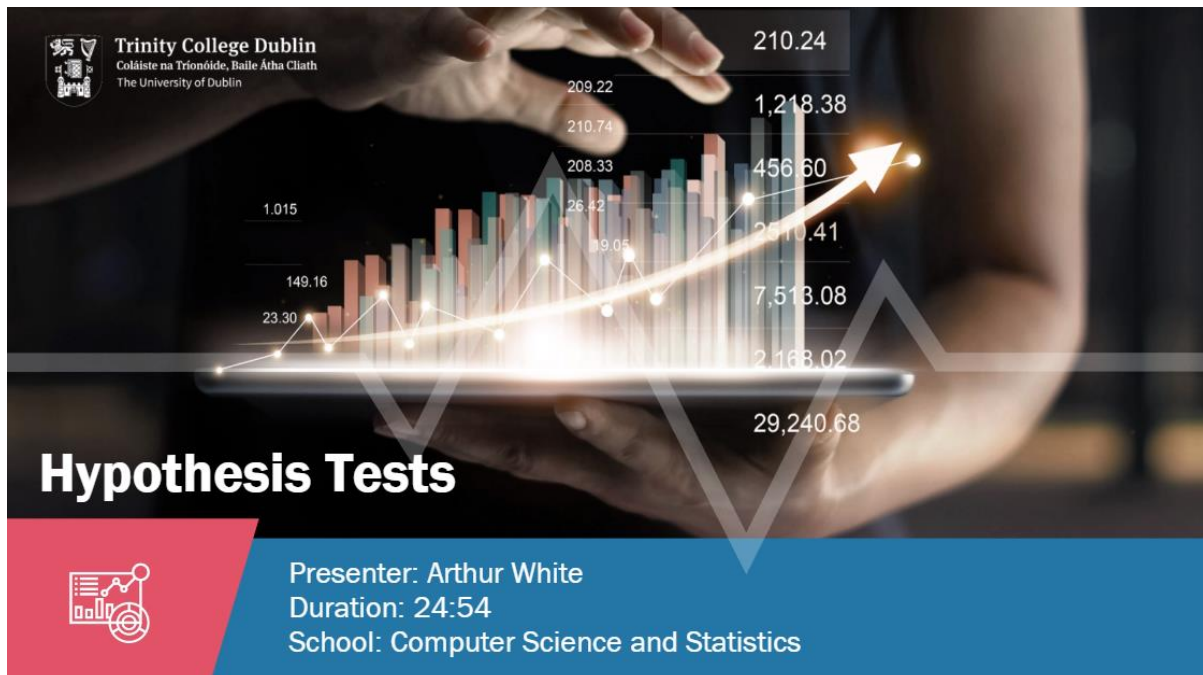


## Hypothesis Tests

Slide 1:	Introduction .....	2
Slide 2:	Motivating Examples of Hypothesis Tests .....	2
Slide 3:	Section 1: Background Concepts From Probability .....	4
Slide 4:	Statistically Significant Data.....	4
Slide 5:	Probability Review .....	5
Tab 1:	Answer to Question A.....	6
Tab 2:	Answer to Question B .....	6
Slide 6:	Differences Between Probability Problems and Hypothesis Tests.....	7
Slide 7:	Section 2: Performing Hypothesis Tests.....	8
Slide 8:	Example One: One Sample Tests.....	8
Tab 1:	Examine the Data .....	9
Tab 2:	Test the Hypothesis .....	10
Slide 9:	Hypothesis Testing Procedure.....	13
Tab 1:	Step 1: Specify the Null Hypothesis .....	14
Tab 2:	Step 2: Specify the Test Statistic .....	14
Tab 3:	Step 3: Specify the Significance Level .....	15
Tab 4:	Step 4: Compute the Test Statistic.....	16
Tab 5:	Step 5: Reject or Accept $H_0$ .....	16
Slide 10:	Paired Comparisons.....	17
Slide 11:	Example Two: Paired Test.....	18
Tab 1:	Examine the Data .....	19
Tab 2:	Test the Hypothesis .....	20
Slide 12:	Independent Groups Test.....	21
Slide 13:	Example Three: Two Sample $t$ -test .....	22
Tab 1:	Examine the Data .....	23
Tab 2:	Test the Hypothesis .....	23
Slide 14:	Conclusion .....	25
Slide 15:	Summary.....	26

## Slide 1: Introduction



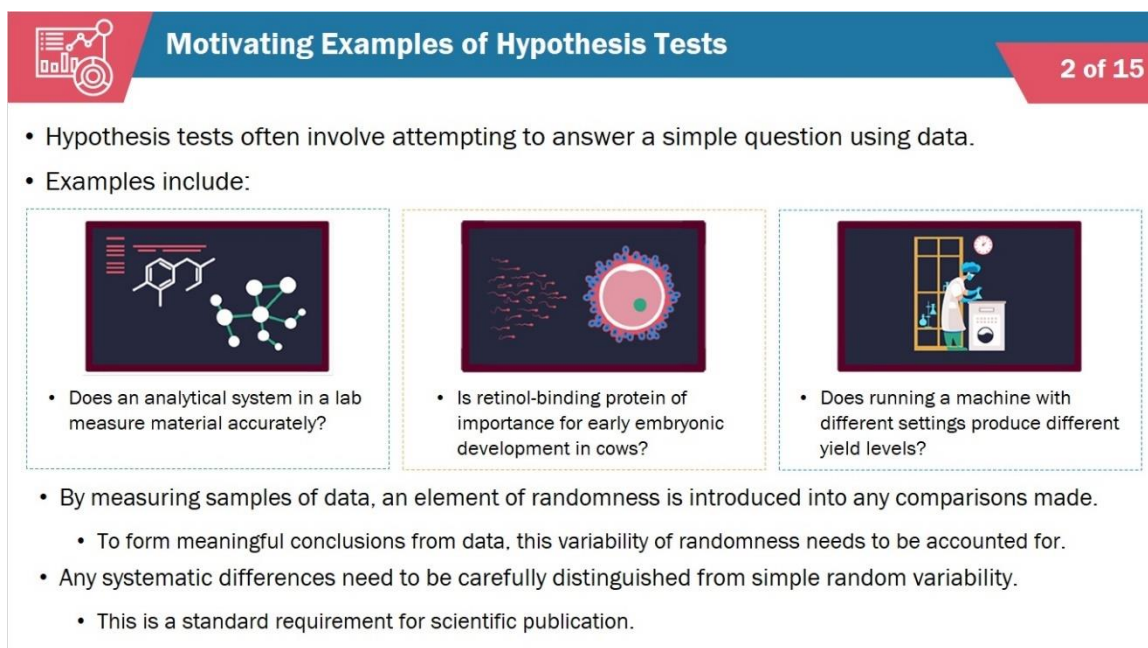
**Hypothesis Tests**

Presenter: Arthur White  
Duration: 24:54  
School: Computer Science and Statistics

Hello and welcome to this presentation on Hypothesis Tests. My name is Arthur White and I will lead you through this presentation.

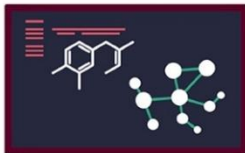
During this presentation, I will introduce and describe some simple hypothesis tests. We will examine what is meant by the terms hypothesis test and statistical significance; what calculations are required to perform an hypothesis test; and how to determine which test is most appropriate for your data. We will explore the principles behind hypothesis testing and go through several examples. These examples will cover the most popular kinds of hypothesis test that are used to perform simple comparative analyses.

## Slide 2: Motivating Examples of Hypothesis Tests

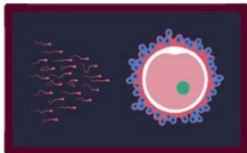


**Motivating Examples of Hypothesis Tests** 2 of 15


- Hypothesis tests often involve attempting to answer a simple question using data.
- Examples include:



- Does an analytical system in a lab measure material accurately?



