please get a handout from the back!

Stats 101/101G/108 Workshop

Confidence Intervals: *Proportions* [CIP]

we will start at 9.35am...

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~~ 3035.288 (~~third~~ second 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**)

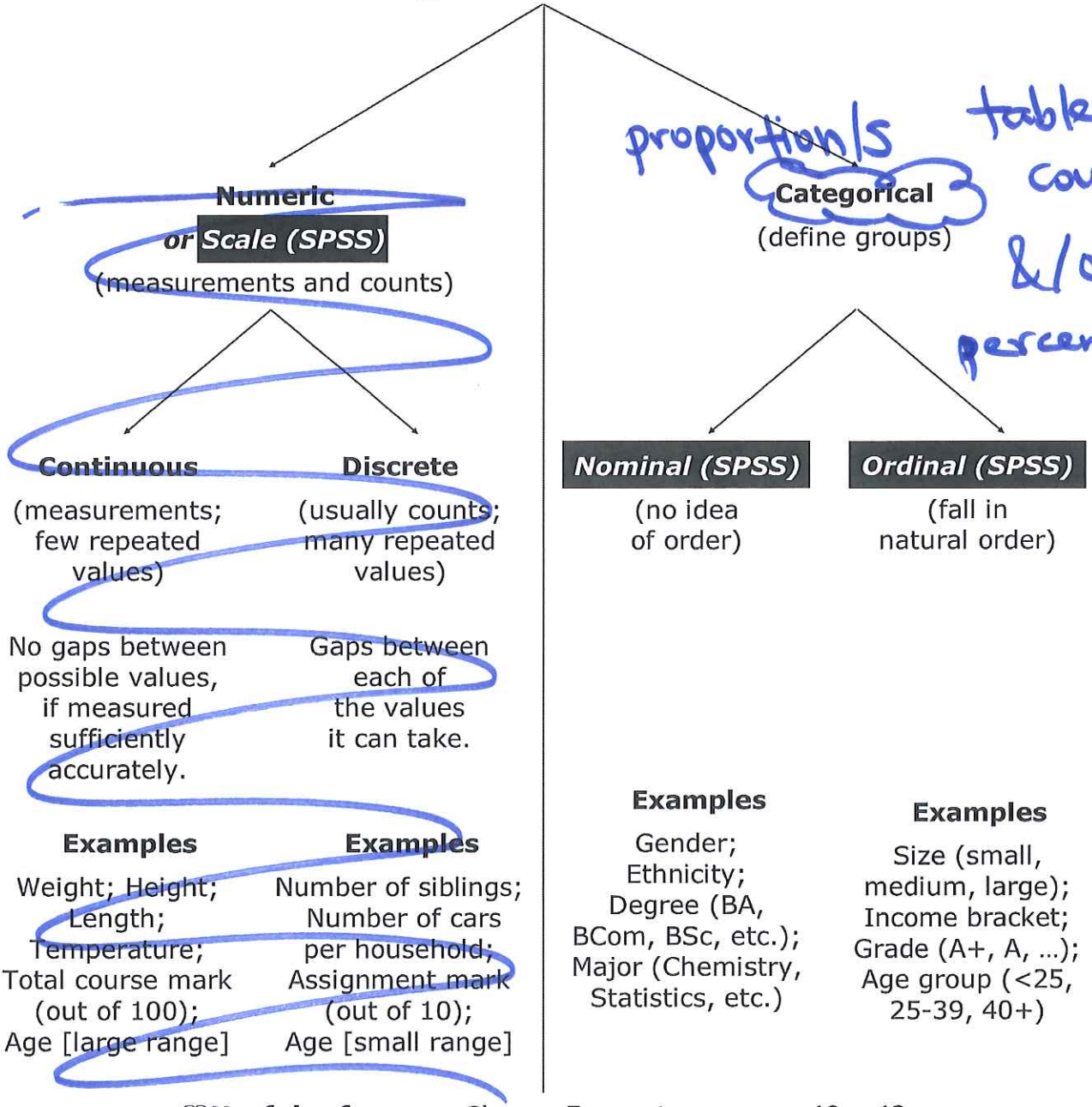
How many decimal places should I use?

