Confidence Intervals (2)

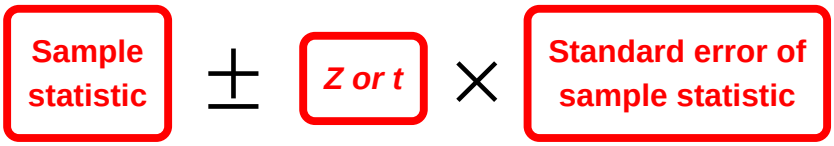
QMET103

Library, Teaching and Learning

Lincoln University
Te Whare Wānaka o Aoraki
CHRISTCHURCH•NEW ZEALAND
New Zealand's specialist land-based university

General

Remember: three values are used to construct all confidence intervals:



Decisions and Parameters to identify:

- Whether differences are between *means* or between *proportions*
- Whether samples are independent or not
- For each independent sample, the means and standard deviations, or the proportions
- Whether the given standard deviations (or variances) are from **population or sample**
- The level of confidence required
- For *t* score, the size of each sample and hence degrees of freedom:

$$n_1 + n_2 - 2$$

IDENTIFY the **sample statistic**

For two sample confidence intervals, this is either

- mean difference, $\bar{x}_d$: calculate the differences between the pairs of data, and process these differences as *one sample*

OR

- difference between means, $(\bar{x}_1 - \bar{x}_2)$: calculate the means of the two samples and calculate the difference between these two means

OR

- difference between proportions, $(p_1 - p_2)$: calculate the difference between the two proportions

SELECT the relevant **Z** or ***t*** score

- The value of *t* depends on the level of confidence and the sample size.

  It is used when the standard deviation is from a **sample** – ie σ is **unknown**

- The value of Z depends on the level of confidence only.
  Use it
  - when the standard deviation is from a **population**– ie σ is **known**,

  or
  - for **proportion**

CALCULATE the **standard error**

You will find these formulae on your formula sheet.

For difference between **MEANS**

- se(mean difference), $se(\bar{x}_d) = \frac{s_d}{\sqrt{n}}$:

  From the processed differences, use the s.d of the differences divided by the square root of the sample size

- se(difference between means), population sd known: $se(\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$
- se(difference between means), population sd NOT known (sample sd known):

$$se(\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)} \text{ where } s_p^2 = \frac{\left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right]}{(n_1 + n_2 - 2)}$$

$s_p^2 = pooled\ \ variance$. This must be calculated *first,* then used to calculate the standard error.

For difference between **PROPORTIONS**

- se(difference between proportions), for *INDEPENDENT samples:*

$$se(p_1 - p_2) = \sqrt{\left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)} \text{ or}$$

$$se(p_1 - p_2) = \sqrt{\left(\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)} \text{ where } \bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

- se(difference between proportions), for NON *INDEPENDENT samples:*

$$se(p_1 - p_2) = \sqrt{\left(\frac{p_1 + p_2 - (p_1 - p_2)^2}{n}\right)}$$

Examples

Two dependent samples, results paired

An Insurance Company obtained estimates of the cost of car repairs at a certain garage. The insurance company randomly selected five cars needing repairs and obtained the actual cost of the finished repair. The data are below:

Car	1	2	3	4	5
Estimate ($)	180	1000	65	320	200
Actual ($)	165	1054	68	362	234
From these paired results, the difference for each pair is obtained and processed:					
Difference	15	-54	-3	-42	-34

This is now treated as a single sample, with $\bar{x} = -23.6$ $s = 28.66$.

95% Confidence interval now = $-23.6 \pm 2.7764 \times \frac{28.66}{\sqrt{5}}$ = [$-59.18, -$11.98]

(No conclusion, since interval contains zero.)

Two independent samples, with population spread given $(\sigma \text{ or } \sigma^2 \text{ known})$

That is, ***sample*** means and ***population*** standard deviations or variances are known.

Weekly expenditure of students of two Universities are studied. Variances of the two populations are 90.1 and 97.7, respectively. From population 1, a sample of 15 students is selected with sample mean being $204.20. From population 2, a sample of 10 students is selected and the mean is $184.60. What is the 95% Confidence Interval for the difference between the two population means?