- Is retinol-binding protein of importance for early embryonic development in cows?



- Does running a machine with different settings produce different yield levels?

- By measuring samples of data, an element of randomness is introduced into any comparisons made.
  - To form meaningful conclusions from data, this variability of randomness needs to be accounted for.
- Any systematic differences need to be carefully distinguished from simple random variability.
  - This is a standard requirement for scientific publication.

Let's consider some motivating examples for these kinds of tests. As the name suggests, hypothesis tests often involve attempting to answer a simple question using data. Here are some examples:

- Does an analytical system in a lab measure material accurately?
- Is retinol-binding protein of importance for early embryonic development in cows?
- Does running a machine with different settings produce different yield levels?

These examples are related to chemistry, agricultural science and pharmacology respectively and are all real case studies; the latter two analyses were performed by former research students here in Trinity College. We will explore these examples in more detail later in this presentation.

Of course, these are only three specific examples. Hypothesis testing is extremely general and can be applied to almost any empirical discipline. In my own research, I have performed hypothesis tests in fields as diverse as education and immunology; one of the first recorded cases of a hypothesis test was a medical application. Students should be able to come up with examples from their own areas of research.

From a technical perspective, the key idea here is that by measuring samples of data, we introduce an element of randomness into any comparisons that we make. For example, measurements made by an analytical lab system may be challenging to make precisely, and a small amount of error may be introduced with each data point; or there may naturally be variability between observations in a population, such as in the bovine study; or observations may vary from day to day, such as student performance in an exam.

To form meaningful conclusions from data, this variability or randomness needs to be accounted for. Any systematic differences that we believe we observe need to be carefully distinguished from simple random variability. This is a standard requirement for scientific publication.

In this presentation, we will discuss various examples of hypothesis tests that are widely used across most disciplines where empirical data are assessed. These tests will illustrate the underlying concepts and ideas behind hypothesis tests. These examples will all have several important features in common.

## Slide 3: Section 1: Background Concepts From Probability




3 of 15

# Background Concepts From Probability

Before proceeding any further, let's revise some important background concepts from probability in the context of hypothesis testing. The calculations that follow assume some knowledge of simple probability that you will have covered in earlier sessions. Specifically, you should already be familiar with how to estimate the area under a normal curve, using either statistical tables or software.


## Slide 4: Statistically Significant Data



Statistically Significant Data

4 of 15


- The key idea of a statistical hypothesis test is to identify whether an observed result is “unusually” different than the one expected, simply because of chance variation.
- If the observed data is unusual, the result is termed as **statistically significant**.
  - The observed difference suggests that a systematic difference from the hypothesised value exists.



The key idea of a statistical hypothesis test is to identify whether an observed result is “unusually” different than one we would expect to see simply because of chance variation. If the observed data really is unusual in this respect, then we term the result as **statistically significant**. The term “signify” here has a technical meaning closer to “identify” than “important”, as we have identified a systematic difference in the data

distinct from chance variation. In other words, the observed difference suggests that a systematic difference from the hypothesised value exists.

## Slide 5: Probability Review



### Probability Review


5 of 15

#### Answer to Question A

#### Answer to Question B

#### Introduction

- Ideas from probability are used to determine when observed values can be considered unusual.
- The heights of a population of women follow a normal distribution  $X \sim N(\mu, \sigma^2)$ , where:
  - $\mu = 162.4$  cm
  - $\sigma = 6.3$  cm
- Consider the following questions:
  - Is a woman from this population whose height is 178.1 cm unusually tall?
  - Can we determine cut-off values for this population to determine what heights correspond to being unusually tall or short?

 Click each tab to learn more. Then, click Next to continue.

We use ideas from probability to determine when observed values can be considered unusual. Consider the following hypothetical example. Suppose there is a population of women whose heights are normally distributed with mean,  $\mu = 162$ cm and standard deviation,  $\sigma = 6.3$ cm. Consider the following questions: a) is a woman from this population whose height is 178.1 cm unusually tall? and b) can we determine cut off values for this population to determine what heights correspond to being unusually tall or short?

Click the tabs to learn the answer to each question. When you are ready, click “Next” to continue.



Tab 1: Answer to Question A

✕

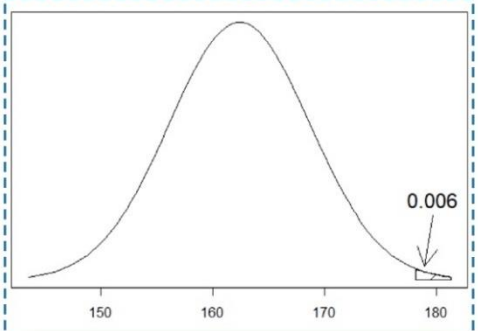
## Answer to Question A

### Determine the Probability

**? Is a woman from this population whose height is 178.1 cm unusually tall?**

- Determine the probability of encountering a woman taller than 178.1 cm:
 

$$P(X > 178.1) = P(Z > \frac{178.1 - 162.4}{6.284}) = P(Z > 2.5) = 0.006$$
- The probability is less than one percent.
  - This implies that meeting someone of this height every day would not be a frequent occurrence.



Heights Curve

To answer the first question, we ask what is the probability of encountering a woman taller than this height. If  $P(X > 178.1) = P(Z > \frac{178.1 - 162.4}{6.284}) = P(Z > 2.5) = 0.006$  This is simply the area under the curve of a standard normal distribution to the right of 2.5. We use tables or software to find that this value is 0.006. Since this value is less than one percent, clearly meeting a woman of this height would not be an everyday occurrence.

Tab 2: Answer to Question B

✕

## Answer to Question B

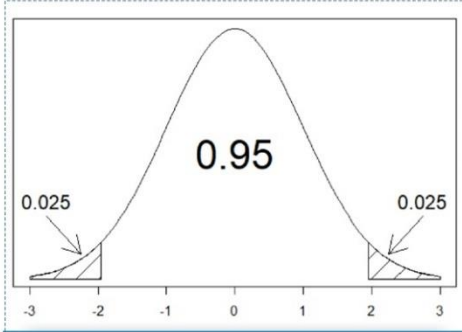
### Determine the "Unusual" Value

**? Can we determine cut-off values for this population to determine what heights correspond to being unusually tall or short?**

- Decide what value constitutes "unusual".
  - Take 5% to be unusual for this example.
- Calculate the corresponding cut-off values of a standard normal  $Z \sim N(0, 1)$ .
  - The tables show a cut-off value of  $Z = \pm 1.96$
- Calculate the women's heights using:  $\mu \pm 1.96\sigma$ 

$$\mu \pm 1.96\sigma$$

  - Calculation =  $16.24 \pm 1.96(6.3)$
  - Cut-off values are 150.08 cm and 174.72 cm.




Standard Normal

To answer the second question, we first need to decide what value constitutes "unusual". If we take a value of 5%, i.e., about a one in twenty occurrence, as being unusual, then we can find the corresponding cut-off values of a standard normal that

have 2.5% probability to their left and right respectively. Again, using software or tables it is elementary to show that this corresponds to a cut-off value of  $Z = \pm 1.96$ . For the women's heights, this becomes  $\mu \pm 1.96 \sigma$ .

So, values less than about 150.1 cm and above 174.7 cm are unusually short or tall. Notice that all we had to do to determine whether a height of 178 cm was unusual was to compare it to this reference value; as  $178 > 174.7$  we would form the correct conclusion.