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:

Types of Variables



📖**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

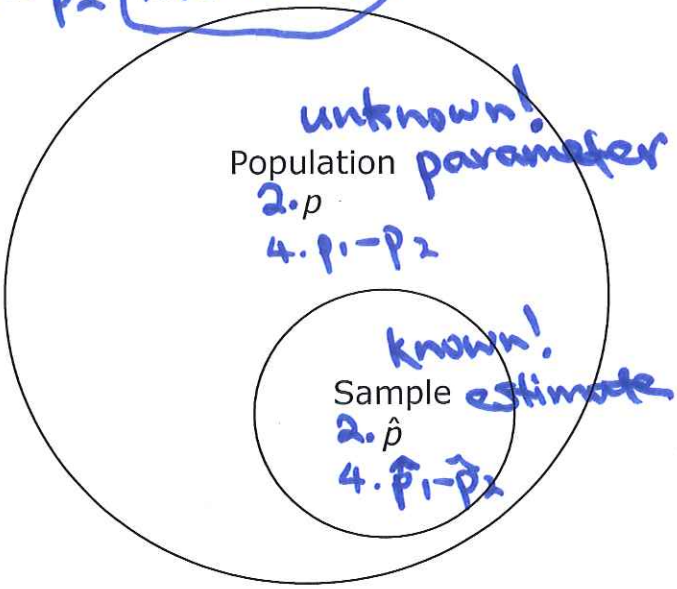
- This means we are unable to establish the **parameter** of interest within our population, such as:
  - ✓ Population proportion, p

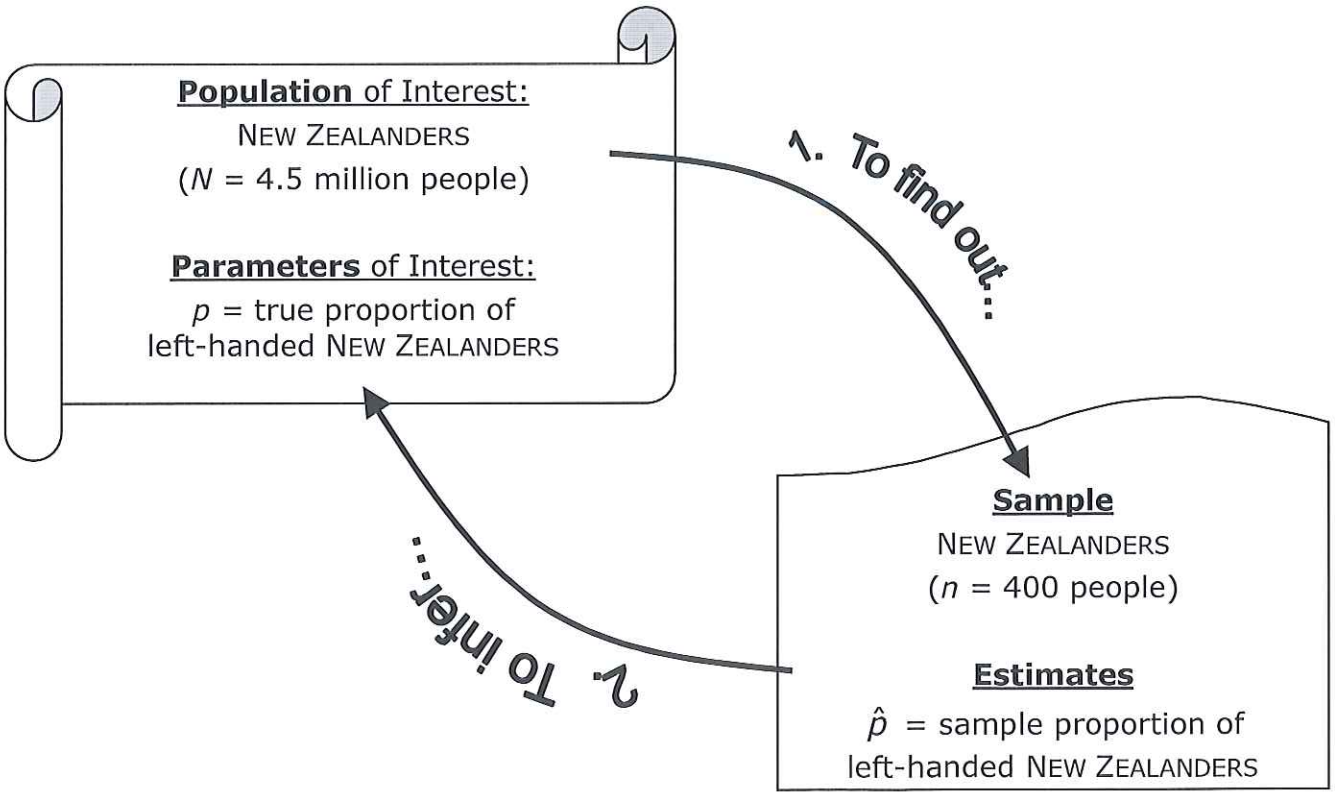
- This means that the **parameter** of interest is an ***unknown numerical characteristic*** for that particular population.

