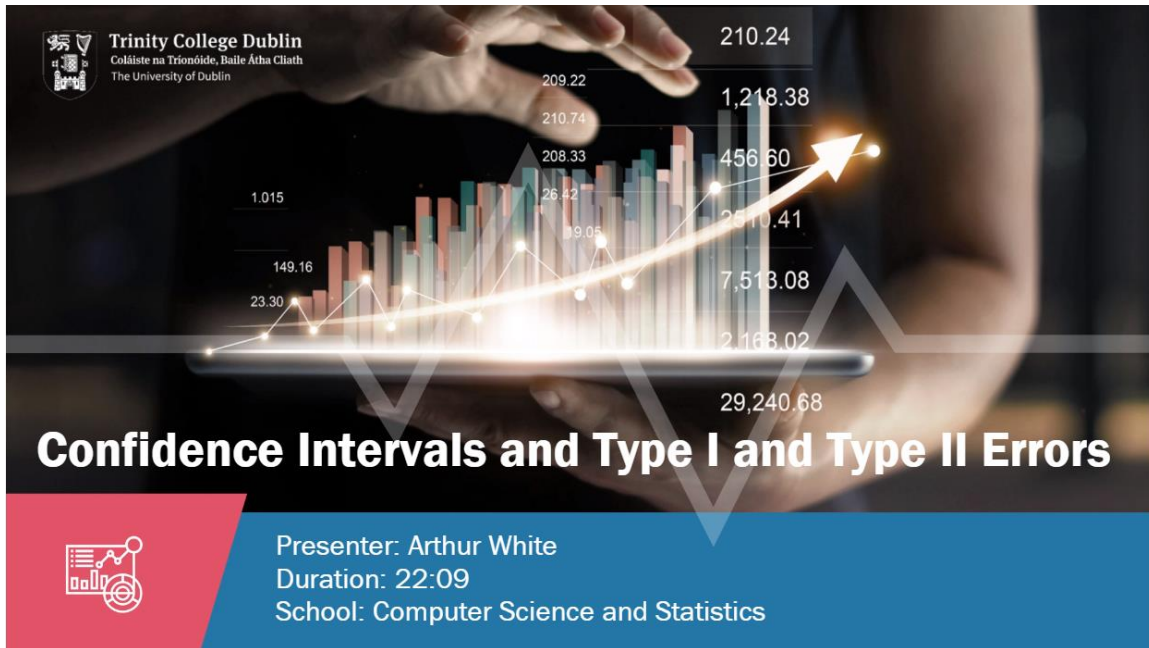


## Confidence Intervals and Type I and Type II Errors

Slide 1:	Introduction .....	2
Slide 2:	Section 1: Background and Motivation .....	3
Slide 3:	Review of Single Sample Hypothesis Test .....	3
Tab 1:	Boxplot.....	4
Slide 4:	Conclusions from the Single Sample Hypothesis .....	5
Slide 5:	Confidence Intervals: Motivating Questions .....	5
Slide 6:	Establishing the Confidence Interval for Lab B.....	6
Slide 7:	Confidence Intervals: Concepts .....	7
Slide 8:	Simulated Data .....	8
Slide 9:	Section 2: Confidence Intervals for Other Tests .....	9
Slide 10:	Paired Comparison in Animal Fertility Study .....	9
Tab 1:	Study Data.....	10
Tab 2:	Confidence Interval.....	11
Slide 11:	Confidence Interval for Independent Groups.....	12
Tab 1:	Study Data.....	13
Tab 2:	Confidence Interval.....	13
Slide 12:	Section 3: Interpreting Hypothesis Results .....	15
Slide 13:	The Importance of Wording when Interpreting Hypothesis Results .....	15
Slide 14:	Potential Errors that can Arise .....	16
Slide 15:	Type I and Type II Errors .....	17
Slide 16:	Defining Type I and Type II Errors .....	18
Slide 17:	Conclusion .....	19
Slide 18:	Summary.....	20

## Slide 1: Introduction



Trinity College Dublin  
Coláiste na Tríonóide, Baile Átha Cliath  
The University of Dublin

**Confidence Intervals and Type I and Type II Errors**

Presenter: Arthur White  
Duration: 22:09  
School: Computer Science and Statistics

Hello and welcome to this presentation, my name is Arthur White and I will lead you through this presentation.

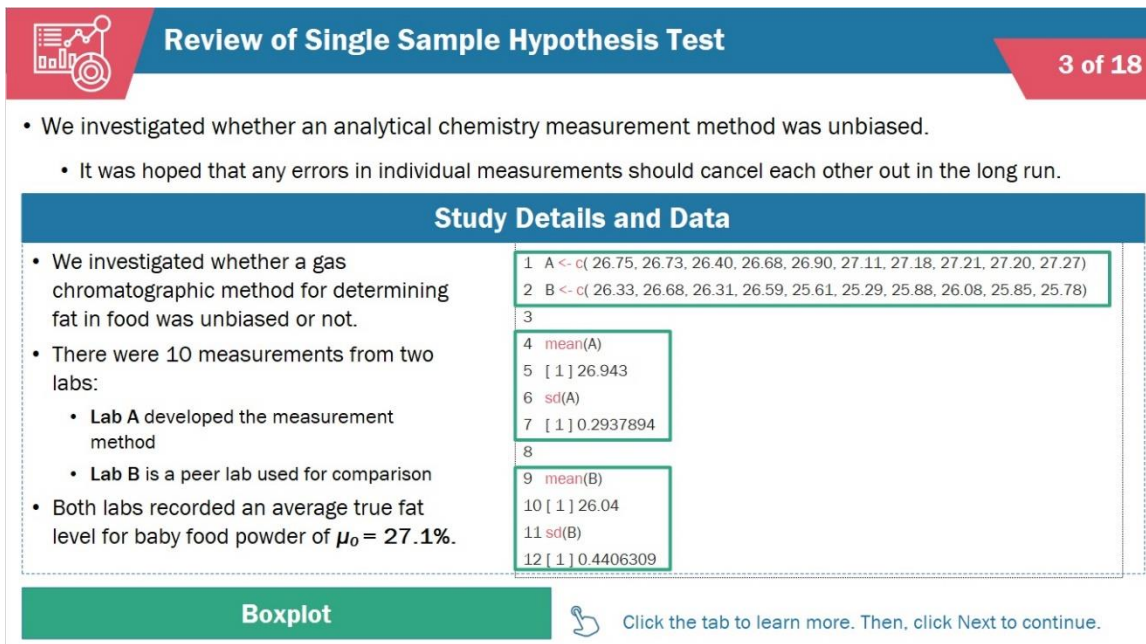
During this presentation, I will introduce the statistical concept of the confidence intervals. This idea is closely related to the ideas regarding hypothesis tests that were introduced in the last session. We will consider several popular kinds of confidence interval and apply them to the same data examples that we saw in the last session. We will carefully examine some of the technical details behind confidence intervals; what calculations are involved when estimating confidence intervals; and clarify how to interpret the examples we apply to data. We will also return to hypothesis testing and introduce the concepts of Type I and Type II error. By the end of this presentation, you should have a much clearer and more nuanced understanding of how to appropriately interpret the results of a hypothesis test.

