Stats 101/101G/108 Workshop

Confidence Intervals: *Proportions* [CIP]

2019

by Leila Boyle

Stats 101/101G/108 Workshops

The Statistics Department offers workshops and one-to-one/small group assistance for Stats 101/101G/108 students wanting to improve their statistics skills and understanding of core concepts and topics.

Leila's website for Stats 101/101G/108 workshop hand-outs and information is here: **www.tinyURL.com/stats-10x**

Resources for this workshop, including pdfs of this hand-out and Leila's scanned slides showing her working for each problem are available here: **www.tinyURL.com/stats-CIP**

Leila Boyle
Undergraduate Statistics Assistance, Department of Statistics
Room 303.320 (third floor of the Science Centre, Building 303)
l.boyle@auckland.ac.nz; (09) 923-9045; 021 447-018

Want help with Stats?

Stats 101/101G/108 appointments

Book your preferred time with Leila here: **www.tinyURL.com/appt-stats**, or contact her directly (see above for her contact details).

Stats 101/101G/108 Workshops

Workshops are run in a relaxed environment, and allow plenty of time for questions. In fact, this is encouraged ☺

Please make sure you bring your calculator with you to all of these workshops!

- **Preparation at the beginning of the semester:**

  Multiple identical sessions of a preparation workshop are run at the beginning of the semester to get students off to a good start – come along to whichever one suits your schedule!

  - Basic Maths and Calculator skills for Statistics **www.tinyURL.com/stats-BM**

- **First half of the semester**

  Five theory workshops are held during the first half of the semester:

  - Exploratory Data Analysis **www.tinyURL.com/stats-EDA**
  - Proportions and Proportional Reasoning **www.tinyURL.com/stats-PPR**
  - Observational Studies, Experiments, Polls and Surveys **www.tinyURL.com/stats-OSE**
  - Confidence Intervals: *Means* **www.tinyURL.com/stats-CIM**
  - Confidence Intervals: *Proportions* **www.tinyURL.com/stats-CIP**

- **Second half of the semester**

  Five theory workshops and one computing workshop are held during the second half of the semester:

- **Statistics Theory Workshops**

  - Hypothesis Tests: *Proportions* **www.tinyURL.com/stats-HTP**
  - Hypothesis Tests: *Means (part 1)* **www.tinyURL.com/stats-HTM**
  - Hypothesis Tests: *Means (part 2)* **www.tinyURL.com/stats-HTM**
  - Chi-Square Tests **www.tinyURL.com/stats-CST**
  - Regression and Correlation **www.tinyURL.com/stats-RC**

- **Computer Workshop:** Hypothesis Tests in *SPSS* **www.tinyURL.com/stats-HTS**

- **Useful Computer Resource:**

  If you haven't used SPSS before, you may find it useful to work your way through this self-paced workshop: **www.tinyURL.com/stats-IS**

A note about rounding numbers

Often students ask me about rounding numbers – how to do it and how much by.

When you do a calculation, you may end up with an answer that has many (5 or more) decimal places associated with it. People don't deal too well with numbers to this level of accuracy; rounding helps us make a number a little simpler while still keeping its value relatively close to what it was. The result is less accurate, but easier to use, interpret and understand.

How to round numbers

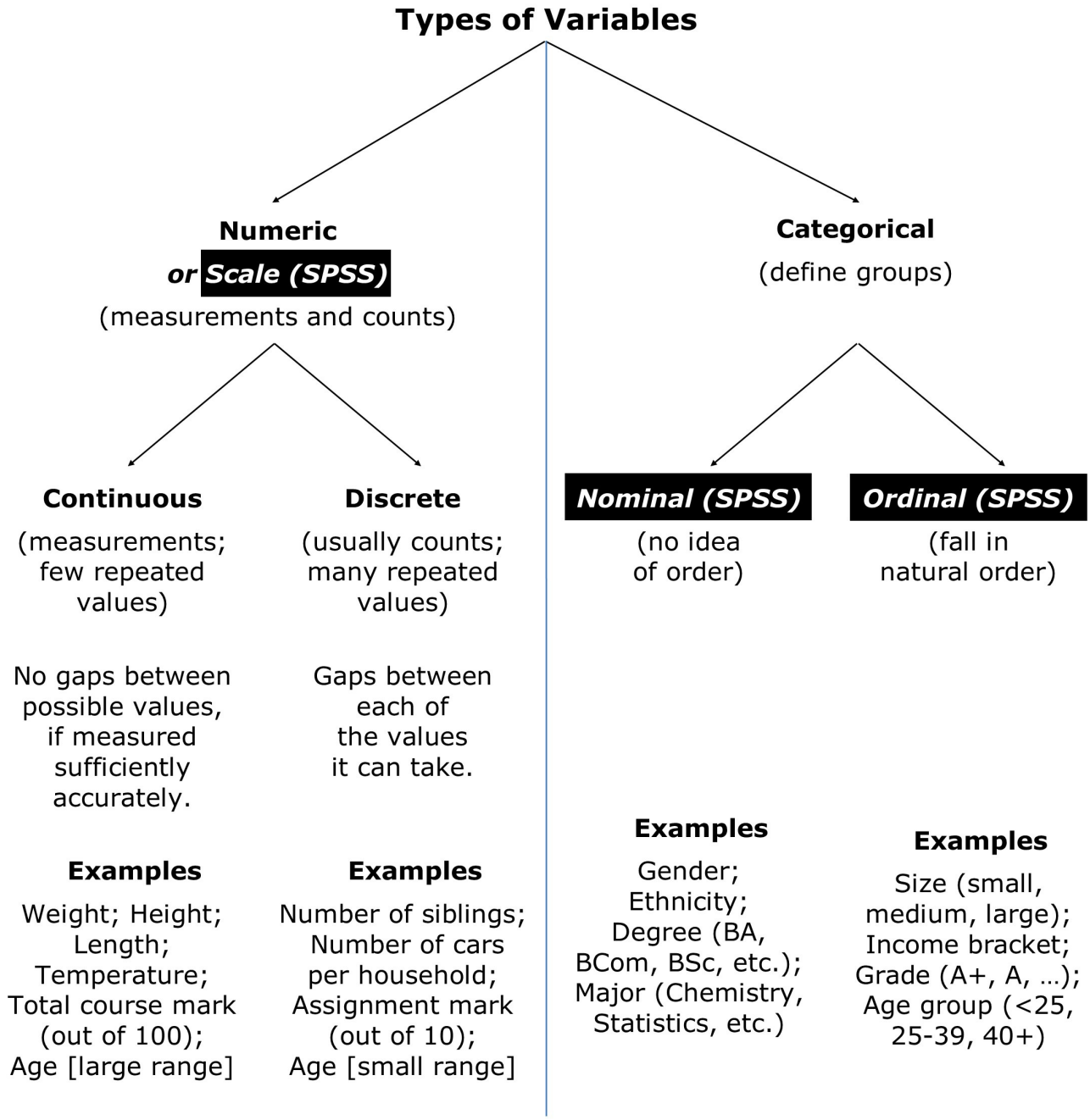
- Decide which is the last digit to keep
- Leave it the same if the next digit is less than 5 (this is called **rounding down**)
- Increase it by 1 if the next digit is 5 or more (this is called **rounding up**)

How many decimal places should I use?

When you are doing calculations in an assignment or test/exam context, you should get some guidance from the question itself. For example, a multi-choice test question may have five answers that are all rounded to one decimal place (1dp for short) so just round your answer when you do the calculation to 1dp.

If you are going to use the value you come up with in a later calculation, go for more accuracy rather than less, as the more you round a number, the more it will affect the results of later calculations that use that value. My default amount of rounding tends to be 4dp – this is reasonably accurate without going over the top!

Recall from Chapter 1:



📖**Useful reference:** Chance Encounters, pages 40 – 42

- A **proportion** is a number between 0 and 1 that estimates the likelihood of an event occurring.
- Proportions can be presented as **fractions**, **decimals** or **percentages**.
- When doing calculating the limits for Normality-based confidence intervals by hand, make sure proportions are in their **decimal** representation. You can convert the limits to **percentages** for interpretation, if you wish.