## Slide 6: Differences Between Probability Problems and Hypothesis Tests




Differences Between Probability Problems and Hypothesis Tests

6 of 15


- For all case studies in the presentation, we will:
  - Standardise an observed value
  - Compare it to the cut-off points of a reference distribution

Technical Differences	
	<b>Hypothesis Test</b> <ul style="list-style-type: none"><li>The focus is on the sample mean.</li><li>The estimated sample standard deviation <math>s</math> is used.</li></ul>
	<b>Probability Problem</b> <ul style="list-style-type: none"><li>The focus is on a single observation.</li><li>The true population standard deviation <math>\sigma</math> is known.</li></ul>

 Even though the calculations will change, the underlying concept remains the same.

The probability exercises performed in this example should be familiar to you from previous sessions. We will apply this idea in all of the case studies that follow: standardising an observed value, then comparing it to the cut-off points of a reference distribution. However there are two important technical differences to consider when looking at hypothesis tests: 1) that we will focus on the sample mean, and not a single observation; and 2) that the true population standard deviation,  $\sigma$ , will not be known in practice, and we will instead be using the estimated sample standard deviation  $s$ . These differences will mean that our calculations will have to change somewhat. But it is worth remembering that the underlying concept is still the same.

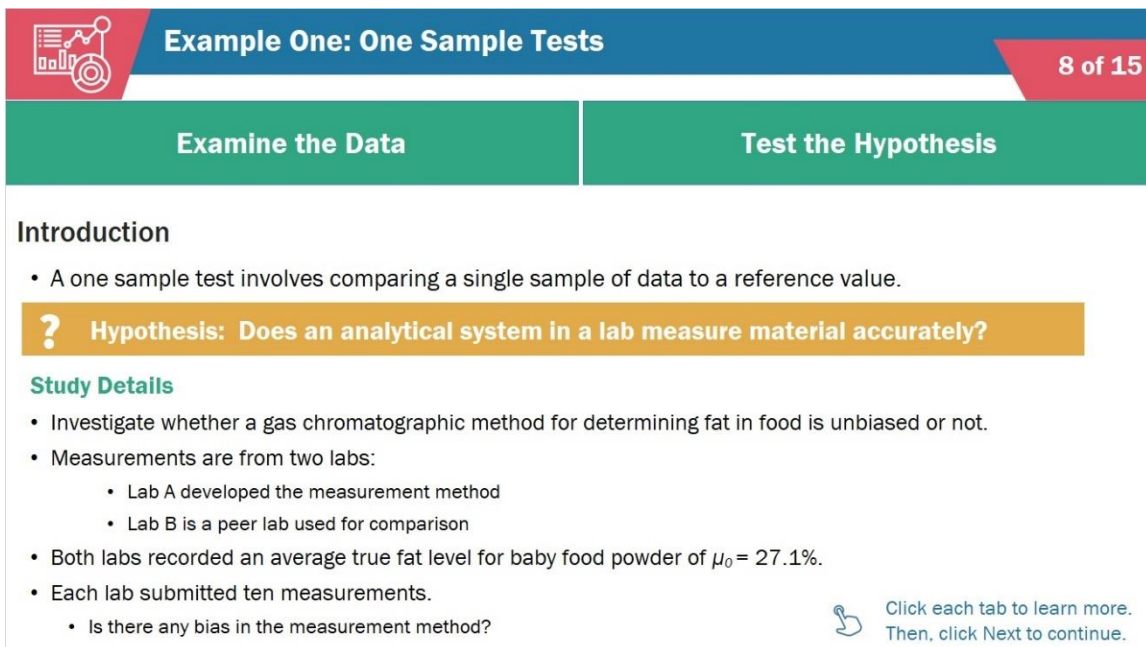
## Slide 7: Section 2: Performing Hypothesis Tests




Slide 7 features a blue header with a red icon of a bar chart and a target. The main content area has a dark green background with a white diagonal line. The text "Performing Hypothesis Tests" is written in white. A red banner in the top right corner says "7 of 15".

At the start of this presentation, I mentioned three different hypotheses. Each hypothesis requires a different type of hypothesis test that we will look at in detail over the following slides.

## Slide 8: Example One: One Sample Tests



Slide 8 features a blue header with a red icon of a bar chart and a target. The main content area has a green background with a white diagonal line. The text "Example One: One Sample Tests" is written in white. A red banner in the top right corner says "8 of 15".

Examine the Data	Test the Hypothesis
<p><b>Introduction</b></p> <ul style="list-style-type: none"><li>A one sample test involves comparing a single sample of data to a reference value.</li></ul> <p><b>? Hypothesis: Does an analytical system in a lab measure material accurately?</b></p> <p><b>Study Details</b></p> <ul style="list-style-type: none"><li>Investigate whether a gas chromatographic method for determining fat in food is unbiased or not.</li><li>Measurements are from two labs:<ul style="list-style-type: none"><li>Lab A developed the measurement method</li><li>Lab B is a peer lab used for comparison</li></ul></li><li>Both labs recorded an average true fat level for baby food powder of <math>\mu_0 = 27.1\%</math>.</li><li>Each lab submitted ten measurements.<ul style="list-style-type: none"><li>Is there any bias in the measurement method?</li></ul></li></ul> <p> Click each tab to learn more. Then, click Next to continue.</p>	

Our first example involves comparing a single sample of data to a reference value. This is known as a one sample, or single sample hypothesis test.

We look at the following hypothesis: Does an analytical system in a lab measure material accurately?



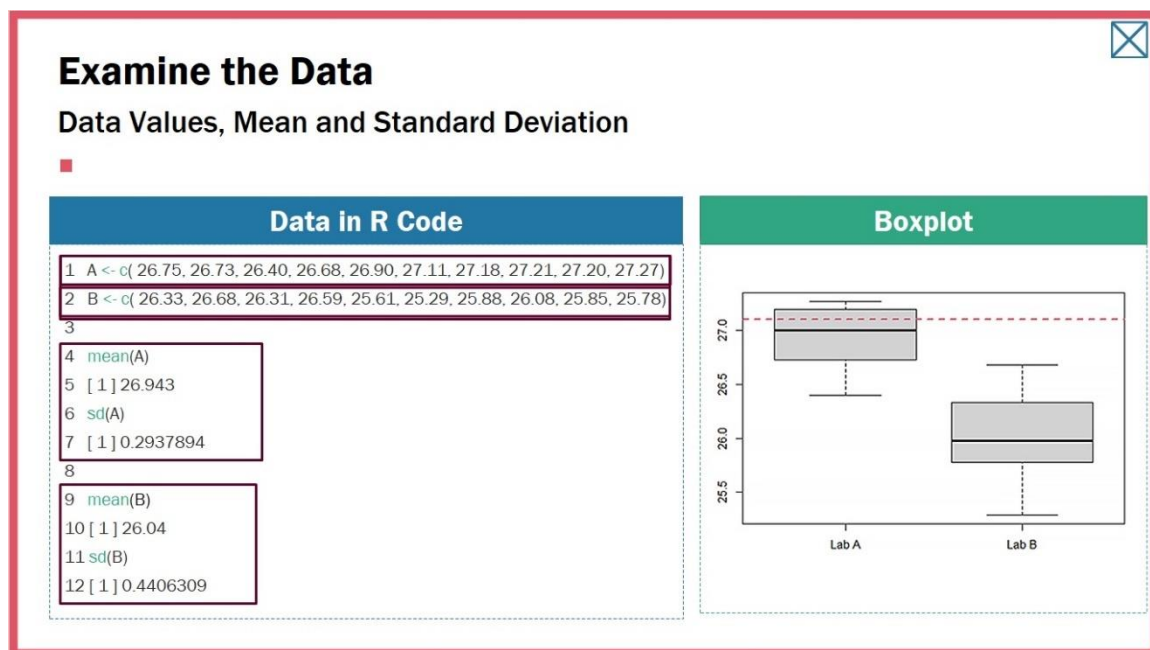
This relates to an analytical chemistry measurement method concerned with measuring quantities in an unbiased way. We must determine whether or not the measurement method does indeed lead to unbiased results. By unbiased, we mean that individual measurements may be larger or smaller than the ‘true’ value for the quantity being measured, but the deviations can be considered to be purely chance fluctuations.

