

## Module 7: Testing the Null: Data on Trial

### Overview and Learning Objectives

This module will introduce the use of Hypothesis Testing for gathering statistical evidence from samples in order to draw inferences about the population. Hypothesis testing is very much like a court room trial where evidence is accumulated to reach a verdict. This module will focus on the use of t-tests for comparing means. The learning objectives associated with this module are:

- Explain the process and logic of Null Hypothesis Significance Testing (NHST).
- Define one-tailed and two-tailed hypothesis testing.
- State and test the assumptions behind the different t-tests.
- Determine when a one-sample t-test should be applied.
- Use technology to compute and interpret a one-sample t-test.
- Identify and distinguish between the two-sample and paired sample research designs for continuous variables.
- Use technology to compute a two-sample (independent samples) t-test and paired-samples (dependent samples) t-test.
- Interpret a two-sample and paired-samples t-test.

### Video

#### CONTENTS

- 1** Module 7: Testing the Null: Data on Trial
  - 1.1** Overview and Learning Objectives
  - 1.2** Video
  - 1.3** Introduction to Hypothesis Testing - The One-sample t-test
    - 1.3.1** Body Temp
    - 1.3.2** [Hypothesis Testing - Data on Trial](#)
    - 1.3.3** A Worked Example
    - 1.3.4** The one-sample t-test in R
    - 1.3.5** The Language of Hypothesis Testing - Reporting Your Results
  - 1.4** Two-sample t-tests - Body Temperatures Revisited
    - 1.4.1** Testing the Assumption of Normality
    - 1.4.2** Central Limit Theorem
    - 1.4.3** Homogeneity of Variance
    - 1.4.4** Two-sample t-test - Assuming Equal Variance
    - 1.4.5** Two-sample t-test - Assuming Unequal Variance
    - 1.4.6** Example Write-up
  - 1.5** Paired Samples t-test
    - 1.5.1** Paired Samples t-test Visualisation
    - 1.5.2** Example Write-up

#### THE LADY TASTING TEA

## Lady Tasting Tea - Inferential Statistics and Experiment...



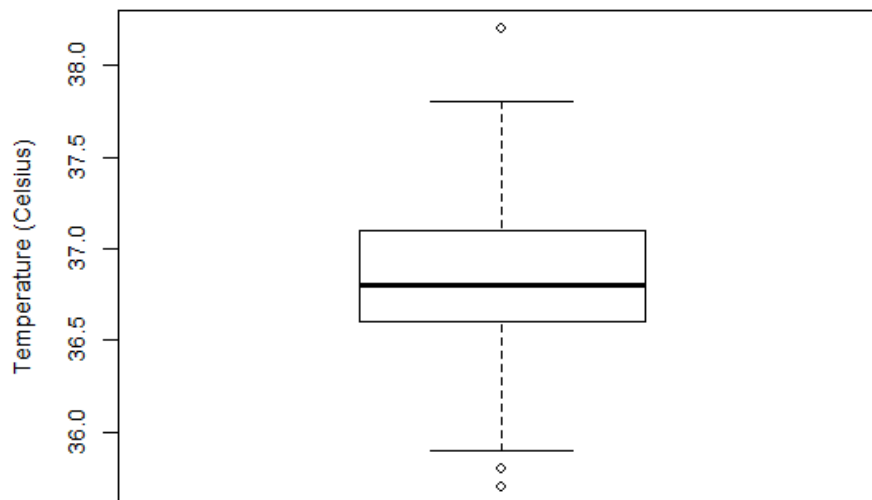
### Introduction to Hypothesis Testing - The One-sample t-test

This module introduces the important concepts of hypothesis testing in the context of one and two sample inference. To motivate and introduce the use of hypothesis testing, we will take a look at an example considering normal body temperature.

#### Body Temp

Prior to 1990 it was thought that the average oral human body temperature of a healthy adult was 37°Celsius (C). Investigators at that time were interested to know if this mean was correct. They gathered a sample of 130 adults and measured their oral body temperature. The dataset, [Body\\_temp.csv](#), can be downloaded from the [data repository](#). The descriptive statistics and a box plot of the data produced using R are reported below.

```
> library(mosaic)
> favstats(Body_temp, data = Body_temp)
   min   Q1 median   Q3  max  mean    sd   n missing
35.7 36.6  36.8 37.1 38.2 36.81 0.4074 130      0
> boxplot(Body_temp$Body_temp, ylab = "Temperature (Celsius)")
```



The sample mean is found to be 36.81°C. The mean was lower, but we know that samples bring with them sampling error. The researcher needs a way to determine if there was sufficient evidence from the sample to support the idea that the mean body temperature was not equal to 37°C.

## Hypothesis Testing - Data on Trial

**Null hypothesis significance testing**, or hypothesis testing for short, is an inferential decision making method used extensively in statistics and science. It is based solely on the idea of a **Null Hypothesis**, which we will denote as  $H_0$ . In hypothesis testing,  **$H_0$  is always assumed to be true**. It is a statistical hypothesis and should not be confused with other related terms including a research question or research hypothesis.


By assuming  $H_0$  is true, we can use our understanding of sampling distributions to calculate the probability of observing a particular sample statistic, or one more extreme. If this probability is really small or unlikely, we reject  $H_0$ . By rejecting  $H_0$  we can provide evidence to support  $H_A$ . Another way of writing this is to say the results are statistically significant. It's important to keep in mind at this stage that Hypothesis testing is probabilistic in nature. It cannot be used as a way to "prove" anything. It is merely a tool for accumulating probabilistic evidence. Science still needs to do its thing (e.g. results checked, validity assessed, replication, convergent evidence from multiple lines of enquiry etc.).

If the sample result is consistent or highly probable assuming the  $H_0$  is true, we will decide to "**fail to reject  $H_0$** ". If we fail to reject  $H_0$ , the study fails to find statistically significant evidence to support  $H_A$ . Or, in other words, the results of the hypothesis test are **not statistically significant**.