## Slide 2: Section 1: Background and Motivation



Let's start by motivating some of the ideas for this session.

## Slide 3: Review of Single Sample Hypothesis Test



- We investigated whether an analytical chemistry measurement method was unbiased.
  - It was hoped that any errors in individual measurements should cancel each other out in the long run.

Study Details and Data	
• We investigated whether a gas chromatographic method for determining fat in food was unbiased or not.	1 A <- c( 26.75, 26.73, 26.40, 26.68, 26.90, 27.11, 27.18, 27.21, 27.20, 27.27)
• There were 10 measurements from two labs:	2 B <- c( 26.33, 26.68, 26.31, 26.59, 25.61, 25.29, 25.88, 26.08, 25.85, 25.78)
• Lab A developed the measurement method	3
• Lab B is a peer lab used for comparison	4 mean(A)
• Both labs recorded an average true fat level for baby food powder of $\mu_0 = 27.1\%$ .	5 [ 1 ] 26.943
	6 sd(A)
	7 [ 1 ] 0.2937894
	8
	9 mean(B)
	10 [ 1 ] 26.04
	11 sd(B)
	12 [ 1 ] 0.4406309

**Boxplot** [Click the tab to learn more. Then, click Next to continue.](#)

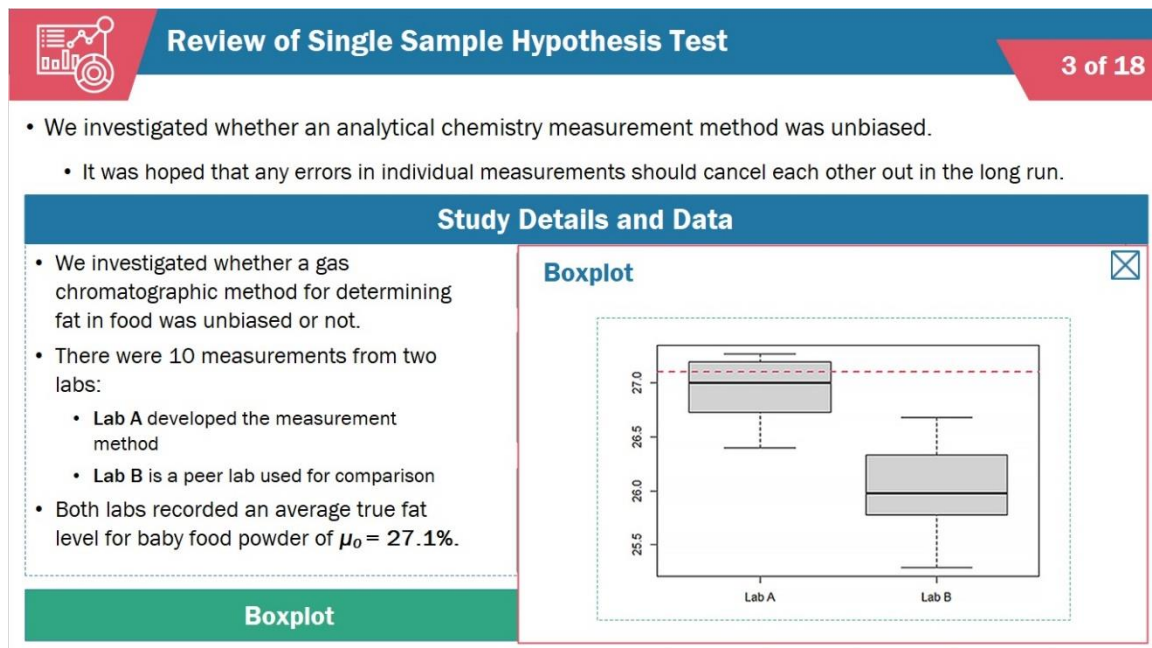
You will recall the following example from our previous session. We investigated whether an analytical chemistry measurement method was unbiased. While there may have been errors in individual measurements using the method, it was to be hoped that these errors should cancel each other out on average. In other words, we were attempting to identify whether a systematic component of the measurement deviations from the true value could be identified.

Specifically, we investigated whether a gas chromatographic method for determining fat in food was unbiased or not. We had ten measurements from two labs, Lab A and Lab B, where Lab B was a peer lab used for comparison. Both labs recorded measurements for a reference material, baby food powder, where the average true fat level was certified to be  $\mu_0 = 27.1\%$ .

Here again are the data for Lab A and Lab B, along with the mean and standard deviations for each lab.


Click the tab to review the boxplot for this data. When you are ready click “Next” to continue.

## Tab 1: Boxplot



In the boxplot for the data, while the data from Lab A are intuitively “close” to the true reference value 27.1% (the red dashed line), data from Lab B are consistently below this value.

## Slide 4: Conclusions from the Single Sample Hypothesis



Conclusions from the Single Sample Hypothesis

4 of 18


Lab A Results and R Code	Lab B Results and R Code
<pre>1 &gt; t.test(A, mu = 27.1) 2 3 One Sample t - test 4 5 data: A 6 t = -1.6899, df = 9, p-value = 0.1253 7 alternative hypothesis: true mean is not equal to 27.1 8 95 percent confidence interval: 9 26.73284 27.15316 10 sample estimates: 11 mean of x 12 26.943 #</pre>	<pre>1 &gt; t.test(B, mu = 27.1) 2 3 One Sample t - test 4 5 data: B 6 t = -7.6073, df = 9, p-value = 3.302e-05 7 alternative hypothesis: true mean is not equal to 27.1 8 95 percent confidence interval: 9 25.72479 26.35521 10 sample estimates: 11 mean of x 12 26.04 #</pre>

- We failed to reject  $H_0: \mu_0 = 27.1\%$ .

- We rejected  $H_0: \mu_0 = 27.1\%$  as the method is biased downwards.

We described how to apply a single sample hypothesis test to the data from Lab A and Lab B. For Lab A, we observed a test statistic of -1.69, and we failed to reject the hypothesis that the analytical system in Lab A is biased. For Lab B, we observed  $t = -7.6$ , a statistically significant result. Hence, we concluded that the analytical system in Lab B was biased downwards.

