STATS 101/108 WORKSHOP

TEST PREP 3: CHAPTERS 7 AND 8

FRIDAY 10 SEPTEMBER, 2010

Students **MUST REGISTER** for all workshops with
The Student Learning Centre, 3rd Floor, Information Commons

Student Learning Centre

Topics we teach and can provide advice on include:

- ✓ Essay writing
- ✓ Computer skills
- ✓ Reading and notetaking
- ✓ Memory and concentration
- ✓ Report writing
- ✓ Test and examination skills
- ✓ Thesis and dissertation writing
- ✓ Tutorial skills
- ✓ Research skills
- ✓ Time and stress management
- ✓ Mathematics
- ✓ **Statistics**
- ✓ Oral presentation and seminar skills
- ✓ Language learning
- ✓ Specific learning disabilities
- ✓ Motivation and goal setting
- ✓ Survival skills (in the University system)

Programmes within SLC include:

- Te Puni Wananga
  Maori university tutors committed to enhancing Maori students' success

- Fale Pasifika
  Pacific Island tutors committed to enhancing success for Pacific Island students

- Students with impairments
  Learning assessments are available for students with specific learning disabilities; academic assistance is available for these students and those with mental health impairments.
  If you have any special learning requirements, please feel free to discuss this with Leila in person or via email.

- Academic English Conversation Groups
  Improve your academic English; develop communication skills including critical/creative thinking and clear expression of ideas and opinions. Weekly class held at the SLC on Thursdays, 3-5pm (during semester)

Statistical help available at the SLC

The Student Learning Centre (SLC) offers help for STATS 101/108 by offering:

- one-on-one tutoring help, and
- a number of workshops

One-on-one help

The SLC employs tutors specifically to help students with one-on-one assistance for STATS 101/108. One-on-one tutoring must be booked at SLC reception on the third floor of the Information Commons in person or by calling 373-7599 X 88850. Enquire at the SLC reception for available times.

Note: SLC tutors are not allowed to help students complete their assignments.

SLC STATS 101/108 Workshops

Any questions regarding STATS 101/108 workshops should be forwarded to:

Leila Boyle
SLC Statistics Co-ordinator
l.boyle@auckland.ac.nz

Workshops are run in a relaxed environment, typically set at a pace for those students that find the Statistics Department's tutorials too fast. All workshops allow plenty of time for questions. In fact, this is encouraged ☺

1) Saturday Workshops

These five 3-hour workshops are held on Saturdays throughout the semester to help students with different sections of the course.

2) Computer Workshops: Excel / PASW (SPSS)

These three computer-based workshops introduce students to the skills needed for Excel and PASW (SPSS) use in STATS 101/108 assignments.

3) Pre-test Workshops

These three workshops will cover the basics that you need for the test.

4) Pre-exam Workshops

These six workshops will cover the basics that you need for the exam.

Note: All workshops concentrate on questions reviewing the basic concepts, rather than questions on finer details. They are designed to assist students to achieve a pass; they are not designed to cover all material.

The timetable for these workshops is available with this handout. Currently the SLC website is still partly down so online enrolments are not available until further notice. Please enrol in each of your preferred classes at the Student Learning Centre by:

- **Going to the SLC in person**
- **Emailing slc@auckland.ac.nz with your name, ID number and the workshop/s you wish to attend.**
- **Calling the SLC reception on 373-7599 ext. 88850 and enrol over the phone. Make sure you know which workshop/s you want to enrol in and have your ID number handy.**

Useful Websites

- SLC webpage: **www.slc.auckland.ac.nz** (The SLC website currently has all functionality except online enrolment! Download an undergraduate brochure and enrol in workshops in person or by emailing/phoning the SLC Reception as per above instructions).
- Cecil: **https://cecil.auckland.ac.nz/**
- **Leila's website for STATS 101/108 SLC workshop handouts & information: www.stat.auckland.ac.nz/~leila**

Revision Notes

Chapter 7 – Sampling Distributions of Estimates

Look at blue pages for good notes, 26 statements (2 of which are false!) and test/exam questions for practice

- 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 **parameters** of interest within our population, such as:
  - ✓ Population mean, μ
  - ✓ Population standard deviation, σ