Notice how we have not used the term "accept". Some statistics instructors will use this term, but this terminology leads to the misconception that hypothesis testing somehow "proves" your results statistically. In this course, the decision from hypothesis tests will always come down to either "reject  $H_0$ " or "fail to reject  $H_0$ ".



You might find it helpful to compare the decisions from hypothesis testing to a jury trial. This is outlined in the following slide show:

### Hypothesis Testing - Data on Trial



The following table outlines the possible outcomes from a jury trial:

		Unknown Reality	
		Guilty	Innocent
Jury Decision	Guilty	Correct Decision (True Positive)	Type I Error ( $\alpha$ ) (False Positive)
	Innocent	Type II Error ( $\beta$ ) (False Negative)	Correct Decision (True Negative)

 ©2015 Dr James Baglin ([james.baglin@rmit.edu.au](mailto:james.baglin@rmit.edu.au))

Slide 1

## A Worked Example

The one-sample  $t$ -test is used to test whether there is evidence taken from a sample mean to suggest that the population mean is different to a previously assumed value. The one-sample  $t$ -test assumes the data are normally distributed and the population standard deviation is unknown. If you have a large sample, the normality assumption is not generally needed (see CLT in Module 5). The first step of hypothesis testing is to state the statistical hypotheses. We can summarise these as follows:

$$H_0: \mu = 37^\circ\text{C}$$

$$H_A: \mu < 37^{\circ}\text{C}$$

where  $\mu$  denotes a population mean. The alternate hypothesis is stated as a **lower one-tailed test**, note the use of the symbol  $<$ . This tells you the investigators are predicting that mean body temperature will be lower. One-tailed hypothesis tests predict the alternate hypothesis to be in a specific direction. We could also use  $>$  (upper tail) or  $\neq$  (not equal to).

When we use  $\neq$ , the test is called a **two-tailed test** or **non-directional hypothesis**. In practice, almost all hypothesis tests are two-tailed. If you're asked to conduct hypothesis testing and the type is not specified, assume you need to conduct a two-tailed test. The type of hypothesis tests we use (i.e. one-tailed or two-tailed) becomes important in the next step.

Next, we need to define a rule to determine if a sample result is unusual enough to reject  $H_0$ . Recall from previous modules that we have explored the unusualness of sample means by looking at sampling distributions. The hypothesis testing rules introduced in this module build on these concepts.

There are three related ways for doing this - the critical value method, the p-value approach and the confidence interval approach. All methods are based on the idea of a **significance level**, denoted  $\alpha$ . The significance level, typically set at  $\alpha = 0.05$ , is the line in the sand we use to judge "unusualness".

The significance level represents the minimum acceptable probability of committing a Type I Error assuming  $H_0$  is true (see the previous slideshow). A Type I Error would be committed by rejecting  $H_0$  when in fact  $H_0$  is true in the population. How is it possible to get it wrong? Remember, samples are imperfect representations of the populations. Sampling error produces uncertainty. We may never actually know if we have made a Type I error as we never really know for sure what's happening in the population. That's the very reason why we are doing the research in the first place.

We will work through using each one of the methods for testing the Null hypothesis using the Body temperature data. The good news is that they will all tell us the same thing. However, in practice, we stick to the p-value and confidence interval methods. The critical value method is a remnant of times before computers. However, it will be demonstrated for completeness.

## Critical Value Approach

The idea of the critical value approach is to build a model of a sampling distribution assuming  $H_0$  is true. Using this sampling distribution, the point at which a sample result has a less than  $\alpha$  (e.g.  $\alpha = 0.05$ ) probability of occurring is recorded as a **critical value**. We then compare the sample mean **test statistic** to the critical value taken from the sampling distribution under the  $H_0$  to see if it falls below the significance level  $\alpha$  threshold. This will make sense as we work through an example.

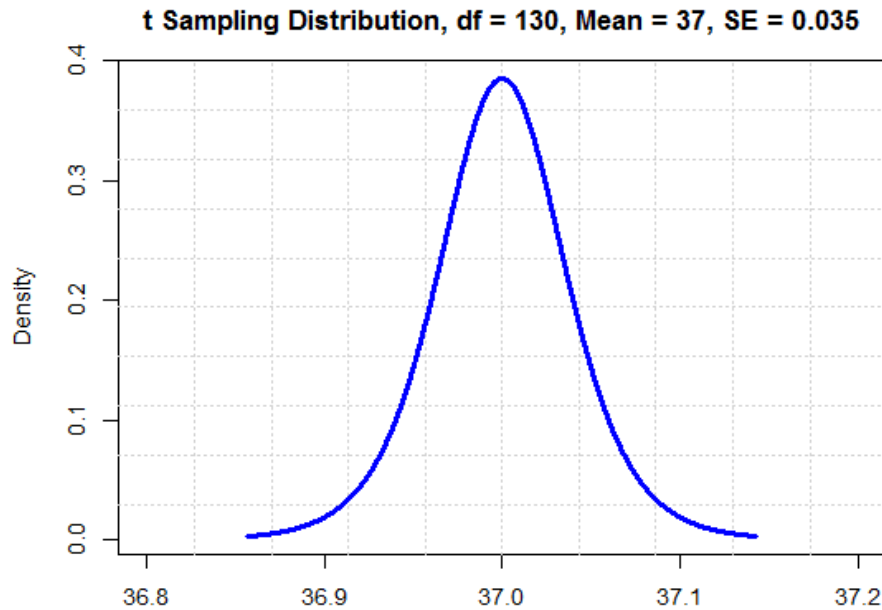
Referring back to sampling distributions, we know the sampling distribution of a population mean with a sample size greater than 30, in this example  $n = 130$ , will be approximately normally distributed with a mean equal to the population mean,  $\mu = 37$ , and a standard error of the mean, SEM:

$$\frac{\sigma}{\sqrt{n}}$$

However, we have one problem. We don't know the population standard deviation,  $\sigma$ . Therefore, we need to use the t-distribution introduced back in Module 6. We can approximate the sampling distribution using the SEM as follows:

$$\frac{s}{\sqrt{n}}$$

Note the use of the sample standard deviation,  $s$ , in place of  $\sigma$ . The sampling distribution looks like the following:



As the investigator was using a one-tailed test, we need to know the point in the sampling distribution where there is a less than 5% chance of a sample mean falling below  $\mu = 37$ . We can find this **critical value**,  $t^*$  using R.