## Slide 5: Confidence Intervals: Motivating Questions




Confidence Intervals: Motivating Questions

5 of 18

### Reflect on the Conclusions from the Single Sample Hypothesis

- We concluded that the analytical system in Lab B was biased.
- By how much is it biased?
  - Small inaccuracies in measurement accuracy may be permissible in some contexts.
  - Very large bias may indicate that the system is simply not fit for purpose.
- The result for Lab A was not significant, so we failed to reject  $H_0$ .
- How should this be interpreted?
  - Is the Lab A method truly unbiased?

- Confidence intervals are a statistical method to address these types of problems.
- Sample  $\bar{x}$  is only an estimate of the population mean, or long run mean  $\mu$ .
  - $\bar{x}$  is random, so the true value of  $\mu$  is uncertain.
- Quantify the uncertainty by constructing an interval estimate about  $\mu$ .
  - This consists of an upper and lower bound for its value using  $\bar{x}$  and its associated standard error.



It is worth reflecting for a moment on the specific conclusions that we formed in our analysis. For Lab B, we concluded that the analytical system in Lab B was biased. A very reasonable follow up question is to ask, by how much is it biased? Small inaccuracies in measurement accuracy may be permissible in some contexts; the nutritional




components of some food stuffs may not be quite as important as they are for baby food! On the other hand, very large bias may indicate that the system is simply not fit for purpose. We should also consider again the conclusions we made following the outcome of the significance test for Lab A. In this case the result was not significant, so we failed to reject the null. How exactly should we interpret this statement? Does this mean that the Lab A measurement method is truly unbiased?

Confidence intervals are a statistical method to address these types of problems. Specifically, for any data sample, we know that the sample mean  $\bar{x}$  is only an estimate of the population mean, or long run mean,  $\mu$ . This means that  $\bar{x}$  is random and hence we are never completely certain what the true value of  $\mu$  is. We quantify this uncertainty by constructing an interval estimate about  $\mu$ . This consists of an upper and lower bound for its value using  $\bar{x}$  and its associated standard error.

Starting with the data from Lab B, we will discuss how to estimate confidence intervals for each of the data examples and associated hypothesis tests covered in our previous session. These tests will illustrate the underlying concepts and ideas behind hypothesis tests. These examples will all have several important features in common.

### Slide 6: Establishing the Confidence Interval for Lab B



Establishing the Confidence Interval for Lab B

6 of 18

- The hypothesis test for Lab B was performed using  $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ .
- This value was compared to the critical values where  $t_c = \pm 2.26$ .

Compute the Confidence Interval

- Confidence Interval =  $\bar{x} \pm t_c \frac{s}{\sqrt{n}}$
- Note:  $\mu_0 = 27.1\%$  is not used.

**Calculation:**

- Confidence Interval =  $26.04 \pm 2.26 \frac{0.441}{\sqrt{10}} = 26.04 \pm 0.32$ 
  - Say with 95% confidence that the long run mean for Lab B measurements lies between 25.72% and 26.36% fat.
  - Measurement bias is between 0.74% and 1.38%.

Lab B - R Code

```
1 > t.test(B, mu = 27.1)
2
3 One Sample t-test
4
5 data: B
6 t = -7.6073, df = 9, p-value = 3.302e-05
7 alternative hypothesis: true mean is not equal to 27.1
8 95 percent confidence interval:
9 25.72479 26.35521
10 sample estimates:
11 mean of x
12 26.04 #
```

**! A confidence interval will not contain  $\mu_0$  when  $H_0$  is rejected.**

Recall the constituent elements of the hypothesis test we performed for the data from Lab B: we took the difference between the sample mean  $\bar{x}$  and null mean  $\mu_0$ , and divided by the standard error. Also recall the standard error is just the sample standard deviation  $s$  divided by the square root of the sample size,  $n$ . This quantity,  $t$ , was then compared to the critical values,  $t_c = \text{plus/minus } 2.26$ .

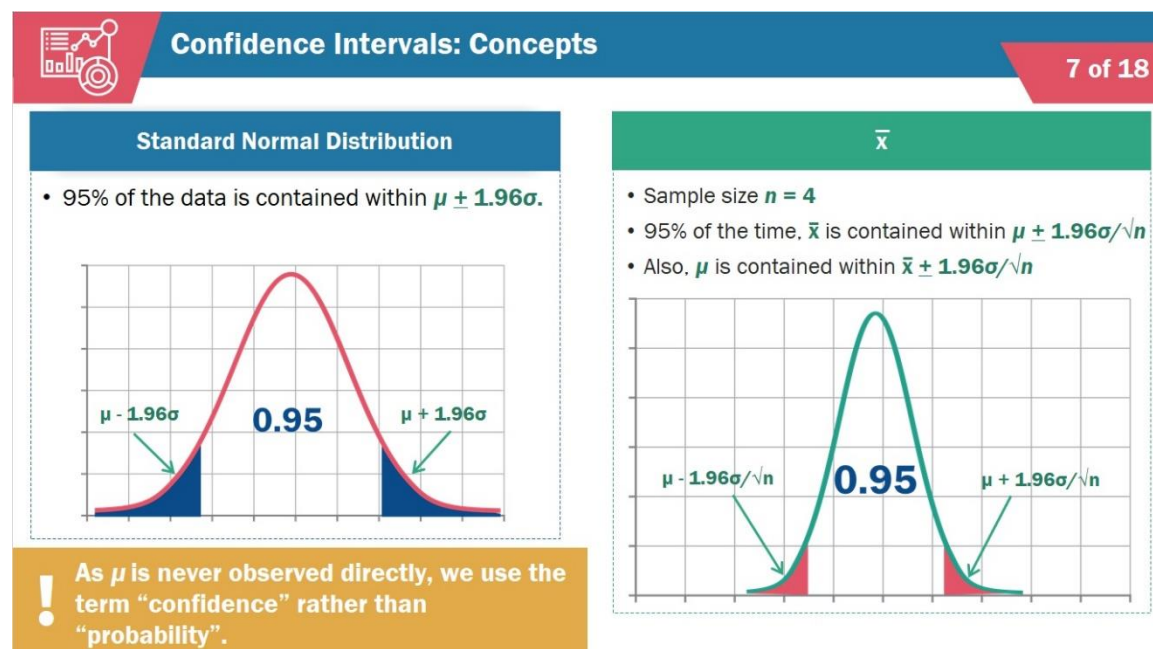
To compute the confidence interval, no new terms are needed. We simply re-arrange the existing terms to construct the interval  $\bar{x}$  plus/minus  $t_c \times$  the standard error. Notice the null mean  $\mu_0$  is not used here; we don't need to make any assumption about the value of the true mean when constructing the confidence interval.

In this case, we compute the confidence interval to be 26.04 plus/minus 0.32. Speaking formally, we say with 95% confidence that the long run mean for Lab B measurements lies somewhere between 25.72 and 26.36 percentage fat. Notice that 27.1%, the true reference value, is not contained in this interval. It will always be the case that a confidence interval will not contain  $\mu_0$  when  $H_0$  is rejected. These statements are in fact equivalent. One can think of the confidence interval as the set of possible long-run values which would not be rejected by a t-test.