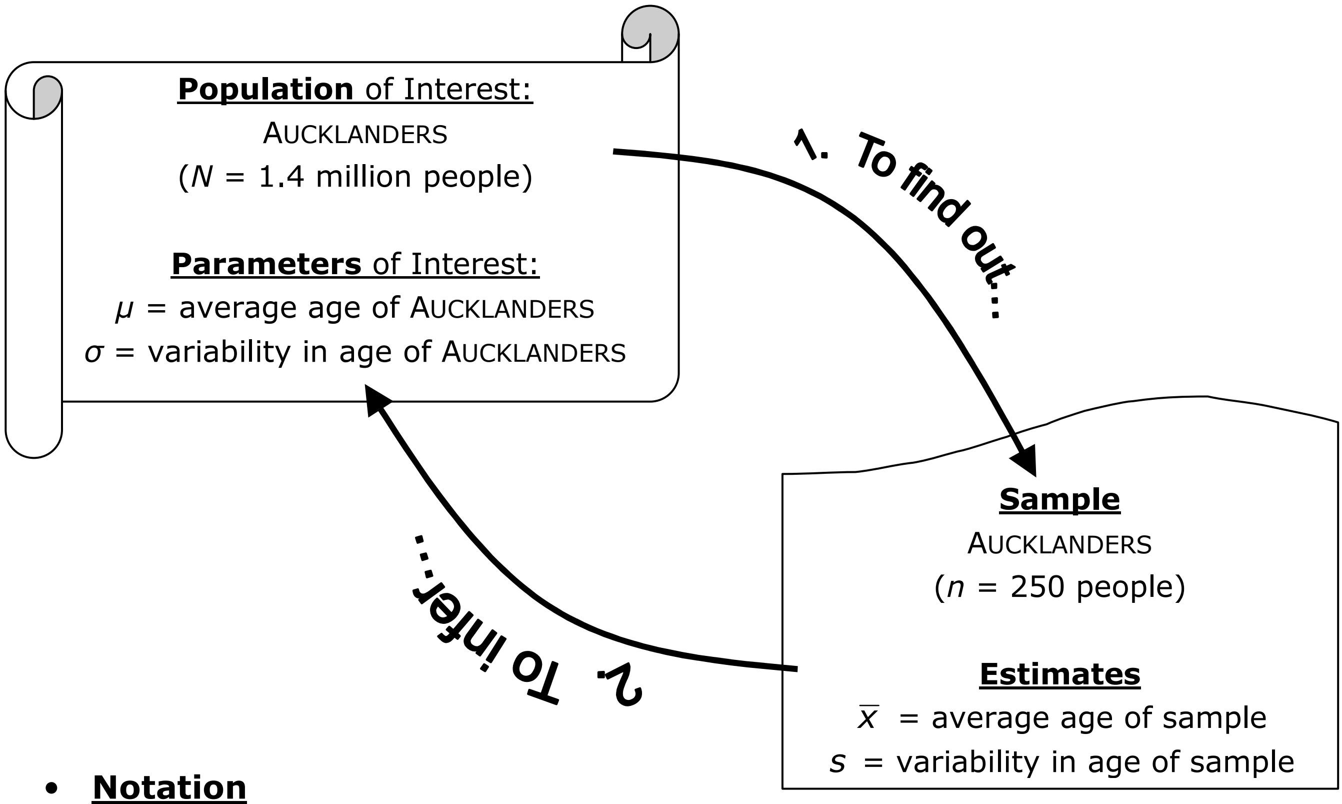
- This means that the **parameter** of interest is an ***<u>unknown</u> numerical characteristic*** for that particular population.

- To estimate an ***<u>unknown</u> 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 estimates** of the above **population parameters** are:
  - ✓ Sample mean, $\bar{x}$
  - ✓ Sample standard deviation, $sd(\bar{x})$ or σ_{n-1} or s

- Usually ^HATS or ‾BARS are used to distinguish between **sample estimates** and **population parameters**.

- Random variables X_1, X_2, … X_n, form a random sample from a distribution if:
  - ✓ they all have the same distribution; and
  - ✓ they are independent of one another.

- The big question which we will answer in Chapter 7 is "But how can we trust the sample estimates ($\bar{x}$ and s) from our single sample of size n?"



- **Notation**

  In statistics we use **CAPITAL letters** to refer to the **variable of interest** for the population and **small letters** to specify the **actual "number" observed** for that variable in our particular sample.

Variable of Interest **CAPITAL** LETTER	Actual "number" **small** LETTER
X	x
$\bar{X}$	$\bar{x}$

- **The sample mean – $\bar{X}$**

  If X_1, X_2, … X_n, form a random sample from a distribution where $E(X_i) = \mu$ and $sd(X_i) = \sigma$, then,

  ✓ The **expected value** of the sample mean, $E(\bar{X})$ is calculated by:

  $$\mu_{\bar{X}} = E(\bar{X}) = \mu$$

  ✓ The **standard deviation** of the sample mean, $sd(\bar{X})$ is calculated by:

  $$\sigma_{\bar{X}} = sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

- **<u>The Central Limit Theorem</u>**

  ✓ The Central Limit Theorem tells us that the larger the sample size, the closer the distribution of $\overline{X}$ comes to a Normal distribution. Even if the distribution of X is non-Normal, the distribution of $\overline{X}$ will be **approximately** Normal for a sufficiently large sample size n.

  ✓ If X is from a "well-behaved" distribution (i.e. symmetric, no outliers) the Central Limit Theorem works reasonably fast. n = 10 may be sufficient!

  ✓ In **general**, n = 30 works well for most distributions *except* distributions that are severely skewed or have large outliers.

  ✓ If the distribution is **severely skewed**, n = 50 should be sufficient.