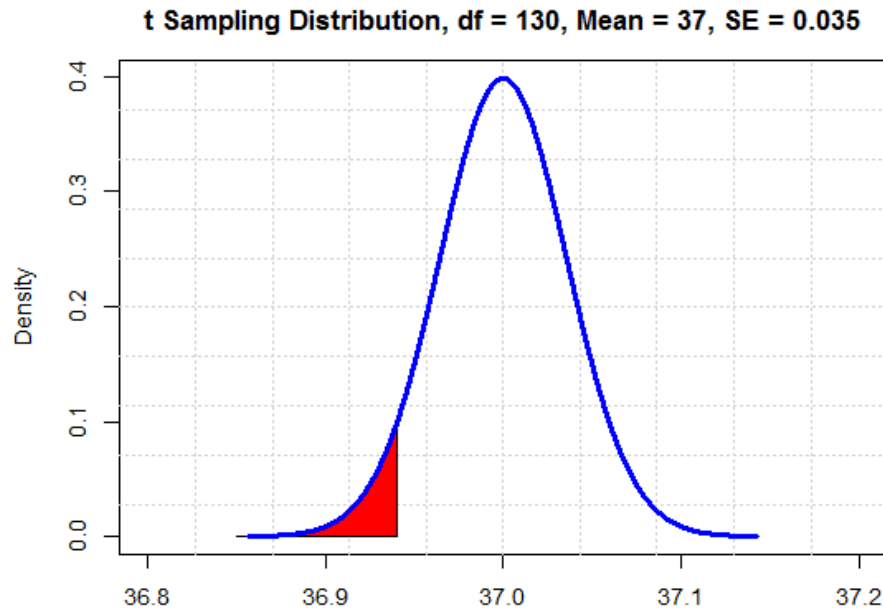
```
> qt(0.05,130-1,lower.tail = TRUE)
[1] -1.66
```

This means the  $t^*$ mean sits 1.66 SEs below the population mean of 37. We can write a short R program to convert this back to a critical mean.

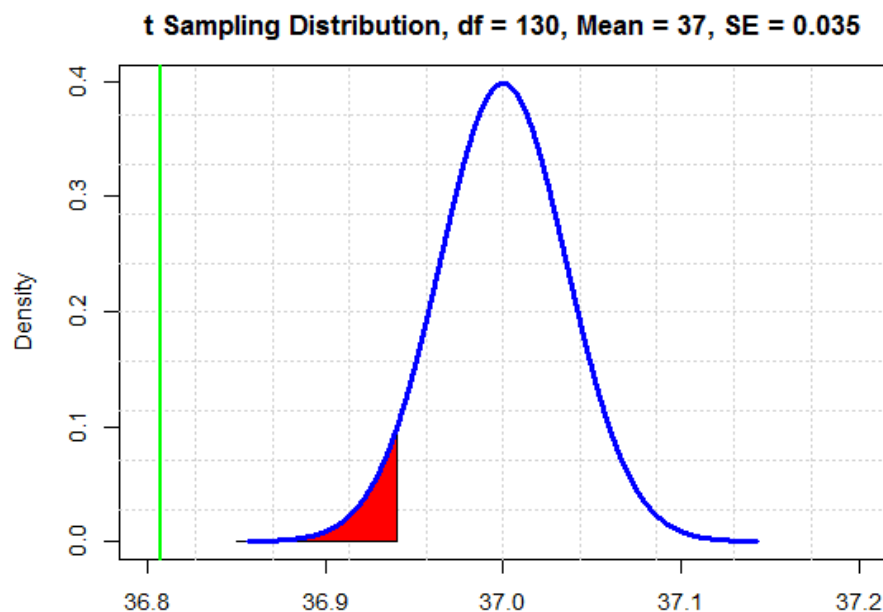
```
> t<-qt(0.05,130-1,lower.tail = TRUE) #save t-crit
> mu <- 37 #Assign mu
> s <- sd(Body_temp$Body_temp) #Assign sd
> n <-length(Body_temp$Body_temp) #Assign n
> se <-s/sqrt(n) #Calculate se
> mu + (t*se) #Determine critical mean
[1] 36.94079988
```

The  $t$  critical mean was found to be 36.94. This means  $\Pr(\bar{x} < 36.94) = 0.05$ . In other-words, assuming the population mean = 37, sampling 130 people from the population and finding a mean to be less than 36.94 would happen less than 5% of the time.

Visually, the value of 36.94 is the point where there is .05 probability in the lower tail of the sampling distribution under  $H_0$ . This looks like the following plot where the  $t$  distribution represents the hypothetical sampling distribution. This area beyond 36.94 is referred to as the **rejection region**.



If the sample mean from the study falls in the rejection region, the decision will be to reject  $H_0$ . Where did the sample mean  $\bar{x} = 36.8$  fall?



The green line in the image above marks the location of the investigator's sample mean,  $\bar{x} = 36.81$ . It is well within the tail of the rejection region of the sampling distribution of the mean under  $H_0$ . This is telling us, assuming  $H_0: \mu = 37$  is true, the probability of observing  $\bar{x} = 36.81$  is less than  $\alpha = 0.05$ . This is a statistically significant result.

