Stats 101/101G/108 Workshop

Hypothesis Tests: *Means* [HTM]

2020

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-HTM**

Leila Boyle
Undergraduate Statistics Assistance, Department of Statistics
Room 303S.288 (second floor of the Science Centre, Building 303S)
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

One computing workshop, four exam prep workshops and four drop-in sessions are held during the second half of the semester.

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!

No booking is required – just turn up to any workshop! You are also welcome to come along virtually on Zoom if you prefer. Search your emails for "Leila" to find the link – email Leila at l.boyle@auckland.ac.nz if you can't find it.

- **Computer workshop: *Hypothesis Tests in SPSS***
  www.tinyURL.com/stats-HTS
  Computing for Assignment 3 **–** covers the **computing** you need to do for **Questions 3 and 4** (iNZight plots & SPSS output). There are **six identical sessions**:
  - Friday 16 October, 3-4pm
  - Monday 19 October, 10-11am
  - Monday 19 October, 2-3pm
  - Tuesday 20 October, 4-5pm
  - Wednesday 21 October, 11am-midday
  - Wednesday 21 October, 3-4pm

- **Exam prep workshops**
  - ***Chi-Square Tests*** **www.tinyURL.com/stats-CST**
    Exam revision for Chapter 9 **–** Saturday 24 October, 1-4pm, LibB15 (useful exam prep and also useful for the **Chapter 9 Quiz** due at 11pm on Wednesday 28 October!)
  - ***Regression and Correlation*** **www.tinyURL.com/stats-RC**
    Exam revision for Chapter 10 **–** Saturday 31 October, 9.30am-12.30pm, LibB10 (useful exam prep and also useful for the **Chapter 10 Quiz** due at 11pm on Wednesday 4 November!)
  - ***Hypothesis Tests: Proportions*** **www.tinyURL.com/stats-HTP**
    Exam revision for Chapters 6 & 7 *(with a focus on proportions)* **–** Tuesday 3 November, 9.30am-12.30pm, LibB10 (useful exam prep)
  - ***Hypothesis Tests: Means*** **www.tinyURL.com/stats-HTM**
    Exam revision for Chapter 6, 7 & 8 *(with a focus on means)* **–** Tuesday 3 November, 1-4pm, LibB10 (useful exam prep)

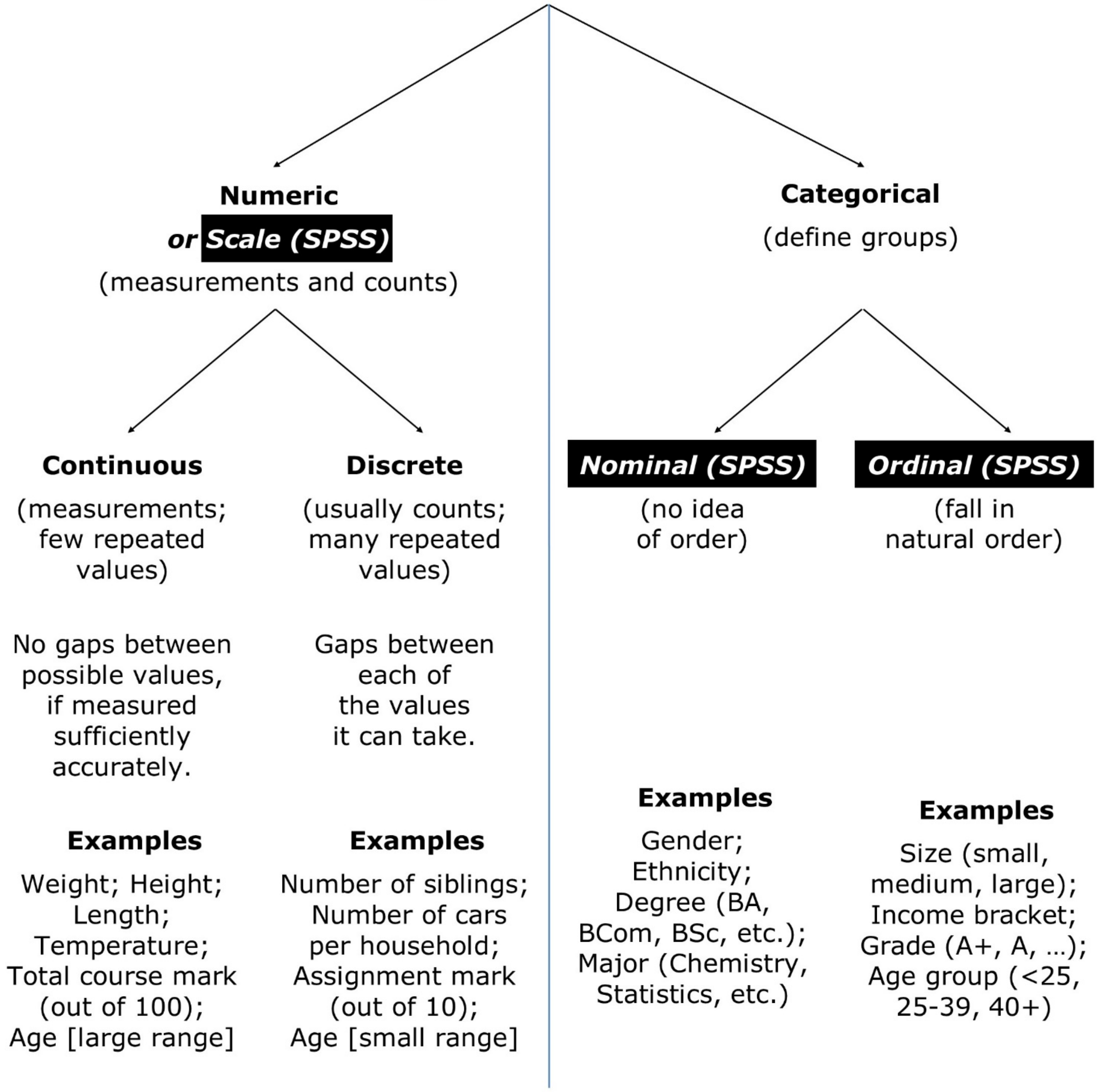
- **Drop-in sessions**
  - Saturday 17 October, 9.30am-4pm, LibB10
  - Saturday 24 October, 9.30am-12.30pm, LibB15
  - Monday 26 October, 9.30am-4pm, LibB10
  - Saturday 31 October, 1-4pm, LibB10

Hypothesis Tests: *Means* [HTM]

This material builds on a couple of workshops already held in the first half of this semester, which you may or may not have attended.

For more practice on how to **quantify the size of a single mean or difference between two means**, see the ***Confidence Intervals: Means*** workshop materials. If you want to learn more about how to **explore our sample data**, see the ***Exploratory Data Analysis*** workshop.

Types of Variables



Numeric ***or Scale (SPSS)*** (measurements and counts)

- **Continuous** (measurements; few repeated values)

  No gaps between possible values, if measured sufficiently accurately.

  Examples

  Weight; Height; Length; Temperature; Total course mark (out of 100); Age [large range]

- **Discrete** (usually counts; many repeated values)

  Gaps between each of the values it can take.

  Examples

  Number of siblings; Number of cars per household; Assignment mark (out of 10); Age [small range]

Categorical (define groups)

- ***Nominal (SPSS)*** (no idea of order)

  Examples

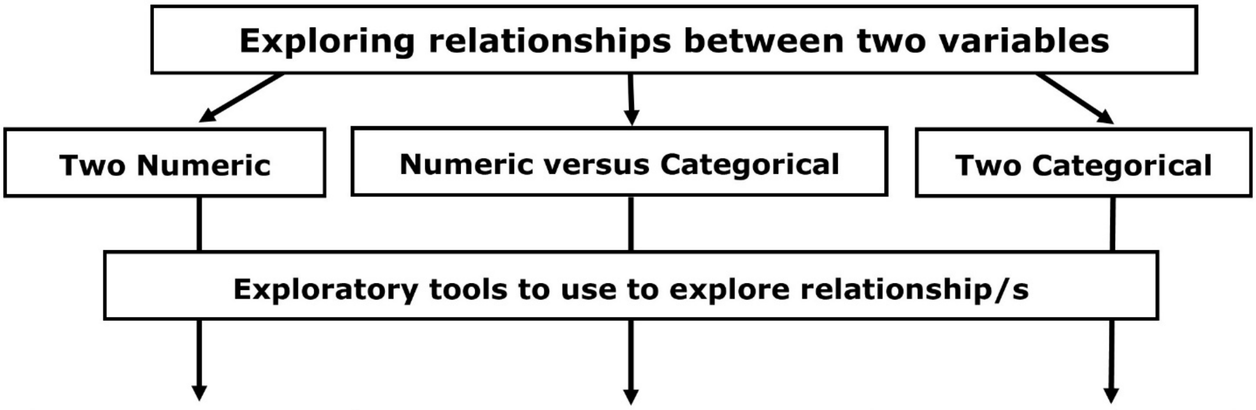
  Gender; Ethnicity; Degree (BA, BCom, BSc, etc.); Major (Chemistry, Statistics, etc.)

- ***Ordinal (SPSS)*** (fall in natural order)

  Examples

  Size (small, medium, large); Income bracket; Grade (A+, A, …); Age group (<25, 25-39, 40+)

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



Exploring relationships between two variables

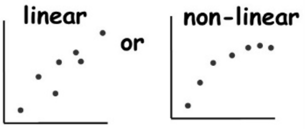
Two Numeric | **Numeric versus Categorical** | **Two Categorical**

Exploratory tools to use to explore relationship/s

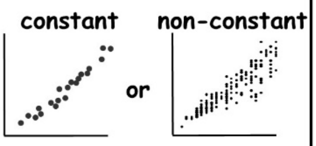
Scatter plot	Side-by-side plots on the same scale: any n { • Dot plots $n \geq 20$ { • Stem-and-leaf plots • Box plots $n \geq 50$ { • Histograms	Two-way table of counts and/or Bar graphs of proportions (on rows/columns)

Features to look for:

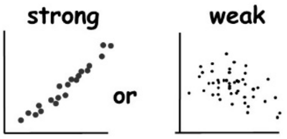
- Trend



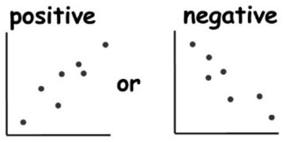
- Scatter



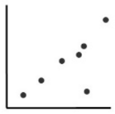
- Strength of relationship



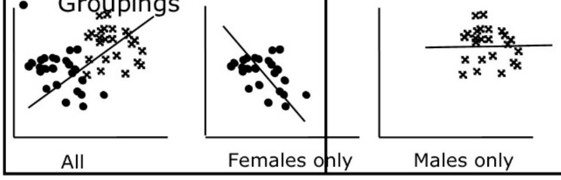
- Association



- Outliers



- Groupings



Compare the groups by looking at:

- Any group differences:
  - averages (centres) → **medians**, **means**
  - variability (spread) → **IQRs**, **ranges**
  - shapes
    - symmetric/skewed → **left/negative**, **right/positive**
    - modes → **unimodal**, **bimodal**, **trimodal** } **not box!**
- Details of individual groups:
  - outliers, gaps, clusters, groupings

Think about reasons *why* these differences, similarities and features are seen

- Most common and least common combinations
- Differences in distributions (e.g. row and/or column bar graphs)

Recall that two **numerical summaries** of **centre** are:

- Sample mean, $\bar{x}$ (also known as the average or expected value)
  $\sum \frac{x_i}{n}$ – affected by outliers

- Median (= Med – also known as the 50th percentile) = middle number of the ordered data – not affected by outliers

t-tests by Hand – One and Two Mean/s

We use statistics to find 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.

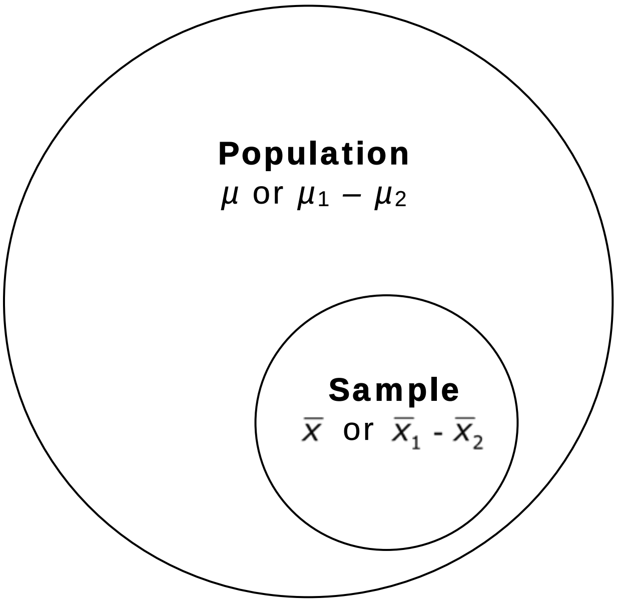
We are usually unable to survey the entire population (take a census) as it is too large and/or there are:

- budget constraints
- time limits
- logistical barriers

This means we are unable to establish the **parameters** of interest within our population, such as:

1. Population mean, μ
3. Difference in population means, $\mu_1 - \mu_2$

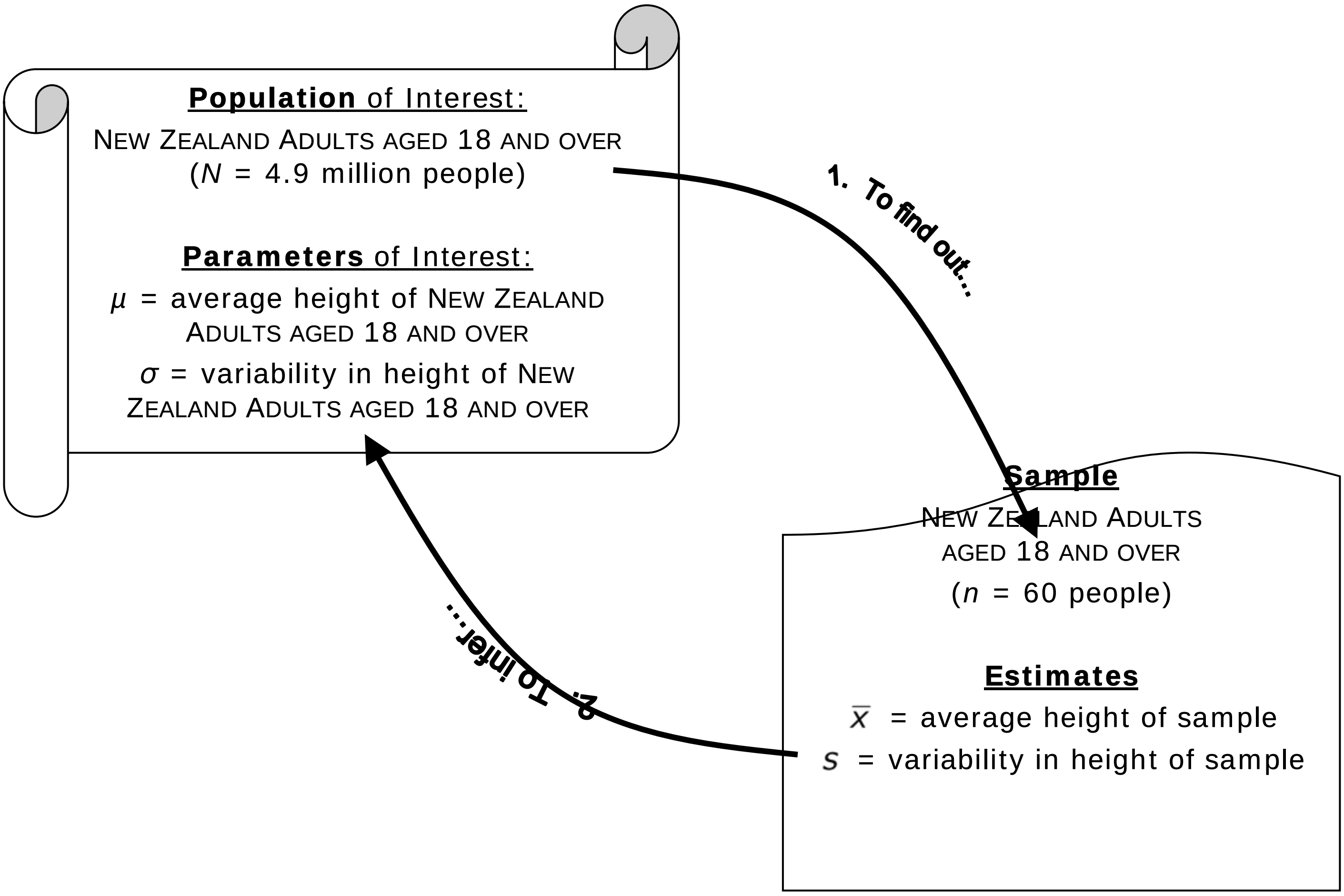
This means that the **parameter** of interest is an ***<u>unknown</u> numerical characteristic*** for that particular population.



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:

1. Sample mean, $\bar{x}$
3. Difference in sample means, $\bar{x}_1 - \bar{x}_2$

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



Population of Interest:

NEW ZEALAND ADULTS AGED 18 AND OVER
(N = 4.9 million people)

Parameters of Interest:

μ = average height of NEW ZEALAND ADULTS AGED 18 AND OVER

σ = variability in height of NEW ZEALAND ADULTS AGED 18 AND OVER

Sample

NEW ZEALAND ADULTS AGED 18 AND OVER
(n = 60 people)

Estimates

$\bar{x}$ = average height of sample

s = variability in height of sample

We use **sample** data to make inferences (draw conclusions) about **population** parameters by carrying out hypothesis tests and constructing confidence intervals.