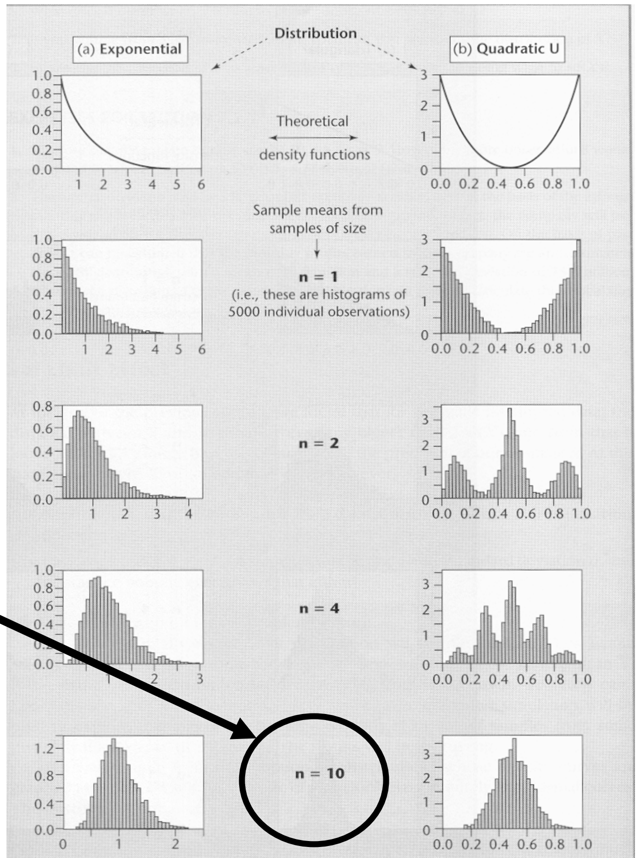


FIGURE 7.2.4 Histograms of means calculated from random samples from the Exponential and Quadratic U distributions.

Chance Encounters, C.J. Wild & G.A.F. Seber, p286

  ✓ If X is from a Normal distribution, then $\overline{X}$ is **exactly** Normally distributed.

- $\overline{X}$ is an **unbiased** estimator of μ because $E(\overline{X}) = \mu$.

- **<u>Standard errors</u>**

  However, as $sd(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$, it is not a useful measure of the precision of $\overline{x}$, because we do not know the value of σ. Therefore, we have to use the ***standard error*** of the sample mean to estimate the precision of $\overline{x}$ as an estimate of μ:

$$\text{The standard error of } \overline{x} = \frac{\text{sample standard deviation}}{\sqrt{\text{sample size}}} = \text{se}(\overline{x}) = \frac{s}{\sqrt{n}}$$

- **Student's *t*-distribution**

  ✓ Parameter: Degrees of Freedom (*df*).

  ✓ Bell shaped and centred at 0 like the (Standard) Normal (0,1) distribution but it's more variable.

  ✓ 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.

  ✓ The random sample from a Normal distribution : $T = \frac{\overline{X} - \mu}{se(\overline{X})}$ is exactly distributed as Student($df = n - 1$)

  ✓ Methods based on this distribution works very well even for small samples that are from very non-Normal distributions.

  ✓ By t_{df} (*prob*), we mean the number *t* such that when $T \sim$ Student(*df*), $\text{pr}(T \geq t) = prob$; that is, the tail area above *t* (that is to the right of *t* on the graph) is *prob*:

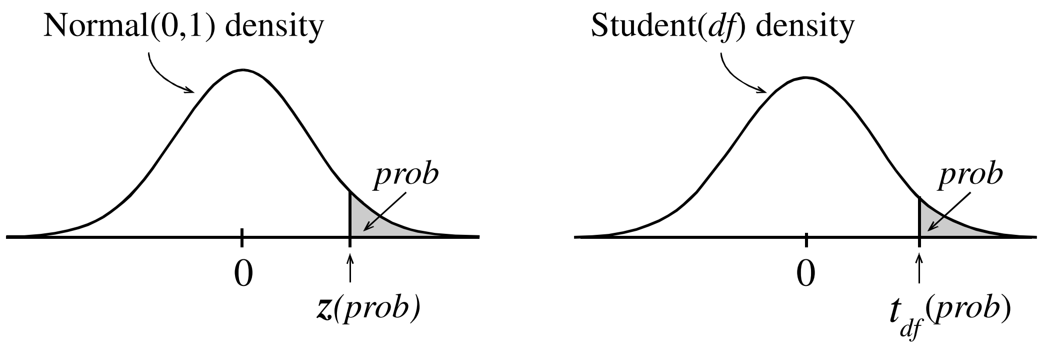


Figure 7.6.2 The *z(prob)* and *t(prob)* notations.

Chapter 8 – Confidence Intervals

Look at blue pages for good notes, 26 statements (2 of which are false!) and test/exam questions for practice

- 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.