For this example, the measurement bias is estimated as being somewhere between 0.74% and 1.38%. In this case, expert judgement would be needed to decide to what extent this amount of bias is of practical relevance.

Let's return to the Lab B hypothesis output provided by R. Notice that along with the items such as t statistic and p-value identified previously, a 95% confidence interval is also provided with the output. This is the case for most statistical software applications.

### Slide 7: Confidence Intervals: Concepts




It is worth thinking about some of the probability concepts underlying the construction of confidence intervals. To begin with, consider a single normally distributed random variable. We know that a strong majority, i.e., 95% of the data, is contained within about 2 standard deviations of the mean. More precisely, the probability that  $X$  is within  $\mu \pm 1.96$  times  $\sigma = 0.95$ .

Now consider the sample mean  $\bar{x}$  of normally distributed data, with, e.g., sample size  $n = 4$ . We know that in this case, 95% of the time,  $\bar{x}$  is contained within  $\mu \pm 1.96$  times the standard error,  $\sigma/\sqrt{n}$ . Of course, in practice, we never observe  $\mu$  or  $\sigma$ , only the sample estimates  $\bar{x}$  and  $s$ . So it makes sense to try and turn this idea around – if  $\bar{x}$  is contained within  $\mu \pm 1.96$  times the standard error 95% of the time, then  $\mu$  has to be contained within  $\bar{x} \pm 1.96$  times the standard error as well.

While this concept is hopefully quite straightforward, there are some technical details we need to be careful about here. We can't make probability statements regarding  $\mu$  using  $\bar{x}$  once we observe our data. The issue is that, while in principle at least, we get to observe our data many times, the population mean  $\mu$  is a hypothetical quantity that we never observe directly and hence we can't make probability statements about this term in our usual way. We can only do so using something called a Bayesian framework and such approaches are beyond the scope of this course. Hence, we use the term confidence rather than probability.

## Slide 8: Simulated Data

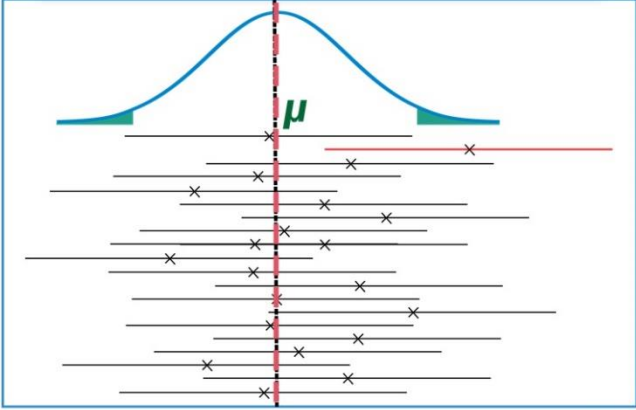


### Simulated Data

8 of 18

- Simulate  $\bar{x} \sim N(\mu, \sigma^2/n)$ , twenty times.
- Construct  $CI \bar{x} \pm 1.96\sigma/\sqrt{n}$ 
  - As a 95% confidence interval is being constructed, it should be correct nineteen out of the twenty times.

#### Confidence Interval of Simulated Data



Let's make this concept more concrete with an illustrative example using simulated data. Here we are going to simulate 20 different data sets, each with a common mean and standard deviation. In each case, we will construct a confidence interval based on the observed data. Because we are constructing a 95% confidence interval, we should be right about 19 times in 20, i.e., about one interval out of 20 should "miss" the true value. Here is a graph illustrating all 20 such intervals, below the true distribution centred at the mean  $\mu$ . The vertical red dashed line indicates the true value of the mean. The assorted horizontal black lines indicate the confidence intervals, with an x marking where the estimated value of the sample mean is. Notice that the black lines overlap with the dashed red line. As we expected, one horizontal line out of the twenty, marked red, does not overlap with the true mean  $\mu$ . Notice that the sample mean is positioned in the tail of the distribution in this case. This is entirely consistent with our discussion of how such intervals are to be interpreted mentioned earlier.



## Slide 9: Section 2: Confidence Intervals for Other Tests




9 of 18

# Confidence Intervals for Other Tests

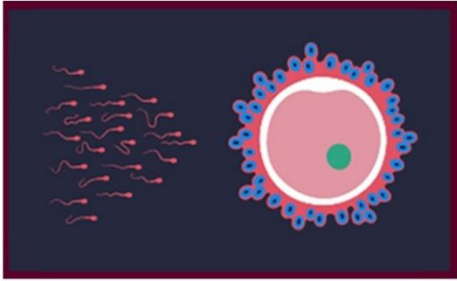

It is straightforward to compute confidence intervals for other kinds of tests. Here we will briefly describe how to compute confidence intervals for paired tests and independent two sample tests covered in the previous session.

## Slide 10: Paired Comparison in Animal Fertility Study



Confidence Interval for Paired Test

10 of 18

Study Data	Confidence Interval
<p><b>Introduction</b></p> <ul style="list-style-type: none"><li>Computing confidence intervals for tests follows the same fundamental procedure:<ul style="list-style-type: none"><li><b>(Point estimate <math>\pm</math> Critical value) <math>\times</math> Standard error</b><ul style="list-style-type: none"><li>These quantities are calculated for hypothesis tests.</li></ul></li></ul></li></ul> <p><b>Study Details</b></p> <ul style="list-style-type: none"><li>The concentrations of retinol-binding protein (RBP) in uterine fluid from both the ipsi and contra sides of the uterus for 16 cows were measured to determine if there was a difference between them on the day of ovulation.</li><li>If more RBP is produced on the ipsi side, this would suggest a direct role for RBP in implantation.</li></ul>	 <p> Click each tab to learn more. Then, click Next to continue.</p>

In much the same way that hypothesis tests follow the same fundamental procedure, the formula for computing confidence intervals has the same essential form. It is simply the point estimate plus/minus the critical value times the standard error. Note that these quantities will have already been calculated whenever a hypothesis test has been performed.

Let's look at the data example concerning a paired study of cows. Retinol-binding protein (RBP), a major secretory product of the endometrium, was assumed to be of importance for early embryonic development in cows. The concentrations of RBP in uterine fluid from both the ipsi and contra sides of the uterus for 16 cows were measured to determine if there was a difference between them on the day of ovulation. We were interested in whether more RBP was produced by cows on the *ipsi* or *contra* sides of the uterus. Specifically, more RBP being produced on the *ipsi* side would suggest a direct role for RBP in implantation.

Click the tabs to see the particular data for the study and how to calculate the confidence interval. When you are ready, click "Next" to continue.