- A **significance test,** tests one possible value for the parameter, called the **hypothesised** value. We determine the strength of evidence provided by the data against the null hypothesis, H_0.

- 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).

A significance test determines the **strength** of the evidence **against** the **hypothesised** value, while a confidence interval determines the **size** of the effect or difference.

Significance testing and confidence intervals are methods used to deal with the **uncertainty** about the true value of a parameter caused by the **sampling variation** in estimates.

Step-by-Step Guide to Performing a Hypothesis Test by Hand

1. State the **parameter** of interest (symbol and words).

For example, is it μ, p, $\mu_1 - \mu_2$, or $p_1 - p_2$?

2. State the **null hypothesis**, ***H*₀**. **e.g.** H_0: parameter = *hyp. val.*

3. State the **alternative hypothesis**, ***H*₁**. **e.g.** H_1: parameter ≠ *hyp. val.*

or H_1: parameter > *hyp. val.*

or H_1: parameter < *hyp. val.*

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

5. Calculate the **test statistic**:

For example, for a ***t*-test statistic**:

see back page for Formulae Sheet

- Use: $\boxed{t_0 = \dfrac{\textit{estimate} - \textit{hypothesised value}}{\textit{std error}}}$

- Use the estimate from Step 4 and the hypothesised value from Steps 2&3.
- Use the appropriate standard error. (Will be provided)
- Calculate t_0.

6. Estimate the ***P-value***. (Will be provided)

7. **Interpret** the *P-value*. (see page 10)

8. Calculate the **confidence interval**.

For example, for a **Normality-based confidence interval**:

- Use: $\boxed{\textit{estimate} \pm t \times \textit{se(estimate)}}$

- Use the estimate from Step 4 and the standard error from Step 5.
- Use the appropriate t-multiplier. (Will be provided)

9. **Interpret** the confidence interval using plain English.

10. Give an overall **conclusion**.

- 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***~~

Step 1

The **parameter** of interest we are investigating depends on the problem type:

Parameter
1. Single mean $\boldsymbol{\mu}$:
~~2. Single proportion $\boldsymbol{p}$:~~
3. Difference between two means $\boldsymbol{\mu_1 - \mu_2}$: (independent samples)
~~4. Difference between two proportions $\boldsymbol{p_1 - p_2}$:~~

Steps 2 & 3

The null hypothesis, H_0

✓ It is our best guess as to what we think the parameter of interest is – a single plausible value.

✓ The hypothesised value is **not** the parameter of interest. Remember that the parameter of interest is an unknown quantity.

✓ General form: H_0: *parameter = hypothesised value (some number)*

1. $H_0: \mu =$ 3. $H_0: \mu_1 - \mu_2 =$

✓ It's the **boring** thing – **there is no** effect or difference.

The alternative hypothesis, H_1

✓ Specifies the type of departure from H_0 that we expect to detect.

✓ Corresponds to the research hypothesis.

✓ There are three different types:

- o H_1: *parameter ≠ hypothesised value (some number)*
- o H_1: *parameter > hypothesised value (some number)*
- o H_1: *parameter < hypothesised value (some number)*

1. $H_1: \mu$ 3. $H_1: \mu_1 - \mu_2$

✓ When do we use a 1-sided alternative hypothesis?
* if in doubt * data * research

✓ It's the **interesting** thing – **there is an** effect or difference.

Step 4 (and Step 8)

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

Parameter	**Estimate**
1. Single mean $\boldsymbol{\mu}$:	*estimate* $= \overline{x}$
3. Difference between two means $\boldsymbol{\mu_1 - \mu_2}$: (independent samples)	*estimate* $= \overline{x}_1 - \overline{x}_2$

Step 5 (and Step 8)

- The **standard error** can be found from the *t*-procedures tool.

In the exam situation, the standard error will be provided.

- The **degrees of freedom** are based on the sample size(s):

Degrees of Freedom
1. $df = n - 1$
3. $df = minimum(n_1 - 1, n_2 - 1)$

***e.g.* $\boldsymbol{n_1 = 50}$ and $\boldsymbol{n_2 = 30}$:**

$$
\begin{aligned}
df &= minimum(n_1 - 1, n_2 - 1)\\
&= min(50 - 1, 30 - 1)\\
&= min(49, 29)\\
&= \_\_\_\_\_
\end{aligned}
$$

- **The *t*-test statistic, $\boldsymbol{t_0}$:**
  - ✓ tells us how many standard errors the estimate is away from the hypothesised value.
  - ✓ is calculated using: $\boxed{t_0 = \dfrac{\textit{estimate} - \textit{hypothesised value}}{\textit{std error}}}$
  - ✓ is **positive**, if the estimate is **above** the hypothesised value.
  - ✓ is **negative**, if the estimate is **below** the hypothesised value.
  - ✓ is a **measure** of **difference/distance/discrepancy** between the estimate and the hypothesised value in terms of standard errors.

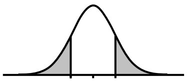
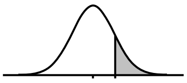
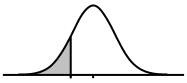
see back page for Formulae Sheet

Step 6

- The ***P*-value:**
  - ✓ is the conditional probability of observing a test statistic as extreme as that observed or more so, given that the null hypothesis, H_0, is true.
  - ✓ is the probability that sampling variation would produce an estimate that is at least as far from the hypothesised value than the estimate we obtained from our data, assuming that the null hypothesis is true.

- ✓ measures the strength of evidence **against** H_0.
- ✓ is calculated using the t-test statistic and the appropriate Student's t-distribution for the t-test.

In the exam situation, the *P*-value will be provided.

Alternative hypothesis	***P-value* ≈ area of shaded region**
H_1: parameter ≠ hypothesised value (2-sided)	 2-tailed test
H_1: parameter **>** hypothesised value (1-sided)	 1-tailed test
H_1: parameter **<** hypothesised value (1-sided)	 1-tailed test

Student(*df*) or Re-randomisation distribution

A. Which **one** of the following statements about a *P-value* is **false**?

(1) The larger a *P-value*, the stronger the evidence against the null hypothesis.

(2) A *P-value* measures the strength of evidence against the null hypothesis.

(3) A relatively large test statistic results in a relatively small *P-value*.

(4) A *P-value* is the conditional probability of observing a test statistic as extreme as that observed or even more so, if the null hypothesis were true.

(5) A *P-value* says nothing about the size of an effect or difference.

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

(1) A small *P-value* provides evidence of the size of an effect.

(2) Statistical significance is the same as practical significance.

(3) Practical significance depends on the size of the effect.

(4) A small *P-value* provides no evidence against H_0.

(5) A confidence interval estimates the strength of an effect.

Step 7

- The ***P*-value** measures the strength of evidence against the null hypothesis, H_0. We interpret the *P-value* as a description of the **strength of evidence against the null hypothesis, H_0**. The **smaller** the *P-value*, the **stronger** the evidence against H_0:

P-value	Evidence against H_0
> 0.10	None
≈ 0.07	Weak
≈ 0.05	Some
≈ 0.01	Strong
≤ 0.001	Very Strong

- An alternative approach often found in research articles and news items is to describe the test result as (statistically) significant or not significant. A test result is said to be significant when the *P-value* is "small enough"; usually people say a *P-value* is "small enough" if it is less than 0.05 (5%):

Testing at a 5% level of significance:

P-value	Test result	Action
< 0.05	Significant	Reject H_0 in favour of H_1
> 0.05	Nonsignificant	Do not reject H_0

Testing can be done at any level of significance; 1% is common but 5% is what most researchers use.

The level of significance can be thought of as a false alarm error rate, i.e. it is the proportion of times that the null hypothesis will be rejected when it is actually true (which can result in action being taken when really no action should be taken).

Thus, a statistically significant result means that a study has produced a "small" *P-value* (usually < 5%).

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

(1) A small *P-value* provides evidence of the size of an effect.
(2) Statistical significance is the same as practical significance.
(3) Practical significance depends on the size of the effect.
(4) A small *P-value* provides no evidence against H_0.
(5) A confidence interval estimates the strength of an effect.

D. Which **one** of the following statements about hypothesis testing is **false**?

(1) The *P-value* is the probability that, if the null hypothesis were true, sampling variation would produce an estimate that is further away from the hypothesised value than our data estimate.
(2) We cannot establish an hypothesised value for a parameter, we can only determine whether there is evidence to reject a hypothesised value.
(3) H_0 is typically a sceptical reaction to a research hypothesis.
(4) The *P-value* measures the strength of evidence against the null hypothesis.
(5) The larger the *P-value,* the stronger the evidence against the null hypothesis.

E. Suppose the hypothesis test H_0: μ = 100 versus H_1: $\mu \neq$ 100 obtained a *P-value* = 0.001. Which **one** of the following statements is **true**?

(1) The *P-value* is very small, therefore H_0 is false.
(2) We would reject H_0 at the 1% level of significance.
(3) A 95% confidence interval for μ contains the value 100.
(4) A 99% confidence interval for μ contains the hypothesised value.
(5) We will accept that H_0 is true.

F. Which **one** of the following statements about significance tests is **false**?

(1) Formal tests can help determine whether effects we see in our data may just be due to sampling error.
(2) A test statistic is a measure of discrepancy between what we see in our data and what we would expect to see if H_0 was true.
(3) The *P-value* says nothing about the size of an effect.
(4) The data should be carefully examined in order to determine whether the alternative hypothesis needs to be one-sided or two-sided.
(5) The *P-value* describes the strength of evidence against the null hypothesis.

Normality-based (Chapter 6) Confidence Interval:

Step 8

$$\boxed{\textit{estimate} \pm \overbrace{t \times \mathrm{se}(\textit{estimate})}^{\_\_\_\_\_\_\_\_\_\ \textit{of}\ \_\_\_\_\_\_\_\_\_}}$$

The ***t*-multiplier** 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. $estimate = \hat{p}_1 - \hat{p}_2$	$df = \infty$

In the exam situation, you will be given the appropriate *t*-multiplier for a 95% confidence interval.

G. Which **one** of the following statements about confidence intervals is **false**?

(1) For a given level of confidence, increasing the sample size generally decreases the width of a confidence interval.

(2) For a given level of confidence, decreasing the standard error decreases the width of a confidence interval.

(3) For a given sample size, increasing the confidence level increases the width of a confidence interval.

(4) For a given sample size, increasing the confidence level increases the precision of a confidence interval.

(5) For a given level of confidence, decreasing the sample size generally decreases the precision of a confidence interval.

Step 9

- 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 CI's, 90%, 95%, 99%...

- Increasing the confidence level will **increase** the width of the interval.

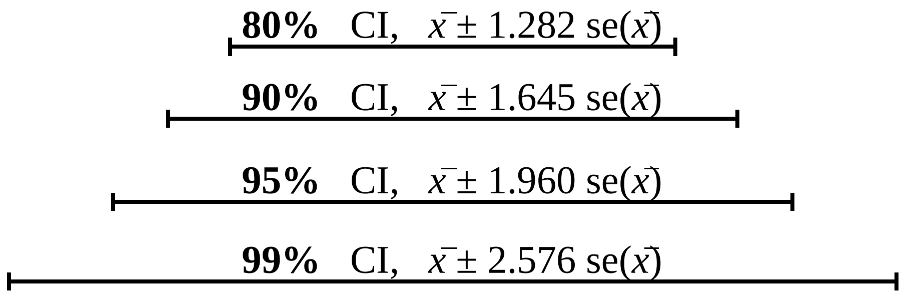

80% CI, $\bar{x} \pm 1.282\ \mathrm{se}(\bar{x})$
90% CI, $\bar{x} \pm 1.645\ \mathrm{se}(\bar{x})$
95% CI, $\bar{x} \pm 1.960\ \mathrm{se}(\bar{x})$
99% CI, $\bar{x} \pm 2.576\ \mathrm{se}(\bar{x})$

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 precision of the confidence interval we **need 4 times** as many observations.

- To triple the precision 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, true difference between means) of the population.

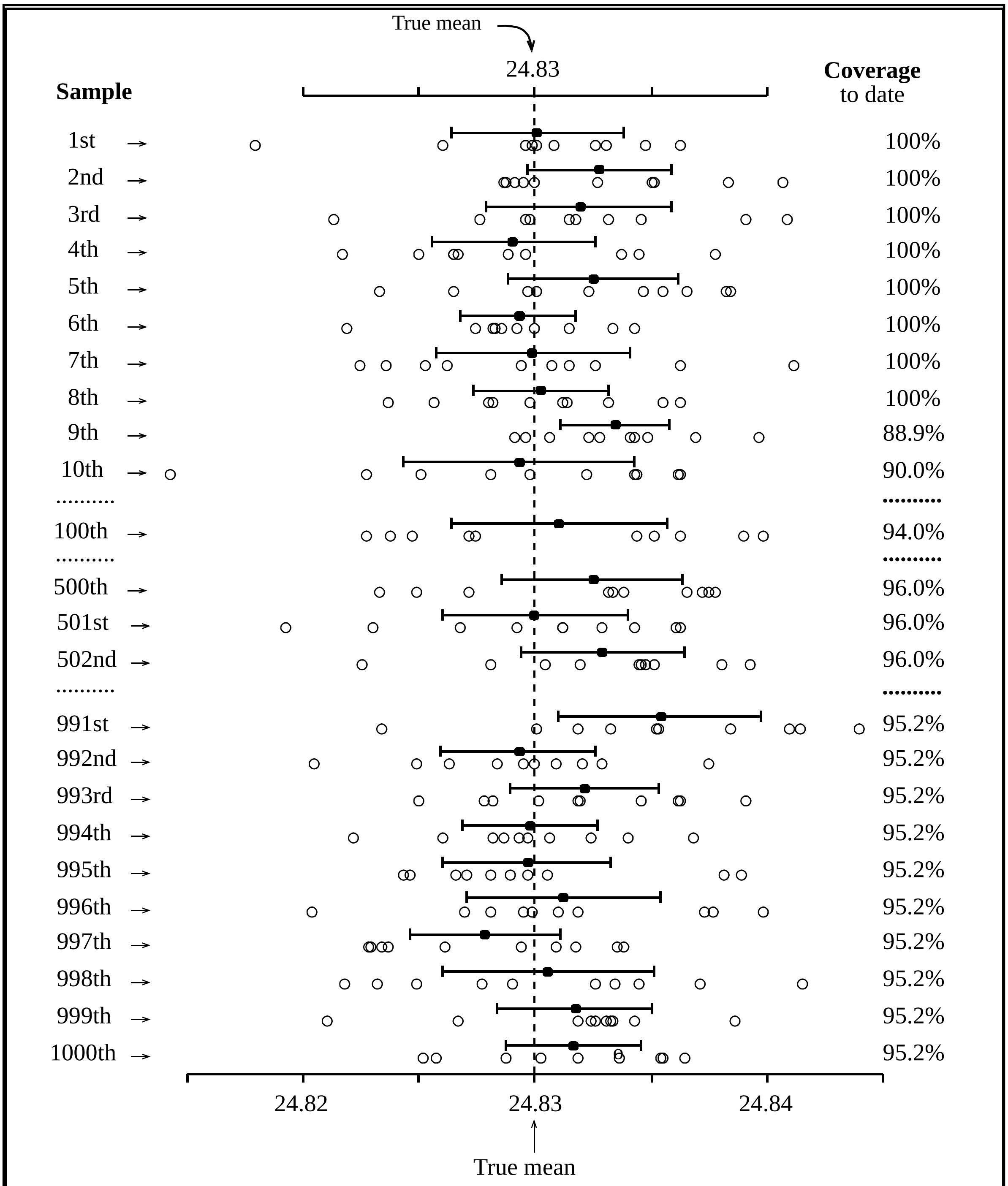


Figure 8.1.2 Samples of size 10 from a Normal(μ=24.83, s=.005) distribution and their 95% confidence intervals for µ..

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

Interpreting the CI limits → Step 9 for story type 3:

- CIs for the difference between two means:

Examples:

✓ If the CI contains 0 (i.e. one negative and one positive number), there may be no difference between the two means. (-7, 19)

✓ If CI is positive, then μ_1 is higher/larger than μ_2. (7, 19)

✓ If CI is negative, then μ_1 is lower/smaller than μ_2. (-19, -7)

H. Suppose that a 95% confidence interval for the difference in true mean HOSP.RATE level between *small cars* and *medium cars*, $\mu_{\text{Small}} - \mu_{\text{Medium}}$, is given by (-0.05, 0.9). Which one of the following statements is **true**?

(1) There is a significant difference between the true means at the 5% level.

(2) There is a significant difference between the sample means at the 5% level.

(3) It is likely that mean HOSP.RATE for *small cars* is much smaller than the mean HOSP.RATE for *medium cars*.

(4) With 95% confidence the true mean HOSP.RATE for *small cars* is somewhere between 0.05 units smaller and 0.9 units bigger than the mean HOSP.RATE for *medium cars*.

(5) The difference between the sample means will be outside this interval 5% of the time.

