2	44.70	31.60	22.33	14.09	9.925	6.965	4.303	2.920	1.886	0.816
3	16.33	12.92	10.22	7.453	5.841	4.541	3.182	2.353	1.638	0.765
4	10.31	8.610	7.173	5.598	4.604	3.747	2.776	2.132	1.533	0.741
5	7.976	6.869	5.893	4.773	4.032	3.365	2.571	2.015	1.476	0.727
6	6.788	5.959	5.208	4.317	3.707	3.143	2.447	1.943	1.440	0.718
7	6.082	5.408	4.785	4.029	3.499	2.998	2.365	1.895	1.415	0.711
8	5.617	5.041	4.501	3.833	3.355	2.896	2.306	1.860	1.397	0.706
9	5.291	4.781	4.297	3.690	3.250	2.821	2.262	1.833	1.383	0.703
10	5.049	4.587	4.144	3.581	3.169	2.764	2.228	1.812	1.372	0.700
11	4.863	4.437	4.025	3.497	3.106	2.718	2.201	1.796	1.363	0.697
12	4.716	4.318	3.930	3.428	3.055	2.681	2.179	1.782	1.356	0.696
13	4.597	4.221	3.852	3.372	3.012	2.650	2.160	1.771	1.350	0.694
14	4.499	4.140	3.787	3.326	2.977	2.624	2.145	1.761	1.345	0.692
15	4.417	4.073	3.733	3.286	2.947	2.602	2.131	1.753	1.341	0.691
16	4.346	4.015	3.686	3.252	2.921	2.583	2.120	1.746	1.337	0.690
17	4.286	3.965	3.646	3.222	2.898	2.567	2.110	1.740	1.333	0.689
18	4.233	3.922	3.610	3.197	2.878	2.552	2.101	1.734	1.330	0.688
19	4.187	3.883	3.579	3.174	2.861	2.539	2.093	1.729	1.328	0.688
20	4.146	3.850	3.552	3.153	2.845	2.528	2.086	1.725	1.325	0.687
21	4.110	3.819	3.527	3.135	2.831	2.518	2.080	1.721	1.323	0.686
22	4.077	3.792	3.505	3.119	2.819	2.508	2.074	1.717	1.321	0.686
23	4.047	3.768	3.485	3.104	2.807	2.500	2.069	1.714	1.319	0.685
24	4.021	3.745	3.467	3.091	2.797	2.492	2.064	1.711	1.318	0.685
25	3.996	3.725	3.450	3.078	2.787	2.485	2.060	1.708	1.316	0.684
26	3.974	3.707	3.435	3.067	2.779	2.479	2.056	1.706	1.315	0.684
27	3.954	3.690	3.421	3.057	2.771	2.473	2.052	1.703	1.314	0.684
28	3.935	3.674	3.408	3.047	2.763	2.467	2.048	1.701	1.313	0.683
29	3.918	3.659	3.396	3.038	2.756	2.462	2.045	1.699	1.311	0.683
30	3.902	3.646	3.385	3.030	2.750	2.457	2.042	1.697	1.310	0.683
35	3.836	3.591	3.340	2.996	2.724	2.438	2.030	1.690	1.306	0.682
40	3.788	3.551	3.307	2.971	2.704	2.423	2.021	1.684	1.303	0.681
45	3.752	3.520	3.281	2.952	2.690	2.412	2.014	1.679	1.301	0.680
50	3.723	3.496	3.261	2.937	2.678	2.403	2.009	1.676	1.299	0.679
60	3.681	3.460	3.232	2.915	2.660	2.390	2.000	1.671	1.296	0.679
70	3.651	3.435	3.211	2.899	2.648	2.381	1.994	1.667	1.294	0.678
80	3.629	3.416	3.195	2.887	2.639	2.374	1.990	1.664	1.292	0.678
90	3.612	3.402	3.183	2.879	2.632	2.368	1.987	1.662	1.291	0.677
100	3.598	3.390	3.174	2.871	2.626	2.364	1.984	1.660	1.290	0.676
∞	3.481	3.291	3.090	2.807	2.576	2.326	1.960	1.645	1.282	0.674

t-Distribution: Values of |t| corresponding to two-tailed and one-tailed p-values for Student's t- distribution