What would happen if we used a non-directional two-tailed hypothesis test, i.e.  $H_A: \mu \neq 37$ ? What would the rejection regions look like then?

For two-tailed hypothesis testing, the rejection regions are split between above and below  $H_0$ . We still need to maintain an overall significance level of 0.05, so because  $\alpha/2$ , we need to find the  $t$  critical values associated with  $0.05/2 = 0.025$  in the upper and lower tail of the sampling distribution under  $H_0$ . We can find this  $t^*$  value in R:

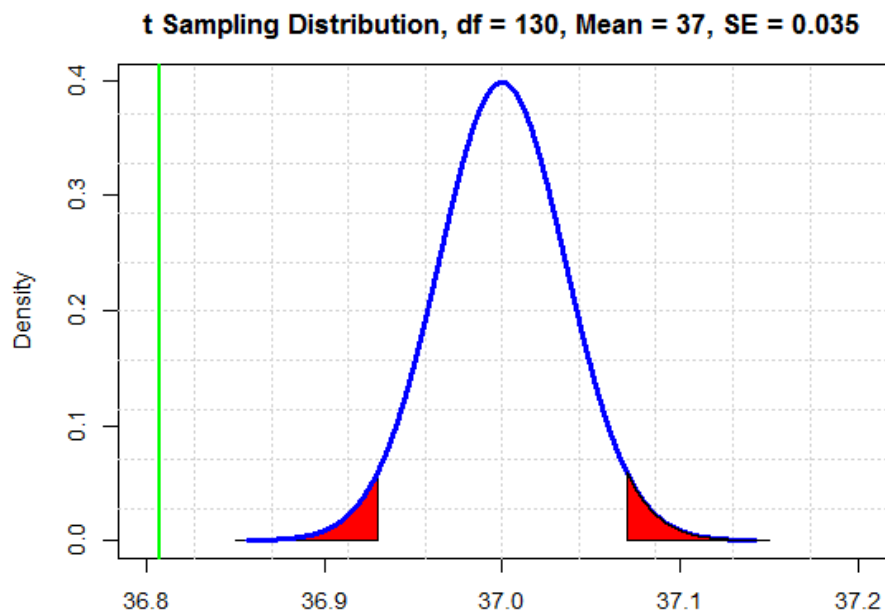
```
> qt(0.05/2, 130-1, lower.tail = TRUE)
[1] -1.978524
```

As the t distribution is symmetric, we know the  $t^*$  values lie at -1.98 and +1.98 SE from the mean. Using R we can readily locate the critical means:

```
> t<-qt(0.025,130-1,lower.tail=FALSE) #save t*
> mu <- 37 #Assign mu
> s <- sd(Body_temp$Body_temp) #Assign sd
> n <-length(Body_temp$Body_temp) #Assign n
> se <-s/sqrt(n) #Calculate se
> mu + (t*se) #Determine lower critical mean
[1] 37.07
> mu - (t*se) #Determine upper critical mean
[1] 36.93
```

The lower t critical mean is equal to 36.93 and the upper critical region starts at 37.07.

Now we have the critical regions for a two-tailed test, we can visualise them and plot the location of the sample mean.



As you can see, the sample mean ( $\bar{x} = 36.81$ ) still falls well within the rejection regions even for a two-tailed test. You will notice, when looking back at the one-tailed test, that the lower rejection region of the two-tailed test was further from the mean than the one-tailed test (36.93 vs. 36.94). This is the trade-off for two-tailed tests. As the significance level is split between two tails of the sampling distribution, the critical regions are further from the Null hypothesised mean, meaning it will be slightly harder to reject.

### p-value Approach

The critical value method tells us whether a sample result is less than the significance level, but it doesn't tell us by how much. The p-value method overcomes this problem.

**A p-value,  $p$ , represents the probability of observing a sample statistic, or one even more extreme, under the assumption that  $H_0$  is true.** You may recognise the p-value to be an example of a [conditional probability](#). For the body temp example using a lower, one-tailed hypothesis test we would write the p-value as follows:

$$p = \Pr(\bar{x} < 36.81 \mid \mu = 37)$$

To calculate the one-sided p-value, we need to convert the mean into a test statistic,  $t$ .

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

```
> m<-mean(Body_temp$Body_temp)
> mu<-37
```

```
> s<-sd(Body_temp$Body_temp)
> n<-length(Body_temp$Body_temp)
> se<-s/sqrt(n)
> t<-(m-mu)/se
> t
[1] -5.381848
```

This test statistic is interpreted as the sample mean falling 5.38 SE below the population mean. Now we can restate the p-value in terms of a SE of the mean using the t statistic. By doing this, we can look up a t distribution which has a mean of 0 and df = n - 1. The p-value becomes:

$$p = \Pr(t < -5.38 | t = 0)$$

We now calculate the exact p-value using the pt() function:

```
> pt(t,df=n-1)
[1] 1.680393e-07
```

The p-value is very small. In fact, it's so small that it rounds to 0.000. When the p-value is this small, we write  $p < .001$ .

Things get a little strange if we use a two-tailed test. Because we need to take into account the mean also falling 5.38 SE above the mean, the p-value for a two-tailed test becomes:

$$p = \Pr(t < -5.83 | t = 0) + \Pr(t > 5.83 | t = 0)$$

Now here's a neat trick. As the t-distribution is symmetric, the two probabilities to be added are exactly the same. Therefore, a short cut to a two-tailed p-value can be calculated as:

$$p = \Pr(t < -5.83 | t = 0) * 2$$

Using R:

```
> pt(t,df=n-1)*2
[1] 3.360785e-07
```

The answer is still really, really small, so we write  $p < .001$ .