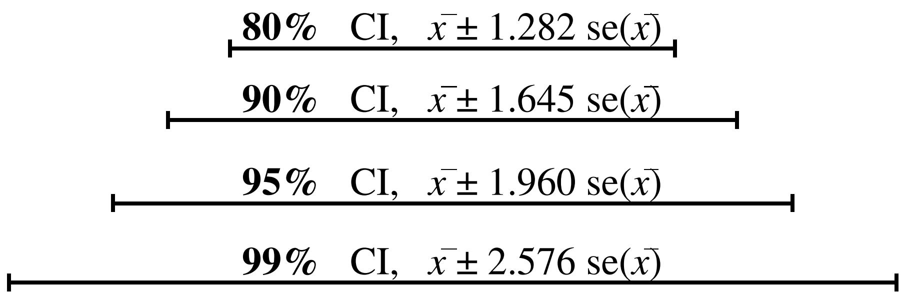


Figure 8.1.3 The greater the confidence level, the wider the interval

- You can do all kind of CI's, 90%, 95%, 99%...
- Increasing the confidence level will **increase** the width of the interval.
- 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.

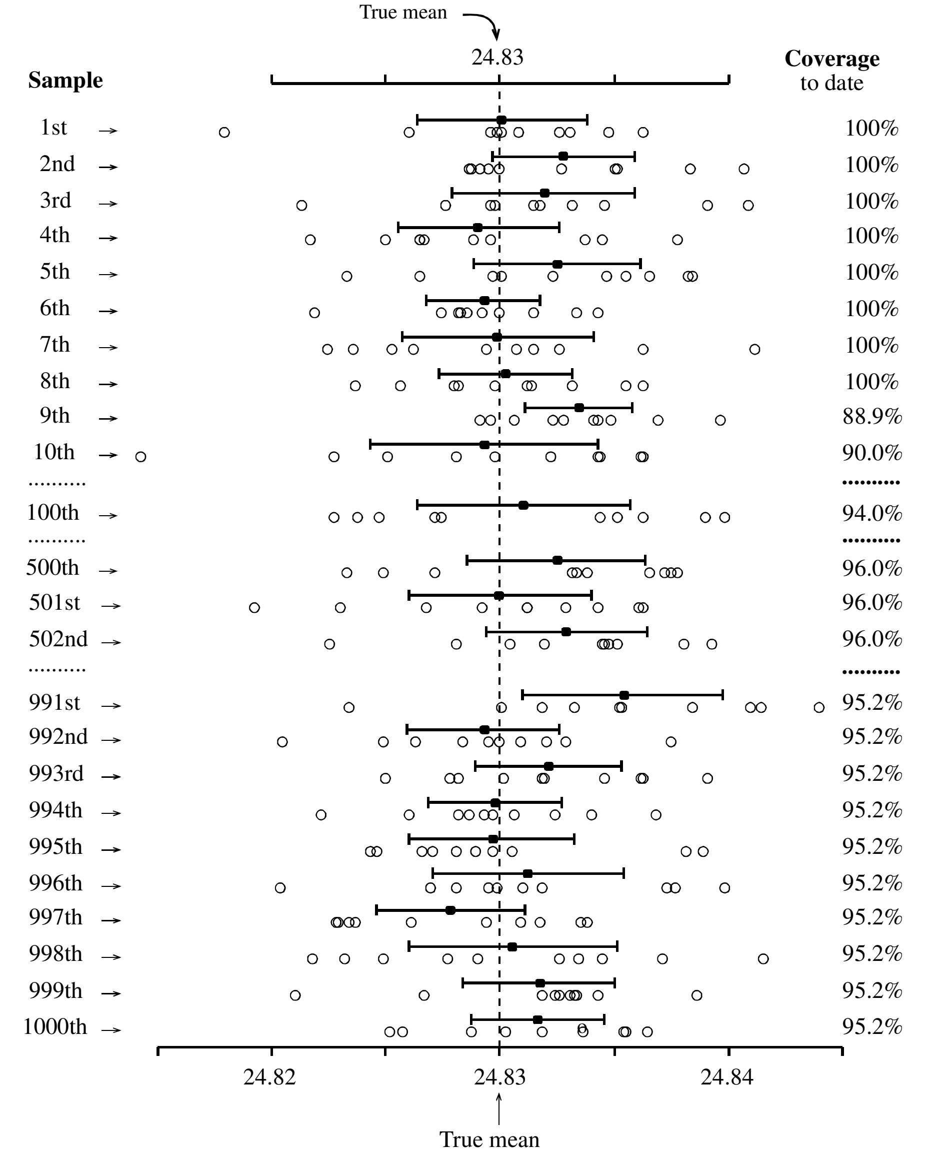


Figure 8.1.2 Samples of size 10 from a Normal(μ=24.83, s=.005) distribution and their 95% confidence intervals for μ..

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, ***estimate* ± *t* × se(*estimate*)**, from the Formula Sheet
4. Find the appropriate **standard error** from the Formula Sheet.
5. Find the appropriate ***df***.
6. Find the appropriate ***t*-value.** This is called the '*t*-multiplier'. Need to know the confidence level and *df*.
7. Calculate the **confidence limits**, the end points of the confidence interval.
8. Interpret the interval using plain English. Use the confidence limits to construct an answer to the original question in plain English.

- There are four different types of problem covered in Chapter 8:
  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***

- 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* = $\hat{p}_1 - \hat{p}_2$

- The ***t*-multiplier** (means) / ***z*-multiplier** (proportions) is based on:
  - ✓ Whether we are investigating means or proportions
  - ✓ The desired level of confidence
  - ✓ The degrees of freedom

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

i.e. for proportions, assume the degrees of freedom is infinity, hence replace *t* with *z* score (i.e. the standard Normal).