Understanding Confidence Intervals

- Statistics is concerned with finding out about the real world and aspects of it specific to our area of interest. Statistical tools allow us to deal with the **uncertainty** present in all samples due to **sampling variation** which occurs because we are unable to survey the entire population of interest.

- We are usually unable to survey the entire population (take a census) as it is too large and/or there are:
  - ✓ budget constraints
  - ✓ time limits
  - ✓ logistical barriers

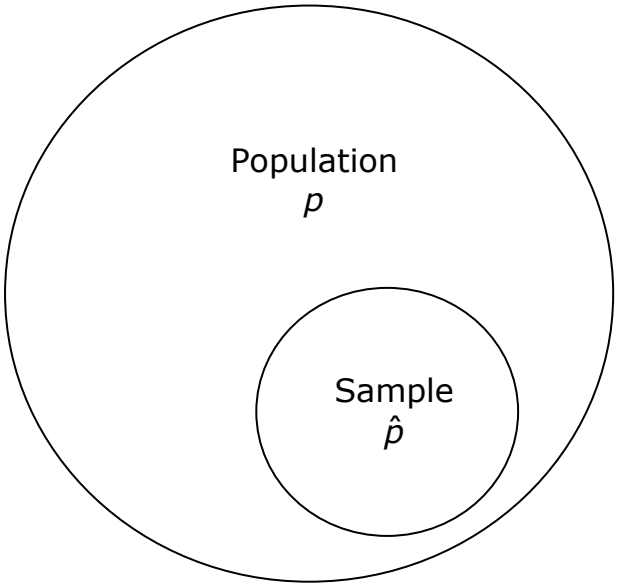
- This means we are unable to establish the **parameter** of interest within our population, such as:
  - ✓ Population proportion, p

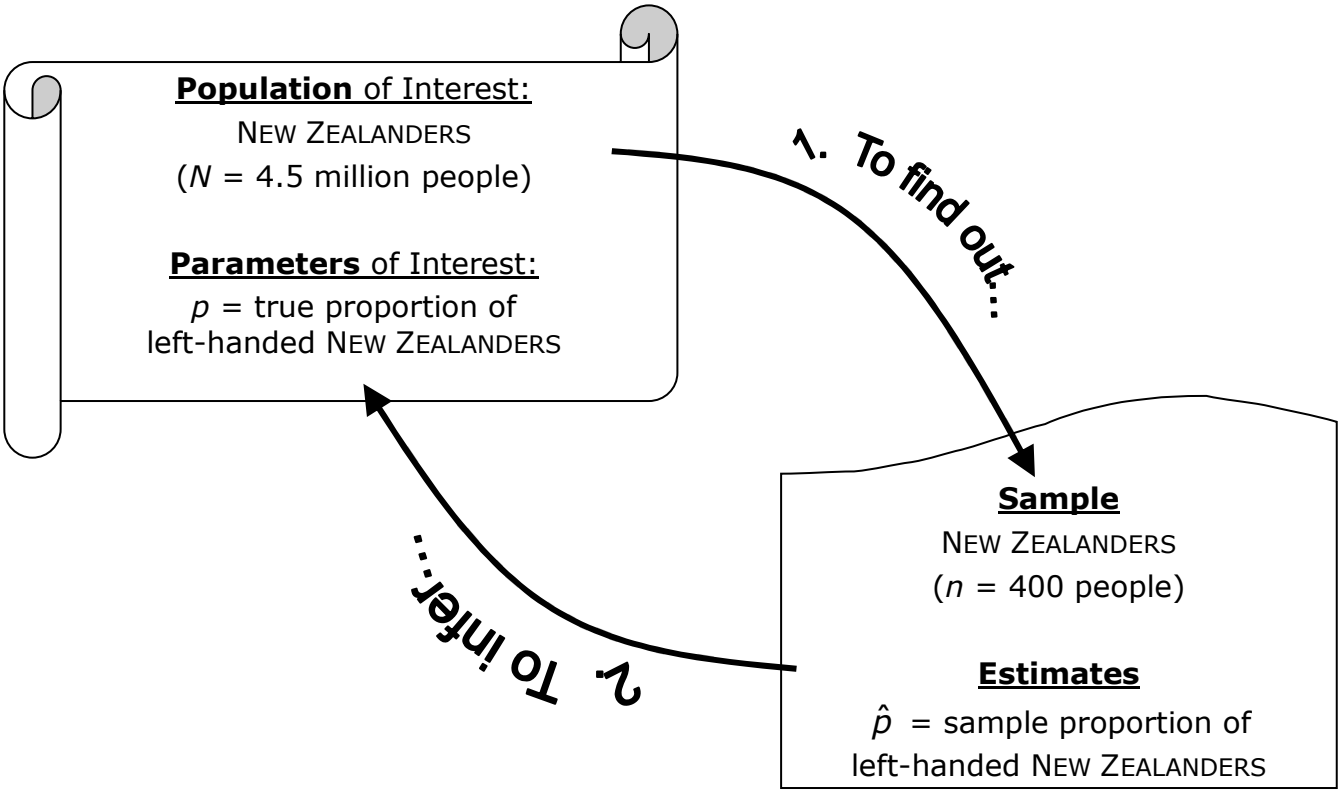
- This means that the **parameter** of interest is an ***unknown* numerical characteristic** for that particular population.

- To estimate an ***unknown* numerical characteristic* (parameter)** for our population of interest, we take a sample and find a sample **estimate** from it (that is, we make a **statistical inference**). The **sample estimate** of the above **population parameter** is:
  - ✓ Sample proportion, $\hat{p}$

- Usually $\hat{}_{HATS}$ or $\bar{}_{BARS}$ are used to distinguish between **sample estimates** and **population parameters**.

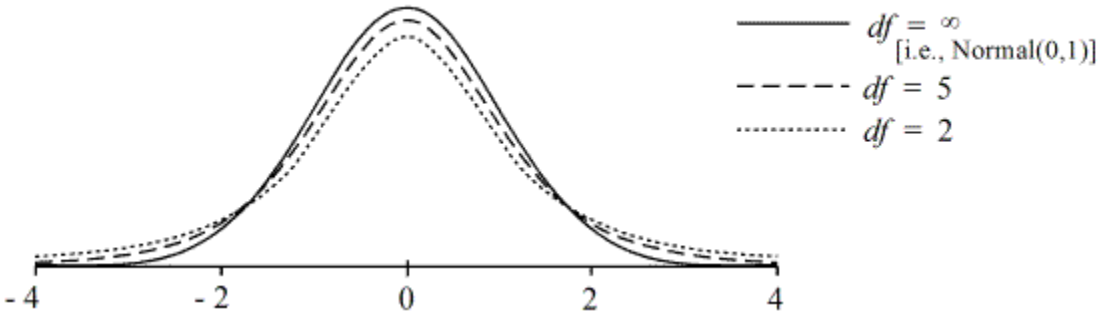
- The process of using sample data to make useful statements about a population is a type of **statistical inference** called **sample-to-population inference**, e.g., when using sample data to estimate an unknown population proportion.





Student's *t*-distribution

- ✓ Parameter: Degrees of Freedom (*df*).
- ✓ Smooth symmetric, bell-shaped curve centred at 0 like the Standard Normal distribution [$Z \sim$ Normal ($\mu = 0$, $\sigma = 1$)] but it's more variable (is more spread out).



- ✓ As *df* becomes larger, the Student (*df*) distribution becomes more and more like the Standard Normal distribution.
- ✓ Student's *t*-distribution ($df = \infty$) and Normal (0,1) are the same distribution.
- ✓ Methods based on this distribution work very well even for small samples that are from very non-Normal distributions.

Confidence Intervals (Normality-based)

- A **confidence interval** gives a range of plausible values for the parameter of interest that is consistent with the data (at the specified level of confidence). It determines the **size** of the effect or difference.
- You can do all kind of CIs, 90%, 95%, 99%...
- Increasing the confidence level will **increase** the width of the interval.