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

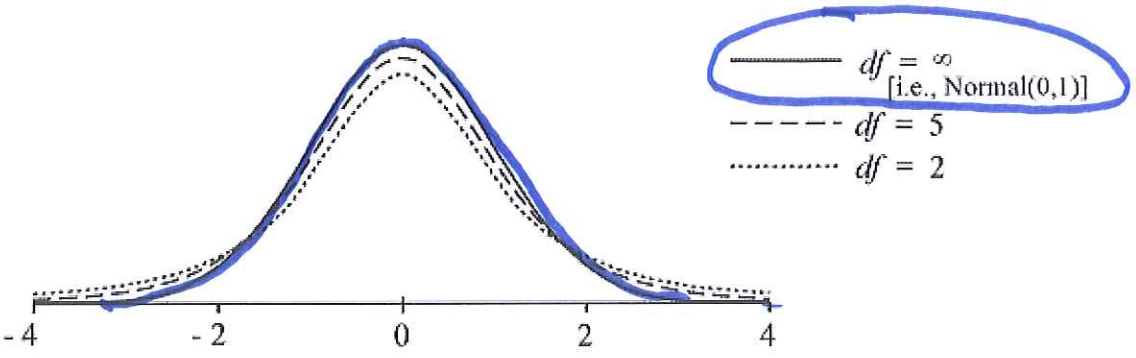
- The process of using sample data to make useful statements about a population is a type of **statistical inference** called **sample-to-population inference**, e.g., when using sample data to estimate an unknown population proportion.





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.

→ Ch 6 (norm-based) CIs

→ skewed, odd (not symmetric)

(Ch 6)

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%... 2p 4. p1-p2
- Increasing the confidence level will **increase** the width of the interval.
- Increasing the sample size (n) will make the confidence interval more precise.

