Confidence Intervals

QMET201
CONFIDENCE INTERVALS provide an interval estimate of the unknown population parameter.

What is a confidence interval?
Statisticians have a habit of hedging their bets. They always insert qualifiers into reports, warn about all sorts of assumptions, and never admit to anything more extreme than probable. There's a famous saying:
“Statistics means never having to say you’re certain.”

Statements must be qualified, of course, because we are always dealing with imperfect information. In particular, it is often necessary to make statements about a population using information from a sample. No matter how carefully this sample is selected to be a fair and unbiased representation of the population, relying on information from a sample will always lead to some level of uncertainty.

So, a confidence interval is an interval within which we can estimate, with some confidence, that the true population parameter lies.

Introduction
Suppose we were interested in answering a simple research question such as:
“What is the mean number of digits that can be remembered?”

Having specified the population of people to be: ‘Lincoln University students’, we take a sample of 10. The number of digits remembered for these 10 students is: 4, 4, 5, 5, 5, 6, 6, 7, 8, 9. From these results we find the estimated value of $\mu$ to be 5.9.

But this will certainly not be a perfect estimate. It is bound to be at least either a little too high or a little too low.

For the estimate of $\mu$ to be of value, we need to have some idea of how precise it is. That is, how close to $\mu$ is the estimate likely to be?
An excellent way to specify the precision is to construct a confidence interval.

The wider the interval, the more confident you can be that it contains the parameter. In the above example, a 95% confidence interval would range from 4.71 to 7.09.

$$4.71 \leq \mu \leq 7.09$$

That is, we can be 95% certain that the true mean number of digits the population can remember is somewhere between 4.7 and 7.1.

A 99% confidence interval is, however, wider and extends from 4.19 to 7.61.

$$4.19 \leq \mu \leq 7.61$$

That is, we can be 99% certain that the true mean number of digits the population can remember is somewhere between 4.2 and 7.6.

Interpretation:

Suppose that a very large number of samples are taken, with a 95% conf. interval being calculated using each sample. Then 95% of these ‘95% confidence intervals’ will include the true population mean somewhere in their interval.
So, having calculated a 95% conf. interval for the population mean, we say we are 95% confident that the interval contains the true population mean.
(Other frequently used confidence intervals are 90% CI or 99% CI).
How is it calculated?

1. From the sample, calculate the sample statistic - eg \( \bar{x} \)

2. Look up a \( t \)-score from the Student’s \( t \)-distribution table (*see below), based on the level of confidence required and the sample size.

3. Calculate the standard error of the sample statistic. This is the standard deviation of the distribution of the sample statistic.
   The correct formula for each statistic must be used.
   
   For a mean, the s.e.(sample statistic) = se(mean) = \( \frac{s}{\sqrt{n}} \).

**Formula for a confidence interval:**

\[
C.I = \text{sample statistic} \pm t\text{-score} \times \text{s.e.(sample statistic)}
\]

Thus confidence interval for the population mean, \( \mu \) is calculated as:

\[
C.I. = \bar{x} \pm t\text{-score} \times \frac{s}{\sqrt{n}}
\]

*To find \( t \):*

Use the \( t \)-distribution table. This differs from the Z table with the probability given in the tail, and these are tabulated in the top row. Sample size is taken into account by degrees of freedom, tabulated in the left hand column. The relevant \( t \)-score is found in the “body” of the table. Part of this is shown:-

Use the given level of confidence to calculate \( p(\text{lying in the tail}) \). This is the column to used to find \( t \) and is located in the top row.

For example, if 90% C.I. required, \( p(\text{lying in the tail}) = 0.05 \)

For 95% C.I., \( p(\text{lying in the tail}) = 0.025 \).

That is, subtract % from 100%, halve and change to a decimal.

\( t \)-distribution critical values are in the ‘body’ of the table:

<table>
<thead>
<tr>
<th>df</th>
<th>0.25</th>
<th>0.15</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.02</th>
<th>0.01</th>
<th>0.005</th>
<th>0.0025</th>
<th>0.001</th>
<th>0.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Degrees of freedom for a one sample test are \( n - 1 \).

The correct \( t \) value is at the intersection of the relevant probability column and df row,
Example: For a set of data, \( \bar{x} = 85, s = 30, \text{ and } n = 30 \), find a 95% C.I. for \( \mu \)

\[
se(\text{mean}) = \frac{30}{\sqrt{30}},
\]

degrees of freedom \((n-1) = 29\) and 95% \(\Rightarrow\) upper tail = 0.025.