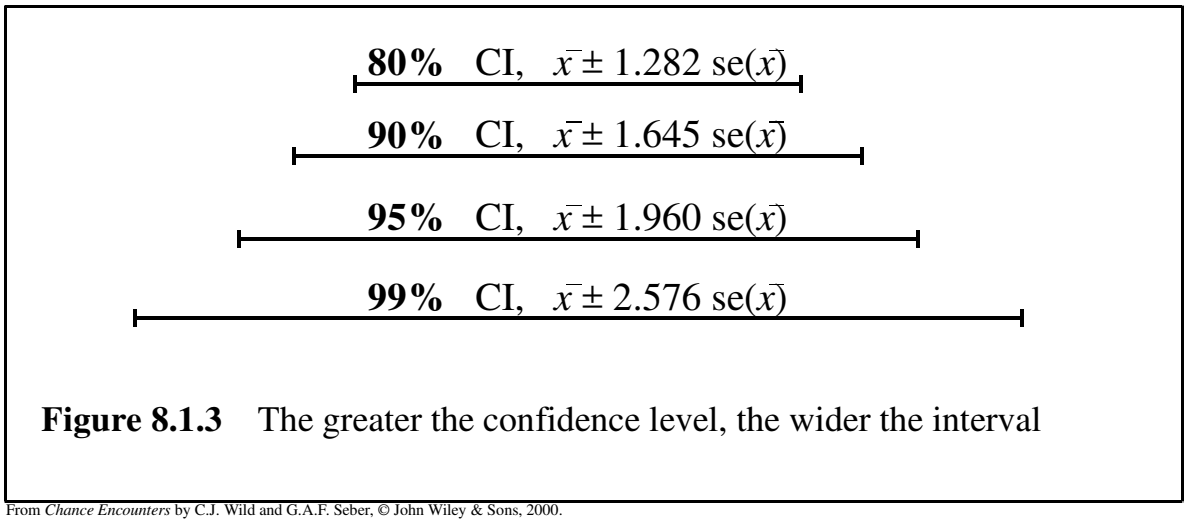


Figure 8.1.3 The greater the confidence level, the wider the interval

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

- Increasing the sample size will make the confidence interval more precise.
- To double the accuracy of the confidence interval we **need 4 times** as many observations.
- To triple the accuracy of the confidence interval we **need 9 times** as many observations.
- 95% confidence interval
  - ✓ Range of plausible values for the parameter of interest that contains the **true value** of our parameter of interest for 95% of samples taken.
  - ✓ 5% of samples taken will not have the parameter within the calculated confidence interval.
  - ✓ We do not know if the sample we have taken is one of the 95% that contains the true unknown parameter. All we can say is that 95% of the time it will.

✓ If you take 1000 samples, based on the same sampling protocol, then you can expect approximately 950 of these samples will contain the true value (e.g. true mean) of the population.

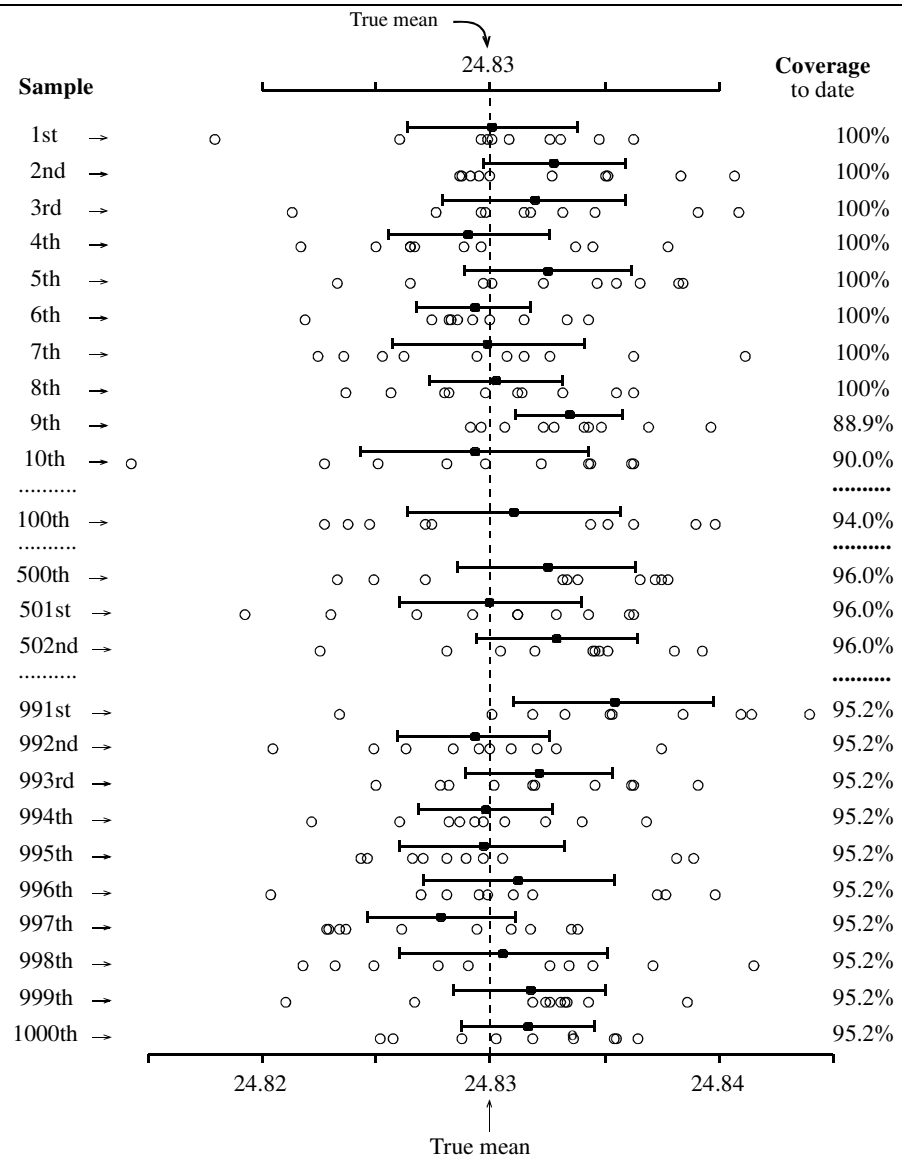


Figure 8.1.2 Samples of size 10 from a Normal(μ=24.83, s=.005) distribution and their 95% confidence intervals for μ..

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 1999.

Step-by-Step Guide to Producing a Confidence Interval by Hand

1. State the **parameter** to be estimated. (Symbol and words)
Is it μ, p, $\mu_1 - \mu_2$, or $p_1 - p_2$?

2. State the **estimate** and its value

3. Write down the **formula** for a CI, from the Formula Sheet — ***estimate* ± *t* × se(*estimate*)**

4. Use the appropriate **standard error**. (Will be provided)

5. Use the appropriate ***t*-multiplier**. (Will be provided)

6. Calculate the **confidence limits**. (End points of the confidence interval)

7. Interpret the interval using plain English.
Use the confidence limits to construct an answer to the original question in plain English.

see back page for Formulae Sheet

- There are four different types of problem:
  1. Single mean 2. Single proportion. 3. Difference between two means
  4. Difference between two proportions:
     Situation (a) ***Proportions from two independent samples***
     Situation (b) ***One sample of size n, several response categories***
     Situation (c) ***One sample of size n, many yes/no items***

estimate ± _t_ × se(_estimate_)

First piece of CI formula / Step 2: **_estimate_ ± _t_ × se(_estimate_)**

- The **estimate** is based on the **parameter** of interest we are investigating:

Parameter	**Estimate**
1. Single mean $\boldsymbol{\mu}$:	$estimate = \bar{x}$
2. Single proportion $\boldsymbol{p}$:	$estimate = \hat{p}$
3. Difference between two means $\boldsymbol{\mu_1 - \mu_2}$: (independent samples)	$estimate = \bar{x}_1 - \bar{x}_2$
4. Difference between two proportions $\boldsymbol{p_1 - p_2}$:	$estimate = \hat{p}_1 - \hat{p}_2$

Second piece of CI formula / Step 5: ***estimate* ± *t* × se(*estimate*)**

The ***t*-multiplier** is based on:

- Whether we are investigating means or proportions
- The desired level of confidence
- The degrees of freedom
- The **degrees of freedom** are based on the problem type:

Estimate	**Degrees of Freedom**
1. $estimate = \bar{x}$	$df = n - 1$
2. $estimate = \hat{p}$	$df = \infty$
3. $estimate = \bar{x}_1 - \bar{x}_2$	$df = minimum(n_1 - 1, n_2 - 1)$
4. $estimate = \hat{p}_1 - \hat{p}_2$	$df = \infty$

In the test/exam situation, you will be given the *t*-multiplier for a 95% confidence interval for a single proportion or a difference between proportions (it's 1.96!)