So what do we do with this p-value? We compare it to the significance level and apply the following rule:

$$\begin{aligned} \text{If } p < \alpha, & \text{ reject } H_0 \\ \text{If } p \geq \alpha, & \text{ fail to reject } H_0 \end{aligned}$$

In both the one-tailed and two-tailed hypothesis tests,  $p < .001 < \alpha = 0.05$ , therefore we reject  $H_0$ . As you can see, the probability of observing  $\bar{x} = 36.81$ , or a sample mean more extreme, assuming  $H_0: \mu = 37$  is true, is extremely unlikely. Therefore,  $H_0$  is rejected and we find statistical evidence to support  $H_A$ . The conclusions you reach from hypothesis testing will be the same across all methods (critical value, p-value and confidence intervals) for testing  $H_0$ . Different methods just give you more or less information. The critical value is rarely used in practice.

## Confidence Interval Approach

If we use a confidence interval to test  $H_0$  for a one-sample t-test, we will automatically use a two-tailed hypothesis test. That's because most confidence intervals divide the significance level by 2 in their calculations. One-sided confidence intervals can be calculated, but are not typically supported by most statistical software. So let's test  $H_0$  using a confidence interval. First, we calculate the 95% CI for the sample mean  $\bar{x} = 36.81$ . Recall, when the population standard deviation is unknown, the 95% CI is calculated as:

$$\bar{x} \pm t_{n-1, 1-(\alpha/2)} \frac{s}{\sqrt{n}}$$

In R, we can calculate the CI using the following code:

```
> confint(t.test( ~ Body_temp, data = Body_temp))
mean of x      lower      upper      level
```



```
36.80769 36.73699 36.87839 0.95000
```

We find 95% CI for the sample mean,  $\bar{x} = 36.81$  [36.74, 36.87].

Now we apply the following rule:

**If the 95% CI does not capture  $H_0$ , reject  $H_0$**   
**If the 95% CI captures  $H_0$ , fail to reject  $H_0$**

We recall  $H_0: \mu = 37$ . Is  $\mu = 37$  captured by the 95% CI [36.74, 36.87]? No. Therefore, our decision should be to reject  $H_0$ .

As I have explained previously, all three methods will arrive at the same decision, assuming you use the same significance level and tails. The critical value method is the crudest method, the p-value is better and the 95% CI approach is preferred.

## The one-sample t-test in R

You might be a little overwhelmed at this point with all the calculations. Don't worry. You're not expected to remember all these. You should focus on understanding the concepts and logic of the hypothesis test. We can use R to easily perform the one-sample t-test in a split second. We use the `t.test()` function for this purpose. Here's an example of the lower-tailed hypothesis test:

```
> t.test(~ Body_temp, data=Body_temp ,mu = 37, alternative="less")

One Sample t-test

data:  data$Body_temp
t = -5.3818, df = 129, p-value = 1.68e-07
alternative hypothesis: true mean is less than 37
95 percent confidence interval:
 -Inf 36.86689
sample estimates:
mean of x
 36.80769
```

Here's an example of the two-tailed hypothesis test:

```
> t.test(~ Body_temp, data=Body_temp ,mu = 37, alternative="two.sided")

One Sample t-test

data:  data$Body_temp
t = -5.3818, df = 129, p-value = 3.361e-07
alternative hypothesis: true mean is not equal to 37
95 percent confidence interval:
 36.73699 36.87839
sample estimates:
mean of x
 36.80769
```

Now that makes things a lot easier! Why did we not do this sooner? You need to understand how these values are calculated, the concepts reported and the logic of hypothesis testing. Without this knowledge, you won't understand how to interpret the R output.

## The Language of Hypothesis Testing - Reporting Your Results

When we report the results of hypothesis testing for professional or scientific writing, we won't need to mention statistical hypotheses or our decision to reject or fail to reject  $H_0$ . Instead we use the term "statistical significance" to refer to the rejection of  $H_0$ . When we fail to reject  $H_0$ , we would write that the test was NOT statistically significant. Read the following example summarising the results of the one-sample t-test.

A two-tailed, one-sample t-test was used to determine if the mean oral body temperature readings were significantly different from the previously assumed oral body temperature population mean of 37°C. The 0.05 level of significance was used. The sample's mean oral body temperature was  $M = 36.81^\circ\text{C}$ ,  $SD = 0.407^\circ\text{C}$ . The results of the one-sample t-

test found the mean rectal body temperature to be statistically significantly lower than the population oral mean temperature,  $t(129) = -5.38$ ,  $p < .001$ , 95% CI [36.74, 36.87].

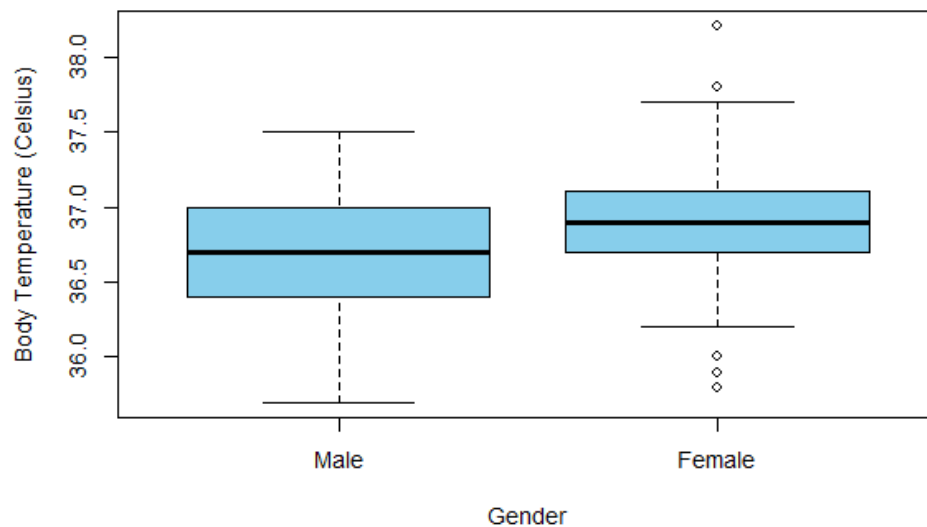
As you can see, there is no mention of a statistical hypothesis or the rejection of  $H_0$ .