**Tab 1: Study Data**

### Study Data (1/2)

Data in R

1 > data

		ipsi	Contra	Difference
2		8085	6644	1441
3	1	8544	5818	2726
4	2	9002	8942	60
5	3	7786	6939	847
6	4	9498	8594	904
7	5	5906	5488	418
8	6	7078	6124	954
9	7	9766	8137	1629
10	8	7109	6907	202
11	9	7802	6154	1648
12	10	8213	7709	504
13	11	7184	8235	-1051
14	12	9824	9711	113
15	13	7136	6514	622
16	14	7216	6907	309
17	15	7708	5413	2295
18	16			

20 > apply(data[,2:4], 2, mean)

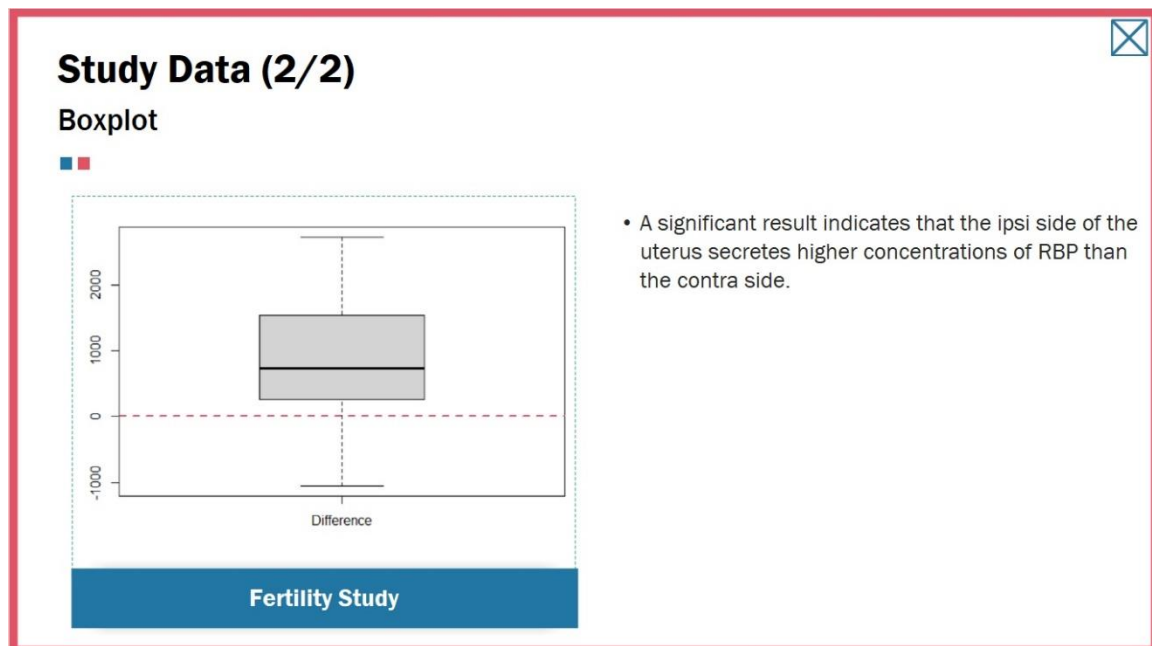
	ipsi	Contra	Difference
21			
22	7991.0625	7139.7500	851.3125

23 > apply(data[,2:4], 2, sd)

	ipsi	Contra	Difference
24			
25	1103.5417	1284.2325	933.3102

As the data is being collected from both sides of the uterus for each cow, we performed a paired hypothesis test. Here are the data for the ipsi and contra sides of each cow, as well as the difference, along with the means and standard deviations.

Tab 1.1: Boxplot



A boxplot of the paired differences is shown on the slide. A majority of observations are clearly positive, and we observed a significant result, indicating that on average, the ipsi side of the uterus secretes higher concentrations of RBP than does the contra side.

Tab 2: Confidence Interval

## Confidence Interval

✕

### Compute Confidence Interval and Confirm Against R Output

**Compare Test Statistic to Critical Values**

**Hypothesis Test R Output**

- The test statistic is calculated using:
  - $\bar{d}$  = Sample mean of the differences
  - $s_d$  = Standard error of the differences
  - $n$  = Sample size
  - $t_c$  = Reference value

**95% Confidence Interval:**

$$\bar{d} \pm t_c * \frac{s_d}{\sqrt{n}} = 851 \pm 2.13 * \frac{933.3}{\sqrt{16}}$$

$$= 851 \pm 497$$

- The long-run mean ipsi contra RBP difference lies between 354 pg/μg and 1,348 pg/μg protein.
- The wide interval reflects the high variability in the data.

```


1 > t.test(data$Ipsi, data$Contra, paired = TRUE)
2
3 Paired t - test
4
5 data : data$Ipsi and data$Contra
6 t = 3.6486, df = 15, p-value = 0.002377
7 alternative hypothesis: true difference in means is not equal to 0
8 95 percent confidence interval:
9  353.9866 1348.6384
10 sample estimates:
11 mean of the differences
12      851.3125
            
```

For the paired t-test, we compute the test statistic using  $\bar{d}$ , the sample mean of the differences,  $s_d$ , the standard error of the differences, and the sample size  $n$ . This is then compared to a reference value  $t_c$ . In this case the 95% confidence interval is simply  $\bar{d} \pm t_c \times s_d/\sqrt{n}$ , a very similar calculation to our last example. So we can say with 95% confidence that the long-run mean ipsi-contra RBP difference lies between 354 and 1348 pg/μg protein. Again, expert knowledge is needed to interpret

this in a meaningful way; however notice that the interval is wide in comparison to the previous example, in the sense that it covers a difference in order of magnitude. This wider interval may be a reflection of the variability inherent in the data – each cow is a distinct biological entity and so we might expect to see more fluctuations in our data than for the other more lab analytic examples we have seen.



Let's return to the hypothesis test output provided by R. You should be able to identify the 95% confident interval included in the output. This output should now be more or less completely clear to you.