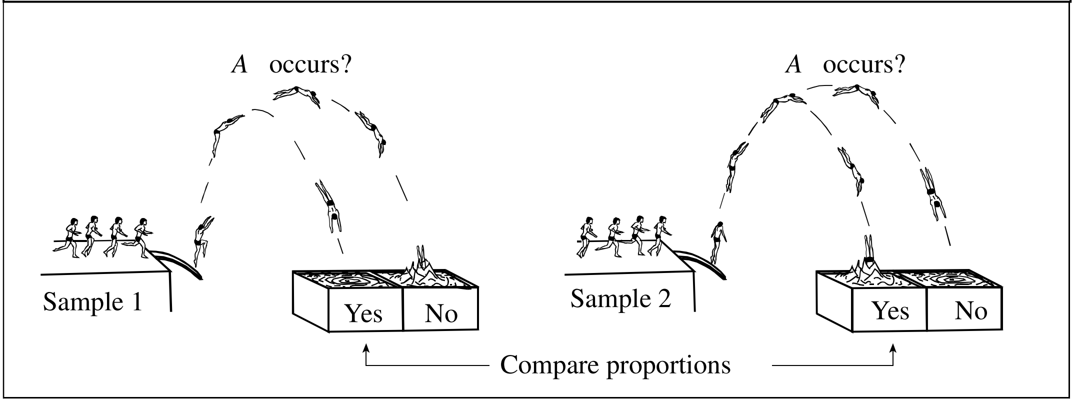
- The **standard error** is based on the estimate, the number of samples and sample size(s).

Estimate	se(estimate)
1. $estimate = \overline{x}$	$se(\overline{x}) = \frac{s}{\sqrt{n}}$
2. $estimate = \hat{p}$	$se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
3. $estimate = \overline{x}_1 - \overline{x}_2$	$se(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
4. $estimate = \hat{p}_1 - \hat{p}_2$	**Situation (a)** ***Proportions from two independent samples*** $se(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ **Situation (b)** ***One sample of size n, several response categories*** $se(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 + \hat{p}_2 - (\hat{p}_1 - \hat{p}_2)^2}{n}}$ **Situation (c)** ***One sample of size n, many yes / no items*** $se(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\min imum(\hat{p}_1 + \hat{p}_2, \hat{q}_1 + \hat{q}_2) - (\hat{p}_1 - \hat{p}_2)^2}{n}}$ where $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$

- CIs for the difference between two means/proportions:
  - ✓ If the CI contains 0 (i.e. one negative and one positive number), there may be no difference between the two means/proportions.
  - ✓ If CI is positive, then μ_1/p_1 is higher/larger than μ_2/p_2.
  - ✓ If CI is negative, then μ_1/p_1 is lower/smaller than μ_2/p_2.

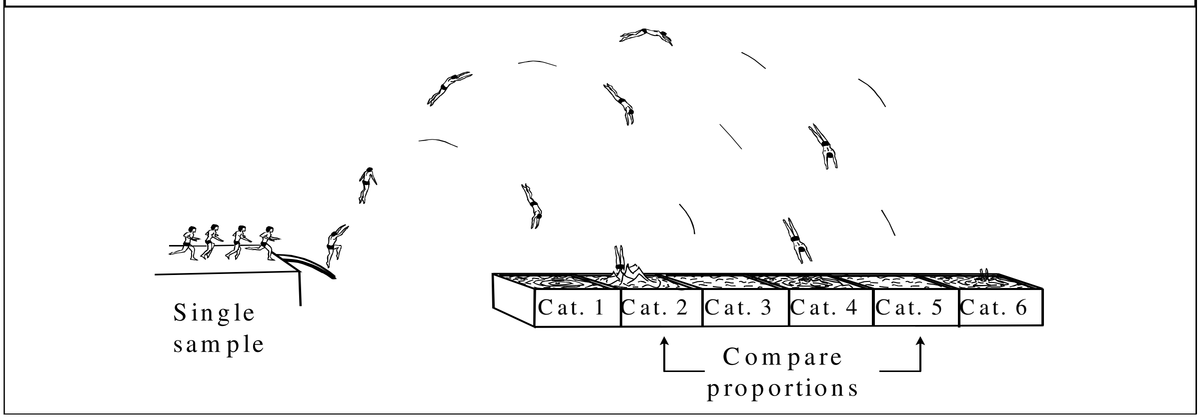
- **3 sampling situations for the difference between two proportions**