Third piece of CI formula / Step 4: ***estimate* ± *t* × se(*estimate*)**

- The **standard error** can be found from the *t*-procedures spreadsheet. In the test/exam situation, the standard error will be provided.

Interpreting the CI limits → Step 7 for story type 4:

- CIs for the difference between two means/proportions:

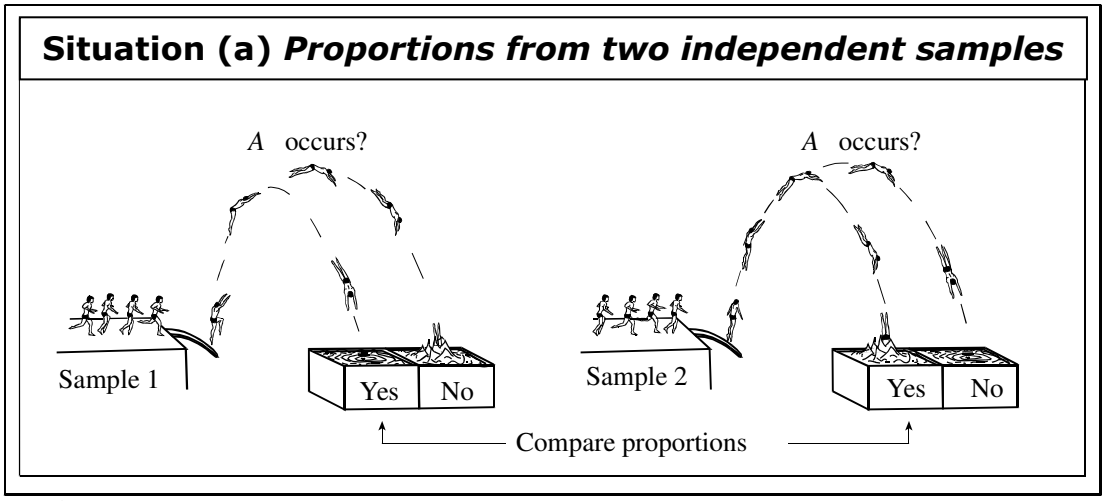
Examples:

✓ If the CI contains 0 (i.e. one negative and one positive number), there may be no difference between the two proportions. (-.07, .03)

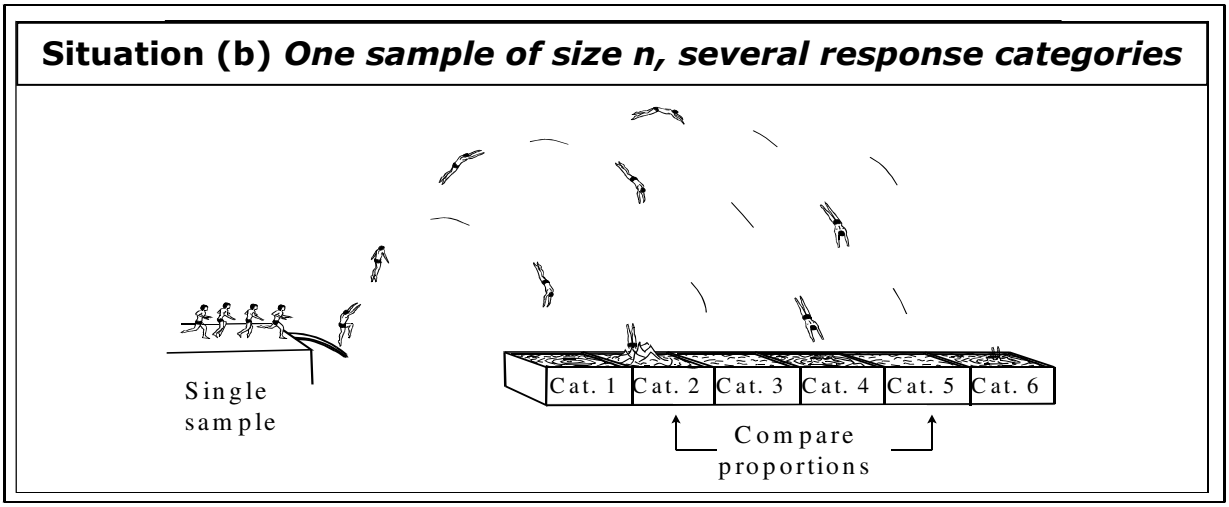
✓ If CI is positive, then p_1 is higher/larger than p_2. (.03, .07)

✓ If CI is negative, then p_1 is lower/smaller than p_2. (-.07, -.03)

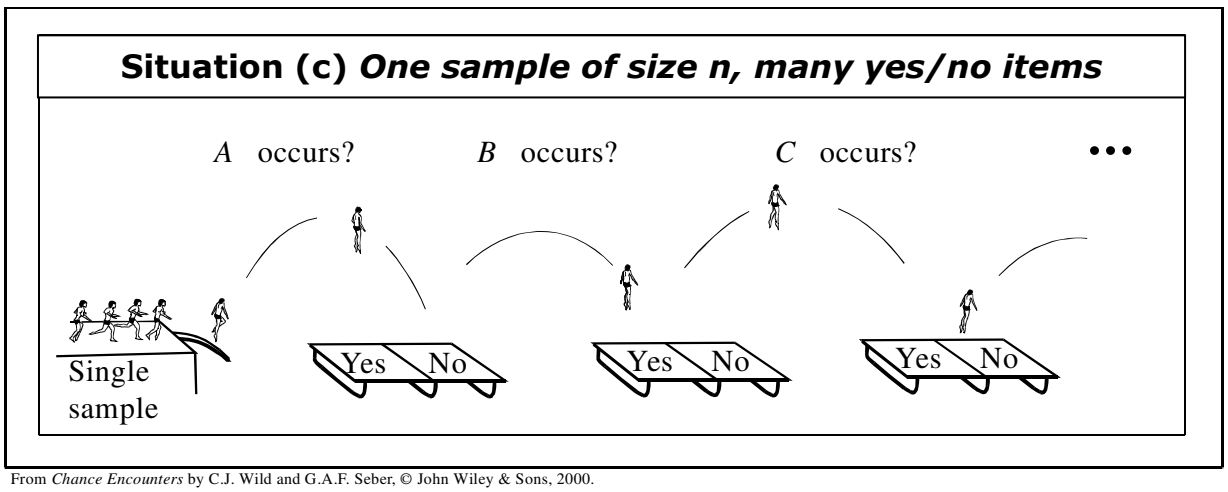
- **3 sampling situations for the difference between two proportions**



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 1999.



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Check you understand! – Practice Questions

Questions 1 to **3** refer to the following information.

CNN/Time sought information on how young American adults viewed their parents' marriage. In a telephone poll, one of the questions they asked of six hundred and two (602) 18-29 year old Americans was "Would you like to have a marriage like the one your parents have?". Forty-four percent (44%) responded "Yes".

1. *CNN/Time* were interested in determining what proportion of the 18-29 year old American population would answer "Yes" to this question. Which **one** of the following statements is **false**?