So from table: \( t\)-score:

<table>
<thead>
<tr>
<th>df</th>
<th>(t_{.100})</th>
<th>(t_{.050})</th>
<th>(t_{.025})</th>
<th>(t_{.010})</th>
<th>(t_{.005})</th>
<th>(t_{.001})</th>
<th>(t_{.005})</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>1.311</td>
<td>1.699</td>
<td>2.045</td>
<td>2.462</td>
<td>2.756</td>
<td>3.396</td>
<td>3.659</td>
</tr>
</tbody>
</table>

That is, \( t_{.025} = 2.045 \).

Hence, the C.I. \[
= 85 \pm 2.045 \times \frac{30}{\sqrt{30}}
\]

\[
= (73.79, 96.20)
\]

That is, we can be 95% confident that the true population mean \((\mu)\) lies between approximately 74 and 96.

*Note on use of calculator:*

The \(\pm\) in the formula means you must do two calculations. Use the “replay” key on your calculator for this. For the lower value in the C.I., use the ‘minus’ key, then use the > or < to scroll across until the cursor is over the −; change to + and press =.

You now have the upper value of the interval.

**Practice Questions – One sample**

1. A machine manufactures bolts to a set length. A random sample of 20 bolts is checked and found to have a mean length of 75.2 mm and standard deviation of 2.5 mm. Find the 99% confidence interval for the mean length of the bolts.

2. 60 people were asked to measure their pulse rates after completing a 3 km run. The mean was 105 beats and the standard deviation was 8 beats. Construct a 95% confidence interval for the mean of the population of people.

3. A type of golf ball is tested by dropping it onto a hard surface from a height of 1 metre. The height it bounces is known to be normally distributed. If a sample of 100 balls are tested and the mean height of the bounces = 82 cm and standard deviation = 3.6 cm. Find a. 90% b. 95% and c. 99% confidence intervals for the mean bounce height of the golf ball.

4. A sample of stalactites (a type of rock formation) found in a glow worm cave produced the following lengths in cm:

<table>
<thead>
<tr>
<th>9.6</th>
<th>16.9</th>
<th>15.1</th>
<th>14.3</th>
<th>15.9</th>
<th>17.2</th>
<th>13.0</th>
<th>17.1</th>
<th>15.4</th>
<th>16.2</th>
<th>4.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.3</td>
<td>21.2</td>
<td>15.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Assuming that this sample came from a normal population, calculate a 95% confidence interval for the mean length of stalactites in the cave.

5. A doctor conducts a small survey with a random sample of his patients, measuring their cholesterol levels.

Here is his data (the measurements are in m.mol/L):

<table>
<thead>
<tr>
<th>3.6</th>
<th>6.9</th>
<th>5.1</th>
<th>4.2</th>
<th>5.5</th>
<th>7.2</th>
<th>3.0</th>
<th>5.8</th>
<th>4.9</th>
<th>9.9</th>
<th>7.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4</td>
<td>6.2</td>
<td>4.5</td>
<td>6.3</td>
<td>8.2</td>
<td>5.7</td>
<td>4.4</td>
<td>7.9</td>
<td>3.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Find an 80% confidence interval for the mean cholesterol level of his patients.
Two-samples
Need to first decide whether the data is from matched or independent samples.

**Matched Pairs: C.I. for the mean difference $\mu_d$**
In 2 sample tests, if the data from the samples can be paired or matched, then the difference between each pair can be calculated. The two samples are not independent because the same ‘individuals’ are being compared from one ‘treatment’ to another. Find the difference for each ‘individual’ and use these values to find a CI for the mean difference. Symbols used for sample mean difference and standard deviation of the differences are $\bar{x}_d$ and $s_d$ respectively and the confidence interval for $\mu_d$ is calculated as:

$$C.I. = \bar{x}_d \pm t \times \frac{s_d}{\sqrt{n}}$$

For example (p.110 Harroway): To measure the average difference between English and Mathematics marks in a school, the following data was collected.

<table>
<thead>
<tr>
<th>Pupil</th>
<th>English</th>
<th>Maths</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>64</td>
<td>54</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>66</td>
<td>54</td>
<td>12</td>
</tr>
<tr>
<td>C</td>
<td>89</td>
<td>70</td>
<td>19</td>
</tr>
<tr>
<td>D</td>
<td>77</td>
<td>62</td>
<td>15</td>
</tr>
</tbody>
</table>

From the data, $\bar{x}_d = 14$, $s_d = 3.915$. Hence a 95% confidence interval for the average difference between English and Maths marks is

$$14 \pm 3.182 \times \frac{3.915}{\sqrt{4}} \text{ ie } 7.8 < \mu_d < 20.2.$$ 

That is, we can be 95% confident that the true mean difference is between 8 and 20.