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

(1) A two-sided test of H_0: *parameter = hypothesised value* has *P-value* less than 0.05 if the *hypothesised value* lies within a 95% confidence interval for the *parameter*.

(2) The larger the *P-value,* the stronger the evidence against H_0.

(3) The larger the test statistic, $|t_0|$, for a two-sided test, the larger the *P-value* will be.

(4) Tests of hypotheses can only deal with random errors and sampling variation. They are ineffective when confronted with data that has systematic bias.

(5) An extremely small *P-value* means that the actual effect differs markedly from that claimed in the null hypothesis.

Questions J to **N** refer to the following information.

Lai *et al.* (2017) measured illegal drugs in water processed by wastewater treatment plants to estimate drug use in Auckland, New Zealand. Random sampling and testing of wastewater occurred between 2 May and 18 July, 2014, at two treatment plants.

A plant's 'catchment area' refers to where the water it treats comes from:

- Plant 1 catchment area: Auckland City, Papakura, Waitakere and Manukau
- Plant 2 catchment area: North Shore

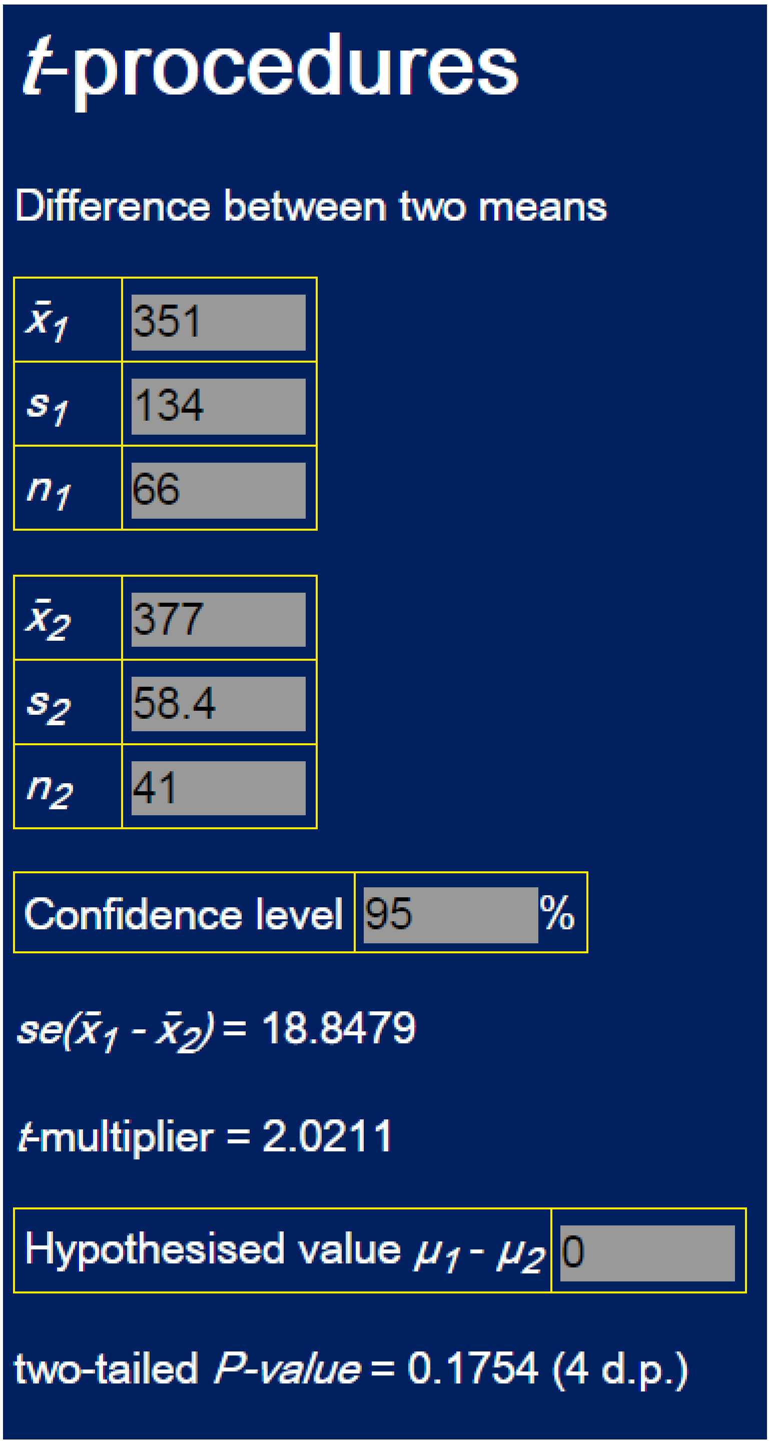
Methamphetamine consumption in milligrams per day per 1000 people was estimated for each plant's catchment area from the samples of wastewater collected at each plant.

Let:

μ_1 be the underlying mean methamphetamine consumption in milligrams per day per 1000 people in the wastewater treatment plant 1 catchment area between 2 May and 18 July, 2014,

and

μ_2 be the underlying mean methamphetamine consumption in milligrams per day per 1000 people in the wastewater treatment plant 2 catchment area between 2 May and 18 July, 2014.



t-procedures

Difference between two means

$\bar{x}_1$	351
s_1	134
n_1	66

$\bar{x}_2$	377
s_2	58.4
n_2	41

Confidence level 95 %

$se(\bar{x}_1 - \bar{x}_2) = 18.8479$

t-multiplier = 2.0211

Hypothesised value $\mu_1 - \mu_2$ 0

two-tailed *P-value* = 0.1754 (4 d.p.)

Figure 1: Screen-shot of the *t*-procedures tool

A two-tailed *t*-test for no difference between μ_1 and μ_2 was conducted. The *t*-procedures tool with the means, standard deviations and sample sizes for the estimated methamphetamine consumption for the two wastewater catchment areas is shown to the right where $\bar{x}_1$ is the sample mean for the plant 1 catchment area and $\bar{x}_2$ is the sample mean for the plant 2 catchment area.

J. Which **one** of the following is a **correct** pair of hypotheses for this *t*-test?

(1) $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 > 0$

(2) $H_0 : p_1 - p_2 = 0$ versus $H_1 : p_1 - p_2 \neq 0$

(3) $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$

(4) $H_0 : \overline{x}_1 - \overline{x}_2 = 0$ versus $H_1 : \overline{x}_1 - \overline{x}_2 \neq 0$

(5) $H_0 : p_D - p_S > 0$ versus $H_1 : p_D - p_S = 0$

K. The value of the *t*-test statistic, t_0, is approximately:

(1) −1.38

(2) −26

(3) 2.02

(4) 26

(5) 1.38

L. To the nearest whole number, which **one** of the following is the **correct** margin of error for the 95% confidence interval for $\mu_1 - \mu_2$?

(1) 74

(2) 19

(3) 37

(4) 38

(5) 76

M. If the confidence level in the *t*-procedures tool was changed from 95% to 99%, which **one** of the following statements would be **true**?

(1) The standard error would increase.

(2) The *t*-multiplier would increase.

(3) The *P-value* would change.

(4) The confidence interval would be narrower.

(5) The margin of error would decrease by 4%.

N. Which **one** of the following is a **correct** interpretation of the *P-value* shown in Figure 1, page 17?

(1) There is a 0.1754 chance that the null hypothesis is true.

(2) If the alternative hypothesis was true, the probability of getting a test statistic at least as extreme as the observed test statistic is 0.1754.

(3) There is a 0.1754 chance that the alternative hypothesis is true.

(4) If the null hypothesis was true, the probability of getting a test statistic at least as extreme as the observed test statistic is 0.1754.

(5) The probability of getting an estimate of at least −26 is 0.1754.

Practical significance versus ***Statistical significance***

Statistical significance

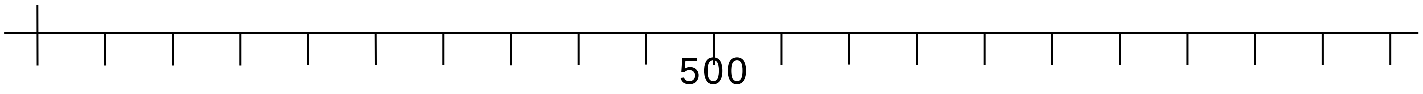
- Relates to having evidence of the **existence** of an effect or difference.
- Determined by examining the ***P*-value** of your significance test.
- To be statistically significant at the 5% level, the *P-value* must be greater than / less than 0.05 (5%).

Practical significance

- Depends on the **size** of the effect or difference.
- Determined by examining the **confidence interval** in relation to the context of the question/s (i.e. the story).

For example: Jam jar example
(please see Lecture Workbook, Chapter 7, page 19):

$$H_0\colon \mu = 500$$
vs
$$H_1\colon \mu \neq 500$$



The link between the* P-value *and the confidence interval

Recall that a confidence interval for a parameter gives a range of plausible (believable) values for the unknown true parameter value.

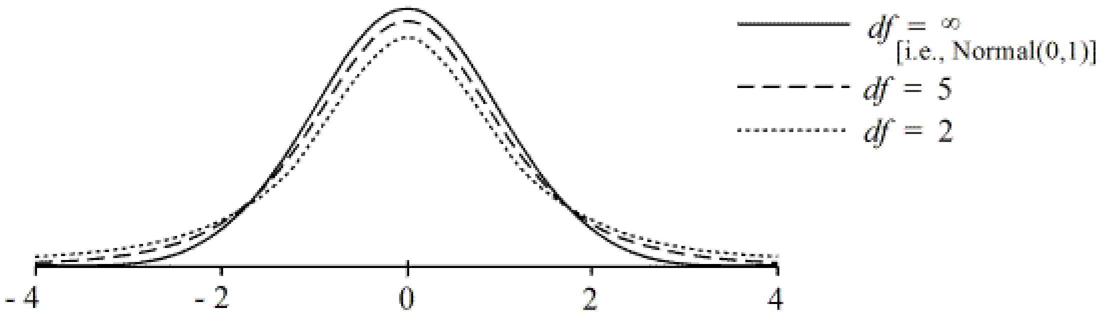
If a 2-tailed test has a *P-value* less than 5% then the test is significant at the 5% level of significance and the hypothesised value is not plausible (not believable) and it will / will not be in the 95% confidence interval. Conversely, if the hypothesised value is not in the 95% confidence interval it is not a plausible value and so the test is significant at the 5% level of significance and the *P-value* will be less than / greater than 5%.

If a 2-tailed test has a *P-value* greater than 5% then the test is not significant at the 5% level of significance and the hypothesised value is plausible (is believable) and so it will / will not be in the 95% confidence interval. Conversely, if the hypothesised value is in the 95% confidence interval it is a plausible value and so H_0 will be not rejected at the 5% level and the *P-value* will be less than / greater than 5%.

Note: The same relationship applies to 90% confidence intervals and *P-values* less than 10% (tests at the 10% level of significance), or 99% confidence intervals and *P-values* less than 1% (tests at the 1% level).

Student's *t*-distribution (background understanding)

- ✓ The parameter is the degrees of freedom, *df*.
- ✓ Smooth symmetric, bell-shaped curve centred at 0 like the Standard Normal distribution [Z ~ Normal (μ = 0, σ = 1)] but it's more variable (it's 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* = ∞) and Normal (0,1) are the same distribution.

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

(1) In a *t*-test for no difference between two means, being able to demonstrate that the difference was of practical significance (importance) would almost always imply statistical significance.

(2) In hypothesis testing, large samples can lead to small *P-values* without the results having any practical significance (importance).

(3) In hypothesis testing, statistical significance does not imply practical significance (importance).

(4) In a hypothesis test for no difference between two means, a very small *P-value* always indicates a very large difference in the means.

(5) In hypothesis testing, a non-significant test result does not imply that the null hypothesis is true.

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

(1) The Student's *t*-distribution has tails which become fatter as the degrees of freedom increase.

(2) An estimate is more precise if it has more variability.

(3) Student(*df* = ∞) and Normal(μ = 0, σ = 1) are identical distributions.

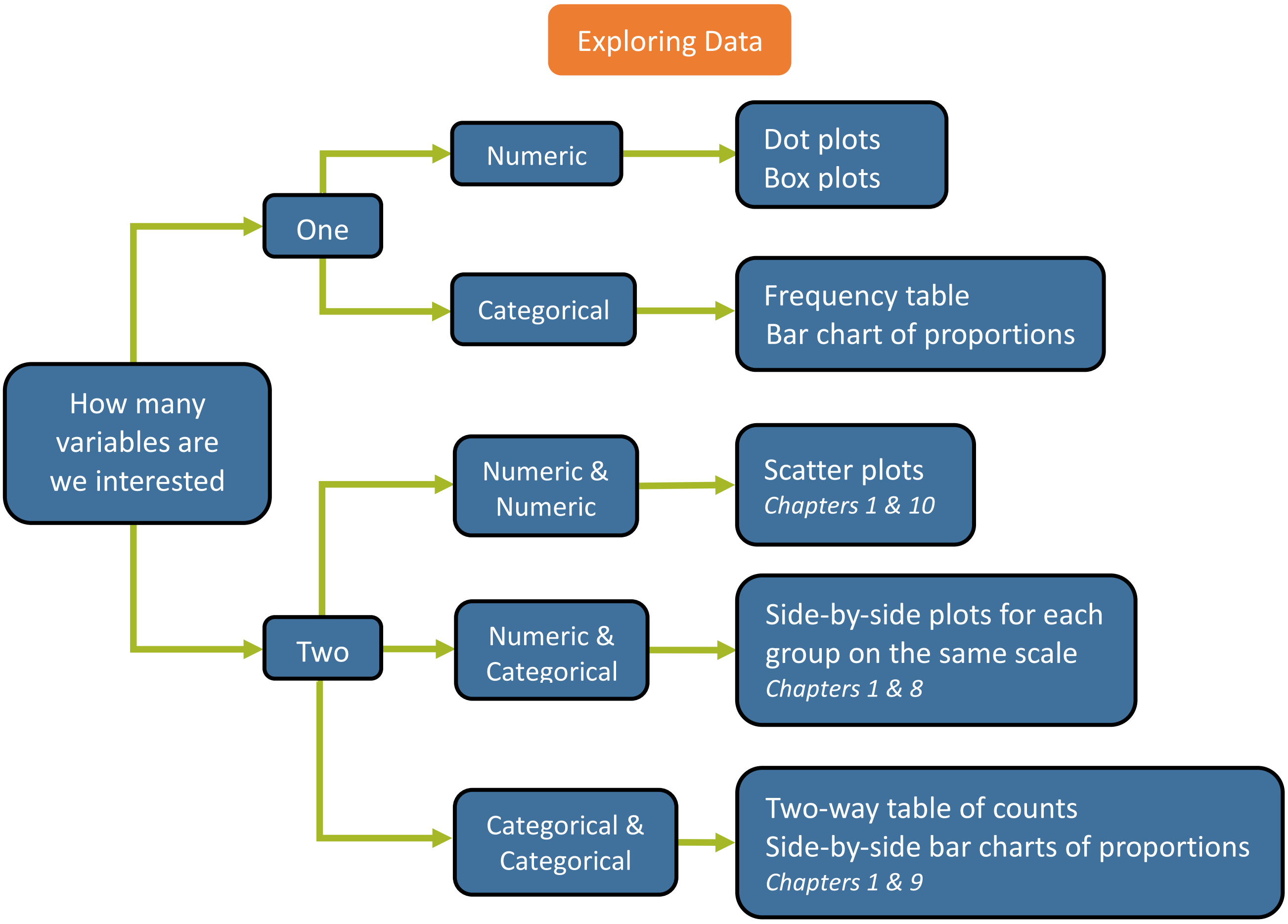
(4) If greater confidence in a confidence interval calculated from our data is desired, then a narrower interval needs to be used.

(5) A parameter is a numerical characteristic which can be calculated from a sample.

One, Two and Three or More Means; Using SPSS

We will now use SPSS to consider:

- Two independent samples
- Paired data comparisons
- One sample
- More than 2 samples



Code	Form of analysis
A	One sample *t*-test on a mean
B	One sample *t*-test on a proportion
C	One sample *t*-test on a mean of differences
D	Two sample t-test on a difference between two means
E	*t*-test on a difference between two proportions
F	One-way analysis of variance *F*-test

Checking the Normality assumption

✓ You should make sure data don't show separation into clusters or have a multi-modal nature and then apply the 15-40 rule as follows:

Sample Size Guidelines – "15 – 40 Guide"

Small (total $n \le 15$ or so)	Medium ($15 <$ total $n < 40$)	Large (total $n \ge 40$ or so)
no outliers	no outliers	no gross outliers
at most, slight skewness	not strongly skewed	data may be strongly skewed

✓ The one sample/paired *t*-test and CIs are reasonably robust against non-Normality, but sensitive to outliers in small-medium samples. The two sample *t*-test & CIs and the *F*-test & Tukey pairwise CIs are very robust against non-Normality. The *F*-test is reasonably robust with respect to the standard deviations assumption, but the Tukey pairwise CIs are not.

Q. Which **one** of the following statements about the validity of confidence intervals of the form sample mean ± *t* standard errors is **false**?
(1) It is critical that the sample is random.
(2) It is critical that the distribution being sampled is Normal.
(3) It is critical that the observations come from the same distribution.
(4) Outliers and clusters of data can invalidate confidence intervals.
(5) It is critical that observations are independent.