## Slide 11: Confidence Interval for Independent Groups



Confidence Interval for Independent Group Tests

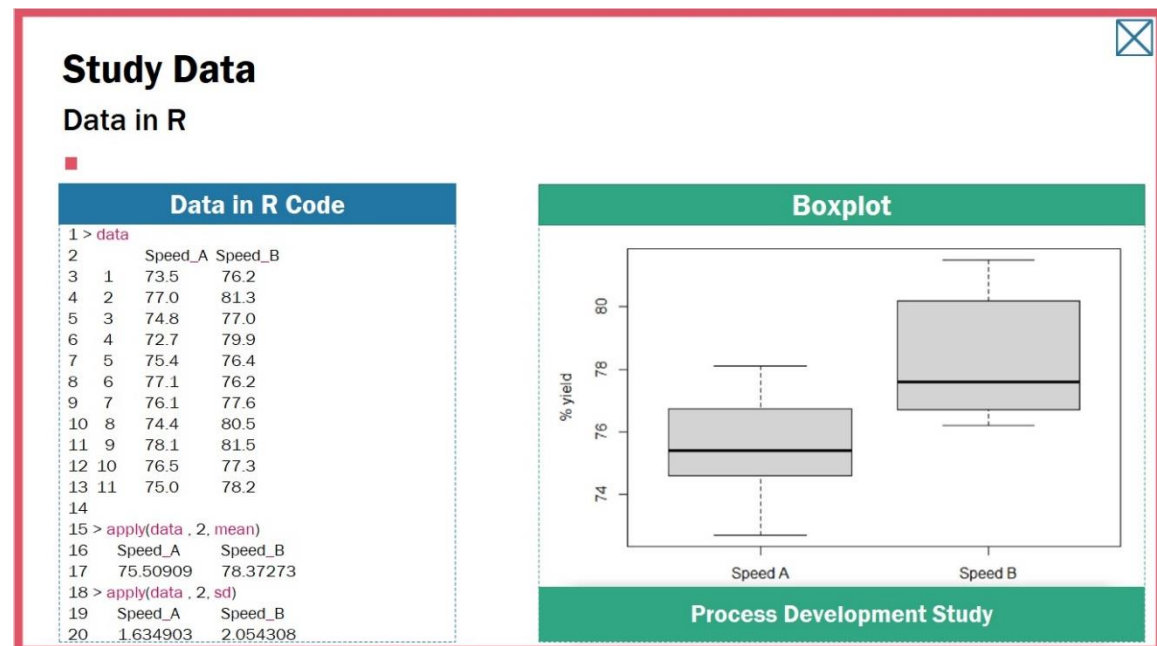
11 of 18

Study Data	Confidence Interval
<div><h3>Introduction</h3><h4>Study Details</h4><ul style="list-style-type: none"><li>How can you optimise a development process for producing small pellets used to fill capsules when making pharmaceutical products?</li><li>The study involved charging a spheroniser with a dough-like material and running it at two speeds (labelled A and B) to determine which gives the higher yield.</li></ul></div> <div></div> <div> Click each tab to learn more. Then, click Next to continue.</div>	

In our final example, we return to the data collected by a research student in the school of pharmacy. Recall that the student was interested in optimising a development process involving producing small pellets that are used to fill capsules when making pharmaceutical products. The study involved charging a spheroniser (essentially a centrifuge) with a dough-like material and running it at two speeds (labelled A and B) to determine which gives the higher yield. The yield is the percentage of the material that ends up as usable pellets (the pellets are sieved to separate out and discard pellets that are either too large or too small).

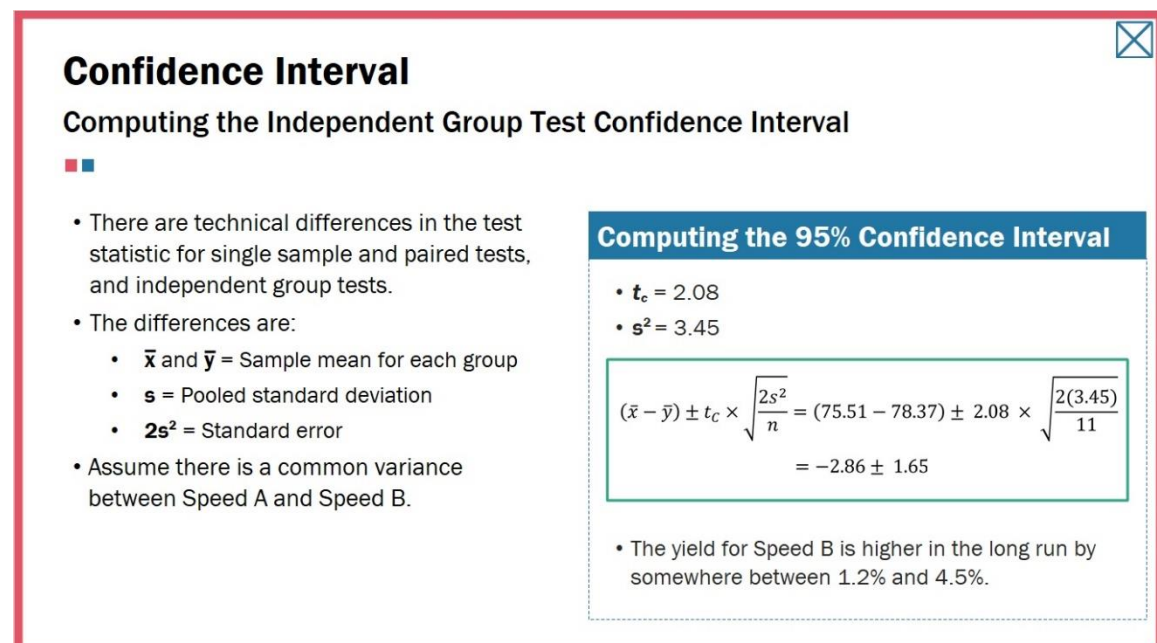
Click the tabs to access the study data and how to compute the confidence interval. When you are ready, click “Next” to continue.

Tab 1: Study Data



The data are shown here. We have eleven observations for each speed type. The order of the runs was randomised, and each run was independent of the others so an independent two sample test was required in this case. Visualising the data, we observed some separation between the yields obtained by each speed, and this difference proved to be statistically significant when a hypothesis test was performed. We concluded that Speed B gave a higher yield on average than Speed A.

Tab 2: Confidence Interval



The independent groups hypothesis test involved computing a similar test statistic to those used in the single sample and paired tests, but with some technical differences. In this case, elements of the test statistic include the sample means for each group,



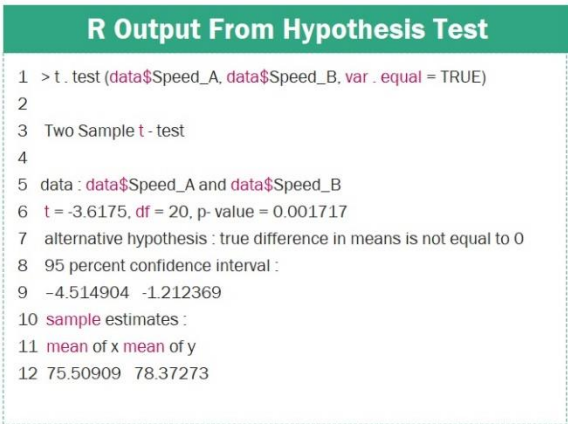
$\bar{x}$  and  $\bar{y}$ , as well as the pooled standard deviation  $s$ . The specific form of the standard error is also a little different as the formula now involves  $2s^2$  and not simply  $s^2$ , as before. We also make the extra assumption that there is a common variance in the data for Speed A and Speed B.