**Two-sample tests: Independent Samples (unpaired)**

**C.I. for the difference between two means: $\mu_1 - \mu_2$**
Decide firstly whether the variances of the 2 samples are to be pooled. Can only do this when the 2 s.d.’s are not significantly different.

**Rule to decide whether to ‘pool’ the variance:**
- Calculate $F = \frac{s_1^2}{s_2^2}$: larger variance over smaller variance
- As a rough guide if $F < 4$, it’s OK to pool

**If the variances can be pooled:**
- The difference between the two sample means is $\bar{x}_1 - \bar{x}_2$
- The Standard Error of the difference between the means (SED) is

$$\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \text{ where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

(Read $s_p^2$ as “pooled variance”, $s_1^2$ and $s_2^2$ as first and second sample variance respectively.)
- the $t$- score uses $(n_1 + n_2 - 2)$ degrees of freedom
Hence the C.I. is now: \[
\left( \bar{x}_1 - \bar{x}_2 \right) \pm t \times \sqrt{\frac{s^2_p}{n_1} + \frac{s^2_p}{n_2}}
\]

**Example:** P.109 Example 10.2 (Harroway) Effect of diet on cholesterol levels - 2 diets to be compared
\[\bar{x}_1 = 4.775, \bar{x}_2 = 5.300, s^2_1 = 0.870, s^2_2 = 0.667, n_1 = 8, n_2 = 12\]

Since \[F = \frac{0.87}{0.667} = 1.3\] which is less than 4, variances are pooled,
\[s^2_p = \frac{(7 \times 0.87 + 11 \times 0.667)}{18} = 0.746\] and \[(n_1 + n_2 - 2) = (8 + 12 - 2) = 18\]

90% \(\Rightarrow\) upper tail = 0.025.
Hence t-score from table:

<table>
<thead>
<tr>
<th>df</th>
<th>t.100</th>
<th>t.050</th>
<th>t.025</th>
<th>t.010</th>
<th>t.005</th>
<th>t.001</th>
<th>t.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>1.330</td>
<td>1.734</td>
<td>2.101</td>
<td>2.552</td>
<td>2.878</td>
<td>3.610</td>
<td>3.922</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

and a 90% C.I. \[(5.3 - 4.775) \pm 1.734 \times \sqrt{\frac{0.746}{8} + \frac{0.746}{12}} = (-0.159, 1.209)\]

Can use this interval to decide whether there’s evidence that mean cholesterol is different under the two diets. Since the interval contains zero, we conclude there is no evidence of a significant difference between the two diets.

**If the variances are NOT to be pooled:**
- The sample difference between the means is \(\bar{x}_1 - \bar{x}_2\)
- The Standard Error of the difference between the means (SED) is 
  \[
  \sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}
  \]
- calculate degrees of freedom as given in lecture notes and find the t-score

C.I. is calculated as: \[(\bar{x}_1 - \bar{x}_2) \pm t \times \sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}\]

Suppose sample data gives:
\[\bar{x}_1 = 11.5, \bar{x}_2 = 9.0, s_1 = 10.2, s_2 = 4.58, n_1 = 45, n_2 = 38\]
\[F = \frac{10.2^2}{4.58^2} = 4.96\] which is not < 4. Hence to calculate d.f, use:
\[c = \left(1 + \frac{10.2^2 + 38}{4.58^2 + 45}\right) = 0.19274 \rightarrow memory\]
\[df = \left(0.19274^2 + 37 + (1 - 0.19274)^2 + 44\right)^2 = 63\] (rounded down)

This leads to the 95% confidence interval for the difference between the means
\[= (11.5 - 9.0) \pm 1.9983 \times \sqrt{\frac{10.2^2}{45} + \frac{4.58^2}{38}} = (-0.8818, 5.882)\]

The interval contains zero \(\Rightarrow\) there is no evidence of a difference between the means of the two groups.
Worked Examples

1. Trace metals in drinking water affect the flavour, and unusually high concentrations can pose a health hazard. A paper in Environmental Studies, (1982, 62-6), reported on the zinc levels in six different river locations, looking at the surface concentrations and the bottom concentration. (Data in micrograms.)