   (1) The value of the parameter of interest is an unknown quantity.

   (2) In this context, 0.44 is an estimate for the parameter of interest.

   (3) The parameter of interest depends on the sample and hence is a random quantity.

   (4) A confidence interval for the parameter of interest will give a range of possible values for this parameter.

   (5) The parameter of interest is the proportion of 18-29 year old Americans who would have answered "Yes".

2. Let p be the proportion of the 18-29 year old American population who would have answered "Yes" to this question.

   The 95% confidence interval for p is:

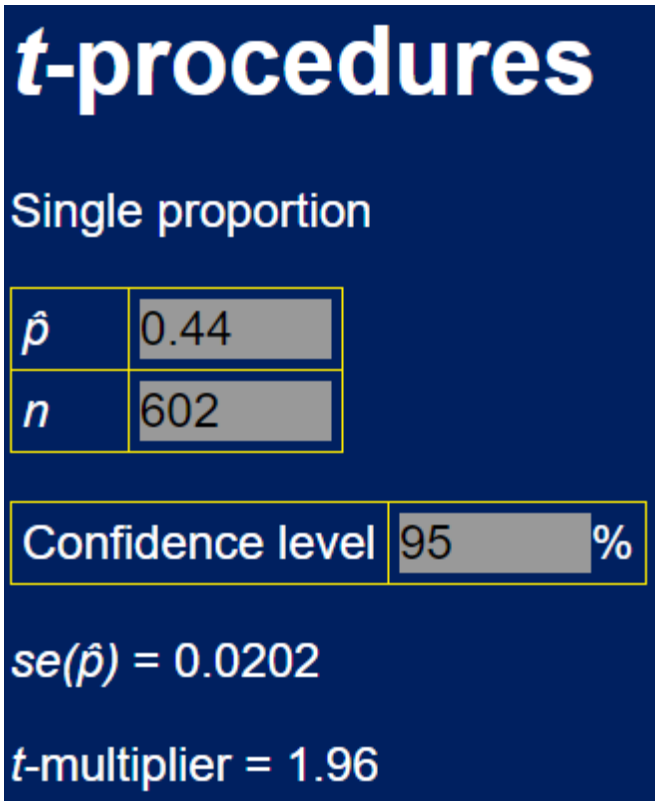
   (1) [0.415, 0.465]

   (2) [0.439, 0.441]

   (3) [0.400, 0.480]

   (4) [0.420, 0.460]

   (5) [0.407, 0.473]



3. If two thousand four hundred (2400) 18-29 year old Americans had been sampled instead of six hundred and two (602) 18-29 year old Americans, then the new 95% confidence interval would be approximately:

   (1) twice as wide.

   (2) one-quarter as wide.

   (3) half as wide.

   (4) four times as wide.

   (5) equally as wide.

4. A recent study in the USA investigated the drinking habits of a large sample of pregnant women. It found that 16% of the subjects drank alcohol frequently (seven or more drinks a week).

   Let p be the population proportion of pregnant women in the USA who drink alcohol frequently.

   Which **one** of the following statements is **false**?

   (1) The variability of the sample proportion, $\hat{p}$, increases as we increase the size of the sample on which it is based.

   (2) If a second sample were taken, the population proportion p would not change.

   (3) p is the probability that a randomly chosen pregnant woman in the USA drinks alcohol frequently.

   (4) The variability of the sample proportion, $\hat{p}$, decreases as we increase the size of the sample on which it is based.

   (5) The sample proportion $\hat{p} = 0.16$ is an estimate of p.

5. Which one of the following statements is **true**?

   (1) A point estimate is preferred to a confidence interval because the interval summarises the uncertainty due to sampling variation.

   (2) The standard error used to construct the interval will be identical for all samples of the same size.

   (3) If I plan to do a study in the future in which I will take a random sample and calculate a 90% confidence interval, there is a 90% chance that I will catch the true value of p in my interval.

   (4) The size of the t-multiplier depends on only the sample size and not the desired confidence level.

   (5) The process of using a population parameter to construct an interval for the data estimate is an example of statistical inference.

Questions 6 and **7** refer to the following information.

A survey of 1,146 New Zealanders asked the question "Is it a good time to buy a major household item?" 585 respondents replied "Yes", 332 replied "no" and 229 replied "don't know".

6. The proportion of the sample who think that it is a good time to buy a major household item is approximately:

   (1) 0.20
   (2) 0.29
   (3) 0.49
   (4) 0.51
   (5) 0.95

7. Let p be the proportion of the population who think that it is a good time to buy a major household item.

   The standard error of the estimate, se($\hat{p}$), is approximately 0.0148 and the value of the t-multiplier for constructing a 95% confidence interval for p is approximately 1.960.

   The 95% confidence interval for p is approximately:

   (1) [0.43, 0.50]
   (2) [0.45, 0.57]
   (3) [0.48, 0.54]
   (4) [0.50, 0.52]
   (5) [0.61, 0.67]

8. Which one of the following statements about a confidence interval for a parameter p is **false**?

   (1) A two-standard-error interval will always capture the true value of p.
   (2) Large samples tend to yield narrower 95% confidence intervals than small samples.
   (3) In the long run, if we repeatedly take samples and calculate a 95% confidence interval from each sample, we expect that 95% of the intervals will contain the true value of p.
   (4) If I plan to do a study in the future in which I will take a random sample and calculate a 90% confidence interval, there is a 90% chance that I will catch the true value of p in my interval.
   (5) If a large number of researchers independently perform studies to estimate p, about 95% of them will catch the true value of p in their 95% confidence intervals.

9. When using a t-procedure to construct a confidence interval for a population proportion, the confidence interval is constructed using the formula:

   $$\textbf{\textit{estimate}} \pm \textbf{\textit{t}} \times \textbf{se}(\textbf{\textit{estimate}})$$

   Which one of the following statements is **false**?

   (1) The margin of error is the quantity added to and subtracted from the estimate to construct the interval.
   (2) A confidence interval is preferred to a point estimate because the interval summarises the uncertainty due to sampling variation.
   (3) The size of the multiplier, t, depends only on the desired confidence level and not on the sample size.
   (4) The standard error used to construct the interval will be identical for all samples of the same size.
   (5) Large samples tend to yield narrower 95% confidence intervals than small samples.

Questions 10 refers to the following information.

The *NZ Herald* reported the results of a two-year study at Hong Kong's Kwong Wah Hospital. The study comprised 5450 impotent men who were given Viagra at the hospital. Of these men, 3651 were smokers. Of the 926 impotent men for whom Viagra did not work, 840 were smokers.

Assume this sample of 5450 impotent men is a random sample of all impotent Chinese men and recall that Viagra did **not** work for 926 of the men in this sample. Suppose it is known that Viagra works for 78% of impotent Western males.

10. Which **one** of the following statements is **true**?
    (1) The proportion of Chinese men in the sample for whom Viagra works is **both** an **estimate** value and a **parameter** value.
    (2) The proportion of Chinese men in the sample for whom Viagra works is an **estimate** value whereas the '78%' of impotent Western males for whom Viagra works is a **parameter** value.
    (3) The proportion of Chinese men in the sample for whom Viagra works and the '78%' of impotent Western males for whom Viagra works are **both parameter** values.
    (4) The proportion of Chinese men in the sample for whom Viagra works and the '78%' of impotent Western males for whom Viagra works are **neither estimates nor parameter** values.
    (5) The proportion of Chinese men in the sample for whom Viagra works is a **parameter** value whereas the '78%' of impotent Western males for whom Viagra works is an **estimate** value.