Identify: $\bar{x}_1 = 204.20,\ \bar{x}_2 = 184.60,\ s_1^2 = 90.1,\ s_2^2 = 97.7,\ n_1 = 15,\ n_2 = 10$

C.I.(difference between means)= $(\bar{x}_1 - \bar{x}_2) \pm Z \times \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$ **,**since σ is known.

$$= (204.20 - 184.60) \pm 1.96 \times \sqrt{\left(\frac{90.1}{15} + \frac{97.7}{10}\right)} = (\$11.81,\ \$27.39)$$

That is, we can be 95% confident that the true mean difference in expenditure between the two universities is between $11.81 and $27.39.

Note that in this example, the **variances** were given, not the standard deviations. There is no need therefore to square 90.1 and 97.7 in the s.e. formula.

Two independent samples, with sample spread given $(s \text{ or } s^2 \text{ known})$

That is, ***sample*** means and ***sample*** standard deviations or variances are known.

20 male students were randomly sampled to determine the mean number of hours per week males spend watching television. The mean for this sample was 15.3 hours and the standard deviation 4.2 hours.

A similar survey of 25 females produced a mean of 11.2 hours and a standard deviation of 3.0 hours.

Calculate the 90% confidence interval for the difference in hours spent watching television between the male and female groups. Interpret this confidence interval.

Identify: $\bar{x}_1 = 15.3,\ \bar{x}_2 = 11.2,\ s_1 = 4.2,\ s_2 = 3.0,\ n_1 = 20,\ n_2 = 25$

C.I.(diff between means)= $(\bar{x}_1 - \bar{x}_2) \pm t_{(n_1+n_2-2)} \times \sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}$**,** since σ is unknown

First calculate $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)} = \frac{(19 \times 4.2^2 + 24 \times 3.0^2)}{43} = 12.82$

Now complete the C.I. $= (15.3-11.2) \pm 1.684 \times \sqrt{\left(\frac{12.82}{20} + \frac{12.82}{25}\right)} = (2.20, 5.91)$

That is, we can be 95% confident that the true mean difference in hours spent watching television between males and females is between 2.3 and 5.9 hours.

Two independent samples, proportions given. ie, ***two*** separate samples.
The senior accounts manager of a chain of shops is concerned about the differences between shops in their control of a stock shrinkage (loss due to theft and poor inventory control). She thinks that the difference is due to the type of shopping centre the shops are located in: either in malls or with street frontages. She collects the following data.

	Type of locality	
	Mall	Street frontage
Number of shops	8	10
Stock shrinkage	1%	2%

Calculate the 95% confidence interval for the difference in shrinkage between the shop types $(\pi_{mall} - \pi_{street})$.

$$C.I. = (p_1 - p_2) \pm Z\text{-}score \times \sqrt{\left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)}$$

Identify:

- $p_1 = 0.01$ *and* $p_2 = 0.02$, the probability of "success" from each sample
- the size of each sample $n_1 = 8,\ n_2 = 10$.

Hence

$$C.I. = (0.02 - 0.01) \pm 1.96 \times \sqrt{\left(\frac{0.02 \times 0.98}{10} + \frac{0.01 \times 0.99}{8}\right)} = (-0.10, 0.12)$$

Because there is a zero in the 95% C.I., (the interval goes form positive to negative), no conclusion can be drawn. This implies that the difference in stock shrinkage could be zero.

Two Non- independent samples, proportions given; sometimes referred to as a "same sample" survey; since the proportions come from a single sample.

Surveyors asked a random sample of 120 women, what factor was the most important in deciding where to shop. The results are summarised in the following table:

Factor	***Percentage (%)***
Price and value	*40*
Quality and selection of merchandise	*30*
Service	*15*
Shopping environment	*15*

Calculate a 95% confidence interval for the difference between the population proportion of women who identified 'price and value' as most important and the proportion of women who identified 'quality and selection of merchandise' as most important ($\pi_{price\ and\ value} - \pi_{quality\ and\ selection}$)

In this situation, the proportion in each category is inter–dependent. A change in one value will affect one or more of the others.