If enough measurements were made and averaged, the deviations should cancel and result in the ‘true value’. There may be (hopefully small) errors in individual measurements but these errors should cancel each other out on average.

Specifically, we investigate whether a gas chromatographic method for determining fat in food is unbiased or not. We have measurements from two labs, Lab A and Lab B. The measurement method was developed in Lab A; Lab B is a peer lab used for comparison. Both Labs recorded measurements for a reference material, baby food powder, where the average true fat level is certified to be  $\mu_0 = 27.1\%$ . We have ten measurements of this material from each lab. Can we determine any bias in the measurement method?

Click the tabs to work through this example. When you are ready, click “Next” to continue.

**Tab 1: Examine the Data**



Before performing any tests, it's always a good idea to look at the data first. Here we can see the data values for Lab A and Lab B as they would be entered into R. We also have the mean and standard deviations for each lab. Notice that the data values for Lab A range between 26.4 and 27.2, and that the mean value of 26.9 seems quite close to the true mean of 27.1. More concerningly, all of the observed values for Lab B are below this value.

A boxplot of the data for Lab A and Lab B are also shown. Notice the true value mean of 27.1%, illustrated by the dashed red line, overlaps with the “box”, i.e., the middle 50% of the data, for Lab A. However, the line is completely missed by both the box and whiskers for Lab B.

Tab 2: Test the Hypothesis


## Test the Hypothesis (1/4)

### Test Statistic $t$

- Test the hypothesis that the Lab A measurement is biased.
- The sample mean is  $\bar{x} = 26.9$  which is lower than the true value  $\mu_0 = 27.1$ .
  - This may be down to random chance.
- Use the test statistic  $t$  to formally test whether the difference in mean values is highly unusual in the context of random variation:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

- The standard error measures the variability of  $\bar{x}$ .
- $t$  is the standardised version of our observed sample mean.
- As  $s$  is estimated, we use a Student  $t$ -distribution.
- The distribution is defined in terms of its degrees of freedom,  $\nu$ .
  - $\nu = n - 1$

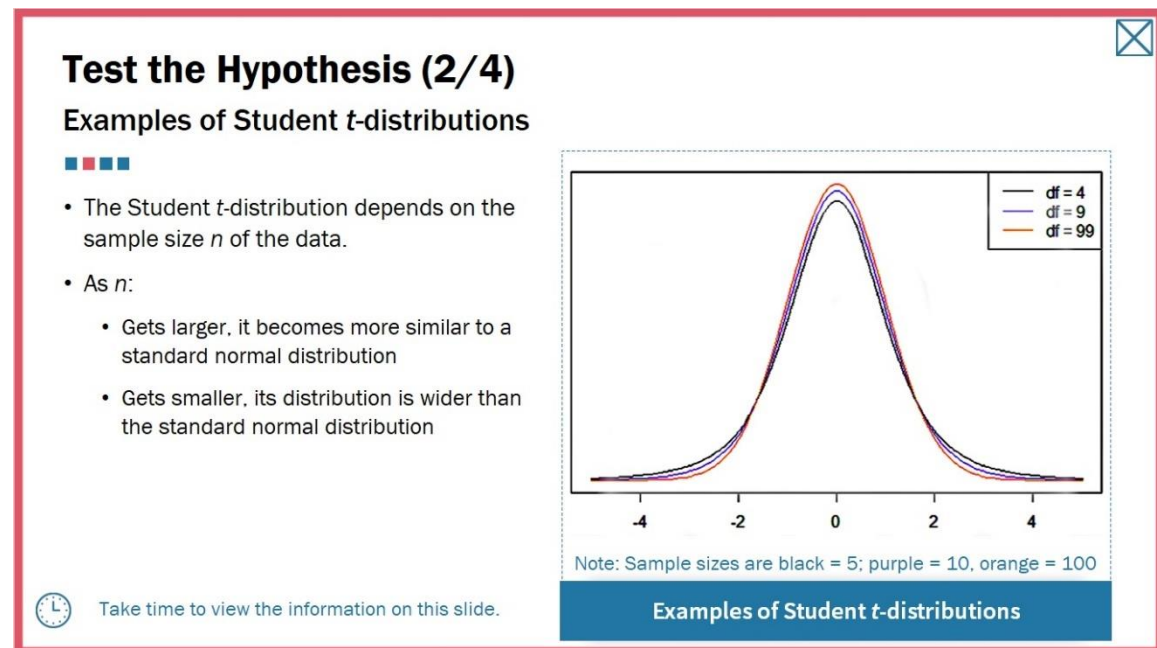


We'll now formally test the hypothesis that the Lab A measurement is biased. While the observed mean of 26.9 is lower than the true value of 27.1, this may be down to random chance. Indeed, it would be highly unusual for the sample mean to equal the true mean exactly even if the method was completely unbiased. To formally test whether the difference in mean values is highly unusual in the context of random variation, we construct the test statistic  $t$ .

This is simply the difference between the sample mean  $\bar{x}$  and the true mean  $\mu_0$  divided by the standard error. The standard error measures the variability of the sample mean  $\bar{x}$ , and is simply calculated as the standard deviation divided by the square root of the sample size of the data.

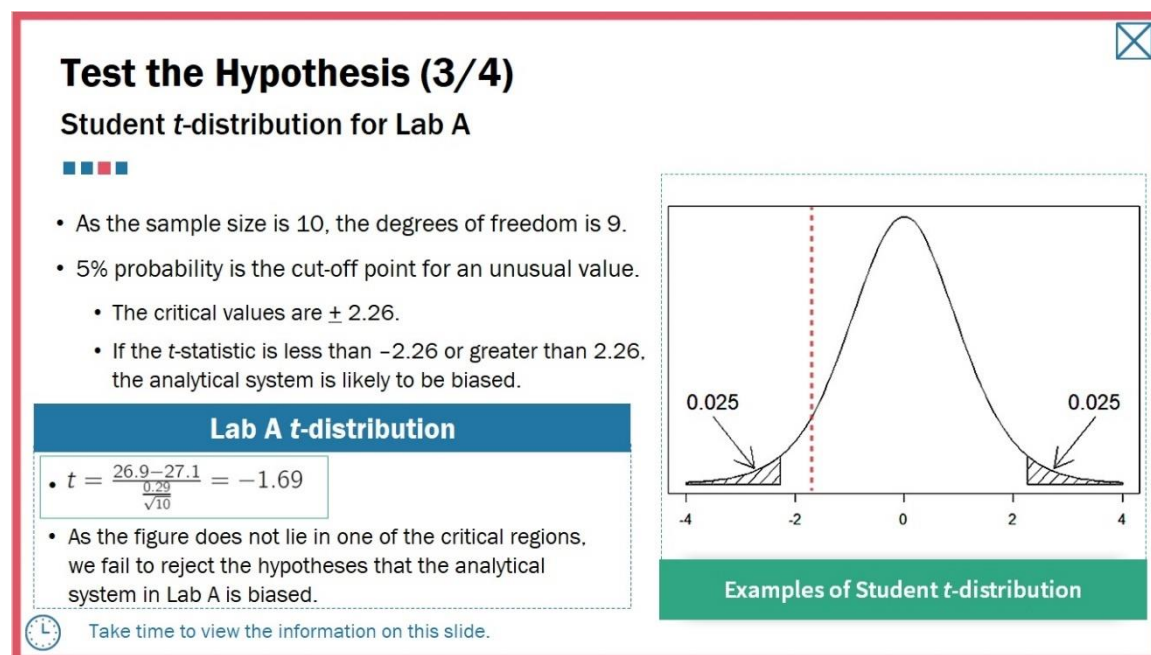
So  $t$  is the standardised version of our observed sample mean. It remains to identify the reference distribution and its corresponding cut-off values. Because  $s$  is estimated, we cannot use a normal distribution like in our previous hypothetical example. Instead we use a Student  $t$ -distribution. This distribution is defined in terms of its degrees of freedom,  $\nu$ . In this case the degrees of freedom equal the sample size,  $n$ , -1.