Quiz 6

According to liposuction4you.com, the maximum amount of fat and fluid that can be removed safely during a liposuction procedure is 6 litres. Suppose that the following data represent the amount of fat and fluid removed during 12 randomly selected liposuction procedures:

Fat and fluid removed (litres) during liposuction					
1.84	2.66	2.96	2.42	2.88	2.86
3.66	3.65	2.33	2.66	3.20	2.24

Calculate a 95% confidence interval to estimate the population mean.

Source: Introductory Statistics 7th edition, Mann, *Wiley*

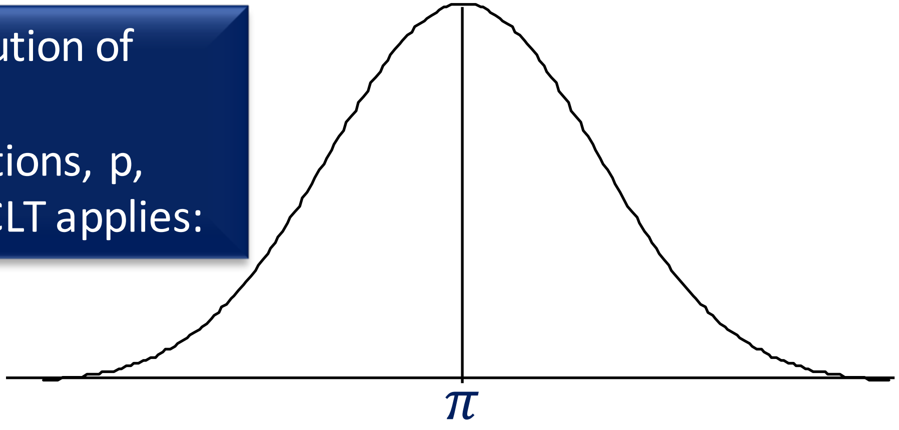
Solution to Quiz 6

Confidence interval for population proportion, π

Estimating a Population Proportion

In the last lecture we saw that a proportion is just a special case of a mean (where the sample consists of observations taking values of 0 or 1 eg. male = 0, female = 1). We saw that in repeated sampling, sample proportions follow an approximately normal distribution provided the Central Limit Theorem applies, that is providing n is large enough.
For the CLT to apply we need **nπ and n(1 – π) both ≥ 5**.
When the CLT applies we are able to calculate a confidence interval for a population proportion.

Distribution of sample proportions, p, when CLT applies:



π represents the population proportion.
The standard error of sample proportions, $se_p = \sqrt{\frac{\pi(1-\pi)}{n}}$

95% Confidence Interval for π

- When we take a sample from some population and find that its sample proportion, *p*, is from a normal distribution (the sample will be large enough for the CLT to apply if both $np \; and \; n(1-p) \geq 5$), we know the probability that *p* is within 1.96 standard errors of π is 0.95 or 95%.
- So, with 95% confidence, we can say that the interval: $\left(p - 1.96\times\sqrt{\frac{p(1-p)}{n}}, p + 1.96\times\sqrt{\frac{p(1-p)}{n}}\right)$ contains π.
- **This is called a 95% confidence interval for π.**

Estimating the Standard Error for Sample Proportions

- The standard error of p is $\sqrt{\frac{\pi(1-\pi)}{n}}$

- **BUT** we used the estimated standard error of p: $\sqrt{\frac{p(1-p)}{n}}$

  to calculate a confidence interval for π

- **WHY?**

- Because we have to **estimate** the true standard error using our sample estimate (p) of the true proportion π. **This means that confidence intervals for π are only approximate.**

Example: What Proportion of Sydney Teenagers Smoke Daily?

In a survey of tobacco use among Sydney teenagers, 995 adolescents were asked whether they:
smoked daily, smoked occasionally, had never smoked or had given up smoking. 216 responded that they smoked daily. We will calculate a 95% confidence interval to estimate the proportion of all Sydney teenagers who smoke daily.