Identify:

- $p_1 = 0.4 \ and \ p_2 = 0.3$, the probability of "success" from each sample
- the size of the sample, $n = 120$ Note, there is only ***one*** sample size.

$$C.I. = (p_1 - p_2) \pm Z \times \sqrt{\frac{p_1 + p_2 - (p_1 - p_2)^2}{n}}$$

$$= (0.40 - 0.30) \pm 1.96 \times \sqrt{\left[\frac{0.4 + 0.3 - (0.4 - 0.3)^2}{120}\right]}$$

$$= (-0.05, 0.25)$$

Again, there is a zero in the 95% C.I., so no conclusion can be drawn. This implies that the difference in preference could be zero.

Practice

1. A real-estate company appraised the market value of 27 homes in Lyttelton and found that the sample mean and standard deviation were \$150,000 and \$17,000 respectively. The real-estate company also appraised the market value of 45 homes in Aranui and found that the sample mean and standard deviation were \$100,000 and \$12,000 respectively.

Calculate the 90% confidence interval estimate for the population difference in market value between the Lyttelton and Aranui areas ($\mu_{Lyttelton} - \mu_{Aranui}$). [5 Marks]

2. The 'Country Taste' bread making company wants to estimate the actual weight of their 700 gm bread. It is known that the government specification for the standard deviation of weight of this bread is 15 gm.
   A random sample of 50 breads is selected, and the sample mean weight is 696 gm.
   Another company, South Taste, also conducted a similar study with a random sample of 60 breads and found that the sample mean is 701 gm.

Calculate the 90% confidence interval estimate for the difference between the two population mean weights $(\mu_{country} - \mu_{south})$.

3. A sample of 50 observation wells are randomly selected from Canterbury region and it is found that nitrate level in 3 wells exceeds the allowable limit.
   Of sample of 60 wells in Waikato area 8% of the wells indicate that their nitrate level is over the acceptable limit.

   Construct a 90% confidence interval for the difference between the two population proportions that exceed the allowable nitrate concentration.

4. *In a Rugby World Cup, a random sample of supporters was asked, "Which country do you think will win the 2003 Rugby World Cup?" The results are summarised:*

Country	Number of supporters who think their country will win
Australia	116
England	13
France	25
New Zealand	140
Western Samoa	50
South Africa	47
Wales	24
Undecided	65
Total	480

Calculate the 95% confidence interval for the difference between the proportion who think Australia will win and the proportion who think that New Zealand will win.

5. A farmer wants to examine the effect a new drench is having on the weight of his sheep. To do this, he weighs 10 sheep prior to drenching and then reweighs the same sheep 2 weeks after drenching. The measurements obtained are given in the table below.

Sheep no.	Weight before (kg)	Weight after (kg)
1	55.6	58.3
2	60.1	62.1
3	46.8	45.7
4	42.6	46.2
5	58.1	31.2
6	54.3	56.8
7	62.9	62.4
8	49.4	53.2
9	58.6	59.4
10	51.3	55.6

Calculate a 95% confidence interval for the difference in sheep weights after drenching.

6. In a study, a group of 42 sedentary men were placed on a diet. After 6 months these men had lost an average of 7.2 kg of bodyweight with a standard deviation of 3.7 kg.

a) Calculate the 95% confidence interval estimate of the population mean weight loss.

b) In a separate study, 47 previously sedentary men were put on an exercise routine. After 6 months, these men had lost an average of 4.0kg with a standard deviation of 3.9kg. Calculate the 95% confidence interval estimate for the population difference in mean weight loss ($\mu_{diet} - \mu_{exercise}$)

c) Interpret the confidence interval calculated in (b).

In a study investigating the best treatment to help people to stop smoking, 244 smokers were given Zyban (antidepressant). After 6 months 85 of the 244 Zyban users had quit smoking.

d) Calculate the 95% confidence interval estimate of the population proportion who had quit smoking.

In addition to the Zyban treatment, a further 244 smokers were given nicotine patches. After 6 months, 52 of these patch users had given up smoking.

e) Calculate the 95% confidence interval estimate for the difference in the population proportion of people giving up smoking using the different treatments (π_{Zyban} - $\pi_{Patches}$).