Situation (a) *Proportions from two independent samples*



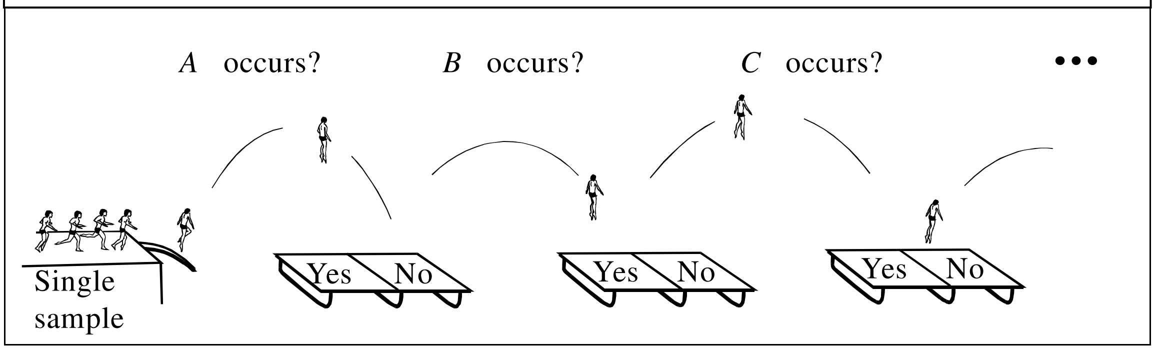
From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 1999.

Situation (b) *One sample of size n, several response categories*



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

Situation (c) *One sample of size n, many yes/no items*



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

Chapters 7 & 8 – Questions

1. Suppose $X_1, X_2, \ldots, X_{11}$ are the weights of 11 randomly selected packets of M & M's which come from a distribution with mean μ = 54 and standard deviation σ = 2. Then the distribution of the sample mean $\overline{X} = (X_1 + X_2 + \cdots + X_{11})/11$ has mean $\mu_{\overline{X}}$ and standard deviation $\sigma_{\overline{X}}$ given by:
   (1) $\mu_{\overline{X}} = 4.91$, $\sigma_{\overline{X}} = 0.182$
   (2) $\mu_{\overline{X}} = 594$, $\sigma_{\overline{X}} = 2.000$
   (3) $\mu_{\overline{X}} = 54$, $\sigma_{\overline{X}} = 0.603$
   (4) $\mu_{\overline{X}} = 54$, $\sigma_{\overline{X}} = 2.000$
   (5) $\mu_{\overline{X}} = 4.91$, $\sigma_{\overline{X}} = 0.603$

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and standard deviation σ. Which **one** of the following statements is **false**?
   (1) $E(X_1) = E(X_2) = \ldots = E(X_n)$.
   (2) $X_1, X_2, \ldots, X_n$ all have the same distribution.
   (3) $X_1, X_2, \ldots, X_n$ are independent of each other.
   (4) $sd(X_1) = sd(X_2) = \ldots = sd(X_n)$.
   (5) $nX_1 = X_1 + X_2 + \ldots + X_n$.

3. A random variable T has a Student's t-distribution with df degrees of freedom. Which **one** of the following statements is **false**?
   (1) The t-distribution for small degrees is flatter with wider tails than the Normal distribution.
   (2) The mean of T is zero.
   (3) The t-distribution gets closer to the Normal(0,1) distribution as df gets larger.
   (4) The t-distribution gets closer to the Normal(0,1) distribution as df gets smaller.
   (5) The distribution of T is symmetric.

4. Let $X_1, \ldots, X_{10}$ be a random sample of size 10 from a distribution with μ = 12 and σ = 2, then the distribution of $\overline{X}$ has expected value $E[\overline{X}]$ and standard deviation $sd[\overline{X}]$ where:
   (1) $E[\overline{X}] = 1.2$, $sd[\overline{X}] = 2$
   (2) $E[\overline{X}] = 12$, $sd[\overline{X}] = 0.2$
   (3) $E[\overline{X}] = 1.2$, $sd[\overline{X}] = 0.63$
   (4) $E[\overline{X}] = 12$, $sd[\overline{X}] = 0.63$
   (5) $E[\overline{X}] = 120$, $sd[\overline{X}] = 6.32$

5. According to Auckland Regional Transport, the mean distance an Aucklander travels to work is 12.6 km. Assume that the distance travelled by an Aucklander to work is well modelled by a Normal distribution with a mean of 12.6 km and a standard deviation of 4.3 km. Let Z be a random variable which has the standard Normal distribution. The probability that the mean distance travelled by a random sample of 10 Aucklanders is greater than 14 km is:

(1) $\Pr\left(Z > \dfrac{10 \times 14 - 10 \times 12.6}{4.3/\sqrt{10}}\right)$

(2) $\Pr\left(Z > \dfrac{14 - 12.6}{\sqrt{10} \times 4.3}\right)$

(3) $\Pr\left(Z > \dfrac{14 - 12.6}{4.3}\right)$

(4) $\Pr\left(Z > \dfrac{14 - 12.6}{\sqrt{10}/4.3}\right)$

(5) $\Pr\left(Z > \dfrac{14 - 12.6}{4.3/\sqrt{10}}\right)$

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

(1) For large random samples, the true value of μ lies inside the interval $\bar{x} \pm 2se(\bar{x})$ for a little more than 95% of all samples taken.

(2) For random samples for a Normal distribution, $T = (\bar{X} - \mu)/se(\bar{X})$ is exactly distributed as Student's *t*-distribution ($df = n - 1$).