## Two-sample *t*-tests - Body Temperatures Revisited

The two-sample *t*-test, or independent samples *t*-test, is used to compare the difference between two population means, e.g. a control group vs. an experimental group, males vs. females, etc. The two-sample *t*-test assumes the populations being compared are independent of each other, that the data for both populations have equal variance and, for small samples, that the data for both populations are normally distributed. These assumptions must be checked prior to interpreting the results of the two-sample *t*-test.

We will work through an example to introduce and demonstrate each stage for performing the two-sample *t*-test. We will revisit the `Body_temp.csv` data. We ask the question: do males and females have different average body temperatures? Here are the descriptive statistics and boxplot from R:

```
> Body_temp$Gender <- factor(Body_temp$Gender, levels = c(1,2),  
                             labels = c("Male","Female")) #Assign correct labels  
> favstats(~ Body_temp | Gender,data = Body_temp)  
  Gender min   Q1 median   Q3  max  mean    sd  n missing  
1  Male 35.7 36.4  36.7 37.0 37.5 36.726 0.38822 65      0  
2 Female 35.8 36.7  36.9 37.1 38.2 36.889 0.41274 65      0
```



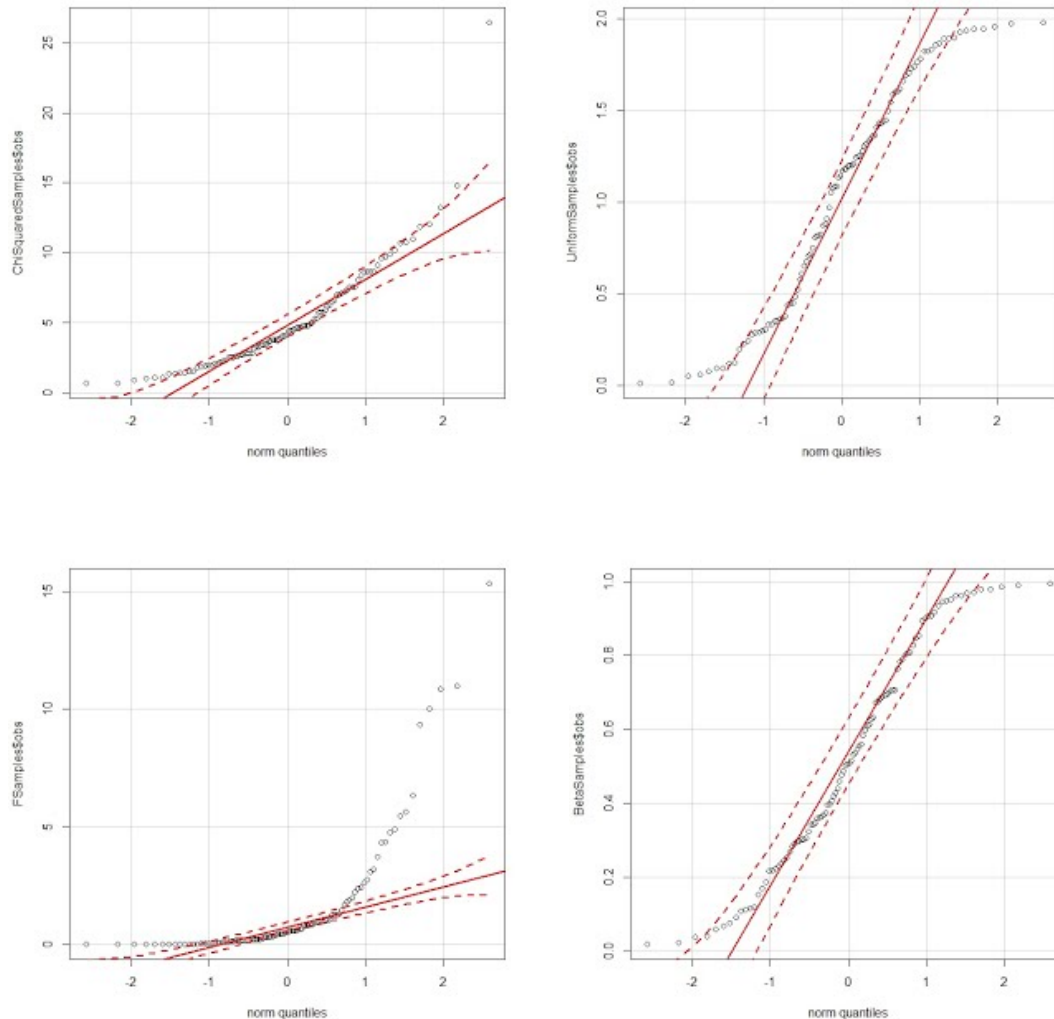
It's a bit close to call, but it looks like females do tend to have higher body temperatures. The two-sample *t*-test will help us consider whether this difference is statistically significant. Let's make a start considering the assumptions behind the two-sample *t*-test.

## Testing the Assumption of Normality

When you have small samples, generally less than 30 in one sample, the two-sample *t*-test assumes the data are drawn from a normal population distribution. This is a hard assumption to test, especially in small samples. The best approach is to look at the data visually and rule out gross departures from normality. If we can satisfy that the data are approximately normal, we can go ahead with the two-sample *t*-test. *T*-tests are generally robust against minor departures from normality. This means they will tend to maintain the desired significance level (e.g. 0.05) even if normality is not met.