Answers

1
$$s_p^2 = \frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{(n_1+n_2-2)}$$
$$= \frac{17000^2(27-1)+12000^2(45-1)}{(27+45-2)}$$
$$= 197\,857\,142.9$$

$$\Rightarrow C.I. = (\bar{x}_1 - \bar{x}_2) \pm t_{(n_1+n_2-2)} \times \sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}$$

$$= (150{,}000 - 100{,}000) \pm 1.6669 \times \sqrt{\left(\frac{197857142.9}{27} + \frac{197857142.9}{45}\right)}$$

$$= (\$44292.26, \ \$55707.73)$$

We can be 90% confident the true mean difference in market value between the Lyttelton and Aranui areas is between $44,000 and $56,000

2 $\sigma_1 = \sigma_2 = 15 \ \ \overline{x_1} = 696, \ \ \overline{x_2} = 701, \ n_1 = 50, \ n_2 = 60$

$$C.I. = (\bar{x}_1 - \bar{x}_2) \pm Z \times \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$$
$$= (701 - 696) \pm 1.645 \times \sqrt{\frac{15^2}{50} + \frac{15^2}{60}}$$
$$= (0.275, \ 9.72)$$

We can be 90% confident the true difference between the two population mean weights $(\mu_{country} - \mu_{south})$ is between 0.275 and 9.72gm.

3. $$p_1 = \frac{3}{50} = 0.06, \ \ p_2 = 0.08$$

These are independent samples with $n_1 = 50, \ n_2 = 60$.

$$C.I. = (0.08 - 0.06) \pm 1.645 \times \sqrt{\left(\frac{0.08 \times 0.92}{60} + \frac{0.06 \times 0.94}{50}\right)}$$

$$= (-0.06 \ \ 0.10)$$

No conclusion can be made about the population difference in proportion exceeding nitrate levels for Waikato and Canterbury.

4. These proportions are non-independent, with $P_1 = 0.2417$; $P_2 = 0.2917$

$$C.I. = (p_1 - p_2) \pm Z \times \sqrt{\frac{p_1 + p_2 - (p_1 - p_2)^2}{n}}$$

$$= (0.2917 - 0.2417) \pm 1.96 \times \sqrt{\left[\frac{\left(0.2917 + 0.2417 - (0.2917 - .02417)^2\right)}{480}\right]}$$

$$= (-0.015,\ 0.115)$$

That is, we can be 95% confident that the population difference in proportion of those who think Australia will win the next world cup and those who think NZ will win the next world cup.

5. Differences in sheep after drenching are:
2.7 2.0 -1.1 3.6 3.1 2.5 -0.5 3.8 0.8 4.3
Mean difference and s.d. are 2.12 and 1.832 respectively. Hence,

$$C.I. = 2.12 \pm 2.262 \times \frac{1.832}{\sqrt{10}} = (0.810, 3.43)$$

6. $n = 42, \bar{x} = 7.2kg, s = 3.7$.

a) For 95% C.I., $t = 2.0105 \Rightarrow C.I.(mean) = 7.2 \pm 2.0105 \times \frac{3.7}{\sqrt{42}}$

$$= (6.052,\ 8.35)kg$$

b) $n_{diet} = 42,\ \bar{x}_{diet} = 7.2kg,\ s_{diet} = 3.7kg$

$n_{exercise} = 47,\ \bar{x}_{exercise} = 4.0kg,\ s_{exercise} = 3.9kg$

For a 95% C.I. df = 87, and $t = 1.9876$

Pooled variance needed = $s_p^2 = \frac{41 \times 7.2^2 + 46 \times 4.0^2}{87} = 32.89$

$$\Rightarrow C.I.(\text{difference between means})$$

$$= (7.2 - 4.0) \pm 1.9876 \times \sqrt{\frac{32.89}{42} + \frac{32.89}{47}}$$

$$= (0.7796,\ \ 5.6203)kg$$

c) We can be 95% confident that the true population mean difference in weight loss is between 0.78 and 5.6kg