11. *The New Zealand Herald's* "Summer of Polls" featured the results of a series of public opinion surveys conducted by *Digipoll*. A total of 650 New Zealanders were asked "Should New Zealand join Australia under a common government?" The responses were:

Yes	14.3%
No	81.5%
Not Sure	4.2%

A 95% confidence interval for the proportion of all New Zealanders who would have answered "Yes" is approximately: (Note, use 1.960 as the *t*-multiplier and 0.0137 as the standard error in your calculation)

(1) (0.120, 0.166)

(2) (0.101, 0.185)

(3) (0.116, 0.170)

(4) (0.129, 0.157)

(5) (0.134, 0.152)

12. The data below are taken from a study by M.C. Gilly about sex roles in advertising which compared television advertisements in Australia, Mexico, and the United States. In this study, a random sample of TV commercials was cross-classified, as shown below. The following table shows the different settings that women in commercials are portrayed in.

Setting	Australia	United States	Mexico	Totals
Home	15	57	31	103
Retail outlet	3	15	19	37
Occupational	0	6	1	7
Outdoors	3	19	14	36
Other	31	72	55	158
Totals	52	169	120	341

The sample estimate of the difference between the proportion of **Mexican** and **Australian** women being portrayed **outdoors** in TV commercials, $\hat{p}_{MEX} - \hat{p}_{AUS}$, is:

(1) 11
(2) 0.306
(3) 0.021
(4) 0.059
(5) 0.03

Questions 13 and **14** refer to the following information.

A poll was taken where 662 voters were interviewed by telephone throughout New Zealand and asked whether *developing the economy* or *protecting the environment* would be more important in the short and in the longer term. Below are the numbers of people who gave each response for the short term.

Short term	National	Labour
Economy	136	80
Environment	66	55
Undecided	36	27

There were 238 National and 162 Labour voters in the poll.

Let p_N be the true proportion of National supporters and let p_L be the true proportion of Labour supporters who think that *protecting the environment* is more important in the *short* term.

13. We estimate the difference in proportions $p_N - p_L$ to be (select **one** only):

(1) –0.62
(2) 6.20
(3) 0.062
(4) –0.062
(5) 0.62

14. The 95% confidence interval for the difference between the proportions $p_N - p_L$ is `(-0.154673, 0.0302824)`. Which **one** of the following interpretations is **true**?

    (1) With a probability of 0.95, the true difference of proportions $p_N - p_L$ lies between –0.1547 and 0.0303.

    (2) In repeated sampling the 95% confidence interval [-0.1547, 0.0303] will contain the true difference in proportions in 95% of the samples taken.

    (3) In repeated sampling the true proportion p_N will be somewhere between 15 percentage points larger and 3 percentage points smaller than p_L.

    (4) With 95% confidence the true proportion p_N is somewhere between 15 percentage points smaller and 3 percentage points smaller larger than p_L.

    (5) With 95% confidence the true proportion p_N is 18.5 percentage points larger than p_L.

Questions 15 to **17** refer to the following information.

In a national youth health and wellbeing survey carried out in New Zealand, a wide range of questions were asked about topics including: Home, School, Injuries and Violence, Health and Emotional Health, Substance Abuse, and Spirituality. Assume that the survey sample is a random sample from all New Zealand secondary school students.

The information on family relationships in Table 4 below was obtained from a sample of 9314 secondary school students who responded to the question: "How do you feel about your family relationships?"

Family relationship	**Male**	**Female**	**Total**
I'm happy with how we get on	2741	2711	5452
My relationships are neither good nor bad	1269	1709	2978
Getting on with my family causes me problems	293	591	884
Total	4303	5011	9314

Table 4: Students' responses to a question on family relationships.

15. The proportion of students in this study who are female and chose the response "I'm happy with how we get on" is:

    (1) $\frac{2711}{4303}$

    (2) $\frac{5011}{9314} + \frac{5452}{9314}$

    (3) $\frac{2711}{5452}$

    (4) $\frac{2711}{5452} + \frac{2711}{5011}$

    (5) $\frac{2711}{9314}$

16. Suppose that a random sample of around 1000 students (instead of 9314) had been used to find the margin of error in the previous question. We would expect this new margin of error to be about:
    (1) three times the margin of error calculated from the 9314 students.
    (2) 4.5 times the margin of error calculated from the 9314 students.
    (3) one ninth the margin of error calculated from the 9314 students.
    (4) nine times the margin of error calculated from the 9314 students.
    (5) one third the margin of error calculated from the 9314 students.

Question 17 refers to the following additional information.

Let p_{MH} be the proportion of **males** who were happy with how they got on with their family and p_{FH} be the corresponding proportion of **females**.

The 95% confidence interval for the difference $p_{MH} - p_{FH}$ is (0.076, 0.116).

17. The **best** interpretation of this confidence interval is:

    With 95% confidence:

    (1) the true proportion of males who were happy with how they got on with their family was somewhere between 7.6 and 11.6 percentage points higher than the corresponding female proportion.
    (2) the true proportion of males who were happy with how they got on with their family was 7.6 percentage points and the corresponding female proportion was 11.6 percentage points.
    (3) the sample proportion of males who were happy with how they got on with their family was somewhere between 7.6 and 11.6 percentage points higher than the corresponding female proportion.
    (4) the sample proportion of males who were happy with how they got on with their family was 7.6 percentage points and the corresponding female proportion was 11.6 percentage points.
    (5) the true proportion of males who were happy with how they got on with their family was somewhere between 7.6 percentage points lower than and 11.6 percentage points higher than the corresponding female proportion.

Questions 18 to **20** refer to the following information:

The *Listener/Heylen* poll from August 6, 1994 reported the following results on what New Zealanders think about the "Ten Commandments" from a sample of 1,000 randomly chosen New Zealanders. For each of three commandments, the percentage of people agreeing that the commandment "fully applies to me" is given. Also reported were the results of a 1985 poll, which asked 1,000 New Zealanders the same questions.

- I am the Lord your God; worship no god but me. (*One God*)
- Do not commit murder. (*Do Not Murder*)
- Do not desire another person's goods. (*Do Not Envy*)

Percentage Agreeing

Year	*One God*	*Do Not Murder*	*Do Not Envy*
1985	39%	85%	53%
1994	32%	89%	62%

18. Consider the 1994 poll. Let p_M denote the proportion that think that *Do Not Murder* fully applies to them and let p_N denote the proportion that think that *One God* fully applies to them. The sample estimate of $\hat{p}_M - \hat{p}_N$ and its associated sampling situation are:

    (1) –0.57 and one sample of size 1000, many yes/no items.

    (2) –0.57 and one sample of size 1000, several response categories.

    (3) 0.57 and one sample of size 1000, many yes/no items.

    (4) 0.57 and one sample of size 1000, several response categories.

    (5) 0.57 and two independent samples, both of size 1000.

19. Let p_1 denote the proportion of New Zealanders that in 1994 thought that *Do Not Envy* fully applied to them. Let p_2 denote the proportion of New Zealanders that in 1985 thought that *Do Not Envy* fully applied to them.

    The standard error of the estimate, se($\hat{p}_1 - \hat{p}_2$), is approximately 0.0220 and the value of the *t*-multiplier for constructing a 95% confidence interval for $p_1 - p_2$ is approximately 1.960.

    The 95% confidence interval for $p_1 - p_2$ is given by:

    (1) (0.047, 0.133)

    (2) (0.035, 0.145)

    (3) (–0.126, –0.054)

    (4) (0.054, 0.126)

    (5) (–0.133, –0.047)

20. A 99% confidence interval for the proportion of New Zealanders who believe that *One God* fully applies to them, p_G, is given by (0.282, 0.358). Which one of the following statements is **true**?

    (1) The interval (0.282, 0.358) will cover the true, but unknown parameter p_G for 99% of samples taken.

    (2) Between 28.2 and 35.8 per cent of New Zealanders believe that *One God* fully applies to them 99% of the time.

    (3) A 95% confidence interval for p_G would be wider than this interval.

    (4) The probability that the interval (0.282, 0.358) covers the sample proportion is 0.99.

    (5) The probability that another interval calculated in the same way from a new sample of 1000 New Zealanders covers p_G is 0.99.

Questions 21 and **22** refer to the following information.