<table>
<thead>
<tr>
<th>Location</th>
<th>Bottom</th>
<th>Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>430</td>
<td>415</td>
</tr>
<tr>
<td>2</td>
<td>266</td>
<td>238</td>
</tr>
<tr>
<td>3</td>
<td>567</td>
<td>390</td>
</tr>
<tr>
<td>4</td>
<td>531</td>
<td>410</td>
</tr>
<tr>
<td>5</td>
<td>707</td>
<td>605</td>
</tr>
<tr>
<td>6</td>
<td>716</td>
<td>609</td>
</tr>
</tbody>
</table>

For each river we have a measure for the top and bottom => we can calculate the difference for each and use the resulting differences to calculate a mean difference and standard deviation, a confidence interval for the mean difference. We can also calculate the t-statistic for an hypothesis test.

<table>
<thead>
<tr>
<th>Location</th>
<th>Bottom</th>
<th>Top</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>430</td>
<td>415</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>266</td>
<td>238</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>567</td>
<td>390</td>
<td>177</td>
</tr>
<tr>
<td>4</td>
<td>531</td>
<td>410</td>
<td>121</td>
</tr>
<tr>
<td>5</td>
<td>707</td>
<td>605</td>
<td>102</td>
</tr>
<tr>
<td>6</td>
<td>716</td>
<td>609</td>
<td>107</td>
</tr>
</tbody>
</table>

Mean of the differences: $\bar{x}_d = 91.67$; sd of the differences: $s_d = 60.69$

Find a 95% Confidence Interval for the mean difference, $\mu_d$: $t_{table}$ with 5 degrees of freedom = 2.5706

Hence, $CI = 91.67 \pm 2.5706 \times \frac{60.69}{\sqrt{6}}$

$= (27.98, 155.36)$

We are 95% certain that the mean difference is between 28 and 155 micrograms.

2. ‘Sportsmedicine’ (1984) reported some physiological measurements on young tennis players.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Mean</th>
<th>Number</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boy’s grip</td>
<td>23.9</td>
<td>7</td>
<td>2.5</td>
</tr>
<tr>
<td>Girl’s grip</td>
<td>22.2</td>
<td>8</td>
<td>4.1</td>
</tr>
</tbody>
</table>

Calculate a 95% Confidence Interval for the difference between the two means, and the t-statistic for an hypothesis test that there is no difference.

Since this is no longer a situation where we have matched pairs, we need to know if we can pool the standard deviation. It is a two sample test.

Calculate $F_{calc} = \frac{S_L^2}{S_S^2} = \frac{4.1^2}{2.5^2} = 2.69$ so we can pool the variances ($2.69 < 4$)
Pooled variance is:

\[ s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \]

\[ = \frac{6 \times 2.5^2 + 7 \times 4.1^2}{(7 + 8 - 2)} = 11.94 \]

Hence, standard error of the difference

\[ = \sqrt{\left( \frac{11.94}{7} + \frac{11.94}{8} \right)} = 1.79 \]

Calculate a 95% confidence interval for \( \mu_{\text{Boys}} - \mu_{\text{Girls}} \):

\( t \)-value = 2.1604 (13 degrees of freedom):

C.I. = \((23.9 - 22.2) \pm 2.1604 \times 1.79\)

\[ = (-2.17, 5.56) \quad (\text{What does this mean?}) \]

The \( t \)-value to use for the hypothesis test would be

Rule for \( t \)-statistic

\[ = \frac{\text{sample statistic} - \text{hypothesized value}}{\text{standard deviation}} \]

\[ = \frac{(23.9 - 22.2) - 0}{1.79} = 0.95 \]

(This last part will be covered in the section on Hypothesis Testing.)

Practice Questions

Two samples

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

<table>
<thead>
<tr>
<th>Car</th>
<th>Estimate</th>
<th>Actual Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$180</td>
<td>$165</td>
</tr>
<tr>
<td>2</td>
<td>$1000</td>
<td>$1054</td>
</tr>
<tr>
<td>3</td>
<td>$65</td>
<td>$68</td>
</tr>
<tr>
<td>4</td>
<td>$320</td>
<td>$362</td>
</tr>
<tr>
<td>5</td>
<td>$200</td>
<td>$234</td>
</tr>
</tbody>
</table>

a) Calculate the mean difference AND the standard error of the difference (estimate-actual).

b) Calculate a 95% confidence interval for the difference between the mean estimated cost and the mean actual cost.