narrower

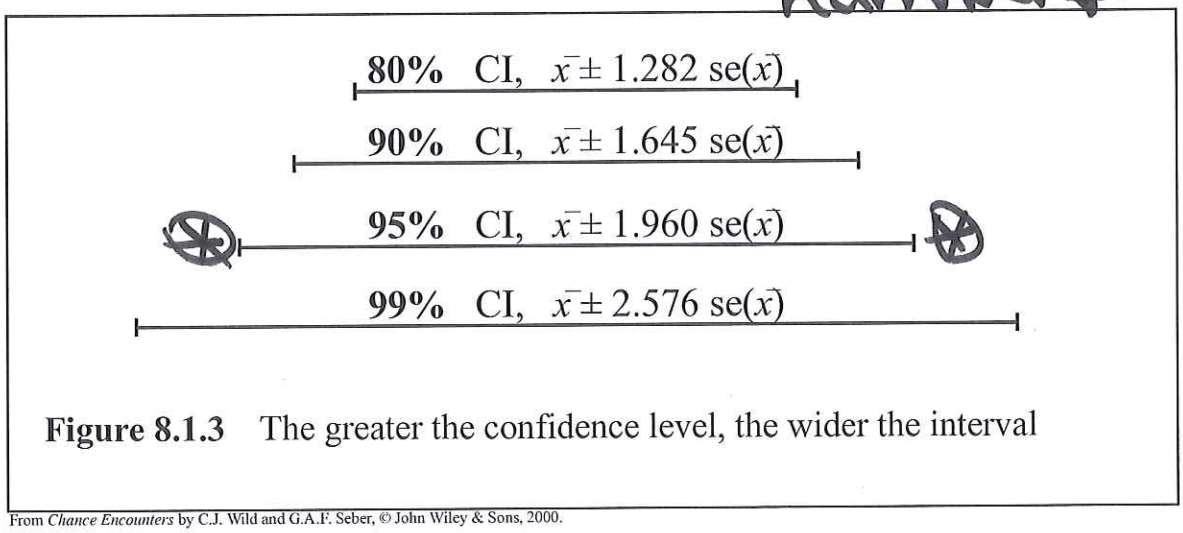


Figure 8.1.3 The greater the confidence level, the wider the interval

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

- To double the accuracy of the confidence interval we **need 4 times** as many observations. → halve the width
- To triple the accuracy of the confidence interval we **need 9 times** as many observations. → third the width
- 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.