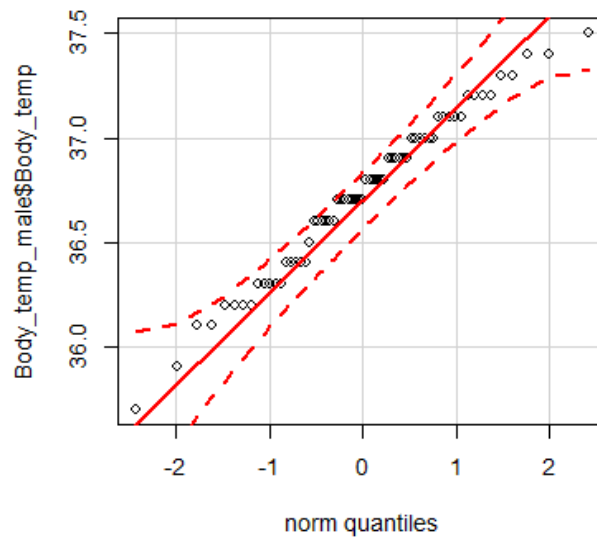
Although not required for the body temperature example due to the large sample sizes (Male  $n = 65$  and Female  $n = 65$ ), we will take a look at visually checking normality using Q-Q plots so you know what to do when  $n < 30$ .

Q-Q plots visualise the data's distribution comparing it to what we would expect to see assuming the data are normally distributed in the population. We can look at this using a normal Q-Q plot. If the data are normally distributed in the population, the data points will fall close to the diagonal line. However, due to sampling error, the data points won't always fall directly on the line, so we tend to check for obvious departures from normality. Non-normality is often characterised by strange 'S'-shaped and other non-linear trends (see the following examples).

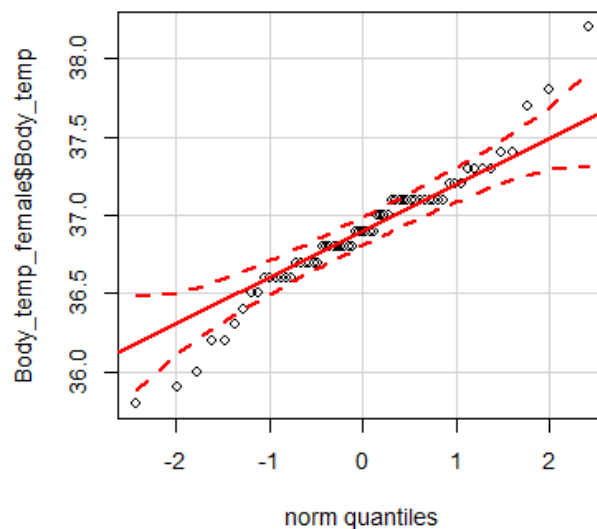


Let's look at the body temperature distributions for males and females using the `qqPlot()` function.

```
> library(car)
> Body_temp_male <- subset(Body_temp, subset=Gender == "Male") #Subset male data
> qqPlot(Body_temp_male$Body_temp, dist="norm")
```



```
> Body_temp_female <- subset(Body_temp, subset=Gender == "Female") #Subset female data
> qqPlot(Body_temp_female$Body_temp, dist="norm")
```



We check to see if the data points fall within the dashed red lines. These red lines correspond to 95% CI for the normal quantiles. If the data points follow a trend of falling outside these red lines, you should be careful about assuming normality. The male data looks fine. All the points appear to be following a normal distribution very closely. The female data on the other hand isn't so clear. You can see points falling outside the tails of the distribution. This suggests the tails are heavier than what we would expect under a normal distribution. You should be cautious about assuming normality for females. Fortunately, due to the large sample size,  $n = 65$ , we don't have to worry too much.

## Central Limit Theorem

Thanks to the CLT, introduced back in Module 5, we know that the sampling distribution of a mean will be approximately normally distributed, regardless of the underlying population distribution when the sample size is large (i.e.  $n > 30$ ). This means that we can proceed with the two-sample t-test if the normality assumption is violated when the sample sizes in each group are greater than 30. In this example, the sample sizes are much greater than 30 in each

group, so we can effectively ignore the issue with normality for the female data. The CLT is extra incentive not to rely on small samples.

## Homogeneity of Variance

Homogeneity of variance, or the assumption of equal variance, is tested using the Levene's test. This test is reported by default when you perform a two-sample t-test in SPSS. The Levene's test has the following statistical hypotheses:

$$H_0: \sigma^2_1 = \sigma^2_2$$

$$H_A: \sigma^2_1 \neq \sigma^2_2$$

where  $\sigma^2_1$  and  $\sigma^2_2$  refer to the population variance of group 1 and 2 respectively. The Levene's test reports a p-value that is compared to the standard 0.05 significance level. We can use the `leveneTest()` function in R to compare the variances of male and female body temperatures:

```
> leveneTest(Body_temp ~ Gender, data = Body_temp)
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1      0.04  0.84
 128
```

The p-value for the Levene's test of equal variance for body temperature between males and females was  $p = 0.84$ . We find  $p > .05$ , therefore, we fail to reject  $H_0$ . In plain language, we are safe to assume equal variance.

If the Levene's test was statistically significant, i.e.  $p < .05$ , this would imply that we need to reject  $H_0$ . In other words, it's not safe to assume equal variance.

The assumption of equal variance is important because it will determine the type of two-sample t-test we will perform. Don't ignore this assumption. Ignoring it can lead to poor inference.

## Two-sample t-test - Assuming Equal Variance

Let's jump straight into R and perform a two-sample t-test assuming equal variance and a two-sided hypothesis test. We use the `t.test()` function.

```
> t.test(Body_temp ~ Gender, data = Body_temp,
         var.equal=TRUE,
         alternative="two.sided")

Two Sample t-test

data:  Body_temp by Gender
t = -2.32, df = 128, p-value = 0.022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.302140 -0.024014
sample estimates:
 mean in group Male mean in group Female
          36.726          36.889
```

There are two important things to note. We use the `var.equal = TRUE` option to perform the equal variance assumed two-sample t-test and the `alternative = "two-sided"` option to specify a two-tailed test. R prints the results of the two-sample t-test. Let's start working through the results of the test.