2. Pine trees were sampled from each of two blocks of pines, 100 km apart and in different rainfall regions. The following summary statistics were obtained:

\[ n_1 = 20 \quad \bar{x}_1 = 12.4 \quad s_1^2 = 4.8 \]

\[ n_2 = 20 \quad \bar{x}_2 = 16.8 \quad s_2^2 = 5.1 \]

a) Calculate the standard error of the difference between the means of the two blocks.

b) What would be the correct \( t \)-value required to calculate a 95% confidence interval for the difference between the means?
c) Calculate the 95% confidence interval for the difference between the two means.

3. To investigate the effect of price on toaster sales, a chain of department stores decided to raise the price of a certain brand of toaster in 20 randomly selected stores (Sample A) and to lower the price in 10 randomly selected stores (Sample B). Monthly revenue (number sales $\times$ sales price) from these toasters was recorded for each of the 30 stores. Results from the two samples are as follows:

<table>
<thead>
<tr>
<th></th>
<th>No. of stores</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample A</td>
<td>20</td>
<td>$842</td>
<td>$217</td>
</tr>
<tr>
<td>Sample B</td>
<td>10</td>
<td>$817</td>
<td>$202</td>
</tr>
</tbody>
</table>

(a) If you considered these to be independent samples, what would be the pooled estimate of the standard deviation for revenue?
(b) Calculate a 95% confidence interval for the difference between the two means.

4. The extent to which X-rays can penetrate tooth enamel had been suggested as a suitable mechanism for differentiating between males and females in forensic medicine. Listed below in appropriate units are the ‘spectropenetration gradients’ for eight female and teeth and eight male teeth:

<table>
<thead>
<tr>
<th>Male (x₁)</th>
<th>4.9</th>
<th>5.4</th>
<th>5.0</th>
<th>5.5</th>
<th>5.4</th>
<th>6.6</th>
<th>6.3</th>
<th>4.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female (x₂)</td>
<td>4.8</td>
<td>5.3</td>
<td>3.7</td>
<td>4.1</td>
<td>5.6</td>
<td>4.0</td>
<td>3.6</td>
<td>5.0</td>
</tr>
</tbody>
</table>

The data give sample means: $\bar{x}_1 = 5.4250, \bar{x}_2 = 4.5125$ and sample variances: $s_1^2 = 0.5536, s_2^2 = 0.5784$

a) Calculate the pooled estimate for the variance common to the male and female populations.
b) Estimate the standard error of the difference between the population means.
c) Construct a 95% confidence interval for the difference between the two population means.
d) Is there any significant difference between the two population means? Give your reason.

(Harroway p.114)

5. Check qn 5 – rest done

To determine which of two seeds was better, an agricultural research station chose 7 two-hectare plots of land randomly in New Zealand. Each plot was split in half, and a coin was tossed to determine in an unbiased way which half would be sown with which seed – A or B. The yields in appropriate units were as follows:

<table>
<thead>
<tr>
<th>Region</th>
<th>Seed A</th>
<th>Seed B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southland</td>
<td>82</td>
<td>88</td>
</tr>
<tr>
<td>South Otago</td>
<td>68</td>
<td>66</td>
</tr>
<tr>
<td>North Otago</td>
<td>109</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Canterbury</td>
<td>95</td>
<td>106</td>
</tr>
<tr>
<td>Waikato</td>
<td>112</td>
<td>116</td>
</tr>
<tr>
<td>Taranaki</td>
<td>76</td>
<td>79</td>
</tr>
<tr>
<td>Hawkes Bay</td>
<td>81</td>
<td>89</td>
</tr>
</tbody>
</table>

a) Which seed do you think is better? To back up your answer, construct a 95% confidence interval for these paired data? Why is it desirable to pair the data?

b) Does your conclusion differ if you treat the data as independent samples?