Source: Australian Business Statistics, Selvanathan et al, *Thomson*

Example: What Proportion of Sydney Teenagers Smoke Daily?

Firstly, we have to make sure the sample is large enough for the CLT to apply (or our interval will not be valid):

n = 995 and p = 216/995: np = 216 and n(1–p) = 779

Since both are ≥ 5 the CLT for a proportion applies and we can calculate a 95% confidence interval for π:

$$p \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

$$= 0.2171 \pm 1.96 \times \sqrt{\frac{0.2171(1-0.2171)}{995}}$$

$$= (0.191, 0.243)$$

We can be 95% confident that the true proportion of all Sydney teenagers who smoke daily is between 0.19 and 0.24.

Quiz 7

Is global warming a major issue for Australians?

A researcher surveys 1255 randomly selected Australians and asks them this question. 714 replied that they believed global warming is a major issue. Construct a 95% confidence interval to estimate the proportion of all Australians who feel that global warming is a major issue.

Solution to Quiz 7

3 Confidence Intervals

CI for μ (σ known): $\bar{y} \pm z_{crit} \times {}^{\sigma}/_{\sqrt{n}}$

CI for μ (σ unknown): $\bar{y} \pm t_{crit} \times {}^{s}/_{\sqrt{n}}$

CI for □ $p \pm z_{crit} \times \sqrt{\frac{p(1-p)}{n}}$

There is a common pattern:

CI for $\boxed{\text{population parameter}} = \boxed{\text{sample statistic}} \pm \boxed{\text{critical value}} \times \boxed{\text{standard error/ estimated standard error}}$

For a 95% confidence interval, the critical value we use is:

$z_{crit} = 1.96$

t_{crit} depends on the sample size

Sample Statistics $\xrightarrow{\text{estimate}}$ Population Parameters

Mean	$\bar{y}$	→	μ
Median	$\tilde{y}$	→	$\tilde{\mu}$
Std. dev	s	→	σ
Std. dev	s^2	→	σ^2
se (mean)	${}^{s}/_{\sqrt{n}}$	→	${}^{\sigma}/_{\sqrt{n}}$
Proportion	p	→	π
se (p)	$\sqrt{\frac{p(1-p)}{n}}$	→	$\sqrt{\frac{\pi(1-\pi)}{n}}$

Homework Questions

Homework Question 1

The world's smallest mammal is the bumblebee bat. It is roughly the size of a large bumblebee. Listed below are the weights (in grams) of a sample of bats. The weights are known to be normally distributed with a standard deviation of 0.25 grams. Provide an appropriate 95% confidence interval for the mean population weight.

1.7	1.6	1.5	2.0	2.3
1.6	1.6	1.8	1.5	1.7
2.2	1.4	1.6	1.6	1.6

Solution to Homework Question 1

Homework Questions 2 and 3

Consider the following samples and provide an appropriate 95% confidence interval for each:

2. A sample of 500 nursing applications included 60 from males. Find a 95% confidence interval to estimate the proportion of males who applied to nursing programs.

3. Noise levels at various area urban hospitals were measured in decibels. The mean noise level in 84 corridors was found to be 61.2 decibels with a standard deviation of 7.9 decibels.

Solution to Homework Questions 2 and 3

Lecture 6 Summary

General form of a confidence interval:

sample estimate □critical value □standard error

CI for μ (σ known): $\bar{y} \pm z_{crit} \times {}^{\sigma}/_{n}$

CI for μ (σ unknown): $\bar{y} \pm t_{crit} \times {}^{s}/_{n}$

CI for π (approximate): $p \pm z_{crit} \times \sqrt{\dfrac{p(1-p)}{n}}$

Textbook References

Further information on the topics discussed in this lecture can be found in:

Modern Statistics: An Introduction
by Don McNeil and Jenny Middledorp
(ISBN 9781486007011).

- Chapter 6: Pages 114 – 125

Note: Sample size calculations are not covered in STAT170.