(3) The precision of an estimate refers to its variability – one estimate is less precise than another if it has more variability.

(4) The Student's *t*-distribution ($df = \infty$) distribution has 'fatter' or 'heavier' tails than the Normal($\mu = 0$, $\sigma = 1$) distribution.

(5) For large random samples, $T = (\bar{X} - \mu)/se(\bar{X})$ is distributed as approximately Normal(μ_X=0, σ_X=1).

7. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and standard deviation σ. Let $\bar{X}$ be the sample mean. Which **one** of the following statements is **false**?

(1) $\bar{X}$ may not be approximately Normally distributed for small samples.

(2) $\bar{X}$ is a random variable.

(3) sd($\bar{X}$) = sd(X).

(4) E($\bar{X}$) = E(X).

(5) $\bar{X}$ is at least approximately Normally distributed for large samples.

8. $X_1,\ldots,X_{25}$ form a random sample from a distribution with mean $\mu = 10$ and standard deviation $\sigma = 9$. The sample mean, $\overline{X}$, has mean $\mu_{\overline{X}}$ and standard deviation $\sigma_{\overline{X}}$ given by:

(1) $\mu_{\overline{X}} = 25 \times 10$, $\sigma_{\overline{X}} = \sqrt{25} \times 10$

(2) $\mu_{\overline{X}} = \frac{10}{25}$, $\sigma_{\overline{X}} = \frac{9}{25}$

(3) $\mu_{\overline{X}} = 10$, $\sigma_{\overline{X}} = 9$

(4) $\mu_{\overline{X}} = 10$, $\sigma_{\overline{X}} = \sqrt{\frac{9}{25}}$

(5) $\mu_{\overline{X}} = 10$, $\sigma_{\overline{X}} = \frac{9}{\sqrt{25}}$

9. Given a simple random sample, which **one** of the following is **false**?

(1) The mean of the distribution of the sample mean, $\mu_{\overline{X}}$, is equal to the population mean, μ.

(2) The sample mean is an unbiased estimate of the population mean since $E[\overline{X}] = \mu$.

(3) The distribution of the sample mean, $\overline{X}$, is approximately Normal for very large samples only if the distribution from which the sample has been drawn is not skewed.

(4) The standard error of the sample mean is an estimate of the standard deviation of the sample mean.

(5) Increasing the sample size by a factor of 4 will double the accuracy of the sample mean as an estimate of the population mean.

10. On most airlines, the maximum allowable weight for checked baggage is 20kg per economy class passenger. Excess baggage is defined to be any baggage over the 20kg limit. Airline staff know that the weight, X, of economy class passengers' checked baggage is approximately Normally distributed with mean μ_X = 15kg and standard deviation σ_X = 3kg.

Let $\overline{X}$ be the mean checked baggage weight for a randomly selected group of 20 economy class passengers. The expected value $E[\overline{X}]$, and standard deviation $sd[\overline{X}]$, are:

(1) $E[\overline{X}]$ = 15kg, $sd[\overline{X}]$ = 0.67kg

(2) $E[\overline{X}]$ = 15kg, $sd[\overline{X}]$ = 0.15kg

(3) $E[\overline{X}]$ = 300kg, $sd[\overline{X}]$ = 0.67kg

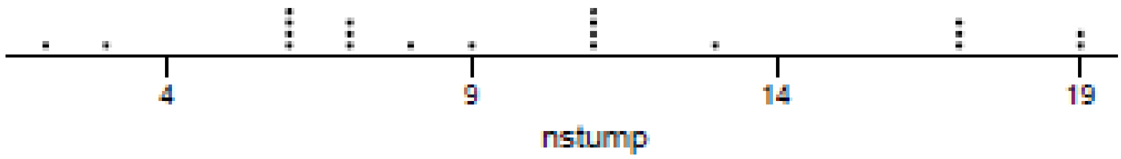
(4) $E[\overline{X}]$ = 300kg, $sd[\overline{X}]$ = 13.42kg

(5) $E[\overline{X}]$ = 15kg, $sd[\overline{X}]$ = 3kg

Questions 11 and **12** refer to the following information.

Data were collected on the number of cricket test match innings in which the batsman was stumped by the wicket keeper. The number of stumping dismissals per year (nstump) were recorded for the years 1970 to 1990 (inclusive). Summary statistics and a dot plot of these data are given below:

	n	$\bar{x}$	s	se($\bar{x}$)
nstump	21	10.14	5.15	1.1238



11. A 95% confidence interval for the cricket data was calculated to be (7.80, 12.49). Which one of the following statements is **false**?
    (1) If many such samples are taken and a 95% confidence interval for μ_{nstump} is calculated from each sample, then statements such as "μ_{nstump} is somewhere between the two confidence limits" are true, on average, 19 times out of 20.
    (2) The technique used to calculate the confidence interval generates an interval which contains the true population mean approximately 95% of the time, in the long run.
    (3) With 95% confidence, the sample mean is somewhere between 7.8 and 12.5.
    (4) In light of the data, the plausible values of μ_{nstump} are between 7.8 and 12.5.
    (5) With 95% confidence, the value of μ_{nstump} is estimated to be 10.1, with a margin of error of 2.3.