Tab 2.1: Examples of Student  $t$ -distributions



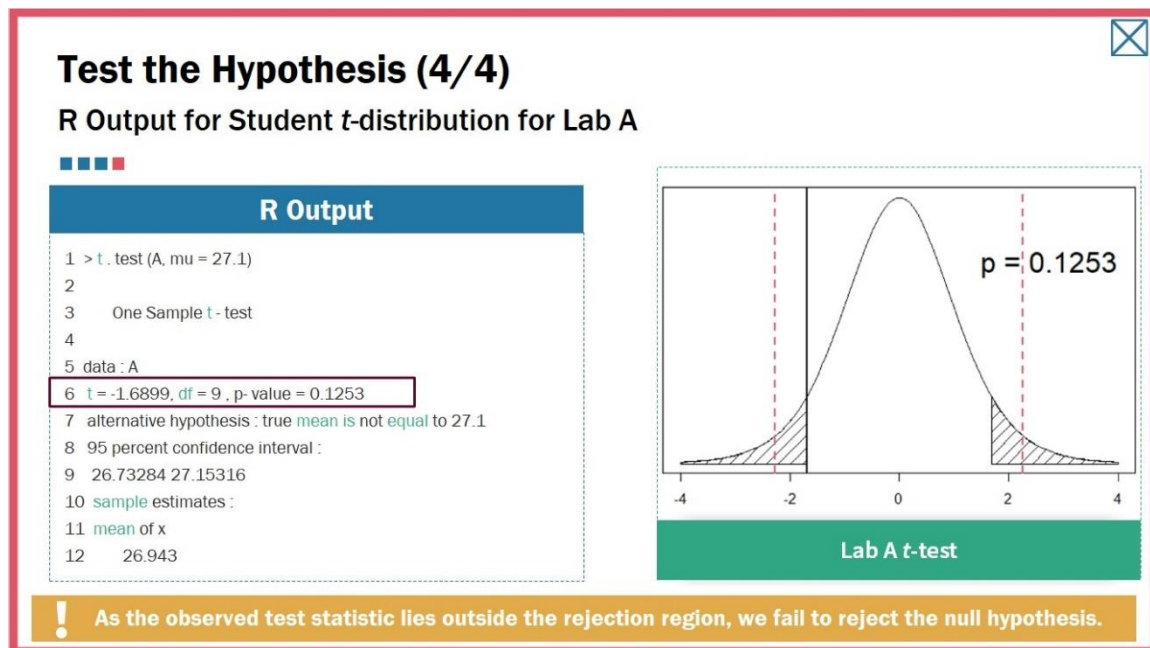
The Student  $t$ -distribution depends on the sample size  $n$  of the data set. Some examples of student  $t$ -distributions are shown here. The black curve has four degrees of freedom, the purple curve has nine degrees of freedom, and the orange curve has 99 degrees of freedom. This corresponds to a sample size of 5, 10 and 100 respectively. As  $n$  gets larger, it becomes more similar to a standard normal distribution. Whereas for smaller sample sizes, its distribution is wider than the standard normal distribution. This is evident in the example shown here. Notice that the distribution of the black curve is *wider in the tails* than the orange and purple curves. In other words, the probability of seeing a value farther from the mean is higher for the purple and orange curves than for the black curve. As the degrees of freedom increase, the curves become very similar to a standard normal distribution.

Tab 2.2: Student *t*-distribution for Lab A




For this data we have a sample size of 10 and hence 9 degrees of freedom. Taking 5% probability as our cut off point for an unusual value, we can use statistical tables or software to find that the corresponding critical values are plus/minus 2.26. This is illustrated in the graph. The area under the curve to the left and right of the tails at these values equals 2.5% and hence totals 5%. If we observe a *t* statistic less than  $-2.26$  or greater than  $2.26$  we will have to conclude that the analytical system is likely to be biased.

In this case we observe a test statistic of  $-1.69$ . This value lies between  $-2.26$  and  $2.26$ , as illustrated by the red dashed line in the figure. In other words, the value does not lie in one of the critical, or rejection regions, of the test. Formally, we fail to reject the hypothesis that the analytical system in Lab A is biased.

Tab 2.3: R Output for Student *t*-distribution for Lab A

This slide shows the output you would see if you performed the test in R. Notice that the output includes the test statistic *t*, the degrees of freedom, denoted *df*, and a *p*-value. The *p*-value is the area under the curve to the right and left of plus/minus the test statistic. This is illustrated in the figure. Notice that the *p*-value here equals 0.125, i.e., it is larger than 5%. This is equivalent to the observed test statistic lying outside the rejection region, and the conclusion that we fail to reject the null hypothesis.

### Slide 9: Hypothesis Testing Procedure

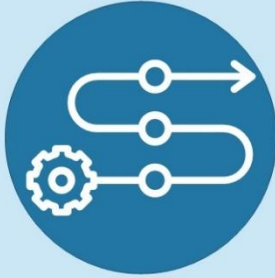



## Hypothesis Testing Procedure

9 of 15

- There is a standard procedure for hypothesis tests that follows five steps.
- Lab B Data:
  - `B <- c(26.33, 26.68, 26.31, 26.59, 25.61, 25.29, 25.88, 26.08, 25.85, 25.78)`
  - Average true fat level for baby food powder recorded is  $\mu_0 = 27.1\%$

- Specify  $H_0$
- Specify the Test Statistic
- Specify the Significance Level
- Compute the Test Statistic
- Reject or Accept  $H_0$




 Click each tab to learn more. Then, click Next to continue.

We have just performed our first hypothesis test. Generically, the steps of a hypothesis test follow a standard procedure. Let's review this testing procedure more formally and apply it to the data from Lab B.



Click the tabs to learn about what happens at each step in the standard procedure.  
When you are ready, click “Next” to continue.

## Tab 1: Step 1: Specify the Null Hypothesis



### Hypothesis Testing Procedure

9 of 15

- There is a standard procedure for hypothesis tests that follows five steps.
- Lab B Data:
  - B <- c( 26.33, 26.68, 26.31, 26.59, 25.61, 25.29, 25.88, 26.08, 25.85, 25.78)
  - Average true fat level for baby food powder recorded is  $\mu_0 = 27.1\%$

**1. Specify  $H_0$**

**2. Specify the Test Statistic**


**3. Specify the Significance Level**

**4. Compute the Test Statistic**

**5. Reject or Accept  $H_0$**


**1. Specify  $H_0$** 

- The null hypothesis is that the machine is unbiased.
  - $H_0: \mu_0 = 27.1\%$
- The alternative hypothesis is the hypothesis if the null hypothesis is untrue.
  - $H_1: \mu_0 \neq 27.1\%$

 Click each tab to learn more. Then, click Next to continue.

In Step 1, we specify our null hypothesis, called  $H_0$ . Here, our null hypothesis is that the machine is unbiased, i.e., that after very many observations, our recorded average fat level would be  $\mu_0 = 27.1\%$ . Our alternative hypothesis, called  $H_1$ , is the hypothesis if our null hypothesis is untrue. In this case that would mean that  $\mu_0$  does not equal 27.1%.