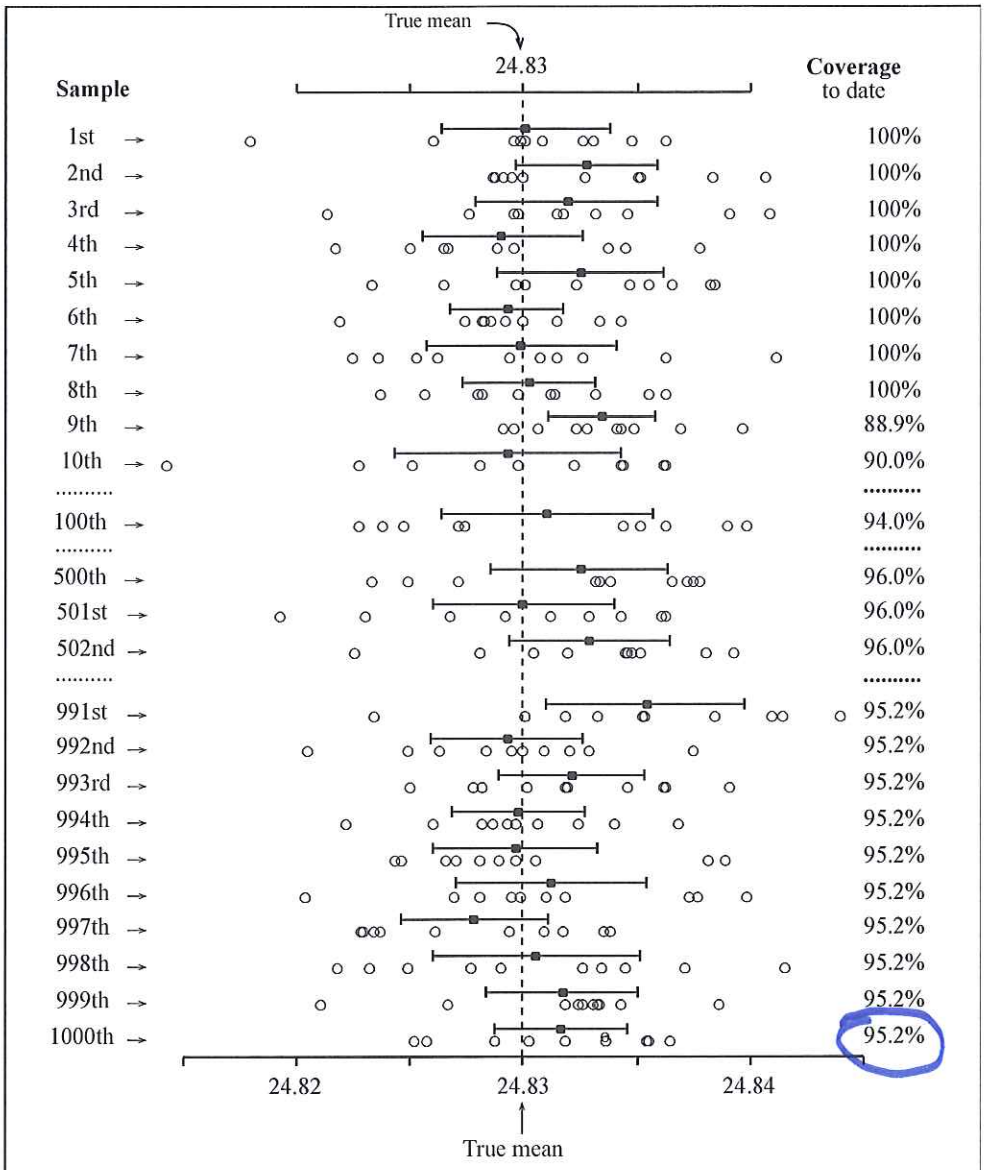


Figure 8.1.2 Samples of size 10 from a Normal(μ=24.83, s=.005) distribution and their 95% confidence intervals for μ..

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

restart on pg 9 @ 11.05am! – have a break – try Q25, Q26, Q28, Q29, Q31

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
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.

margin of error

***estimate* ± *t* × se(*estimate*)**

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

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 95%
- 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:

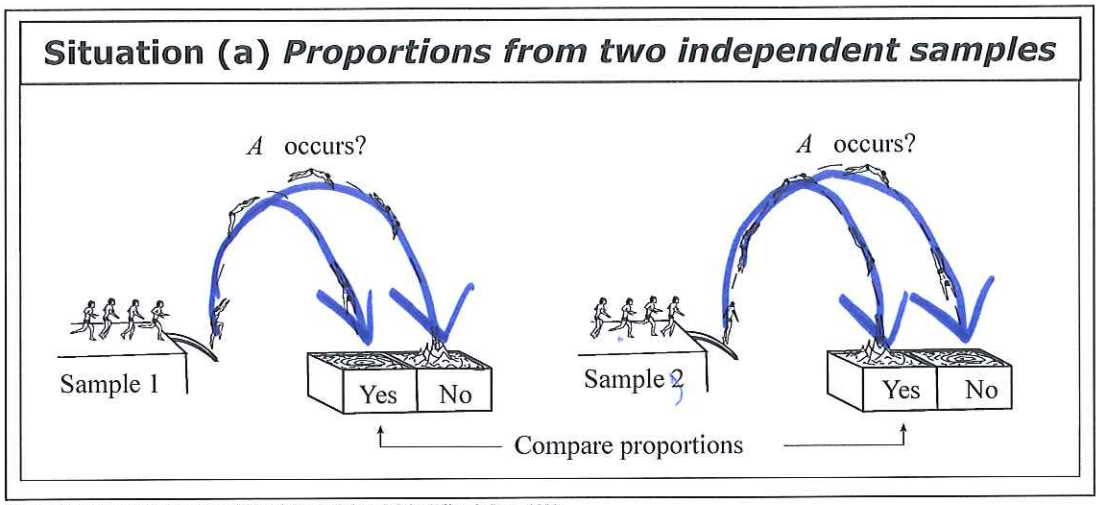
Examples:

✓ If the CI contains 0 (i.e. one negative and one positive number), there may be no difference between the two proportions. (-.07, .03) $p_1 \leq p_2$