R. Which **one** of the following statements about *t*-tests is **false**?
(1) *t*-tests may not be valid if there are outliers present and the sample is not large.
(2) *t*-tests may not be valid when the data show clustering.
(3) In general, *t*-tests are not robust against the Normality assumption.
(4) *t*-tests will generally work well for any large sample.
(5) *t*-tests may not be valid if the data are clearly skewed and the sample is not large.

S. Before conducting formal tests, one should look at plots of the data. Which **one** of the following statements is **false**?
(1) Plots may highlight strange or interesting features of the data which cannot be seen in a formal test.
(2) Summaries of the important features of the data can often be obtained from looking at plots.
(3) Plots are used to check the validity of the assumptions for the formal tests.
(4) Inferences, i.e. conclusions about the population, drawn from plots do not need to be verified by formal tests.
(5) Additional points of interest are often suggested by plots.

2-sample *t*-tests

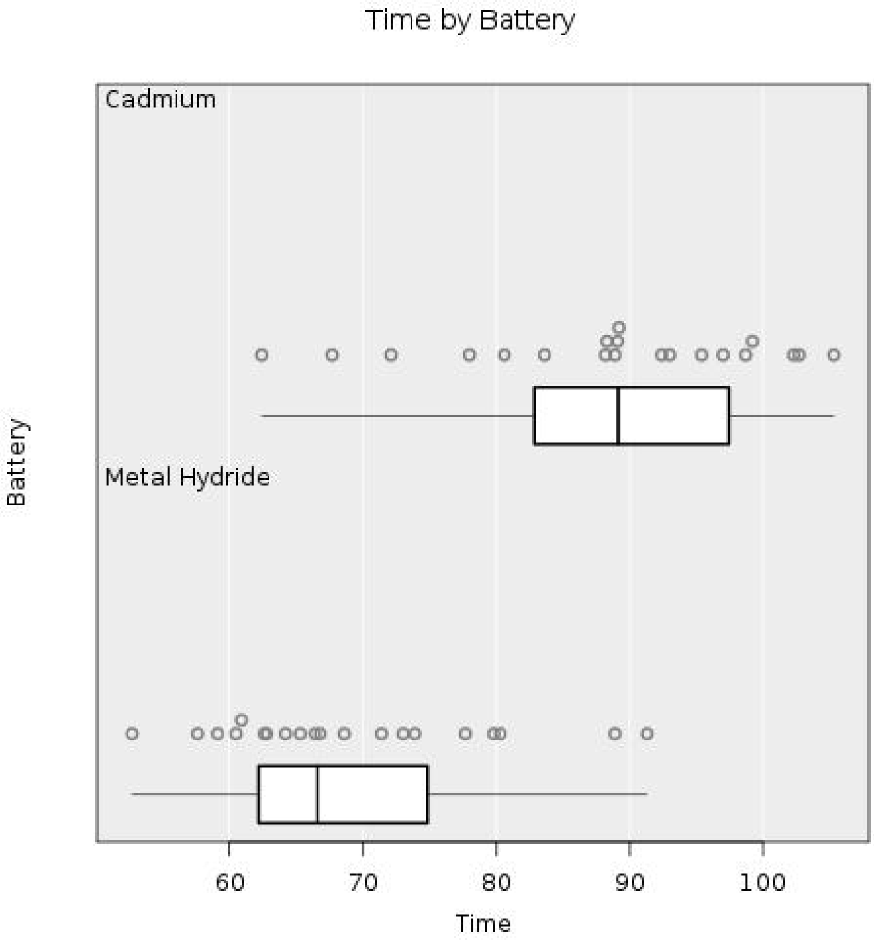
All four datasets used on pages 17 to 26 are available here:

www.tinyURL.com/stats-HTS

as is a hand-out on how to generate the output in SPSS.

- ✓ Two independent samples
- ✓ Parameter = $\mu_1 - \mu_2$
- ✓ **Hypotheses**: H_0: $\mu_1 - \mu_2 = 0$ vs H_1: $\mu_1 - \mu_2 \neq 0$
- ✓ **Assumptions** for 2-sample *t*-tests

1. Observations *within* the samples are **independent** – *CRITICAL!*
2. The two samples/groups are **independent**, i.e., observations *between* samples/groups are independent of each other – *CRITICAL!*
3. **The Normality Assumption**: The population or underlying distributions are Normal. No clusters or multi-modes allowed.



Example: Which cell phone battery is better? A random sample of 40 cellphones of the same make and model were chosen. Half of the cellphones were randomly selected to have a nickel-cadmium battery put in them and the rest had a nickel-metal hydride battery. The talk time (in minutes) before the batteries needed to be recharged was recorded.

Cadmium cellphone batteries last longer on average than metal hydride cellphone batteries. The talk times for both battery types have similar variability. The talk times for cadmium cellphone batteries are `slightly / moderately / strongly negatively (left) skewed / reasonably symmetric / positively (right)` skewed while the talk times for metal hydride cellphone batteries are `slightly / moderately / strongly` `negatively (left) skewed / reasonably symmetric / positively (right) skewed`.

Let μ_C be the underlying mean talk time for cadmium cellphone batteries and μ_M be the underlying mean talk time for metal hydride cellphone batteries, i.e. $\mu_C - \mu_M$ is the difference in underlying mean talk time between the 2 types of battery.

H_0: $\mu_C - \mu_M =$ vs H_1: $\mu_C - \mu_M \neq$

T-Test

Group Statistics

	Battery	N	Mean	Std. Deviation	Std. Error Mean
Time	Cadmium	20	88.7050	11.76066	2.62976
	Metal Hydride	20	69.1900	10.30528	2.30433

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Time	Equal variances assumed	.079	.780	5.581	38	.000	19.51500	3.49651	12.43668	26.59332
	Equal variances not assumed			5.581	37.356	.000	19.51500	3.49651	12.43267	26.59733

We have `no / weak / some / strong / very strong` evidence (*P-value* = .000) against there being no difference in the underlying mean talk time between the two types of cellphone batteries, that is, we have `no / weak / some / strong / very strong` evidence of a difference in the underlying mean talk time between the two types of cellphone battery.

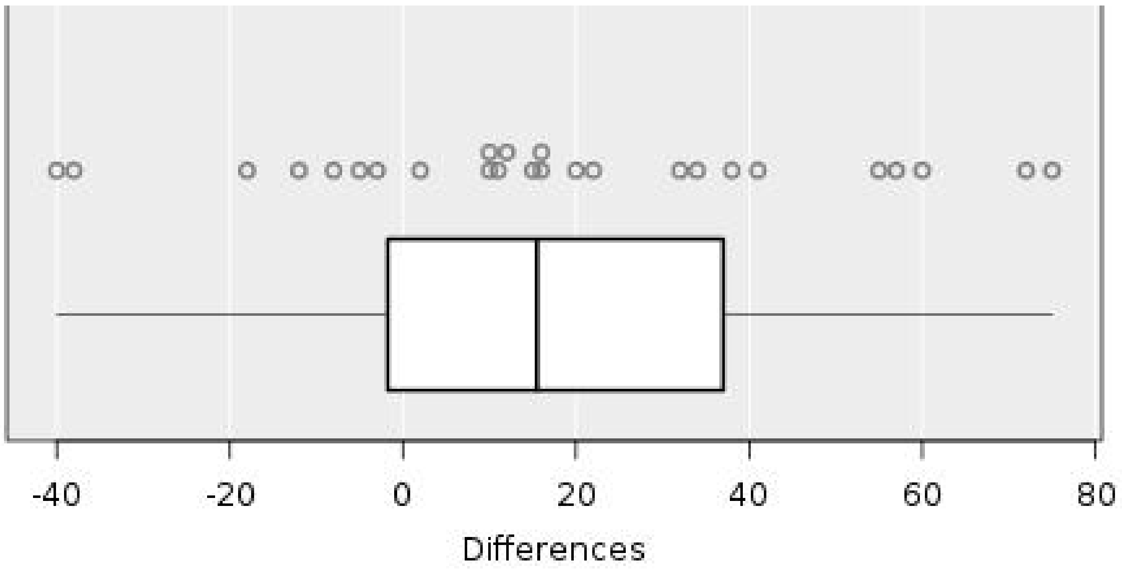
With 95% confidence, we estimate that, the underlying mean talk time for a cadmium cellphone battery is, on average, somewhere between ____________ and ____________ minutes `less / more` than that of a metal hydride battery.

There are no doubts about the validity of the *t*-test. Because a random sample of 40 cellphones of the same make and model was taken, there are no concerns about the independence assumptions. In this two independent samples situation, $n_1 + n_2 = 40$ and, with respect to the Normality assumption, *t*-procedures work well for this number of observations despite the talk times being somewhat skewed.

Paired data: use a paired data *t*-test

- ✓ 1 sample (group) of data
- ✓ Two measurements taken on each experimental unit
- ✓ With related or paired data we analyse the <u>differences</u> and use a 1-sample *t*-test, i.e. we treat the differences as a single sample.
- ✓ Parameter = μ_{diff}
- ✓ **Hypotheses**: H_0: $\mu_{\text{diff}} = 0$
  vs H_1: $\mu_{\text{diff}} \neq 0$
- ✓ **Assumptions** for the paired-*t*-test:
  1. Observations (differences) *within* the sample are **independent** – *CRITICAL!*
  2. **The Normality Assumption**: The population or underlying distribution of the differences is Normal. No clusters or multi-modes allowed.

Example: A market research company is interested in which of two similar electric shavers, model A or model B, is preferred by consumers. 26 men who daily use an electric shaver, but not one of the models of interest are randomly selected to participate in the study. Half the men were randomly allocated to use model A one morning followed by model B the next morning whilst the order was reversed for the remaining men. After every shave, each man completed a questionnaire rating his satisfaction with the shaver. Satisfaction was measured as a score based on the answers to the questionnaire and is given in a range from 1 to 100. (Larger scores indicate greater satisfaction).

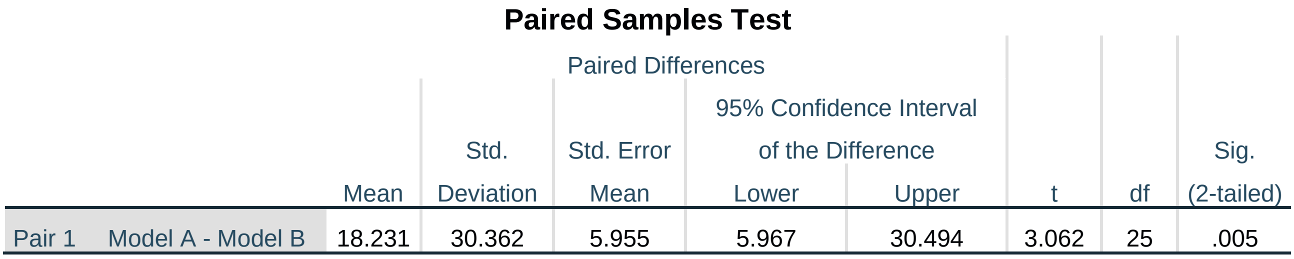


The difference in the satisfaction score is centred above / below zero suggesting that the average satisfaction score for model A is higher / lower than that for model B. The difference in average satisfaction score is slightly / moderately / strongly negatively (left) skewed / reasonably symmetric / positively (right) skewed.

Let μ_{diff} be the underlying mean difference in the average satisfaction score for the model A electric shaver and the average satisfaction score for the model B electric shaver. (Note: Diff = model A – model B)

H_0: $\mu_{\text{diff}} =$ vs H_1: $\mu_{\text{diff}} \neq$

T-Test



Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Model A - Model B	18.231	30.362	5.955	5.967	30.494	3.062	25	.005

We have no / weak / some / strong / very strong evidence (*P-value* = .005) that there is a shaver effect on the average satisfaction score for the model A electric shaver and the average satisfaction score for the model B electric shaver.

With 95% confidence, we estimate that, the average satisfaction score for the model A electric shaver is, on average, somewhere between ____________ and ____________ units less / greater than that for the average satisfaction score for the model B electric shaver.

There are no doubts about the validity of the *t*-test. Because a random sample of 26 men was taken, there are no concerns about the independence assumption. In this paired data situation, *n* = 26 and, with respect to the Normality assumption, *t*-procedures work well for this number of observations as the differences are reasonably symmetric.

1-sample: use a One Sample *t*-test

✓ 1 sample (group) of data

✓ Parameter = μ

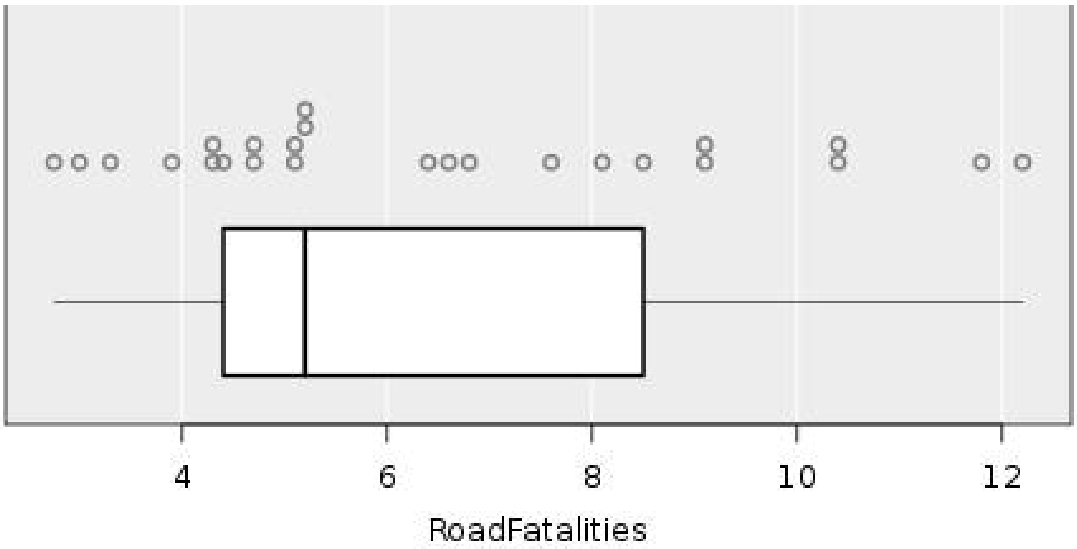
✓ **Hypotheses**: H_0: $\mu = \mu_0$

vs H_1: $\mu \neq \mu_0$

✓ **Assumptions** for 1-sample *t*-tests

1. Observations *within* the sample are **independent** – *CRITICAL!*
2. **The Normality Assumption**: The population or underlying distribution is Normal. No clusters or multi-modes allowed.

Example: Of interest is whether the most recent road fatalities per capita per year for 25 randomly selected countries has changed from the historical average of 10.5 per 100,000 inhabitants per year (for the same 25 countries in the mid-eighties). The data was collected by the World Health Organization (WHO).



The annual road fatalities per 100,000 inhabitants is centred at about 5. Road fatalities ranges from roughly 3 to 12 per 100,000 inhabitants and is `slightly / strongly negatively (left) skewed / reasonably symmetric / positively (right) skewed.`

Let μ be the underlying mean annual road fatalities per 100,000 inhabitants

H_0: $\mu =$ vs H_1: $\mu \neq$

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Road fatalities per 100,000 inhabitants per year	25	6.5180	2.76567	.55313

One-Sample Test

	Test Value = 10.5					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Road fatalities per 100,000 inhabitants per year	-7.199	24	.000	-3.98200	-5.1236	-2.8404

We have no / weak / some / strong / very strong evidence (*P-value* = 0.000) that the underlying mean annual road fatalities per 100,000 inhabitants has changed compared to the historical average of 10.5 per 100,000 inhabitants per year.

With 95% confidence, we estimate that the underlying mean annual road fatalities is somewhere between ___________ and ___________ per 100,000 inhabitants.

There are no doubts about the validity of the *t*-test. Because a random sample of 25 countries was taken, there are no concerns about the independence assumption. In this one-sample situation, n = 25 and, with respect to the Normality assumption, *t*-procedures work well for this number of observations as the data is slightly positively (right) skewed.

T. Thirty observations of the relative return of over-the-counter stocks bought in the week of the 9th to the 13th of May, 1994 are given below.

-0.2940	-0.1092	-0.1053	-0.0707	-0.0563
-0.0541	-0.0423	-0.0398	-0.0396	-0.0390
-0.0381	-0.0323	-0.0221	-0.0169	-0.0139
-0.0081	0.0038	0.0057	0.0156	0.0172
0.0182	0.0192	0.0423	0.0459	0.0476
0.0667	0.0714	0.0780	0.1176	0.1224

Note: $\overline{x}$ = -0.01030 and *s* = 0.0786

Investors would like to know how the market performed. One measure of market performance is the mean relative return for the week.

A 95% confidence interval for the mean relative return is [-0.040, 0.019]. Which **one** of the following statements is **false**?

(1) The *P*-value for testing $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ is larger than 0.05.

(2) There is evidence, at the 5% level, to believe that the mean return is different from zero.

(3) A 99% confidence interval for the mean return would be wider than the 95% confidence interval.

(4) It is plausible that the mean relative return is zero.

(5) An estimate of the mean relative return is –0.01030.