d) $p = \frac{85}{244} = 0.348$ For 95% C.I., Z= 1.96

$$\Rightarrow C.I.(\ proprtion\ ) = 0.348 \pm 1.96 \times \sqrt{\frac{0.348 \times 0.652}{244}}$$

$$= (0.288,\ 0.407)$$

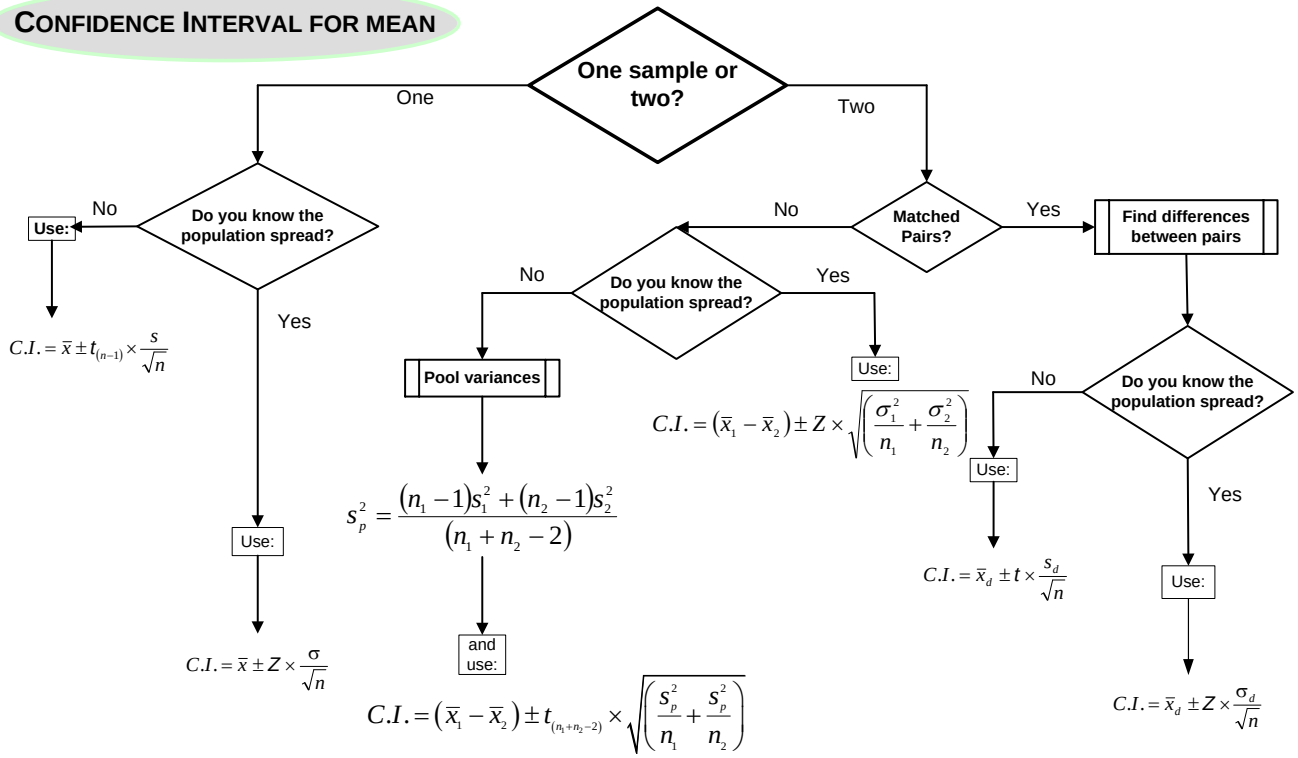
e) $p_{Zyban} = \frac{85}{244} = 0.348,\ p_{patch} = \frac{52}{244} = 0.213$ For 95% C.I., Z= 1.96

$$\Rightarrow C.I.(diff.proportion)$$

$$= (0.348 - 0.213) \pm 1.96 \times \sqrt{\frac{0.348 \times 0.652}{244} + \frac{0.213 \times 0.787}{244}}$$

$$= (0.0562, 0.2138)$$

CONFIDENCE INTERVAL FOR MEAN



CONFIDENCE INTERVAL FOR PROPORTION