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

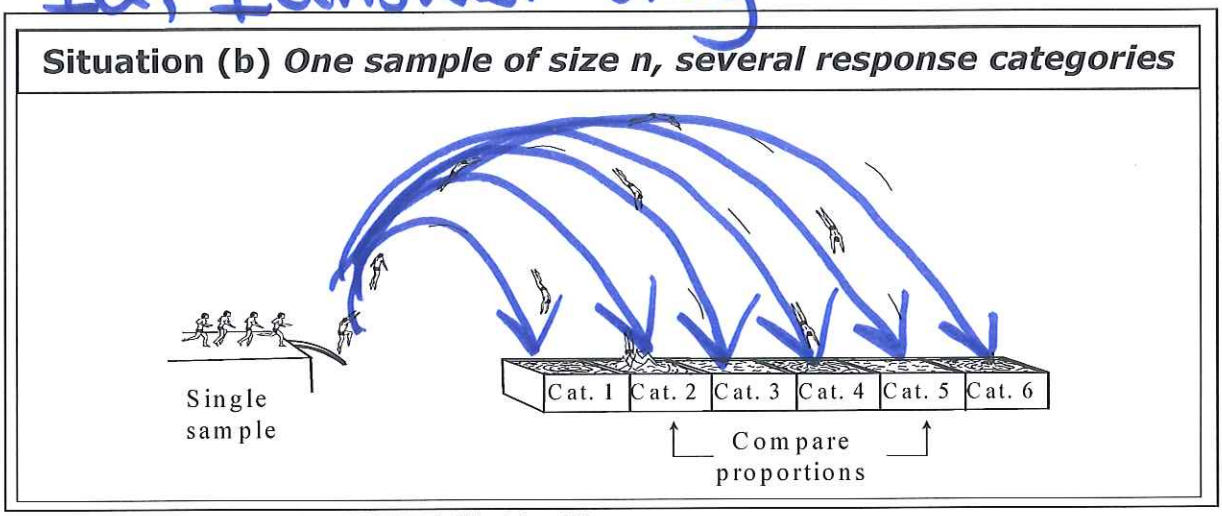
✓ If CI is negative, then p_1 is lower/smaller than p_2. $p_1 < p_2$ (-.07, -.03)

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



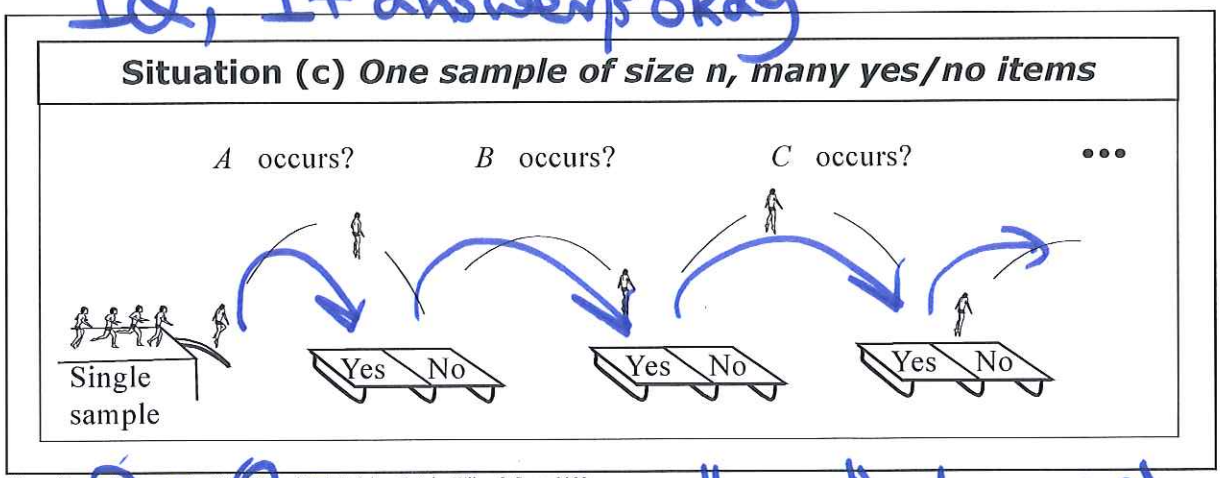
From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 1999.

1Q, 1 answer only



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

1Q, 1+ answers okay



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

2+ Qs, can say "yes" to each, if so wish.