12. Based on the years 1960 to 1969, the mean of nstump is 7.4 and the standard deviation of nstump is 4.35. Which one of the following statements about a new 95% confidence interval for μ_{nstump} based on these data compared to the original confidence interval given above is **true**?
    (1) The t-multiplier used in the new confidence interval is larger.
    (2) The standard error used in the new confidence interval is smaller.
    (3) The new confidence interval is centred around a higher value.
    (4) Both confidence intervals would have the same width because they are both 95% confidence intervals for μ_{nstump}.
    (5) The new confidence interval is narrower than the original confidence interval.

13. The Central Limit Theorem says that the distribution of the mean, $\overline{X}$, of a random sample:
   (1) is a Student's *t*-distribution with *n*–1 degrees of freedom.
   (2) is *approximately* Normally distributed when the mean is large.
   (3) is an *F*–distribution with degrees of freedom, which can be calculated from the sample data.
   (4) can be approximated by the Normal distribution for small samples.
   (5) is *approximately* Normal in large samples.

Questions 14 and 15 refer to the following situation:

The weight gain of women during pregnancy has an important effect on the birth weight of their children. In a study conducted in three countries, weight gains (in kg) of women during the last three months of pregnancy were measured. The results are summarised in the following table:

Country	n	$\overline{x}$	s_x
Egypt (E)	11	4.55	1.83
Kenya (K)	11	3.29	0.851
Mexico (M)	11	2.9	1.8

14. We wish to determine the size of the difference in average weight gain between Egyptian and Kenyan women. Then the value of the standard error of $\overline{x}_E - \overline{x}_K$, $se(\overline{x}_E - \overline{x}_K)$, is (select **one** only):

(1) 0.5990
(2) 0.6085
(3) 0.3703
(4) 0.7739
(5) 0.9288

15. When calculating a 90% confidence interval for $\mu_E - \mu_K$, the value of the *t*-multiplier obtained from a *Student's t*-table is 2.228. Assuming $se(\overline{x}_E - \overline{x}_K) = 0.7$ (it is not), then the **margin of error** for the 90% confidence interval is:

(1) $(4.55 - 3.29) \pm 2.228 \times \frac{0.7}{\sqrt{11}}$

(2) $\pm 2.228 \times \frac{0.7}{\sqrt{11}}$

(3) $(3.29 - 4.55) \pm 2.228 \times 0.7$

(4) $(3.29 - 4.55) \pm 2.\ 228 \times \frac{0.7}{\sqrt{11}}$

(5) $\pm 2.228 \times 0.7$

Questions 16 & 17 refer to the TIME MORGAN POLL "*Living Well Beats Being Green, Say Most Voters*" of July 1994. 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.

16. 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

17. 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 0.1547 larger and 0.0303 smaller than p_L.
    (4) With 95% confidence the true proportion p_N is somewhere between 0.1547 smaller and 0.0303 larger than p_L.
    (5) With 95% confidence the true proportion p_N is 0.1850 larger than p_L.

18. The data below are taken from a study by M.C.Gilly (1988), 'Sex roles in advertising: a comparison of television advertisements in Australia, Mexico, and the United States', *Journal of Marketing* 52, pp 75-85. 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 19 to **21** refer to the following information.

In 1990 *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".

19. *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" in 1990.

20. An approximate 95% confidence interval for the proportion of the 18-29 year old American population who would have answered "Yes" to this question in 1990 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]

21. 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.

Questions 22 to **24** 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%

22. 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 standard error, $se(\hat{p}_M - \hat{p}_N)$, are:

(1) −0.57 and 0.0298

(2) −0.57 and 0.0178

(3) 0.57 and 0.0216

(4) 0.57 and 00298

(5) 0.57 and 0.0178

23. 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. A 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)

24. 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 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) 0.90

26. Assuming that the sample size in 1985 was also 1000, the value of $se(\hat{p}_{1994} - \hat{p}_{1985})$ is (choose **one**):
    (1) 0.0004
    (2) 0.0220
    (3) 0.0009
    (4) 0.0191
    (5) 0.0068

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. Student Job Search is a nationwide organisation that finds part-time jobs for tertiary students. Data is collected on how many weeks work each student found over the summer break. This variable is called Summer Worker Weeks (SWW). Samples were taken from several different student groups. Below is a summary table of SWW for the students in these samples.

SWW For Different Institutions and Ethnic Groups

	Polytechnic	University	Maori	Pacific Islander	Overseas Student	Other
$\bar{x}$	1.38	2.21	1.66	2.34	3.00	2.01
s	0.48	0.53	0.49	2.26	1.70	0.48
n	11	11	11	10	10	11

Suppose we are interested in comparing the SWW for Pacific Island students to the SWW for overseas students. The $se(\bar{x}_{Overseas} - \bar{x}_{Pacific\ Islanders})$ is equal to (Choose **one**):
   (1) 0.63
   (2) 2.58
   (3) 3.96
   (4) 0.89
   (5) 0.40

ANSWERS

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