## Tab 2: Step 2: Specify the Test Statistic



### Hypothesis Testing Procedure

9 of 15

- There is a standard procedure for hypothesis tests that follows five steps.
- Lab B Data:
  - B <- c( 26.33, 26.68, 26.31, 26.59, 25.61, 25.29, 25.88, 26.08, 25.85, 25.78)
  - Average true fat level for baby food powder recorded is  $\mu_0 = 27.1\%$

**1. Specify  $H_0$**

**2. Specify the Test Statistic**

**3. Specify the Significance Level**


**4. Compute the Test Statistic**

**5. Reject or Accept  $H_0$**

**2. Specify the Test Statistic**

- Specify the test statistic to be used.
  - This is determined by the type of test to be used.
  - It is a function of the sample statistics observed in the data.


$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

 Click each tab to learn more. Then, click Next to continue.

In Step 2, we specify the test statistic that we will be using. The test statistic is determined by the kind of test we will be using, and is a function of the sample statistics

we have observed in the data. In this case, this is the sample mean  $\bar{x}$ , the sample standard deviation  $s$  and the sample size  $n$ . It is very important to understand the difference between the observed sample mean  $\bar{x}$  and the population mean  $\mu_0$  specified by the null hypothesis. The sample mean  $\bar{x}$  is subject to random variability; we are investigating whether this variability is consistent with the behaviour we would expect to see under the null hypothesis.

**Tab 3: Step 3: Specify the Significance Level**



## Hypothesis Testing Procedure

9 of 15

- There is a standard procedure for hypothesis tests that follows five steps.
- Lab B Data:
  - `B <- c( 26.33, 26.68, 26.31, 26.59, 25.61, 25.29, 25.88, 26.08, 25.85, 25.78)`
  - Average true fat level for baby food powder recorded is  $\mu_0 = 27.1\%$

1. Specify  $H_0$

2. Specify the Test Statistic


3. Specify the Significance Level

4. Compute the Test Statistic

5. Reject or Accept  $H_0$


### 3. Specify the Significance Level

- Specify the significance level for the test.
- Compute the corresponding critical values.
  - The significance level is denoted by  $\alpha$  and is commonly set at 0.05.
- Significance level and critical values for Lab B:
  - $\alpha = 0.05$
  - $n = 10$
  - Critical values =  $\pm 2.26$

 Click each tab to learn more. Then, click Next to continue.

In Step 3, we specify the *significance level* for the test and compute the corresponding critical values. This is the threshold by which we decide if an observed test statistic is unusual or not. It is usually denoted by alpha. We have to specify alpha in advance of doing the test, otherwise the result will not be considered credible. It is very common to specify  $\alpha = 0.05$ ; we will discuss this choice further in the next lab session. For the data in Lab B, again we have  $n = 10$ , so the critical values are plus/minus 2.26.

Tab 4: Step 4: Compute the Test Statistic

Hypothesis Testing Procedure9 of 15

- There is a standard procedure for hypothesis tests that follows five steps.
- Lab B Data:
  - B <- c( 26.33, 26.68, 26.31, 26.59, 25.61, 25.29, 25.88, 26.08, 25.85, 25.78)
  - Average true fat level for baby food powder recorded is  $\mu_0 = 27.1\%$

1. Specify  $H_0$

2. Specify the Test Statistic

3. Specify the Significance Level

4. Compute the Test Statistic

5. Reject or Accept  $H_0$


4. Compute the Test Statistic

- Compute the test statistic and compare it to the critical values.
- For Lab B:
$$t = \frac{26.04 - 27.1}{\frac{0.44}{\sqrt{10}}} = -7.61 < -2.26$$

Click each tab to learn more. Then, click Next to continue.

In Step 4, we compute the test statistic, and compare to the critical values. In this case we have  $t = -7.6$  which is well below  $-2.26$ . Hence, we can proceed to Step 5.

Tab 5: Step 5: Reject or Accept  $H_0$ 

Hypothesis Testing Procedure9 of 15

- There is a standard procedure for hypothesis tests that follows five steps.
- Lab B Data:
  - B <- c( 26.33, 26.68, 26.31, 26.59, 25.61, 25.29, 25.88, 26.08, 25.85, 25.78)
  - Average true fat level for baby food powder recorded is  $\mu_0 = 27.1\%$

1. Specify  $H_0$

2. Specify the Test Statistic

3. Specify the Significance Level

4. Compute the Test Statistic

5. Reject or Accept  $H_0$

5. Reject or Accept  $H_0$

- Determine whether or not to reject the null hypothesis.
- As  $t = -7.61$  is in the rejection region, we:
  - Reject  $H_0$
  - Conclude that the difference between  $\bar{x}$  and  $\mu_0$  is statistically significant

! Conclusion:

- The analytical system is biased downwards


Click each tab to learn more. Then, click Next to continue.

In Step 5, we determine whether or not to reject the null hypothesis. Since  $t$  is in the rejection region, we reject  $H_0$  and conclude that the difference between the observed mean  $\bar{x}$  and hypothesised mean  $\mu_0$  is *statistically significant*. Note that significant is a technical term and should only be used in this context in a statistical analysis.

It is important to frame the decision to reject or fail to reject in the context of the data being analysed and hence form the appropriate scientific conclusions. In this case,

we conclude that the analytical system is biased downwards. Based on the data from Lab B, these conclusions should not be surprising.













## Slide 10: Paired Comparisons



### Paired Comparisons

10 of 15


- Paired tests compare samples from two different groups rather than comparing a single sample to a reference value.
- Data can be collected from two groups so that a relationship exists between pairs of data points.
- Examples of paired tests:

Student Taking an Exam	Cholesterol Level Check	Twins: Nature or Nurture
 <ul style="list-style-type: none"><li>Student takes exam.</li></ul>	 <ul style="list-style-type: none"><li>Check cholesterol levels.</li></ul>	 <ul style="list-style-type: none"><li>Study identical twins.</li></ul>
 <ul style="list-style-type: none"><li>Student receives training.</li></ul>	 <ul style="list-style-type: none"><li>Give treatment.</li></ul>	 <ul style="list-style-type: none"><li>Twin 1: Nurture A</li></ul>
 <ul style="list-style-type: none"><li>Student re-takes exam.</li></ul>	 <ul style="list-style-type: none"><li>Check cholesterol levels after treatment.</li></ul>	 <ul style="list-style-type: none"><li>Twin 2: Nurture B</li></ul>
 <ul style="list-style-type: none"><li>Compare both sets of results.</li></ul>	 <ul style="list-style-type: none"><li>Compare both sets of results.</li></ul>	 <ul style="list-style-type: none"><li>Compare differences.</li></ul>

Now let's consider a different but related type of hypothesis test, the paired test. In our last example, we looked at comparing a single sample of data to a reference value. In practice, such reference values are rarely available, and instead it is more common to compare samples from two different groups. In some cases, data can be collected from two groups in such a way that a special relationship exists between pairs of data points.

For example, students could take a test, receive training and then be re-tested. In this case it would be of interest to look at how each student's score changed after receiving training. Similarly, in a health setting, patient's cholesterol levels may be measured before and after taking a treatment to determine if any change is observed. More generally, twins are often studied in nature vs. nurture settings. Since their genetic make-up is more or less identical, any differences in outcome can be compared directly. Notice that although the context of these examples is quite different, the statistical elements are essentially the same.

## Slide 11: Example Two: Paired Test

**Example Two: Paired Test**11 of 15

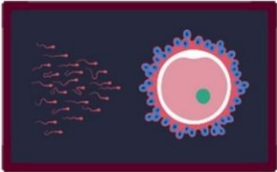
**Examine the Data****Test the Hypothesis**


**Introduction**

**?** Hypothesis: Is retinol-binding protein (RBP) of importance for early embryonic development in cows?

**Study Details**

- 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.
- If more RBP is produced on the ipsi side, this would suggest a direct role for RBP in implantation.