U. Does too much sleep impair intellectual performance? Taub *et al*. (1971) examined this commonly held belief by comparing the performance of 12 subjects on the mornings following (1) two normal nights' sleep and (2) two nights of "extended sleep". In the morning they were given a number of tests of ability to think quickly and clearly. One test was for vigilance where the lower the score, the more vigilant the subject. The following data was collected:

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Normal Sleep	8	9	14	4	12	11	3	26	3	11	10	1
Extended Sleep	8	9	15	2	21	16	9	38	10	11	16	41

To see if the data supports the view that too much sleep can be bad for you, we would test which of the following hypotheses?

(1) $H_0 : \overline{x}_1 - \overline{x}_2 = 0$ versus $H_1 : \overline{x}_1 - \overline{x}_2 < 0$

(2) $H_0 : \overline{x}_{Diff} = 0$ versus $H_1 : \overline{x}_{Diff} \neq 0$

(3) $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$

(4) $H_0 : p_1 - p_2 = 0$ versus $H_1 : p_1 - p_2 \neq 0$

(5) $H_0 : \mu_{diff} = 0$ versus $H_1 : \mu_{diff} \neq 0$

V. In order to study the harmful effects of DDT poisoning, the pesticide was fed to 6 randomly chosen rats out of a group of 12 rats. The other 6 unpoisoned rats comprised of the control group. The following data gives measurements of the amount of tremor detected in the bodies of each rat after the experiment. The more tremor, the more harmful.

Poisoned group:	12.207,	16.869,	25.050,	22.429,	8.456,	20.589
Control group:	11.074,	12.064,	9.351,	6.642,	9.686,	8.182

We wish to test

(1) $H_0 : \mu_{diff} = 0$ versus $H_1 : \mu_{diff} \neq 0$

(2) $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$

(3) $H_0 : \overline{x}_1 - \overline{x}_2 = 0$ versus $H_1 : \overline{x}_1 - \overline{x}_2 \neq 0$

(4) $H_0 : p_1 - p_2 = 0$ versus $H_1 : p_1 - p_2 \neq 0$

(5) $H_0 : \overline{x}_{Diff} = 0$ versus $H_1 : \overline{x}_{Diff} \neq 0$

Question W refers to the following information.

The heights (in cm) of the carapaces (shells) of a sample of 48 painted turtles were recorded. Shown below is a stem-and-leaf plot of this data set.

Units: 3|5 = 35cm

```
3 | 55778888999
4 | 0000111222344
4 | 55566789
5 | 0111113
5 | 567
6 | 01233
6 | 7
```

Figure: Stem-and-leaf plot of carapace height of painted turtles.

W. Based on this sample of size 48, a 95% confidence interval for the underlying mean carapace height of all painted turtles is 44.0cm to 48.8cm. The number of painted turtle carapace heights we would need to sample in order to halve the width of this interval is, approximately:

(1) 24
(2) 192
(3) 12
(4) 96
(5) 7

3 or more samples ⇒ *F*-test for 1-way ANOVA

✓ Three or more independent samples (groups)

✓ **Hypotheses:** H_0 : all the underlying population means are the same

i.e. $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

vs H_1 : at least 2 of the underlying population means are not the same

Note: SHOULD NOT be written using symbols

✓ **Assumptions** for 1-way ANOVA

1. Observations *within* the samples are **independent** – *CRITICAL!*
2. The samples/groups are **independent**, i.e., observations *between* samples/groups are independent of each other – *CRITICAL!*
3. **The Normality Assumption**: The population or underlying distributions are Normal. No clusters or multi-modes allowed.

   Check by looking at plots of each sample and consider sample sizes. Plots should be unimodal and not too skewed for the n_{tot} you have.
4. **Equality of Standard Deviations**

   The standard deviations of the underlying distributions (or populations) are equal.

   Check by using the ratio (fraction) $\dfrac{\textbf{\textit{largest sd}}}{\textbf{\textit{smallest sd}}} < 2$ as a guide

✓ Data not Normal and/or standard deviations not equal? Don't use the *F*-test!

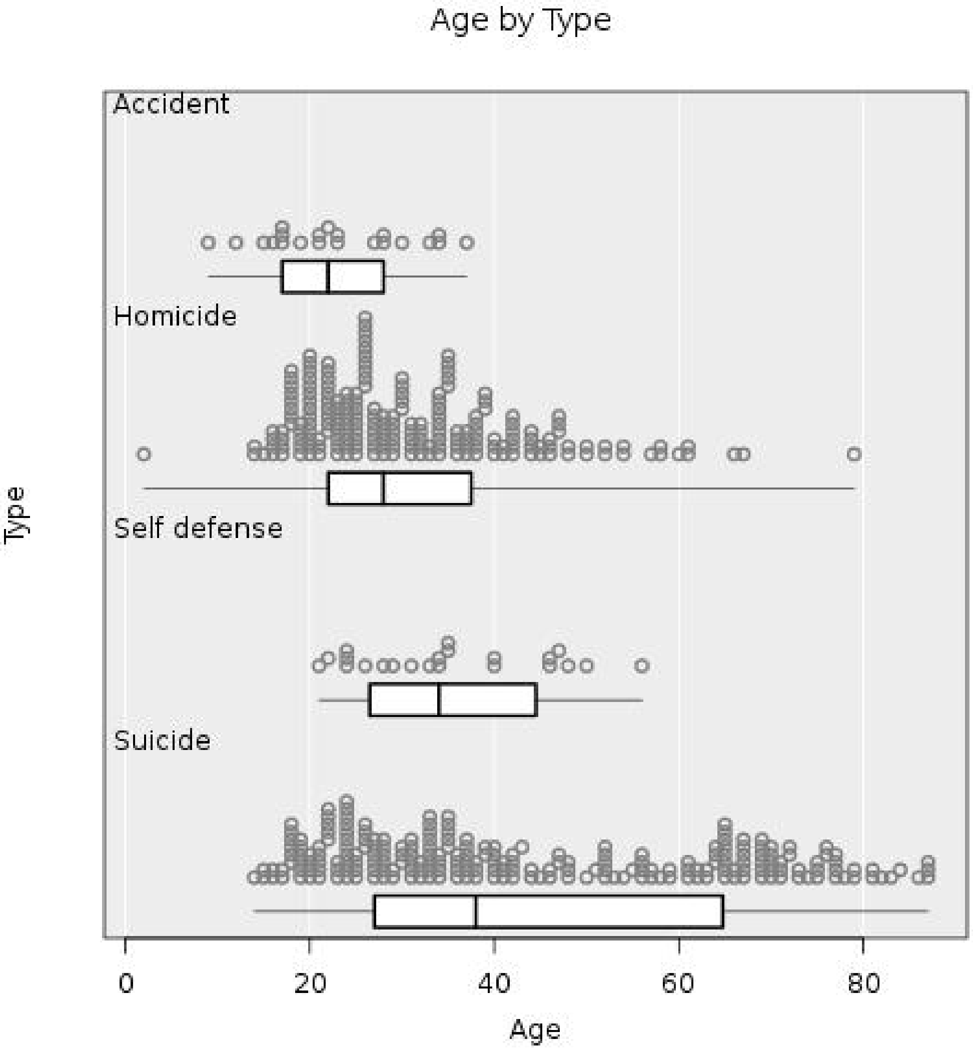
✓ **One-way ANOVA Table**
(SPSS will have the numbers laid out in this way)

	Sum of Squares (SS)	*df*	Mean Square	F	Sig.
Between Groups (BG)	BGSS	$df_1 = k - 1$	s_B^2	$f_0 = \dfrac{s_B^2}{s_W^2}$	Pr($F \geq f_0$)
Within Groups (WG)	WGSS	$df_2 = n_{\text{tot}} - k$	s_W^2		
Total (T)	TSS	$df_1 + df_2$			

where F is the *F*-statistic, $f_0 = \dfrac{s_B^2}{s_W^2}$ where s_B^2 = the **between** group variation

and s_W^2 = the **within** group variation

Example: In 1989, 464 people were killed by a gun in the United States in a single week in May. These deaths have been grouped into four classes: **Accident**; **Homicide**; **Self defense**; and; **Suicide**. The age was also recorded for each person.



The age for suicides is centred highest while the age for accidents is centred lowest. The age for accidents is the least variable. The accident ages are reasonably symmetric while the self defense group is `slightly / moderately negatively (left) / positively (right)` skewed while the other two groups (homicide and suicide) are both `slightly / moderately / strongly negatively (left) skewed / positively (right)` skewed.

Assumptions

The observations within each type of death are independent, that is, the samples are random.

The samples of age for each type of death are independent of each other.

The underlying distributions (populations) of age for each type of death are Normally distributed. That is, the data for each type of death come from a Normal distribution.

The standard deviations of the underlying distributions (populations) of age for each type of death are equal.

There `are / are no` concerns about the validity of the *F*-test.

Descriptives

Age (in years)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Accident	21	23.00	7.797	1.702	19.45	26.55	9	37
Homicide	195	30.86	11.879	.851	29.18	32.54	2	79
Self defense	22	35.14	10.204	2.176	30.61	39.66	21	56
Suicide	222	44.25	20.418	1.370	41.55	46.95	14	87
Total	460	37.17	17.841	.832	35.53	38.80	2	87

Let μ_1 be the underlying mean age for accidents, and similarly define, μ_2, μ_3 and μ_4 for homicide, self defense and suicide respectively.

H_0 : all the underlying population mean ages are the same

i.e. $\mu_1 = \mu_2 = \mu_3 = \mu_4$

vs H_1 : the underlying population mean age is different for at least two of the groups (at least two types of death)

ANOVA

Age (in years)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	23188.885	3	7729.628	28.676	.000
Within Groups	122915.226	456	269.551		
Total	146104.111	459			

We have `no / weak / some / strong / very strong` evidence (*P-value* = .000) against there being no difference in the underlying mean age for the four groups, that is, we have `no / weak / some / strong / very strong` evidence that the underlying mean age is different for at least two of the groups (at least two types of death).

Multiple Comparisons

Dependent Variable: Age (in years)

Tukey HSD

(I) Type of death	(J) Type of death	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Accident	Homicide	-7.862	3.771	.160	-17.58	1.86
	Self defense	-12.136	5.009	.074	-25.05	.78
	Suicide	-21.248*	3.748	.000	-30.91	-11.58
Homicide	Accident	7.862	3.771	.160	-1.86	17.58
	Self defense	-4.275	3.693	.654	-13.80	5.25
	Suicide	-13.386*	1.611	.000	-17.54	-9.23
Self defense	Accident	12.136	5.009	.074	-.78	25.05
	Homicide	4.275	3.693	.654	-5.25	13.80
	Suicide	-9.111	3.670	.064	-18.57	.35
Suicide	Accident	21.248*	3.748	.000	11.58	30.91
	Homicide	13.386*	1.611	.000	9.23	17.54
	Self defense	9.111	3.670	.064	-.35	18.57

*. The mean difference is significant at the 0.05 level.

Assuming the Tukey pairwise comparisons are valid:

We have very strong evidence (P-value = 0.000) that the underlying mean age for suicides is less than / greater than that for accidents and homicides.

With 95% confidence, we estimate, that the underlying mean age for suicides is somewhere between ____________ and ____________ years less than / greater than that for accidents.

With 95% confidence, we estimate, that the underlying mean age for suicides is somewhere between ____________ and ____________ years less than / greater than that for homicides.

We cannot say that suicides have the lowest / highest age on average because there is not a significant difference between suicide and self defense, that is, with 95% confidence, we estimate, that the underlying mean age for suicides is somewhere between ____________ years less than / greater than and ____________ years less than / greater than that for self defense.

Appendix D: Fruitfly Data

Questions X to **BB** refer to the information in this appendix.

It had already been established that increased sexual activity decreases the number of days for which female fruitflies live. Researchers Hanley and Shapiro (1994) designed a study to see if the same were true for male fruitflies. The sexual activity of male fruitflies was manipulated by supplying individual male fruitflies with either one or eight receptive virgin females per day. The lifespan of these males (the number of days they lived for) was compared with the lifespan of males that were supplied daily with one or eight newly inseminated females. Newly inseminated females are not receptive because they will not re-mate for at least two days. There was also a group of males kept with no females.

Thus there were five groups in total and 125 male fruitflies were randomly assigned to one of these five groups. This meant that there were 25 male fruitflies in each group.

The five groups were:

GP1: Male kept alone

GP2: Male supplied daily with 1 newly inseminated female unwilling to mate

GP3: Male supplied daily with 1 receptive virgin female willing to mate

GP4: Male supplied daily with 8 newly inseminated females unwilling to mate

GP5: Male supplied daily with 8 receptive virgin females willing to mate

The *underlying* mean lifespan for **GP1** (μ_{GP1}) is defined to be the mean lifespan (in days) if all of the 125 male fruitflies used in the study had been kept alone. The *underlying* mean lifespans for **GP2**, **GP3**, **GP4** and **GP5** (μ_{GP2}, μ_{GP3}, μ_{GP4}, and μ_{GP5}) are defined in a similar manner.

An *F*-test for one-way analysis of variance was conducted to compare the lifespan of the fruitflies in the different groups.

Dot/box plots of the lifespan for each group are shown in Figure 9, page 36. Output from the *F*-test is shown in Table 6, page 36.

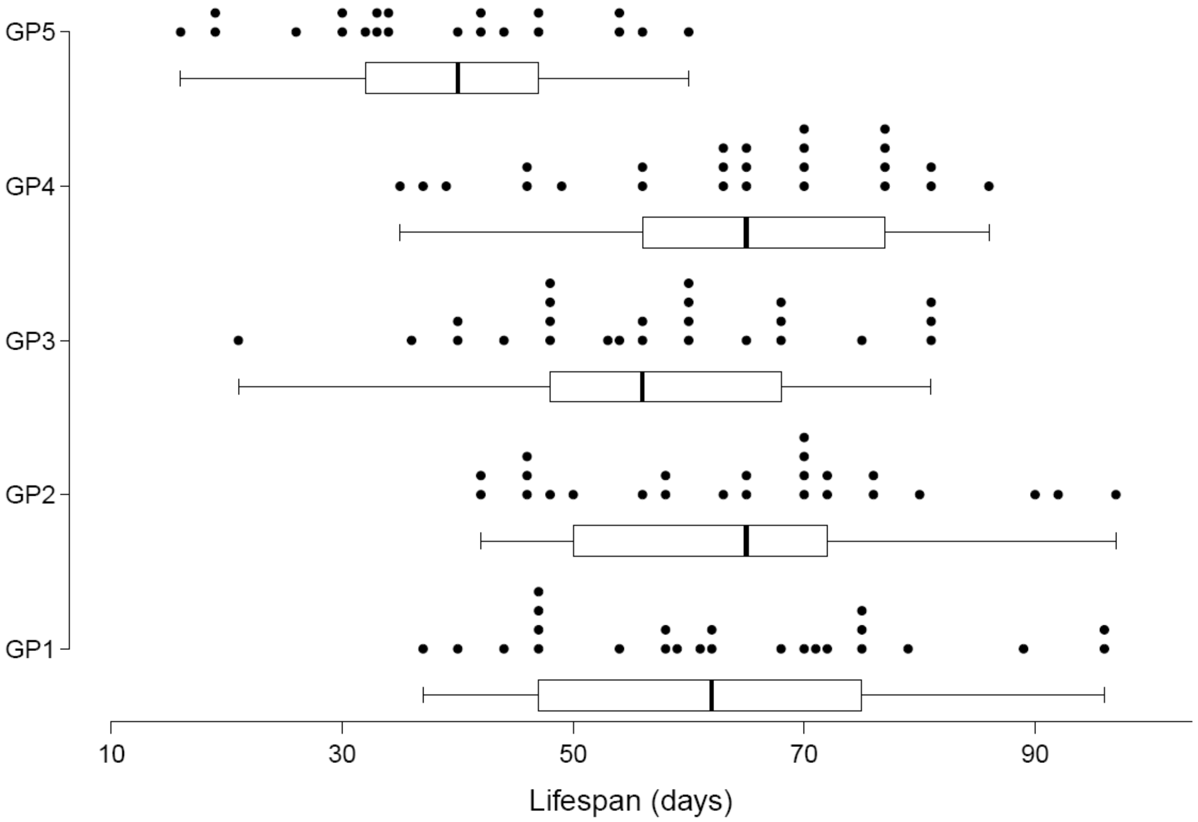
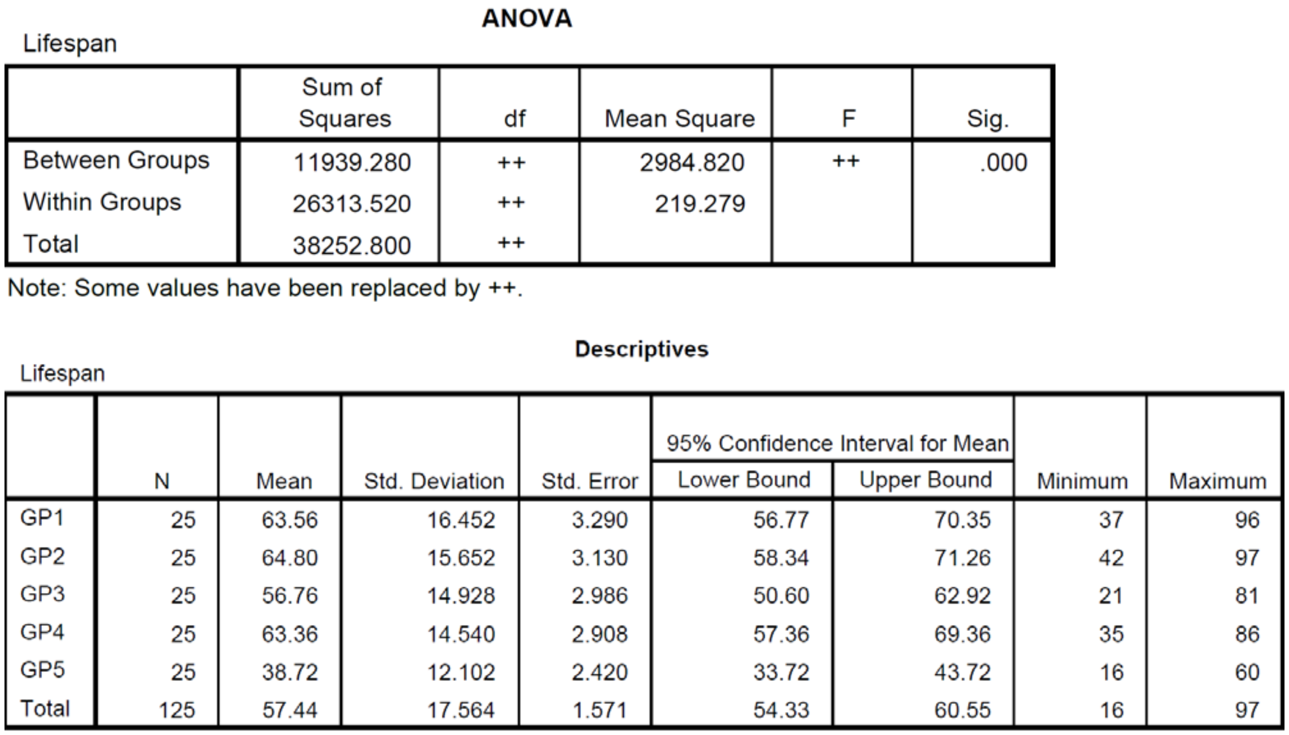


Figure 9: Lifespan of male fruitflies by group



ANOVA

Lifespan

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	11939.280	++	2984.820	++	.000
Within Groups	26313.520	++	219.279		
Total	38252.800	++			

Note: Some values have been replaced by ++.