Death Penalty Survey Results

"Should convicted murderers be put to death?"		
	Australia	N.Z.
Yes	46%	42%
No	39%	41%
Can't Say	15%	17%

"Should an Australian or New Zealander convicted of drug trafficking in Malaysia or Sri Lanka be put to death?"		
	Australia	N.Z.
Yes	76%	64%
No	19%	26%
Can't Say	5%	10%

[Polls of 1307 Australians & 1010 New Zealanders]

21. We are interested in the difference between the proportion of Australians supporting the death penalty for N.Z. and Australian drug traffickers in South East Asia and the proportion of Australians supporting the death penalty for convicted murderers. This sampling situation is:

    (1) two independent samples of size 1307 and 1010 respectively
    (2) two independent samples, each of size 1307
    (3) one sample of size 1010, many yes/no items
    (4) one sample of size 1307, many yes/no items
    (5) one sample of size 1307, several response categories

22. The sampling situation associated with the estimate of the difference between *the proportion of New Zealanders who said 'Yes'* and *the proportion of New Zealanders who said 'No'* to supporting the death penalty for N.Z. and Australian drug traffickers in South East Asia is:

    (1) one sample of size 1010, many yes/no items
    (2) two independent samples of size 1307 and 1010 respectively
    (3) one sample of size 1307, several response categories
    (4) one sample of size 1010, several response categories
    (5) one sample of size 1307, many yes/no items

Questions 23 and **24** refer to the following information.

A survey of 2171 men and 2412 women in Auckland found that 10% of men abstained from drinking alcohol compared with 16% of women. We wish to compare the proportion of female abstainers, p_{female}, with the proportion of male abstainers, p_{male}.

23. The sampling situation is **best** described as:

    (1) two independent samples.
    (2) one sample, several response categories.
    (3) one sample, many yes/no items.
    (4) two samples, several response categories.
    (5) two samples, many yes/no items.

24. Based on the data, a 95% confidence interval for $p_{\text{female}} - p_{\text{male}}$ is (0.041, 0.079). Which **one** of the following statements is **false**?

    (1) Based on the data, a 99% confidence interval would be wider than 0.038.
    (2) The point estimate of $p_{\text{female}} - p_{\text{male}}$ is 0.06.
    (3) We are confident that the proportion of female abstainers is larger than the proportion of male abstainers.
    (4) Zero is a plausible value for $p_{\text{female}} - p_{\text{male}}$.
    (5) Based on the data, a 95% confidence interval for $p_{\text{male}} - p_{\text{female}}$ is $(-0.079, -0.041)$.

Questions 25 to **28** refer to the following information:

The *Listener*, 16 July 1994, reported the results of a survey carried out on a random sample of 1000 New Zealand residents who were older than 15 years. 55% did not want marijuana to be made legal; 29% thought it should be made legal and 16% had no firm view. In a similar poll in 1985, 66% did not think that the drug should be legalised whereas 20% thought that it should be legalised and 14% had no firm view.

25. A sample estimate of the difference, $\hat{p}_{1994} - \hat{p}_{1985}$, between the 1994 and 1985 proportions of New Zealand residents who thought that marijuana should be made legal is (choose **one**):

    (1) 0.15
    (2) 0.11
    (3) 0.09
    (4) −0.11
    (5) −0.09

26. The sampling situation is **best** described as:

    (1) two independent samples.
    (2) one sample, several response categories.
    (3) one sample, many yes/no items.
    (4) two samples, several response categories.
    (5) two samples, many yes/no items.

27. A 95% confidence interval for $p_{1994} - p_{1985}$ is `(0.0525088, 0.127491)`. Which **one** of the following statements can be made with 95% confidence?

    (1) p_{1994} may be bigger than p_{1985} by at least 0.053 and at most 0.127.
    (2) p_{1994} may be smaller than p_{1985} by up to 0.053 or bigger by up to 0.127.
    (3) p_{1994} may be bigger than p_{1985} by up to 0.053 or smaller by up to 0.127.
    (4) p_{1994} may be smaller than p_{1985} by at least 0.053 and at most 0.127.
    (5) None of these statements is true because we cannot tell which proportion is larger from an interval for a difference.

28. The sampling situation when comparing the proportion in 1994 who wanted marijuana legalised with the proportion in 1994 who had no firm view, is **best** described, **for the purpose of determining the correct standard error formulae to be used**, as (give **one** answer only):

    (1) Situation (b): Single sample, several response categories.
    (2) Situation (a): Two independent samples.
    (3) A single sample with a single proportion since 0.29 + 0.14 < 0.55.
    (4) Situation (c): Single sample, two or more Yes/No items.
    (5) A single sample cross-classified by two factors.

29. A genetic researcher is interested in the hair colour of children with blue eyes. She uses some data collected on children with blue eyes in Caithness, and with blue eyes from Aberdeen, Scotland. The data are given in the table below.

			Hair			
City	Fair	Red	Medium	Dark	Black	**Totals**
Caithness	326	38	241	110	3	718
Aberdeen	1368	170	1041	398	1	2978
Totals	1694	208	1282	508	4	3696

A table of counts for the Caithness and Aberdeen data.

Let $\hat{p}_C$ be the sample proportion of blue-eyed children with fair hair in Caithness, and $\hat{p}_A$ be the sample proportion of blue-eyed children with fair hair in Aberdeen. The sampling situation when comparing the difference between these proportions, is:

(1) two independent samples, both of size 3696

(2) one sample of size 3696, many yes/no items

(3) one sample of size 3696, several response categories

(4) two independent samples of size 718 and 2978 respectively

(5) one sample of size 718, several response categories

Question 30 and **31** refer to the following information.
The table below represents the beer market for the New Zealand brewing companies DB group and Lion Nathan Ltd taken from a random sample of 2400 beer drinkers.

Estimated Brand Market Shares (%)

Lion Nathan		**DB Group**	
Lion Red	25.9	DB Draught	12.1
Lion Brown	6.3	DB Bitter	4.5
Speight's	6.2	DB Export	3.2
Australian brands	6.2	DB Export Dry	1.9
Steinlager (green)	4.9	Double Brown	1.8
Steinlager (blue)	4.1	Imported Beer	1.5
Waikato	4.1	Tui	1.2
Canterbury Draught	4.0		
TOTAL	**61.7**	**TOTAL**	**26.2**

Note: Home brew, independent brands and other imported brands make up the remaining 12.1 per cent of the market. Each drinker was classified according to the brand of beer they most commonly consumed.

A 99% confidence interval for p, the proportion of beer drinkers who drink Lion Red (as a proportion of the total market share) is (0.236, 0.282).

30. Which of the following statements is **true**?

(1) 99% of all beer drinkers have between a 23.6% and 28.2% chance of drinking Lion Red.

(2) The probability that p is between 0.236 and 0.282 is 0.99.

(3) A 95% confidence interval for p will be wider than this interval.

(4) 99% of all such samples would give an interval that contains the true proportion p.

(5) The interval (0.236, 0.282) will cover the true, but unknown parameter p for 99% of samples taken.

31. Let p_G denote the proportion of people who drink Steinlager Green, and let p_B denote the proportion of people who drink Steinlager Blue. Thus an estimate for the difference in these two proportions, $p_G - p_B$, is given by $\hat{p}_G - \hat{p}_B$. The sampling situation for $\hat{p}_G - \hat{p}_B$ is:

(1) two independent samples, each of size 2400
(2) one sample of size 2400, many yes/no items
(3) two samples of size 2400, many yes/no items
(4) two samples of size 2400, several response categories
(5) one sample of size 2400, several response categories

32. The general formula for a confidence interval for the difference between two proportions is:

$$\hat{p}_1 - \hat{p}_2 \pm t \times se(\hat{p}_1 - \hat{p}_2)$$

Which **one** of the following statements about confidence intervals for the difference between two proportions is **false**?

(1) The value of the t-multiplier depends on the confidence level.

(2) The confidence interval method for proportions works only if the sample size is sufficiently large.

(3) The confidence interval is centred on $\hat{p}_1 - \hat{p}_2$.

(4) The value of the t-multiplier depends on the sample size.

(5) The size of the standard error depends on the sampling situation.