Once these terms, as well as the critical value  $t_c$  have been calculated however, the procedure for estimating the confidence interval is the same as before. In words the 95% CI is given by the difference in the sample means  $\bar{x}$  and  $\bar{y}$   $\pm t_c$  times the standard error,  $\sqrt{2s^2 / n}$ . In this case we have  $-2.86 \pm 1.65$ , so we say with 95% confidence that the yield for Speed B is higher in the long run by somewhere between 1.2 and 4.5%. In this case the evidence points to a reasonable marginal gain in yield if the Speed B settings are used on the spheroniser.

## Tab 2.1: Confirm Result Against R Output

### Confidence Interval

#### Confirm Result Against R Output



```
1 > t.test(data$Speed_A, data$Speed_B, var.equal = TRUE)
2
3 Two Sample t-test
4
5 data : data$Speed_A and data$Speed_B
6 t = -3.6175, df = 20, p-value = 0.001717
7 alternative hypothesis: true difference in means is not equal to 0
8 95 percent confidence interval:
9  -4.514904 -1.212369
10 sample estimates:
11 mean of x mean of y
12 75.50909 78.37273
```


Again, checking the R output for this example, we can see that the format is very similar to the earlier examples we saw, and straightforward to interpret.

## Slide 12: Section 3: Interpreting Hypothesis Results



We have now seen how to compute and interpret confidence intervals for several data examples. Let's now look at how to interpret hypothesis results.

## Slide 13: The Importance of Wording when Interpreting Hypothesis Results



### The Importance of Wording When Interpreting Hypothesis Results


13 of 18

- In example one where we investigated whether an analytical chemistry measurement method was unbiased,  $H_0$  was rejected for Lab B.

#### Innocent Until Proven Guilty

**? We failed to reject  $H_0$  for Lab A. How should we interpret this statement?**

- We do not accept the null hypothesis; we fail to reject it.
  - While there is insufficient evidence to reject the assumption, no evidence has been collected in favour of the proposition either.
  - It is possible that a null hypothesis is false, but the data collected are not sufficient to reject it.




Let's revisit Example 1, where data were collected to assess the measurement accuracy of the lab analytic method. Recall that we rejected the null hypothesis for Lab B: in other words, we concluded that it was unreasonable to assume that the method was unbiased. For Lab A, we failed to reject the null hypothesis. How should we interpret this statement?

When we fail to reject a null hypothesis, we must be careful in our interpretation of the result. Firstly, notice the language that is used here. We do not usually say that we accept the null hypothesis, rather that we fail to reject it. In other words, while we do not have enough evidence to reject or dismiss the assumption, we should be careful to clarify that we have not necessarily collected any evidence in favour of this proposition.

The idea here is very similar to the legal concept of innocent until proven guilty. If we imagine a legal case where someone is being accused of a crime, the burden of proof is on the prosecution to demonstrate beyond reasonable doubt that this person committed the act in question. If the accused is acquitted because e.g., there were insufficient witnesses present at the time the crime was committed, we should not confuse this lack of evidence for an alibi or evidence that the person is truly innocent! Similarly, it is entirely possible that a null hypothesis is in fact false, but that the collected data are not sufficient to reject it.

#### Slide 14: Potential Errors that can Arise



### Potential Errors that can Arise

14 of 18

#### Review the Confidence Interval for Lab A

- $\mu_0 = 27.1\%$
- The confidence interval was 26.73% to 27.15%.
  - The bias of the method is at most 0.4% away from the true reference value.
  - If  $H_0$  was false, the extent of the bias is practically immaterial.
- If the confidence interval was wider at 21.73% to 32.15%:
  - We would fail to reject  $H_0$  as  $\mu_0 = 27.1\%$  is contained in the interval
- The results would be inconclusive as we cannot determine whether the potential bias is positive or negative.
- More data would be required before a verdict could be passed.

#### Potential Errors

• <b>Type I Error</b> - Reject $H_0$ when it is true	• <b>Type II Error</b> - Fail to reject $H_0$ when it is false
--	--


Let's inspect the 95% confidence interval for Lab A. At this point we should be happy to ignore the calculations and focus on the result. Recall that  $\mu_0$  is 27.1%. We can see that the interval ranges from 26.73% to 27.15%. In this case the bias of the method is at most about .4% away from the true reference value. So even if the null hypothesis were false, and the method is biased, it is likely that the extent of the bias is small enough to be practically immaterial. For this reason, it is reasonable to assume the method used by Lab A is unbiased.

Now suppose that the collected data measurements were much more variable, leading to a much wider confidence interval, ranging from 21.73% to 32.15%. In this case there is considerable potential for the method to be biased, by as much as 4 or 5%, a much worse outcome than we saw for the biased measurement system used by Lab B. Notice however that the true reference value of 27.1 is still contained within this interval, so that in this case we would also fail to reject the null hypothesis. In this case, it is entirely

possible that the method used by Lab A is biased; however, because we cannot say whether the potential bias is positive or negative, our results are inconclusive and hence would not be significant. In this case, however, we should be much less reassured by the result and the assumption the method used by Lab A is unbiased would be very much in doubt. The most reasonable conclusion to draw in this case is that more data would be required before a verdict could be passed.

This hypothetical example highlights the two ways in which we can make errors when performing a hypothesis test. Firstly, we can reject the null hypothesis  $H_0$  when it is in fact true. In this case we will have confused simple random variation for something more systematic. We call this a Type I error. Secondly, we can fail to reject the null hypothesis  $H_0$ , even though it is in fact false. This can happen when we are unable to distinguish systemic differences in our data from simple random variation. This is called a Type II error.

### Slide 15: Type I and Type II Errors



## Type I and Type II Errors

15 of 18

	$H_0$ is true	$H_0$ is false
Fail to reject $H_0$	Correct decision	Type II Error
Reject $H_0$	Type I Error	Correct Decision

**Medical Trial Example: Consequences of Errors**

**If a Type I Error was committed:**

- Clinicians would believe that the treatment worked, when it did not
- Distress or harm could be caused to patients and their family if the treatment did not improve their health as much as expected
- Later studies that based their own research aims and hypotheses on this study, would be likely to fail or be unable to replicate the previous study findings

**If a Type II error was committed:**

- Treatment effectiveness would be missed and patients would miss out on a beneficial treatment
- An opportunity to improve population health would be missed
- The research time, effort and costs of all parties involved would be wasted
- A promising future direction for further research would also be missed

This table gives a systematic overview of the types of error that can occur when performing a hypothesis test. The table columns correspond to the true state of nature, i.e., whether or not the null hypothesis is in fact true or false. Of course, in reality this true state in nature is unfortunately never known. Rows in the table correspond to the decisions we make based on the data; that is, failing to reject, or rejecting, the null hypothesis.