Descriptives

Lifespan

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
GP1	25	63.56	16.452	3.290	56.77	70.35	37	96
GP2	25	64.80	15.652	3.130	58.34	71.26	42	97
GP3	25	56.76	14.928	2.986	50.60	62.92	21	81
GP4	25	63.36	14.540	2.908	57.36	69.36	35	86
GP5	25	38.72	12.102	2.420	33.72	43.72	16	60
Total	125	57.44	17.564	1.571	54.33	60.55	16	97

Table 6: *F*-test output

Questions X to **BB** refer to the information in **Appendix D**, pages 35 and 36.

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

(1) The response variable is lifespan (in days).

(2) The groups GP1, GP2 and GP4 can be viewed as control groups.

(3) This study is an experiment with 5 different ‘treatment’ levels.

(4) The units in this study are 125 male fruitflies.

(5) There is blocking in this study design with the female fruitflies blocked as either ‘newly inseminated’ or ‘receptive virgin’.

Y. Which **one** of following is a **correct** set of hypotheses for this *F*-test?

(1) H_0: All five observed group means have the same value.
H_1: At least two observed group means have different values.

(2) H_0: The underlying means are the same for all five groups.
H_1: The underlying means are not all the same for the five groups.

(3) H_0: The underlying means are the same for all five groups.
H_1: Each of the five groups has a different underlying mean.

(4) H_0: The underlying means are not all the same for the five groups.
H_1: The underlying means are the same for all five groups.

(5) H_0: No two groups have the same underlying mean.
H_1: The underlying means are the same for all five groups.

Z. Which **one** of the following statements about this *F*-test is **true**?

(1) There is no concern about the validity of the *F*-test with regards to the Normality assumption because the data does not suggest clusters nor does it appear strongly skewed.

(2) There is concern about the validity of the *F*-test with regards to the equal underlying standard deviations assumption.

(3) We should be wary about using an *F*-test on these data because the observations within each group are not independent.

(4) It is not appropriate to use an *F*-test on these data because the number of male fruitflies in each group is not greater than or equal to 30.

(5) We should be wary about using an *F*-test on these data because the groups are not independent.

Questions AA and **BB** assume that the use of the *F*-test is appropriate.

(Note that this may not be true.)

Refer to information given in Table 6, page 36, to answer these questions.

AA. The value of the *F*-test statistic, f_0, is approximately:

(1) 0.03

(2) 13.61

(3) 1.45

(4) 0.45

(5) 0.69

BB. The degrees of freedom for this *F*-test are:

(1) $df_1 = 5$ $df_2 = 20$

(2) $df_1 = 4$ $df_2 = 121$

(3) $df_1 = 4$ $df_2 = 120$

(4) $df_1 = 4$ $df_2 = 20$

(5) $df_1 = 5$ $df_2 = 120$

CC. Analysis of variance (ANOVA) is (select **one** only):

(1) an overall test of no difference between sample variances.

(2) an *F*-test of no difference between population means.

(3) an overall test of no difference between population variances.

(4) an *F*-test for the equality of population variances.

(5) an *F*-test of no difference between sample means.

DD. Which **one** of the following statements about the one-way analysis of variance *F*-test is **false**?

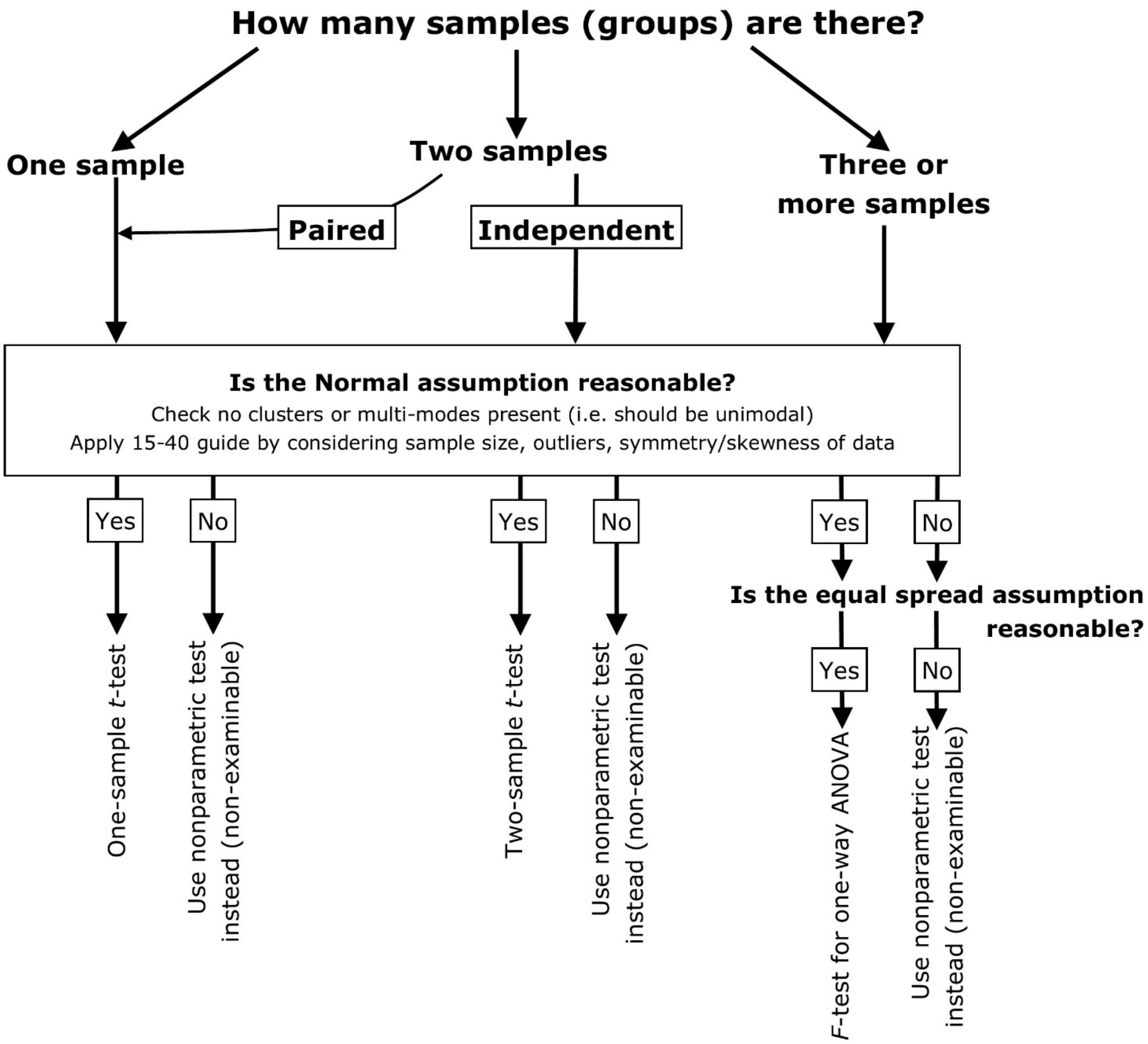
(1) The evidence of differences between the true group means comes from comparing the variability between group means with the variability within the groups.

(2) It should only be used when comparing independent samples.

(3) It provides partial protection against multiple comparisons.

(4) The null hypothesis is that all the true group means are the same.

(5) It is not badly affected by the presence of only one or two outliers.


How many samples (groups) are there?
One sample
Two samples
Paired
Independent
Three or more samples
Is the Normal assumption reasonable?
Check no clusters or multi-modes present (i.e. should be unimodal)
Apply 15-40 guide by considering sample size, outliers, symmetry/skewness of data
Yes
No
Yes
No
Yes
No
One-sample t-test
Use nonparametric test instead (non-examinable)
Two-sample t-test
Use nonparametric test instead (non-examinable)
Is the equal spread assumption reasonable?
Yes
No
F-test for one-way ANOVA
Use nonparametric test instead (non-examinable)

Assumptions	**Checks**
1. Independence — All tests - Single sample assumes indep. between observations. - Paired data assumes indep. between pairs of observations. - Two or more samples assumes indep. between observations *and* samples.	- The design of the experiment/study
2. Normality — All tests - one-sample *t*-test - two-sample *t*-test - *F*-test for one-way ANOVA	- Plot the data - Apply 15-40 guide
3. Equal spread — *F*-test for one-way ANOVA only	- Plot the data - Check the standard deviation ratio: $\frac{\text{largest sd}}{\text{smallest sd}} < 2$

Check you understand!

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

   (1) A statistically significant result is not always practically significant.

   (2) A non-significant hypothesis test does not mean that the null hypothesis is true.

   (3) Large positive *t*-test statistics lead to small *P-values* for two-tailed tests.

   (4) A small *P-value* from a hypothesis test may result from a very large sample, and the results may be of no practical significance.

   (5) A one-tailed *t*-test should be used when the idea for doing the test came about as a result of looking at the data.

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

   (1) In hypothesis testing, statistical significance does not imply practical significance.

   (2) In a hypothesis test for no difference between two means, a very small *P-value* indicates a very large difference in the means.

   (3) In hypothesis testing, a non-significant test result does not imply that H_0 is true.

   (4) In hypothesis testing, large samples can lead to small *P-values* without the results having any practical significance.

   (5) In a hypothesis test for no difference between two means, a two-sided test should be used when the idea of doing the test has been triggered as a result of looking at the data.

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

   (1) Tests of hypotheses can only deal with random errors and sampling variation. They are ineffective when confronted with data that has systematic bias.

   (2) The larger the *P-value,* the weaker the evidence against H_0.

   (3) A two-sided test of H_0: *parameter = hypothesised value* has *P-value* greater than 0.05 if the *hypothesised value* lies within a 95% confidence interval for the *parameter*.

   (4) An extremely small *P-value* means that the actual effect differs markedly from that claimed in the null hypothesis.

   (5) The larger the test statistic, $|t_0|$, for a two-sided test, the smaller the *P-value* will be.

Questions 4 to **9 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

4. We wish to determine the size of the difference in average weight gain between Egyptian and Kenyan women. Given that the data shows no non-Normal features when plotted, the most appropriate procedure to use here is (select **one** only):

   (1) a confidence interval based on the paired t-test.

   (2) a two-independent sample proportion t-test.

   (3) a confidence interval based on the two-independent sample t-test.

   (4) a confidence interval based on the F-test.

   (5) a confidence interval based on the Chi-square test.

5. Suppose a two-independent sample t-test is appropriate here for testing H_0: $\mu_E - \mu_K = 0$, against H_1: $\mu_E - \mu_K \neq 0$. Then the value of the degrees of freedom, df, is (select **one** only):

   (1) 21 (4) 20

   (2) 10 (5) 22

   (3) 11

6. Suppose the *P-value* for the test in Question 16 is 0.1654 (it is not). Which **one** of the following statements is **correct**?

   (1) The data provides no evidence against H_0.

   (2) The data provides strong evidence that H_0 is true.

   (3) The data provides evidence that H_0 is true.

   (4) The data provides evidence in favour of H_1.

   (5) The data provides strong evidence that H_1 is true.

7. When calculating a 95% confidence interval for μ_E - μ_K , the value of the *t*-multiplier obtained from a *Student's t*-table is 2.2281. The standard error, $\mathrm{se}(\bar{x}_E - \bar{x}_K)$ = 0.6085, then the **margin of error** for the 95% confidence interval is:

(1) $(4.55 - 3.29) \pm 2.2281 \times \frac{0.6085}{\sqrt{11}}$

(2) $\pm 2.2281 \times \frac{0.6085}{\sqrt{11}}$

(3) (3.29 - 4.55) ± 2.2281 x 0.6085

(4) $(3.29 - 4.55) \pm 2.2281 \times \frac{0.6085}{\sqrt{11}}$

(5) ± 2.2281 x 0.6085

8. We wish to determine if there are differences in average weight gain in any of the three countries. The most appropriate procedure to use here is (select **one** only):

(1) a confidence interval for the highest-lowest average weight gain: $\mu_{\text{high}} - \mu_{\text{low}}$.

(2) an *F*-test for H_0: $\mu_E = \mu_K = \mu_M$.

(3) a paired t-test for H_0: $\mu_{\text{diff}} = 0$, where $\mu_{\text{diff}} = \mu_{\text{high}} - \mu_{\text{low}}$.

(4) three paired t-tests.

(5) a Tukey interval for the highest-lowest average weight gain: $\mu_{\text{high}} - \mu_{\text{low}}$.

9. We wish to perform an *F*-test for the weight gain data. The appropriate degrees of freedom are (select **one** only):

(1) $df_1 = 2$ and $df_2 = 30$

(2) $df_1 = 3$ and $df_2 = 33$

(3) $df_1 = 33$ and $df_2 = 3$

(4) $df_1 = 30$ and $df_2 = 2$

(5) $df_1 = 2$ and $df_2 = 33$

Questions 10 to **13** refer to the following information.

HDL cholesterol is known as the “good cholesterol” as it is associated with lower risks of problems like heart disease. The following data were collected on men working in New Zealand companies.

The men were divided into exercise groups by the amount of exercise they reported. Levels of exercise were classified as lowest, medium, and most. Sixty men were randomly sampled from each exercise group, and their HDL cholesterol level was measured.

Box plots and some SPSS output analysing the data follow.

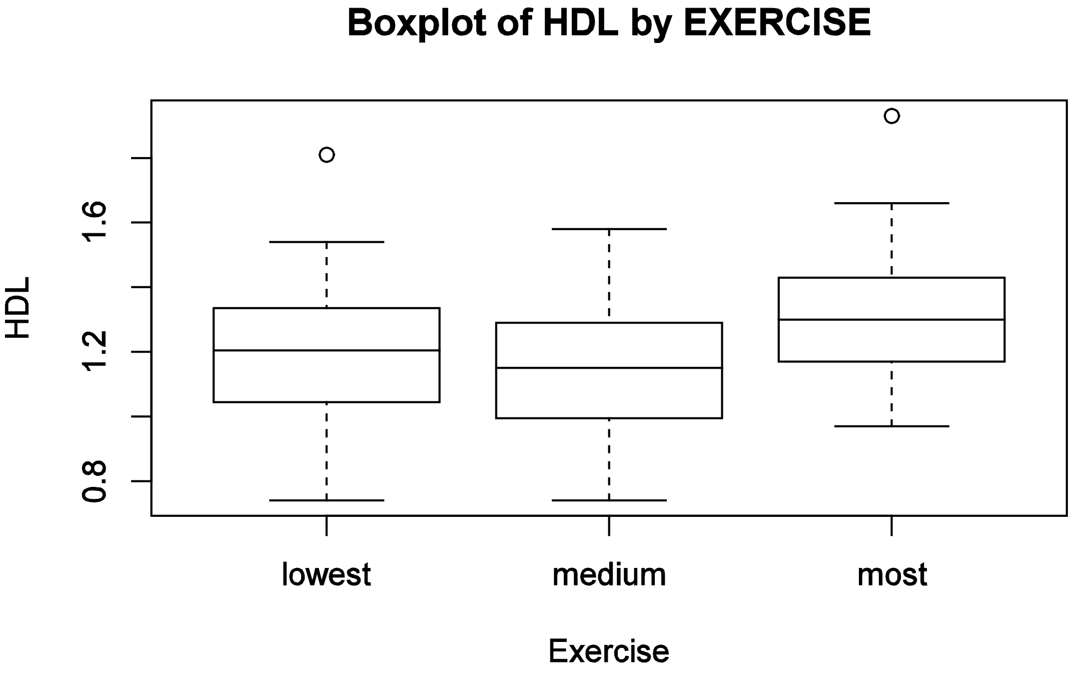


Figure 3: Box plot of HDL by exercise level.

Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
lowest	60	1.1998	.1926					
medium	60	1.1497	.2301					
most	60	1.2998	.1906					
Total	180							

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.7013	*	.3507	*	.000
Within Groups	6.7632	*	.0382		
Total	7.4645	179			

Table 3: SPSS output for the cholesterol data.

10. Which one of the following statements about the *F*–test shown in Table 3 is **false**?

    (1) The outside values in the lowest and most groups show that the Normality assumption of the *F*–test is violated.

    (2) The alternative hypothesis states that at least one of the exercise groups has a different underlying mean HDL level from another.

    (3) The differences in the sample standard deviations of the lowest, medium and most exercise groups do not affect the validity of the *F*–test in practice.

    (4) The box plots in Figure 3 give us no information on the independence of the exercise groups.

    (5) The null hypothesis states that the underlying mean HDL level is the same for each exercise group.

11. The values for the degrees of freedom, df_1 and df_2, at the top of Table 3 are:

    (1) $df_1=2$, $df_2=177$

    (2) $df_1=2$, $df_2=178$

    (3) $df_1=2$, $df_2=59$

    (4) $df_1=3$, $df_2=177$

    (5) $df_1=3$, $df_2=176$

12. The value of the *F*–test statistic, f_0, at the top of Table 3 is nearest to:

    (1) 9.643

    (2) 10.643

    (3) 9.181

    (4) 0.104

    (5) 0.109

13. Which one of the following statements is the **best** interpretation of the *P-value* in the analysis of variance for HDL shown in Table 3?

    (1) There is no evidence that all of the exercise groups have different underlying mean HDL cholesterol levels.

    (2) There is very strong evidence that at least one of the exercise groups has a different underlying mean HDL cholesterol level from another.

    (3) There is no evidence that any of the exercise groups have different underlying mean HDL cholesterol levels.

    (4) There is very strong evidence that all of the exercise groups have different underlying mean HDL cholesterol levels.

    (5) There is some evidence that at least one of the exercise groups has a different underlying mean HDL cholesterol level from another.

Question 14 refers to the following additional information.

A medical study was carried out to test the effectiveness of a new sleeping drug. It was hoped that the drug would be suitable for the New Zealand market. Ten people who had recently been diagnosed with having a particular type of sleeping disorder were used as subjects in the study. They were all given the drug on one night and then, on the next night, they were all given a placebo. The subjects could not tell beforehand, and nor were they told, which was the drug and which was the placebo. On each of the two nights, each subject was measured for the number of hours of sleep. The results are shown below.

	Hours of Sleep		
Patient	**Drug**	**Placebo**	**Difference**
1	6.1	5.2	0.9
2	7.0	7.9	-0.9
3	8.2	3.9	4.3
4	7.6	4.7	2.9
5	6.5	5.3	1.2
6	8.4	4.1	3.0
7	6.9	4.2	2.7
8	6.7	6.1	0.6
9	7.4	3.8	3.6
10	5.8	6.3	-0.5

Patient	**Drug**	**Placebo**	**Difference**
Sample Mean	7.06	5.28	1.78
Sample Std. Dev.	0.85	1.26	1.77

To determine the efficacy of the drug, the researcher wanted to see if there was a difference between the average number of hours of sleep when the drug is taken and the average number of hours of sleep when the placebo is taken.

14. An appropriate test to perform is a:

(1) Paired *t*-test on the differences.

(2) Two sample *t*-test if plots of the two samples are **not** severely non-Normal.

(3) *t*-test on the differences if a plot of the differences displays severe non-Normality.

(4) Two sample *t*-test on the two samples if plots of the two samples are severely non-Normal.

(5) *F*-test on the differences.

15. Which one of the following statements about paired data is **false**?

    (1) For paired data, we analyse the differences.

    (2) Pairing is beneficial when the variability within pairs is small compared with the variability between pairs.

    (3) The one sample *t*-test can be used to analyse the differences in paired data.

    (4) Pairing cannot be used in observational studies.

    (5) The carryover effect occurs when the first treatment alters the effect of the second treatment.

Questions 16 and **17** refer to the following information.

Printed on every packet of “Yummo” corn chips is a weight of 150g. A consumer collects 48 packets of “Yummo” corn chips and finds a mean weight of 148.5g and a standard deviation of 2.1g.

16. The customer wishes to test H_0: μ = 150 versus H_1: $\mu \neq$ 150. The standard error, se($\overline{x}$), is 0.3031. The value of the *t*-test statistic, t_0, and the degrees of freedom *df*, to be used are given by:

    (1) t_0 = -4.95, *df* = 47
    (2) t_0 = 4.95, *df* = 47
    (3) t_0 = -4.95, *df* = 48
    (4) t_0 = -4.95, *df* = 49
    (5) t_0 = 4.95, *df* = 48

17. Suppose the customer finds that the *p-value* for the above test = 0.07 (it is not). The **best** interpretation of this test would be:

    (1) With a *p-value* of 0.07 there is some evidence against the null hypothesis.

    (2) With a *p-value* of 0.07 there is weak evidence against the null hypothesis.

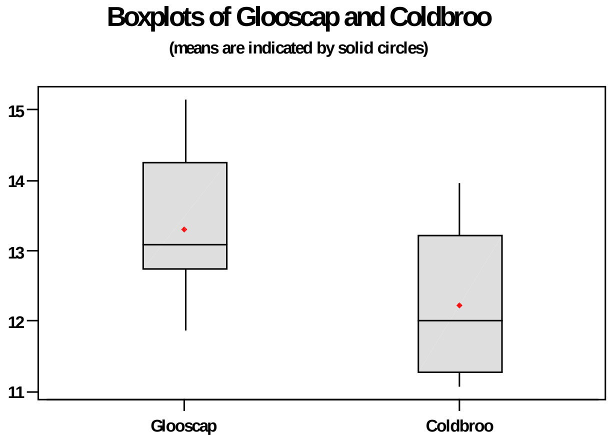
    (3) With a *p-value* of 0.07 there is weak evidence against the null hypothesis that the mean weight of Yummo chips is 150g.

    (4) With a *p-value* of 0.07 there is some evidence against the null hypothesis that the mean weight of Yummo chips is 150g.

    (5) With a *p-value* of .07% there is weak evidence against the null hypothesis that the mean weight of Yummo chips is 150g.

Questions 18 to **21** refer to the following information.

As part of a study to compare the physical education programs at two Canadian schools, running times (in seconds) over a set distance were recorded for two independent samples of sixth grade students taken from each school. (Data source courtesy of Chance Encounters).



Group Statistics

	schcode	N	Mean	Std. Deviation	Std. Error Mean
runtime	Glooscap	12	13.3125	.99133	.28617
	Coldbrook	13	12.2285	.99018	.27463

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
runtime	Equal variances assumed	.008	.930	2.733	23	.012	1.08404	.39661	.26359	1.90449
	Equal variances not assumed			2.733	22.836	.012	1.08404	.39663	.26322	1.90485

18. To test for a difference in the physical education programs of the two schools the null and alternative hypotheses would be:

(1) $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 > 0$

(2) $H_0 : p_1 - p_2 = 0$ versus $H_1 : p_1 - p_2 \neq 0$

(3) $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$

(4) $H_0 : \overline{x}_1 - \overline{x}_2 = 0$ versus $H_1 : \overline{x}_1 - \overline{x}_2 \neq 0$

(5) $H_0 : p_D - p_S > 0$ versus $H_1 : p_D - p_S = 0$

19. Which **one** of the following is **false**?

(1) There are no gross outliers in the data.

(2) The running times for Glooscap are on average, greater than the running times for Coldbrook.

(3) Two-sample t-tests are, in general, more robust to non-Normal features than the one sample t-test.

(4) The data taken from Glooscap and Coldbrook show severely non-normal features.

(5) The range of running times for Glooscap is slightly more than the range of running times for Coldbrook.

20. The **best** interpretation for this test is:

(1) With a *p-value* of 0.12 there is no evidence against the null hypothesis.

(2) With a *p-value* of 0.012 there is no evidence against the null hypothesis that there is a difference in the mean running times.

(3) With a *p-value* of 1.2% there is no evidence against the null hypothesis that there is no difference in the mean running times.

(4) With a *p-value* of 0.012 there is strong evidence against the null hypothesis that there is no difference in the mean running times.

(5) With a *p-value* of 0.012 there is weak evidence against the null hypothesis that there is no difference in the mean running times

21. The 95% confidence interval for the difference between the true mean running times is given in the SPSS output above. Which **one** of the following interpretations is **true**?

(1) With a probability of 0.95, the true difference of means $\mu_1 - \mu_2$ lies between 0.26 and 1.91.

(2) In repeated sampling the 95% confidence interval [0.26, 1.90] will contain the true difference in means in 95% of the samples taken.

(3) With 95% confidence, we estimate that the true proportion p_1 will be somewhere between 0.26 larger and 1.90 smaller than p_2.

(4) With 95% confidence, we estimate that the true mean running time from Glooscap μ_1 is somewhere between 0.26 and 1.90 larger than the true mean running time from Coldbrook μ_2.

(5) With 95% confidence the true mean from Glooscap μ_1 is 1.65 larger than the true mean from Coldbrook μ_2.

Questions 22 and **23** refer to the following information.

Factor V is a protein involved in the forming of blood clots. The higher the level of factor V, the faster the blood clots. The Auckland Blood Transfusion Service is interested in the effects of sterilisation of blood plasma because factor V is known to be unstable and may break down during sterilisation. The table below gives measured levels of factor V in blood samples from 16 blood donors. Both pre- and post-sterilisation measurements are given for each blood sample.

Donor Number	**Pre-Sterilisation**	**Post-Sterilisation**
1	1073	916
2	1064	1030
3	967	923
4	849	892
5	810	628
6	855	759
7	1047	828
8	1008	784
9	957	809
10	829	773
11	821	786
12	1257	1106
13	1095	832
14	1098	863
15	932	783
16	1440	869

Factor V and Blood Sterilisation

	Sample mean	Standard deviation	Sample size
Pre - Sterilisation	1006.37	170.899	16
Post - Sterilisation	848.812	112.195	16
Difference (Pre-Post)	157.562	140.132	16

Summary Statistics

22. Which **one** of the following statements gives the **correct** hypotheses for this test?

(1) H_0: all of the μ's are equal
H_1: none of the μ's are equal

(2) H_0: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 \neq 0$

(3) H_0: $\mu_{\text{diff1}} - \mu_{\text{diff2}} = 0$
H_1: $\mu_{\text{diff1}} - \mu_{\text{diff2}} \neq 0$

(4) H_0: $\mu_{\text{diff}} = 0$
H_1: $\mu_{\text{diff}} \neq 0$

(5) H_0: $p = 0$
H_1: $p \neq 0$

23. The *t*-test statistic for testing whether there is any evidence of an effect of sterilisation is given by:

(1) $\dfrac{\overline{x}_{\text{diff}}}{se\left(\overline{x}_{\text{diff}}\right)}$

(2) $\dfrac{\overline{x}_{\text{diff}}}{se\left(\overline{x}_{\text{Post}} - \overline{x}_{\text{Pre}}\right)}$

(3) $\dfrac{\overline{x}_{\text{Post}} - \overline{x}_{\text{Pre}}}{se\left(\overline{x}_{\text{Post}} - \overline{x}_{\text{Pre}}\right)}$

(4) $\dfrac{\overline{x}_{\text{Post}} - \overline{x}_{\text{Pre}}}{se\left(\overline{x}_{\text{Post}} + \overline{x}_{\text{Pre}}\right)}$

(5) $\dfrac{\overline{x}_{\text{Post}}}{se\left(\overline{x}_{\text{Post}}\right)} - \dfrac{\overline{x}_{\text{Pre}}}{se\left(\overline{x}_{\text{Pre}}\right)}$

24. For an *F*-test to be valid, which of the assumptions listed below are **required**?

I The samples are independent.
II The underlying means (i.e. the population means) are equal.
III The underlying level of variability is the same for each of the groups.
IV The sample sizes are equal.
V The underlying distribution of each group is Normal.

(1) II, III and V
(2) I, III and V
(3) II, III and IV
(4) I, II and V
(5) I, IV and V

25. Which **one** of the following statements gives the **correct** hypotheses for an *F*-test?

(1) H_0: all of the μ's are equal
H_1: none of the μ's are equal

(2) H_0: not all of the μ's are equal
H_1: all of the μ's are equal

(3) H_0: all of the μ's are equal
H_1: at least one of the μ's is different

(4) H_0: none of the μ's are equal
H_1: some of the μ's are equal

(5) H_0: some of the μ's are equal
H_1: not all of the μ's are equal

Questions 26 to **28** refer to the following information.

A certain drug was claimed to have a side effect of increasing the heartbeat rate. An experiment was performed on 8 rats. The number of heartbeats was recorded over a fixed time period immediately before and immediately after each rat received the drug. The data is given below.

26. It would be **inappropriate** to use a two independent sample *t*-test to test the hypothesis that μ_{after} - μ_{before} = 0 mainly because the:
    - (1) Population standard deviations are unknown.
    - (2) Sample sizes are small.
    - (3) Data are related.
    - (4) Samples are independent.
    - (5) Population means are unknown.

27. The value of the *t*-test statistic, t_0, to test the hypothesis that $\mu_{\text{diff}} = 0$, is:

(1) $\dfrac{\overline{x}_{\text{after}}}{se\left(\overline{x}_{\text{after}}\right)} - \dfrac{\overline{x}_{\text{before}}}{se\left(\overline{x}_{\text{before}}\right)}$

(2) $\dfrac{\overline{x}_{\text{diff}}}{se\left(\overline{x}_{\text{after}} - \overline{x}_{\text{before}}\right)}$

(3) $\dfrac{\overline{x}_{\text{after}} - \overline{x}_{\text{before}}}{se\left(\overline{x}_{\text{after}} - \overline{x}_{\text{before}}\right)}$

(4) $\dfrac{\overline{x}_{\text{after}} - \overline{x}_{\text{before}}}{se\left(\overline{x}_{\text{after}} + \overline{x}_{\text{before}}\right)}$

(5) $\dfrac{\overline{x}_{\text{diff}}}{se\left(\overline{x}_{\text{diff}}\right)}$

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	after - before	-69.25	14.60	5.16	-81.46	-57.04	-13.42	7	.000

28. The best description for the paired test on the heartbeats of the rats would be:
    - (1) With 95% confidence, we estimate that on average the heartbeats of the rats is 69.25.
    - (2) With 95% confidence, we estimate that on average the heartbeat of the rats before taking the drug was between 57.04 and 81.46 beats lower than the heartbeat of the rats after taking the drug.
    - (3) With 95% confidence, we estimate that on average the heartbeat of the rats before taking the drug was between 57.04 and 81.46 beats higher than the heartbeat of the rats after taking the drug.
    - (4) With 95% confidence, we estimate that the difference in the rats was between 57.04 and 81.46 heartbeats.
    - (5) With 95% confidence, we estimate that on average the heartbeat of the rats after taking the drug was between 57.04 and 81.46 beats higher than the heartbeat of the rats before taking the drug.

Questions 29 and **30** refer to the following information.

It has already been established that increased reproduction decreases longevity of female fruitflies. Therefore, an experiment was designed to test whether increased reproduction also reduces longevity for male fruitflies. Longevity is the life span (i.e. how long they live). Each male fruitfly was randomly assigned to one of five groups. There were **twenty-five** male fruitflies in each group. This is the variable GP.

The five groups are:

GP1:	Male forced to live alone
GP2	Male lives with one receptive female, i.e. the female is willing to mate.
GP3	Male lives with one non-receptive female.
GP4	Male lives with 8 non-receptive females.
GP5	Male lives with 8 receptive females.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	11939.28	*	2984.82	*	.000
Within Groups	26313.52	*	219.28		
Total	38252.80	*			

29. The degrees of freedom for the test statistic, f_0, for this F-test are:

(1) $df_1 = 5$, $df_2 = 125$

(2) $df_1 = 4$, $df_2 = 120$

(3) $df_1 = 120$, $df_2 = 4$

(4) $df_1 = 4$, $df_2 = 124$

(5) $df_1 = 125$, $df_2 = 5$

30. The value of the test statistic, f_0, for this F-test is:

(1) 79.20

(2) 654,500

(3) 13.61

(4) 0.073S5

(5) 0.4537

31. Which one of the following is **false**?

(1) A *P-value* calculated for a hypothesis formulated after looking at the data provides less convincing evidence than if the study had been designed to investigate the hypothesis.

(2) Formulae for the standard errors of data estimates do not take into account systematic biases in the experiment or survey.

(3) The fact that multiple comparisons have been made from a single set of data can be ignored when reporting the results.

(4) If 100 people independently collect data and calculate a 95% confidence interval for a population mean we expect approximately 95 people to capture the true mean in their interval and 5 to miss it.

(5) If 100 people independently collect data and test a true hypothesis, then just by chance, we expect about 5 to obtain results, which were significant at the 5% level.