 Click each tab to learn more. Then, click Next to continue.

Our second example concerns a study of cows. The hypothesis to be tested is: Is retinol-binding protein (RBP) of importance for early embryonic development in cows? RBP is a major secretory product of the endometrium and is assumed to be of importance for early embryonic development in cows. In a preliminary study, 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. Of interest here is whether more retinol-binding protein is produced by cows on the *ipsi* or *contra* sides of the uterus. If more RBP is produced on the *ipsi* side, which is where the fertilised egg implants, this would suggest a direct role for RBP in implantation.

Click the tabs to work through this example. When you are ready, click “Next” to continue.



**Tab 1: Examine the Data**

✕

## Examine the Data (1/2)

### Data Values, Mean and Standard Deviation

■ ■

#### Data in R Code

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

```
20 > apply(data, 2, mean)
21      ipsi      Contra      Difference
22 7991.0625 7139.7500 851.3125
23 > apply(data, 2, sd)
24      ipsi      Contra      Difference
25 1103.5417 1284.2325 933.3102
```

Take time to view the information on this slide.

Notice that the data are inherently paired here, as the data is being collected from both sides of the uterus for each cow.

Here we can see the measurements for the ipsi and contra sides of each cow, as well as the difference. The means and standard deviations are also reported. You should notice that the standard deviations for the ipsi and contra sides separately are larger than for the individual differences.

**Tab 1.1: Boxplots**

✕

## Examine the Data (2/2)

### Boxplots

■ ■

Fertility Study: Ipsi and Contra Data Samples

Fertility Study: Paired Differences

A boxplot of the RBP levels for both sides is shown here. There appears to be some small differences between the two data samples, although it is not very conclusive. A

19

boxplot of the paired differences is shown in the right figure. This picture of the data is a lot more compelling; the majority of observations are clearly positive.

Tab 2: Test the Hypothesis

### Test the Hypothesis (1/2)

#### Perform a Paired t-test

- Focus on the data set of differences ( $\bar{x} = 851.3$ ;  $s = 933.3$ )
- Proceed as if performing a single sample test.

<b>1. Specify <math>H_0</math>:</b> <ul style="list-style-type: none"><li><math>H_0: \mu_d = 0</math><ul style="list-style-type: none"><li>There is no systematic ipsi-contra difference in RBP levels.</li></ul></li><li><math>H_1: \mu_d \neq 0</math></li></ul>	<b>3. Specify the significance level and compute the critical values:</b> <ul style="list-style-type: none"><li><math>\alpha = 0.05</math></li><li><math>n = 16</math></li><li>Critical value = <math>\pm 2.13</math></li></ul>
<b>2. Specify the test statistic to be used:</b> $t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$	<b>4. Compute test statistic and compare it to the critical values:</b> $t = \frac{851.3 - 0}{\frac{933.3}{\sqrt{16}}} = 3.65 > 2.13$
	<b>5. Determine whether or not to reject the null hypothesis:</b> <ul style="list-style-type: none"><li>Reject the null hypothesis.<ul style="list-style-type: none"><li>Conclude that on average, the ipsi side of the uterus secretes higher concentrations of RBP than the contra side.</li></ul></li></ul>

Performing a paired t-test is straightforward once we know how to perform a single sample test. The idea is simple: we focus on the data set of differences between observation pairs and then proceed as if performing a single sample test. Here our null hypothesis is that the long run mean difference,  $\mu_d$  is 0; that is, there is no systematic difference in the ipsi/contra RBP levels produced in cows on the day of ovulation. Our alternative hypothesis is that this difference is really non-zero, i.e., a systematic difference exists. It is common in paired tests for the null mean to be zero.

The test statistic has a very similar structure to that in the previous example, and is the sample mean of the differences minus the hypothesised overall difference divided by the standard error. If we set  $\alpha = 5\%$ , then given that our sample size is 16, the critical value is plus/minus 2.13.

Our observed test statistic is  $t = 3.65$ , which is greater than 2.13. Hence, we reject the null hypothesis and conclude that on average, the ipsi side of the uterus secretes higher concentrations of RBP than does the contra side.

Tab 2.1: R Output

## Test the Hypothesis (2/2)


### R Output

```
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
```

**Note the p-value is less than 5% which is consistent with the rejection of the null hypothesis.**

Let's check the output you would see if you performed this test in R. Notice that the argument paired is set equal to true. The output is very similar to the last example. Again you should notice the test statistic t equal to 3.65 just as we calculated, the degrees of freedom equal to 15, and the p-value of 0.0024. Notice this value is less than 5% and hence consistent with rejection of the null.

## Slide 12: Independent Groups Test





### Independent Groups Test

12 of 15

- Generally, when you want to compare two groups:
  - There will be no relationship within the data
  - A paired comparison will not be possible
- Data may not be paired as:
  - It may be easier to observe data from one group than the other

#### Examples of Independent Groups




We have just seen an example of how to perform a hypothesis where the data were paired in some way. More generally when we want to compare two groups, no such relationship will exist in the data and a paired comparison will not be possible.

For example, suppose that students were given two different training options before taking a test, one software-based and one textbook-based. In this case it would clearly not be sensible for individual students to take both training methods and then compare results – students cannot “unlearn” a method and then re-learn using a different method as though nothing has happened. It might be possible to pair students by matching them based on some kind of pre-screening test of their ability, however doing such a test may be challenging or infeasible in practice.


There are many other reasons why data may not be paired, among the simplest being that it may be easier to observe data from one group than the other. In any case, it is possible to perform a hypothesis test in this situation by making a small number of technical adjustments.

## Slide 13: Example Three: Two Sample $t$ -test

**Example Three: Two Sample  $t$ -test**13 of 15


**Examine the Data****Test the Hypothesis**


**Introduction**

**Hypothesis: Does running a machine with different settings produce different yield levels?**

**Study Details**

- How can you optimise a development process for producing small pellets used to fill capsules when making pharmaceutical products?
- 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.



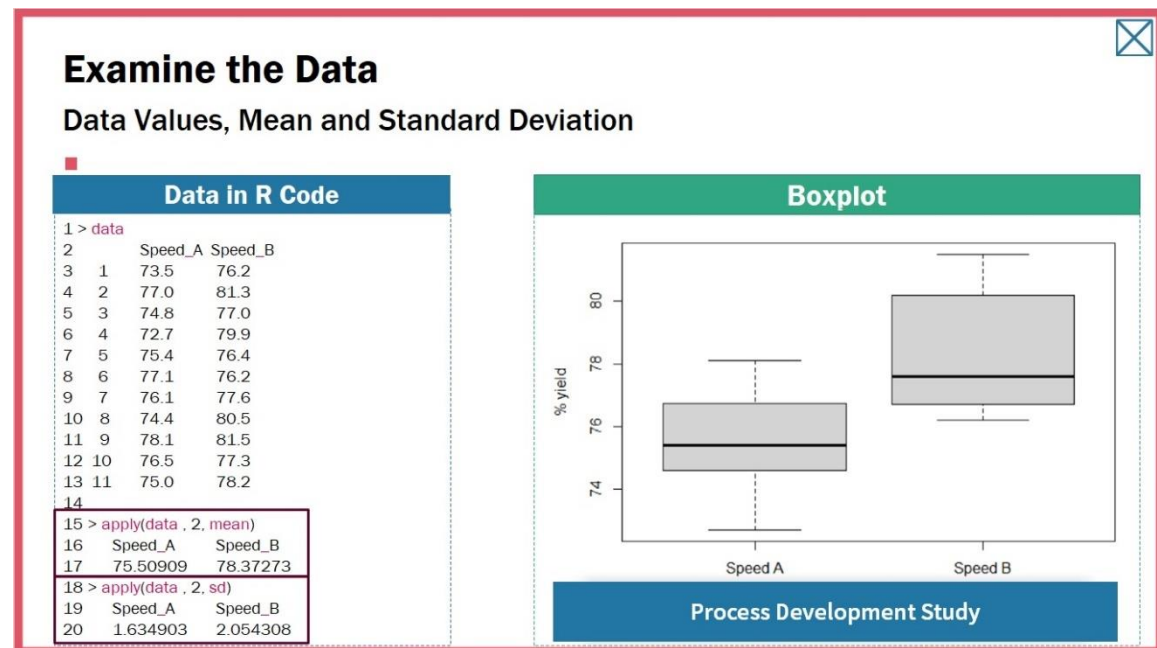
 Click each tab to learn more. Then, click Next to continue.