Answers  One Sample:

1. \[C.I. = 75.2 ± 2.861 \times \frac{2.5}{\sqrt{20}} = (73.6, 76.8)\]

2. \[C.I. = 105 ± 2.009 \times \frac{8}{\sqrt{60}} = (102.9, 107.1)\]

3. (a) \[C.I. = 82 ± 1.664 \times \frac{3.6}{\sqrt{100}} = (81.4, 82.6)\]

(b) \[C.I. = 82 ± 1.999 \times \frac{3.6}{\sqrt{100}} = (81.28, 82.72)\]

(c) \[C.I. = 82 ± 2.639 \times \frac{3.6}{\sqrt{100}} = (81.05, 82.95)\]

4. \[C.I. = 15.17 ± 2.16 \times \frac{4.17}{\sqrt{14}} = (12.76, 17.58)\]

5. \[C.I. = 5.75 ± 1.328 \times \frac{1.768}{\sqrt{20}} = (5.22, 6.28)\]

Answers  Two Samples:

1. Data here is matched pairs. The differences, we get: 15, -54, -3, -42, -34. Using these values as our new data gives us

\[\bar{x} = -23.6, \ s = 28.66 \Rightarrow se(mean) = \frac{28.66}{\sqrt{5}} = 12.82\]

and \[C.I. = -23 ± 2.776 \times 12.82 \Rightarrow (-59.19, 11.99)\]

No conclusion, since interval includes zero.
2. This data is from independent samples. 
Test for pooling variances: \( \frac{S_1^2}{S_2^2} = \frac{5.1}{4.8} < 4 \), so pool variances. 
\[
S_p^2 = \frac{19 \times 5.1 + 19 \times 4.8}{38} = 4.95 \quad \text{Hence}
\]
a) \( \text{se(difference)} = \sqrt{\frac{4.95}{20} + \frac{4.95}{20}} = 0.7035 \)
b) \( d.f. = 38 \)
c) \( C.I. = (16.8 - 12.4) \pm 2.042 \times 0.7035 \)
\( = [2.96, 5.83] \)

3. a) \( S_p^2 = \frac{19 \times 217^2 + 9 \times 202^2}{28} = 45068.82 \) \( \Rightarrow \) \( s_p = 212.294 \)

b) \( C.I. = (842 - 817) \pm 2.048 \times \sqrt{\frac{45068}{10} + \frac{45068}{20}} = (-143.39, 193.39) \)
Includes zero \( \Rightarrow \) no conclusion.

4. a) 0.566 \hspace{1cm} b) 0.3762 \hspace{1cm} c) [0.1056, 1.7194] 

d) Difference as C.I. excludes zero.

5. Finding differences, we get: 
6, -2, 12, 11, 4, 3, 8 which gives us \( \bar{x}_d = 6, \quad s_d = 4.8648; \quad d.f. = 6, \quad t = 2.447 \)

a) \( C.I. = 6 \pm 2.447 \times \frac{4.8648}{\sqrt{7}} = [1.5, 10.5] \)
The CI does not contain zero so evidence of a difference between the two seeds.

b) Analysing as independent samples:
\( \bar{x}_1 = 89, \quad s_1 = 16.77, \quad n_1 = 7 \quad \bar{x}_2 = 95, \quad s_2 = 20.08, \quad n_2 = 7 \Rightarrow \)
\( F = \frac{20.08^2}{16.77^2} = 1.4 \) which is < 4, so pool variances:
\[
S_p^2 = \frac{6 \times 16.77 + 6 \times 20.08}{12} = 342.22
\]
\[
C.I. = (95 - 89) \pm 2.179 \times \sqrt{\left( \frac{342.22}{7} + \frac{342.22}{7} \right)} = (-15.54, 27.54)
\]
The CI does contain zero so there is no evidence of a difference between the two seeds.
When reading a question, note:
- whether one or two samples
- if two samples, whether matched pairs or not
- if not matched pairs, whether variance should be pooled or not

**AND especially check**
- IF INFORMATION GIVES THE VARIANCE(S), STANDARD DEVIATION(S) OR STANDARD ERRORS.

---

**To calculate a Confidence Interval for mean(s)**

**One sample or two?**

- Use: $C.I. = \bar{x} \pm t \times \frac{s}{\sqrt{n}}$
  
  $d.f. = n - 1$

**Matched Pairs?**

- Find differences and then use: $C.I. = \bar{y} \pm t \times \frac{s_p}{\sqrt{n}}$
  
  $d.f. = n - 1$

**Pool variances?**
- i.e. if $\bar{P} = \frac{S_1}{n_1} \times 4$

**Use:**

- $C.I. = (\bar{x}_1 - \bar{x}_2) \pm t \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

**Calculate $c = \frac{1}{\frac{\bar{S}_1}{\text{d.f.}} + \frac{\bar{S}_2}{\text{d.f.}}} \text{ then } 1 - c' = \frac{(1-c)^{\frac{1}{2}(n-1)}}{(n-1)}$**

Use reciprocal key to invert and get d.f.