Question 33 refers to the following information.

Myers *et al.* (2010) conducted a study with 270 customers of a Boston (USA) coffee shop.

Some variables used in the study were:

Gender	The gender of the customer – Female – Male
Coffee type	The type of coffee ordered – Fancy – Not fancy
Waiting time	The time between ordering and receiving coffee, in seconds

Assume the study involves a random sample from the population of all customers of this Boston coffee shop.

Let:

p_{FF} be the proportion of all female customers of this Boston coffee shop who order fancy coffees

and

p_{MF} be the proportion of all male customers of this Boston coffee shop who order fancy coffees.

33. A 95% confidence interval for $p_{FF} - p_{MF}$ is given by (0.089, 0.311).

Which **one** of the following statements is **false**?

(1) Sample size is not taken into account when calculating the standard error of this estimate, se($\hat{p}_{FF} - \hat{p}_{MF}$).

(2) The point estimate of the difference between the proportion of female customers who order fancy coffees and the proportion for male customers is 0.2.

(3) The sampling situation for calculating the standard error of this estimate, se($\hat{p}_{FF} - \hat{p}_{MF}$), is two independent samples.

(4) The margin of error of the confidence interval is 0.111.

(5) We can not be sure that the true value for $p_{FF} - p_{MF}$ is in this confidence interval.

Question 34 refers to the following information.

In 2013, Auckland Transport conducted a survey on cycling in which a random sample of 1048 adult Aucklanders, aged 15 years or over, were surveyed (AT, 2013). Respondents were asked questions related to their usage and attitudes towards cycling through an online survey. One of the questions asked in the survey was:

> *In general, how confident are you/would you be in riding a bicycle in the Auckland area?*

34. Let p_{NC} be the proportion of adult Aucklanders who do not feel confident riding a bicycle in the Auckland area. A 95% confidence interval for p_{NC} is given by (0.63, 0.69).

    Which **one** of the following statements is **false**?

    (1) A claim that the majority of adult Aucklanders do not feel confident riding a bicycle in the Auckland area is believable.

    (2) With 95% confidence we estimate that somewhere between 63% and 69% of adult Aucklanders do not feel confident riding a bicycle in the Auckland area.

    (3) The proportion of adult Aucklanders who do not feel confident riding a bicycle in the Auckland area varies between 63% and 69%.

    (4) It is plausible that 68% of adult Aucklanders do not feel confident riding a bicycle in the Auckland area.

    (5) Approximately 66% of the respondents said they did not feel confident riding a bicycle in the Auckland area.

ANSWERS

1. **(3)**	2. **(3)**	3. **(3)**	4. **(1)**	5. **(3)**	6. **(4)**
7. **(3)**	8. **(1)**	9. **(4)**	10. **(2)**	11. **(3)**	12. **(4)**
13. **(4)**	14. **(4)**	15. **(5)**	16. **(1)**	17. **(1)**	18. **(3)**
19. **(1)**	20. **(5)**	21. **(4)**	22. **(4)**	23. **(1)**	24. **(4)**
25. **(3)**	26. **(1)**	27. **(1)**	28. **(1)**	29. **(4)**	30. **(4)**
31. **(5)**	32. **(4)**	33. **(1)**	34. **(3)**		

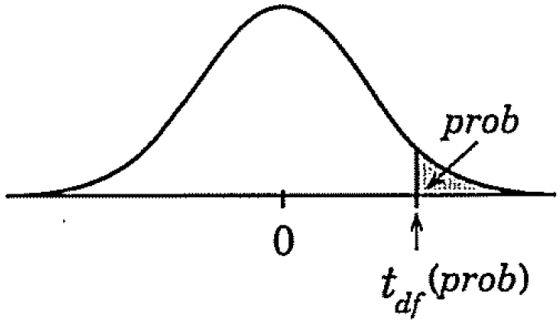
WHAT SHOULD I DO NEXT?

- **Go through the Chapter 6 blue pages. The blue pages relevant to the material in this workshop are the notes on pages 15 to 18, the glossary on page 20, the true/false statements on page 21 (except e. and m.) and the tutorial material on pages 23-25.**
- **Try Chapter 6 questions from three of the past five tests that are relevant to this workshop.**

Student's t–distribution

For fixed $prob$ and df, the tabulated value is the number $(t = t_{df}(prob))$ such that for, $T \sim \text{Student}(df)$,

$pr(T \geq t) = prob.$



[e.g. For $prob = 0.025$ and $df = 23, t_{23}(0.025) = 2.069$]

					prob					
df	.20	.15	.10	.05	.025	.01	.005	.001	.0005	.0001
3	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.21	12.92	22.20
4	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610	13.03
5	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869	9.678
6	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959	8.025
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7.063
8	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041	6.442
9	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781	6.010
10	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587	5.694
11	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437	5.453
12	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318	5.263
13	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221	5.111
14	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140	4.985
15	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073	4.880
16	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015	4.791
17	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965	4.714
18	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922	4.648
19	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883	4.590
20	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.849	4.539
21	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819	4.493
22	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792	4.452
23	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768	4.415
24	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745	4.382
25	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725	4.352
26	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707	4.324
27	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690	4.299
28	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674	4.275
29	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659	4.254
30	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646	4.234
31	0.853	1.054	1.309	1.696	2.040	2.453	2.744	3.375	3.633	4.215
32	0.853	1.054	1.309	1.694	2.037	2.449	2.738	3.365	3.622	4.198
33	0.853	1.053	1.308	1.692	2.035	2.445	2.733	3.356	3.611	4.182
34	0.852	1.052	1.307	1.691	2.032	2.441	2.728	3.348	3.601	4.167
35	0.852	1.052	1.306	1.690	2.030	2.438	2.724	3.340	3.591	4.153
36	0.852	1.052	1.306	1.688	2.028	2.434	2.719	3.333	3.582	4.140
37	0.851	1.051	1.305	1.687	2.026	2.431	2.715	3.326	3.574	4.127
38	0.851	1.051	1.304	1.686	2.024	2.429	2.712	3.319	3.566	4.116
39	0.851	1.050	1.304	1.685	2.023	2.426	2.708	3.313	3.558	4.105
40	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551	4.094
45	0.850	1.049	1.301	1.679	2.014	2.412	2.690	3.281	3.520	4.049
50	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.261	3.496	4.014
60	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460	3.962
80	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416	3.899
100	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390	3.861
∞	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.719

[**Note:** If $df \geq 150$ use $df = \infty$]

FORMULAE

Confidence intervals and t-tests

Confidence interval: $estimate \pm t \times \text{se}(estimate)$

t-test statistic: $t_0 = \dfrac{estimate - hypothesised\ value}{standard\ error}$

Applications:

1. Single mean $\boldsymbol{\mu}$: $estimate = \overline{x}$; $df = n - 1$
2. Single proportion $\boldsymbol{p}$: $estimate = \widehat{p}$; $df = \infty$
3. Difference between two means $\boldsymbol{\mu_1 - \mu_2}$: (independent samples)
   $estimate = \overline{x}_1 - \overline{x}_2$; $df = \min(n_1 - 1, n_2 - 1)$
4. Difference between two proportions $\boldsymbol{p_1 - p_2}$:
   $estimate = \widehat{p}_1 - \widehat{p}_2$; $df = \infty$
   Situation (a): *Proportions from two independent samples*
   Situation (b): *One sample of size n, several response categories*
   Situation (c): *One sample of size n, many yes/no items*

The F-test (ANOVA)

F-test statistic: $f_0 = \dfrac{s_B^2}{s_W^2}$; $df_1 = k - 1$, $df_2 = n_{\text{tot}} - k$

The Chi-square test

Chi-square test statistic: $\chi_0^2 = \displaystyle\sum_{\text{all cells in the table}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

Expected count in cell $(i, j) = \dfrac{R_i C_j}{n}$

$df = (I - 1)(J - 1)$

Regression

Fitted least-squares regression line: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

Inference about the intercept, β_0, and the slope, β_1: $df = n - 2$