32. Which **one** of the following statements about an *F*-test is **false**?

    (1) A decrease in the size of the differences between the group means will result in a decrease in evidence against the hypothesis that the underlying true group means are the same (given the variability/spread within each group remains unchanged).

    (2) The larger the value of the *F*-test statistic, f_0, the smaller the *P-value*.

    (3) An increase in the size of the differences between the group means will result in an increase in evidence against the hypothesis that the underlying true group means are the same.

    (4) An increase in the spread of the data within each group will result in an increase in evidence against the hypothesis that the underlying true group means are the same (given the size of the differences between the group means are the same).

    (5) The value of the *F*-test statistic, f_0, is the ratio of the between-mean variation and the within-group variation.

33. Which **one** of the following statements about one-way analysis of variance *F*-tests is **false**?

    (1) The greater the variability between the sample or group means relative to the variability within the samples or groups then the smaller the *P-value*.

    (2) A very small *P-value* suggests that there is a large difference between at least two of the underlying means.

    (3) If the *P-value* is very small, then observed differences between the sample or group means could be explained as being a result of differences between the underlying means.

    (4) The larger the value of the *F*-test statistic, f_0, the smaller the *P-value*.

    (5) If the *P-value* is very large, then observed differences between the sample or group means could be explained as being just due to chance alone.

Questions **34** to **42** refer to the **Swim Performance Study** information given below.

In 2001, a University of Auckland Sports Science student collected swim times from 58 New Zealand development squad swimmers. **Swim Time** is defined to be the number of minutes taken to swim 200 metres freestyle. Figure 5 below shows a dotplot of these swim times. A confidence interval for the population mean swim time (μ_{Swim}) for the New Zealand development squad is given in Table 7 below.

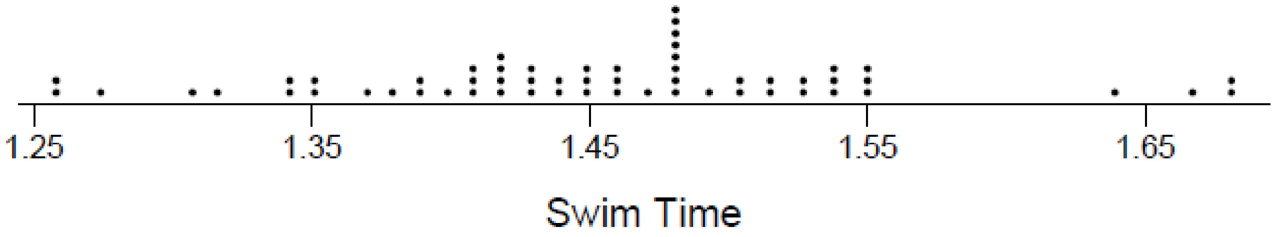


Figure 5: Dotplot of Swim Time (in minutes) for the New Zealand development squad

Summary Statistics and Confidence Interval

	N	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval	
					Lower	Upper
Swim Time	58	1.4543	0.0933	0.0123	1.4298	1.4788

Table 7: Summary statistics and confidence interval for the population mean Swim Time, μ_{Swim} (in minutes) for the New Zealand development squad

The Sports Science student monitored the swim performance for a subsample of 15 New Zealand development squad swimmers. Swim performance was measured by calculating their swim speed in the 200m freestyle as a percentage of the world record swim speed. For example, a swim performance of 100% would mean that the swimmer was as fast as the world record.

Swim performance for each of the 15 swimmers was recorded at the beginning of the study (referred to as **Before**), and at the end of the study (referred to as **After**). The **Differences** in performance for each swimmer were calculated as **After – Before**.

The Sports Science student wished to formally test for no difference between the mean **Before** and the mean **After** swim performance. Results for a two-sample *t*-test testing for no difference between swim performances **Before** and **After** the study are shown below in Table 8, while results for a paired sample *t*-test on the **Differences** are shown in Table 9.

T-Test

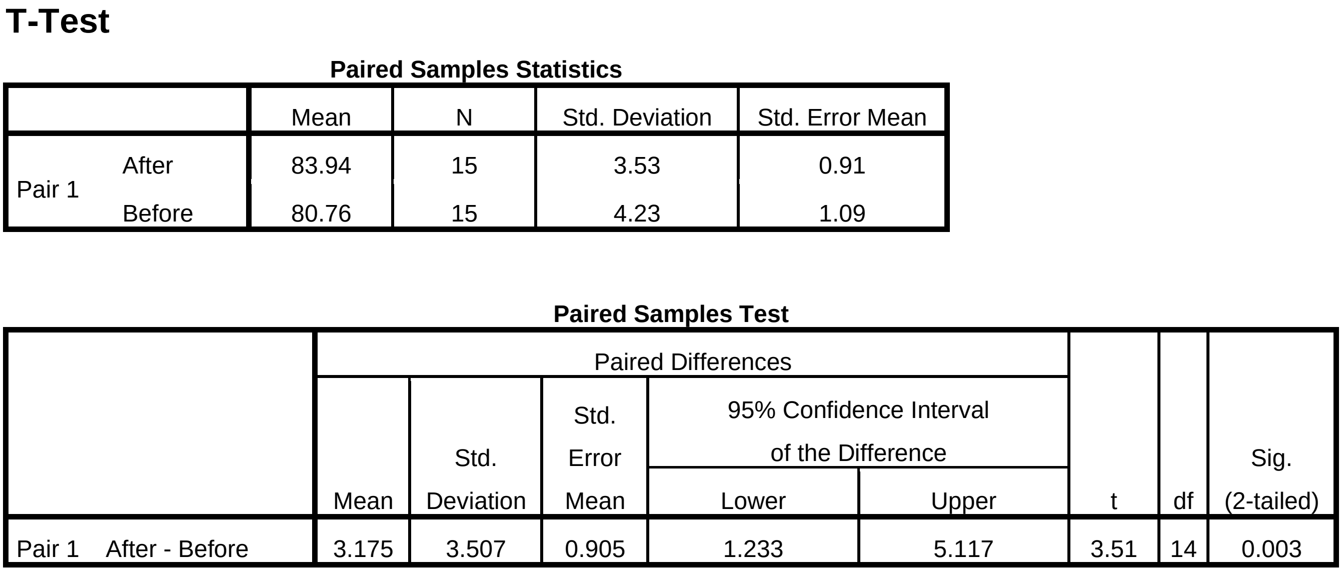
Group Statistics

	Swim performance	N	Mean	Std. Deviation	Std. Error Mean
Swim performance	After	15	83.94	3.53	0.91
	Before	15	80.76	4.23	1.10

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Swim performance	Equal variances assumed	.079	.780	2.23	28	0.034	3.18	1.423	0.25	6.1
	Equal variances not assumed			2.23	27.56	0.034	3.18	1.423	0.25	6.1

Table 8: SPSS output: confidence interval and two-sample *t*-test comparing swim performance **After** with swim performance **Before**



T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	After	83.94	15	3.53	0.91
	Before	80.76	15	4.23	1.09

Paired Samples Test

		Paired Differences							
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
					Lower	Upper			
Pair 1	After - Before	3.175	3.507	0.905	1.233	5.117	3.51	14	0.003

Table 9: SPSS output: confidence interval and paired *t*-test for swim performance **Differences**

The effect of a resting treatment on swim performance for the same subsample of 15 swimmers was also investigated. The resting treatment involved suspending the swimmers in a heated bath in the dark for a number of hours. After recording each swimmer's performance at the beginning of the study (referred to as **Before**) each swimmer was randomly allocated into either the **Control** group (who received no treatment), or the **Rest** group (who received the resting treatment). Each swimmer's performance was also recorded at the end of the study (referred to as **After**). The **Differences** in each swimmer's performance were calculated as **After – Before**.

Two-sample *t*-test results comparing swim performance **Differences** for the **Control** and **Rest** groups are shown in Table 10 below.

T-Test

Group Statistics

	Treatment	N	Mean	Std. Deviation	Std. Error Mean
Swim performance	Control	9	1.76	2.19	0.73
	Rest	6	5.29	4.22	1.72

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Swim performance	Equal variances assumed	.079	.780	-1.88	13	0.11	-3.53	1.871	-8.11	1.1
	Equal variances not assumed			-1.88	12.536	0.11	-3.53	1.871	-8.11	1.1

Table 10: Two sample *t*-test comparing swim performance **Differences** between treatment groups

Questions 34 to 42 refer to the **Swim Performance Study** information given above, on page 55.

34. Which **one** of the following statements is **true**? (Use Table 7, page 55.)
    (1) There is a 95% chance that a randomly selected development squad swimmer has a swim time in the interval from 1.43 to 1.48 minutes.
    (2) With 95% confidence, μ_{Swim} is somewhere between 1.43 and 1.48 minutes.
    (3) μ_{Swim} is estimated to be approximately 1.4543 minutes with a margin of error of 0.0123.
    (4) If many random samples of 58 development squad swimmers' swim times are taken and a 95% confidence interval calculated for each sample, then approximately 18 out of 20 of these confidence intervals will contain μ_{Swim}.
    (5) No valid statement can be made about the population mean swim time since a different sample would lead to a different mean and different confidence interval.

35. Suppose a random sample of 232 swim times (instead of 58) had been used to form a 95% confidence interval for μ_{Swim}. We would expect this new interval to have a width approximately:
    (1) double the width of the confidence interval formed from the 58 swim times.
    (2) the same width as the confidence interval formed from the 58 swim times.
    (3) four times the width of the confidence interval formed from the 58 swim times.
    (4) half the width of the confidence interval formed from the 58 swim times.
    (5) a quarter of the width of the confidence interval formed from the 58 swim times.

36. A confidence interval for the population mean, μ_{Swim}, is found using the formula:

$$\overline{x}_{Swim} \pm t \text{ x } se(\overline{x}_{Swim})$$

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

(1) A confidence interval for μ_{Swim} summarises the uncertainty due to sampling variation.
(2) 95% of the time we carry out such a study, the confidence interval for the population mean, μ_{Swim}, will contain the true sample mean, $\overline{x}_{Swim}$.
(3) A sample of 58 swim times is large enough to allow the sample to consist of related observations.
(4) It is critical that the swim times come from a Normal distribution.
(5) The number of swim times in our sample affects the size of the standard error but does not affects the size of the *t*-multiplier.

37. Suppose the Sports Science student realised that the four swim times greater than 1.6 minutes were all errors. (See Figure 5, page 55.) After removing these values, the new standard deviation was 0.0754. Suppose a new confidence interval for the remaining 54 observations was calculated using the correct *t*-multiplier of 2.006.

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

(1) The new confidence interval would have a smaller mean and be wider than the original confidence interval.
(2) The new confidence interval would be centred around a smaller mean and be narrower than the original confidence interval.
(3) The original and new confidence intervals could not be compared since they would have two different means.
(4) The new confidence interval would be centred around a larger mean and be wider than the original confidence interval.
(5) The new confidence interval would be the same width as the original confidence interval because they are both 95% confidence intervals.

Questions 38 and **39** refer to the **Swim Performance Study** information given above, on pages 49 and 50.

38. Assuming the student interpreted the correct *t*-test, which **one** of the following statements is **false**? (Use Tables 8 and 9 on pages 56 and 57 to answer this question.)

    (1) The test is comparing swim performance at the beginning of the study with swim performance at the end of the study.

    (2) The test is significant at the 5% level of significance.

    (3) The *t*-test statistic is 2.23.

    (4) The test is two-tailed.

    (5) The difference in the means is about 3.2.

39. Suppose Table 8, page 56 shows the correct analysis for the Before/After swim performance comparisons. Note: this may **not** be true.

    How would one **best** explain the results of this SPSS output to someone **unfamiliar** with statistics?

    (1) There is a statistically significant difference between the sample average swim performance before and after the study.

    (2) A 95% confidence interval states that the population mean swim performance of the swimmers in our sample dropped somewhere between 0.25 and 6.1 percentage points during the study.

    (3) We can be 95% confident that the population average swim performance improved somewhere between 0.25 and 6.1 percentage points during the study.

    (4) There is very strong evidence of a difference in population average swim performance at the beginning and end of the study.

    (5) It is a reasonable bet that the population average swim performance at the end of the study was between 0.25 and 6.1 percentage points higher than at the beginning of the study.

Questions 40 to **42** refer to the **Swim Performance Study** information given above, on page 58.

40. In a two-sample *t*-test on the **Differences** for the **Control** and **Rest** treatment groups (Table 10, page 58), which **one** of the following statements is **true**?

    (1) The *P-value* would be smaller if the standard errors of the **Control** and **Rest** groups were larger.

    (2) There is no evidence that the underlying means of the **Control** and **Rest** groups are different.

    (3) The *P-value* is not significant at the 5% level, but the results are practically significant.

    (4) The test is significant at the 5% level of significance.

    (5) The average of the differences was higher for the **Control** group.

41. Using Table 10, page 58, the standard error of the difference between the two independent sample means, se($\overline{x}_{\text{Control}} - \overline{x}_{\text{Rest}}$), is approximately:

    (1) 2.43 (4) 1.56

    (2) 2.45 (5) 1.87

    (3) 0.99

42. Which **one** of the following statements gives the null and alternative hypotheses for the *t*-test shown in Table 10, page 58?

    (1) H_0: $\mu_{\text{Control}} - \mu_{\text{Rest}} = 0$ H_1: $\mu_{\text{Control}} - \mu_{\text{Rest}} > 0$

    (2) H_0: $\mu_{\text{Control}} - \mu_{\text{Rest}} \neq 0$ H_1: $\mu_{\text{Control}} - \mu_{\text{Rest}} = 0$

    (3) H_0 : $\mu_{\text{Control}} - \mu_{\text{Rest}} = 0$ H_1: $\mu_{\text{Control}} - \mu_{\text{Rest}} \neq 0$

    (4) H_0: $\overline{x}_{\text{Control}} - \overline{x}_{\text{Rest}} \neq 0$ H_1: $\overline{x}_{\text{Control}} - \overline{x}_{\text{Rest}} = 0$

    (5) H_0: $\overline{x}_{\text{Control}} - \overline{x}_{\text{Rest}} = 0$ H_1: $\overline{x}_{\text{Control}} - \overline{x}_{\text{Rest}} \neq 0$

ANSWERS

A.	**(1)**	B.	**(3)**	C.	**(3)**	D.	**(5)**	E.	**(2)**	F.	**(4)**
G.	**(4)**	H.	**(4)**	I.	**(4)**	J.	**(3)**	K.	**(1)**	L.	**(4)**
M.	**(2)**	N.	**(4)**	O.	**(4)**	P.	**(3)**	Q.	**(2)**	R.	**(3)**
S.	**(4)**	T.	**(2)**	U.	**(5)**	V.	**(2)**	W.	**(2)**	X.	**(5)**
Y.	**(2)**	Z.	**(1)**	AA.	**(2)**	BB.	**(3)**	CC.	**(2)**	DD.	**(5)**

1.	**(5)**	2.	**(2)**	3.	**(4)**	4.	**(3)**	5.	**(2)**	6.	**(1)**
7.	**(5)**	8.	**(2)**	9.	**(1)**	10.	**(1)**	11.	**(1)**	12.	**(3)**
13.	**(2)**	14.	**(1)**	15.	**(5)**	16.	**(1)**	17.	**(3)**	18.	**(3)**
19.	**(4)**	20.	**(4)**	21.	**(4)**	22.	**(4)**	23.	**(1)**	24.	**(2)**
25.	**(3)**	26.	**(3)**	27.	**(5)**	28.	**(3)**	29.	**(2)**	30.	**(3)**
31.	**(3)**	32.	**(4)**	33.	**(2)**	34.	**(2)**	35.	**(4)**	36.	**(1)**
37.	**(2)**	38.	**(3)**	39.	**(5)**	40.	**(2)**	41.	**(5)**	42.	**(3)**

WHAT SHOULD I DO NEXT?

- Do Assignment 3!
- Go through the Chapter 6, 7 and 8 blue pages. For each chapter, this includes *notes*, a *glossary*, *true/false statements*, *Sample Exam Questions*, and *tutorial* material.
- Attend the optional Chapters 7 & 8 tutorials.
- Do all the problems in this workshop handout and mark them. If you get a question wrong, have a look at the working on Leila's scanned slides at www.tinyURL.com/stats-HTM to see how she did it.
- Try Chapter 6, 7 and 8 questions from three of the past five exams on Canvas (get them from *Modules* → *Past Tests and Exams (with answers)* and use the *Exam questions index* document from there to identify the questions from Chapters 6, 7 and 8!)
- If you get anything wrong and don't know why, get some help. You can post a question on Piazza (search first as it may have already been asked!), or talk to someone about it (your lecturer, an Assistance Room tutor or Leila).

FORMULAE

Confidence intervals and t-tests

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

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

Applications:

1. Single mean $\boldsymbol{\mu}$: $\quad estimate = \overline{x}; \qquad df = n - 1$
2. Single proportion $\boldsymbol{p}$: $\quad estimate = \widehat{p}; \qquad df = \infty$
3. Difference between two means $\boldsymbol{\mu_1 - \mu_2}$: (independent samples)

   $estimate = \overline{x}_1 - \overline{x}_2; \qquad df = \min(n_1 - 1, n_2 - 1)$
4. Difference between two proportions $\boldsymbol{p_1 - p_2}$:

   $estimate = \widehat{p}_1 - \widehat{p}_2; \qquad df = \infty$

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

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: $\quad \hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

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