In this example, our hypothesis is: Does running a machine with different settings produce different yield levels?

A research student in the school of pharmacy was interested in optimising a development process. The process in question involves 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 work through this example. When you are ready, click “Next” to continue.

Tab 1: Examine the Data



The data are shown here. We have eleven observations for each speed type. The order of the runs was randomised. Each run is independent of the others so it would not be valid to perform a paired test in this case. Observing the sample means, Speed B appears to obtain a higher yield, and the standard deviations in yield are similar for both groups.

Visualising the data, we can see that there does appear to be some separation between the yields obtained by each speed. The question now is whether this difference can be explained by chance variation.

Tab 2: Test the Hypothesis

### Test the Hypothesis (1/3)

#### Perform the Test (1)

- The procedure for testing is the same as for the paired test, however, some of the technical details are different.

#### 1. Specify the null hypothesis:

- Make a common variance assumption
  - $Y_1 \sim N(\mu_1, \sigma^2)$  and  $Y_2 \sim N(\mu_2, \sigma^2)$ .
- Then, the null hypothesis is  $\mu_1 = \mu_2$ 
  - $H_0: \mu_1 - \mu_2 = 0$
  - $H_1: \mu_1 - \mu_2 \neq 0$

#### 2. Specify the test statistic to be used:

- The test statistic is adjusted:
$$t = \frac{(\bar{x} - \bar{y}) - 0}{\sqrt{\frac{2s^2}{n}}}$$
- s denotes the pooled standard deviation.
$$s^2 = \frac{s_1^2 + s_2^2}{2}$$

⌚ Take time to view the information on this slide.



We will now test the hypothesis that both speeds obtain the same yield percentage on average. The procedure for testing is the same as before. However, some of the technical details are different. Firstly, we make a common variance assumption. This means that we are assuming that the data for Speed A and Speed B have similar variability, i.e., that they are “spread out” in a similar way.

We denote the long run means for Speed A and B by  $\mu_1$  and  $\mu_2$  respectively. Then our null hypothesis is that  $\mu_1 = \mu_2$ , that is, that the difference in means is zero. Our alternative hypothesis is then that the mean difference is non-zero.

Our test statistic now has an adjusted form to what we have seen before. The numerator is the difference between the sample means minus the hypothesised difference between the population means, i.e, zero. The denominator is the standard error, which is now the square root of  $2s^2$  divided by  $n$ . Here  $n$  is the sample size of each group, i.e., 11.

Here “ $s$ ” denotes the pooled standard deviation; specifically, we take the mean of the sample variances we obtain for each group. We do this because of the common variance assumption that was discussed earlier. Notice that as well as using a pooled variance, our standard error formula now involves  $2s^2$  and not simply  $s^2$ , as before.

**Tab 2.1: Perform the Test (2)**

### Test the Hypothesis (2/3)

#### Perform the Test (2)

##### 3. Specify the significance level and compute the critical values:

- $\nu = 2(n - 1)$  degrees of freedom
- $\alpha = 0.05$
- $n = 11$
- $\nu = 20$
- $t_c = \pm 2.08$


##### 4. Compute test statistic and compare it to the critical values:

$$s^2 = \frac{1.635^2 + 2.054^2}{2} = 3.45$$

• Test statistic is:  $t = \frac{(75.51 - 78.37)}{\sqrt{\frac{2(3.45)}{11}}} = -3.61 < -2.08.$

##### 5. Determine whether or not to reject the null hypothesis:

- Reject the null hypothesis
- The two sample means are said to be statistically significantly different
- Speed B is considered to give the higher yield

 Take time to view the information on this slide.

The  $t$  distribution for this test has 2 times  $n-1$  degrees of freedom. Here, if we set  $\alpha = 0.05$  we get  $\nu = 20$  and hence  $t_c = \text{plus/minus } 2.08$ .

Computing the test statistic we find that  $s^2$  is 3.45 and  $t = -3.61$ . This value is in the rejection region and hence we reject our null hypothesis. The difference in yields obtained by Speeds A and B are statistically significant, and Speed B is considered to give a higher yield on average.

Tab 2.2: Test the Hypothesis: R Output

## Test the Hypothesis (3/3)

### 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
```

**Note the p-value is less than 5% which is consistent with the rejection of the null hypothesis.**


Let's again check our calculations against the output you see when the test is performed in R. The format of the output is again very similar to earlier examples and hopefully by this point you are starting to feel comfortable interpreting it. Here the test statistic  $t$  is -3.62, the degrees of freedom are 20, and the p-value is 0.0017, which is less than 5% and consistent with rejection of the null.

## Slide 14: Conclusion

### Conclusion

14 of 15

- We have covered the following hypothesis tests in detail:
  - One sample tests
  - Paired tests
  - Independent group tests
- If you are planning on performing a hypothesis test on data, you should consider what test should be applied **before** you collect the data.
- Tests for paired and independent groups involve the comparison of two groups.
  - When deciding which to use, you must determine if there is a relationship between pairs of observations in each group or not.



We have now covered three kinds of hypothesis test in detail:

- One sample tests;
- Paired tests;

- Independent group tests.


If you are planning on performing a hypothesis test on data that you have collected, it is worth bearing in mind that the ideal time to consider what test to apply is *before* the data have been collected, particularly if a paired test is to be applied.

It is also worth clarifying the differences between paired and independent group tests. Both tests involve the comparison of two groups; the question is whether a relationship exists between pairs of observations in each group or not.

In Example 2, the bovine fertility study, we took measurements from each side of the uterus of a group of cows. So, the relationship between pairs of observations is clear as they are taken from the same cow.

In Example 3, different material was used for each experiment, and the speed settings were set randomly. So, in this case, it is not possible to match or pair different observations.

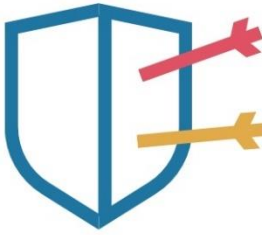
## Slide 15: Summary



### Summary

15 of 15

- Having completed this presentation, you should be able to:
  - Detail the procedure for a hypothesis test
  - List and describe the key elements involved in a test:
    - The test statistic
    - The significance level
    - The degrees of freedom
    - The critical values
    - The p-value
  - Perform the relevant test for a given data set
  - Form the appropriate scientific conclusions after a test
  - Interpret the output of tests performed using statistical software



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:

- Describe the procedure that is being followed when performing a hypothesis test, regardless of the technical details
- List and describe the key elements involved in a test:
  - the test statistic
  - the significance level
  - the degrees of freedom
  - the critical values
  - the p value.

Of the tests we have covered, you should be able to perform the relevant test for a given data set and following the test, be able to form the appropriate scientific conclusions. If a



test was performed using statistical software, you should be able to interpret the output correctly.