Hence, whenever we perform a hypothesis test, one of four possible outcomes occur. If in reality the null hypothesis is true, and we fail to reject  $H_0$ , or if we reject the null and it is in fact false, we have made the correct decision and all is well. These decisions correspond to the top left and bottom right entries in the table. However, if we reject the null when it is fact false we have made the wrong decision and committed a Type I Error, the bottom left entry in the table. If we fail to reject the null hypothesis when it is in fact false we commit a Type II Error, the top right entry in the table [...](#)


While these images should not to be taken too seriously, they help to distinguish between the two types of error at a conceptual level. Clearly Type I and Type II errors are different, and the consequences of making these errors are different. For example, suppose a medical trial was conducted, with the goal of assessing a new form of treatment for a serious medical condition. During the trial data would be collected, and a statistical would be performed to assess whether or not the treatment was beneficial to patients. Following the test, if a Type I error were committed, this would mean that

- Clinicians would believe that the treatment worked, when it did not, or at least it did not perform better than existing standard of care;
- Distress or even harm could be caused to patients and their family, who would receive a treatment that did not improve their health as much as expected, with potentially other side effects to cope with;
- Later studies that based their own research aims and hypotheses on this study would be likely to fail, or be unable to replicate the previous study findings.

On the other hand, if a Type II error were committed, then

- Treatment effectiveness would be missed and patients would miss out on a beneficial treatment;
- Opportunity to improve population health missed;
- The research time and effort, not to mention the costs, of all parties involved in the trial would be wasted;
- A promising future direction for further research would also be missed.

## Slide 16: Defining Type I and Type II Errors



Defining Type I and Type II Errors

16 of 18

Defining Type I and Type II Errors	Power
<ul style="list-style-type: none"><li>• Denote <math>\alpha = \mathbb{P}</math> (Type I error)<ul style="list-style-type: none"><li>• <math>\alpha</math> is usually specified by the researcher as 0.05.</li></ul></li><li>• Denote <math>\beta = \mathbb{P}</math> (Type II error)<ul style="list-style-type: none"><li>• <math>1 - \beta</math> = Power of a statistical test<ul style="list-style-type: none"><li>• If a test is powerful, it will be sensitive to departures from <math>H_0</math>.</li></ul></li></ul></li></ul>	<ul style="list-style-type: none"><li>• Increase the sample size <math>n</math> to make the test more powerful.</li><li>• Determine whether a test is sufficiently powered before collecting data.<ul style="list-style-type: none"><li>• If variability is too high, or the effect being examined is not strong enough, then it may not be feasible to expect a study to deliver conclusive results.</li></ul></li></ul>
<b>The power <math>\beta</math> depends on:</b>	<ul style="list-style-type: none"><li>• True mean <math>\mu_0</math></li><li>• Sample size <math>n</math></li><li>• Sample error <math>s</math></li><li>• Choice of <math>\alpha</math></li></ul>
<b>! Decreasing <math>\alpha</math>, increases <math>\beta</math>.</b> <ul style="list-style-type: none"><li>• Type I and Type II errors cannot be eliminated, only balanced.</li></ul>	

Let's define what we mean by these types of error a little more precisely. Let alpha denote the probability of committing a Type I error. This is specified by the researcher, and is most commonly set to be  $\alpha = 0.05$ , as we discussed in the last session. Let beta denote the probability of making a Type II error. Then we call  $1 - \beta$  the power of a




statistical test. If a test is powerful, then the test will be sensitive to departures from the null hypothesis. While we specify alpha ourselves, it is more challenging to determine beta. This is because its value partly depends on the value of the true population mean  $\mu$ , which of course is never known. Supposing a fixed value for  $\mu$ , the power of a test will also depend on the sample size  $n$ , the sample error  $s$ , and our choice of alpha.

Note that decreasing alpha increases beta. That is, if we reduce the probability of making a Type I error we will increase the probability of making a Type II error. This should make sense intuitively – by decreasing alpha we increase the amount of evidence we need before we will reject the null hypothesis, so that the burden of proof increases, making it more difficult to form concrete conclusions. This is the nature of the relationship between Type I and Type II errors – they cannot be eliminated, only balanced.

The simplest way to make a test more powerful is to increase the sample size. In this way we accumulate more evidence and can more accurately assess the hypothesis we are testing. Note that simple is not the same as easy! It can often be very challenging to collect data – cases can be rare and recording the information correctly can be expensive and time consuming.

This is why it is very important to determine whether a test is sufficiently powered before any data collection has been carried out. If variability is too high, or the effect being examined is not strong enough, then it may simply not be feasible to expect a study to deliver conclusive results, in which case the researcher's, i.e., your own efforts may all go to waste. Such a study is called power analysis, and will be covered in more detail in later sessions.


## Slide 17: Conclusion



### Conclusion

17 of 18

- We have covered:
  - An overview of how confidence intervals are calculated and how they should be interpreted
  - How to calculate confidence intervals for:
    - One sample tests
    - Paired tests
    - Independent group tests
  - Confidence intervals help us to interpret and understand the outcome of a hypothesis test in a more nuanced way.
  - When forming conclusions based on these outcomes, we should be clear about the concepts of Type I and Type II errors.




We have now covered an in-depth overview of how confidence intervals are calculated and how they should be interpreted. We calculated confidence intervals for the three kinds of hypothesis test:

- One sample tests;
- Paired tests
- Independent group tests

Confidence intervals help us to interpret and understand the outcome of a hypothesis test in a more nuanced way. Specifically, when forming conclusions based on these outcomes we should be clear about the concepts of Type I and Type II errors

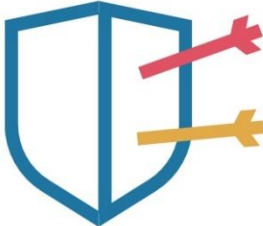
## Slide 18: Summary



### Summary

18 of 18

- Having completed this presentation, you should be able to:
  - State what calculations are involved in computing confidence intervals
  - Interpret the interval in the context of the data being analysed
  - Identify and interpret the relevant elements from standard statistical software output
  - Differentiate between Type I and Type II errors
  - Explain the concept of statistical power and how it relates to these errors



Developed by Trinity Online Services CLG with the School of Computer Science and Statistics, Trinity College Dublin, The University of Dublin

Having completed this presentation, you should be able to:

- State what calculations are involved in computing confidence intervals
- Interpret the interval in the context of the data being analysed
- Identify and interpret the relevant elements from standard statistical software output
- Differentiate between Type I and Type II errors
- Explain the concept of statistical power and how it relates to these errors