The two-sample t-test has the following statistical hypotheses:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \neq 0$$

where  $\mu_1$  and  $\mu_2$  refer to the population means of group 1 and 2 respectively. The null hypothesis is simply that the difference between the two independent population means is 0. The difference between males and females estimated

by the sample was  $36.726 - 36.889 = -0.163$ .

The test statistic  $t = -2.32$  was calculated as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

where  $S_p^2$  is the pooled variance (assuming equal variance):

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

The t-statistic is compared to a two-tailed t-critical value  $t^*$  with df:

$$df = n_1 + n_2 - 2$$

Assuming  $\alpha = 0.05$  and a two-tailed test, we find  $t^*$  using the R function:

```
> qt(0.025, df = 65 + 65 - 2)
[1] -1.9787
```

The two-tailed  $t^*$  was found to be  $\pm 1.98$ .

As the test statistic  $t$  from the two-sample t-test assuming equal variance was  $t = -2.32$ , which was more extreme than  $-1.98$ , we reject  $H_0$ . According to the critical value method, there was a statistically significant difference between male and female body temperature means.

The p-value of the two-sample t-test will tell us the probability of observing a sample difference between the means of  $-0.163$ , or one more extreme, assuming the difference was 0 in the population (i.e.  $H_0$  is true). The two-tailed p-value was reported to be  $p = .02$ .

According to the p-value method, as  $p = .022 < \alpha = 0.05$ , we reject  $H_0$ . There was a statistically significant difference between the means.

The 95% CI of the difference between the means ( $-0.163$ ) was calculated using the following formula in R:

$$\left[ (\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2, 1-\frac{\alpha}{2}} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2, 1-\frac{\alpha}{2}} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \right]$$

which R reports as 95% CI  $[-0.30 -0.02]$ . As this interval does not capture  $H_0 = 0$ , we reject it. Once again, there was a significant difference between the means.

## Two-sample t-test - Assuming Unequal Variance

Had the Levene's test been statistically significant and the assumption of equal variance violated, the two-sample t-test not assuming equal variance should be interpreted. This version of the test uses an adjusted test statistic  $t$ :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

You will notice how the variances are not pooled. The test also uses an adjusted df:

$$df' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

If you don't specify `var.equal` in the `t.test()` function for R, the two-sample t-test not assuming equal variance is reported by default. This test is also known as the Welch two-sample t-test.

```
> t.test(Body_temp ~ Gender, data = Body_temp,
         var.equal=FALSE,
         alternative="two.sided")

Welch Two Sample t-test

data:  Body_temp by Gender
t = -2.32, df = 128, p-value = 0.022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.302145 -0.024009
sample estimates:
 mean in group Male mean in group Female
          36.726          36.889
```

Because the variances were very similar between males and females, the adjusted test statistic  $t = -2.32$  and  $df = 128$  for the Welch test are actually the same as the equal variances assumed two-sample t-test. This is why [some recommend](#) that you should also use the Welch test. If the variances are unequal, the test will make the required adjustment. If not, the test will be similar to a regular two-sample t-test. This means you can skip testing equal variance using the Levene's test. This might make things simpler, but understanding the difference between these two versions of the two-sample t-test will help to make this decision.

## Example Write-up

A two-sample t-test was used to test for a significant difference between the mean body temperature of males and females. While the body temperatures for females exhibited evidence of non-normality upon inspection of the normal Q-Q plot, the central limit theorem ensured that the t-test could be applied due to the large sample size in each group. The Levene's test of homogeneity of variance indicated that equal variance could be assumed. The results of the two-sample t-test assuming equal variance found a statistically significant difference between the mean body temperatures of males and females,  $t(df = 128) = -2.32$ ,  $p = .02$ , 95% CI for the difference in means  $[-0.30 -0.02]$ . The results of the investigation suggest that females have significantly higher average body temperatures than males.

## Paired Samples t-test

When we measure the same sample twice, the measurements are said to be "paired" or "dependent". Many experiments measure the same people or objects before (baseline) and after (follow-up) a treatment. The **paired-samples t-test**, also known as the **dependent samples t-test**, is used to check for a statistically significant mean change or difference in these situations. Consider the following example.

Investigators want to determine if exercise can help reduce stress in prisoners. The investigators have 15 prisoners fill out a stress questionnaire and then have them play a physically active sport for one hour. Following the exercise, the prisoners fill out the same stress questionnaire. The investigator wants to know if the exercise had an effect on the prisoners' stress levels. The `PrisonStress.csv` data are available from the Data Repository.

Here are the descriptive statistics from R:

```
> PrisonStress_Sport <- subset(PrisonStress, subset = (Group == "Sport")) #Filter data for
sport group
> favstats(~PSSbefore, data = PrisonStress_Sport) #Stress before
  min Q1 median  Q3 max  mean    sd  n missing
12 19    23 27.5  44 23.933 7.4878 15      0
> favstats(~PSSafter, data = PrisonStress_Sport) #Stress after
  min Q1 median  Q3 max  mean    sd  n missing
 8 14    21 24  33  20 6.9076 15      0
```

As you can see, stress levels appear to have dropped. We can use the paired-samples t-test to determine if this change can be considered statistically significant. Like the two-sample t-test, the paired-samples t-test assumes the data are normally distributed. This is important in this example because the sample size is only 15. Specifically, the assumption of normality lies in the distribution of the difference scores. A person's difference score can be calculated as:

$$d_i = x_{i1} - x_{i2}$$

where  $i$  refers to a particular person and 1 and 2 refer to the two dependent measurements. The paired-samples t-test assumes that these differences,  $d$ , are normally distributed. In R, we add a difference column to the `PrisonStress` data object:

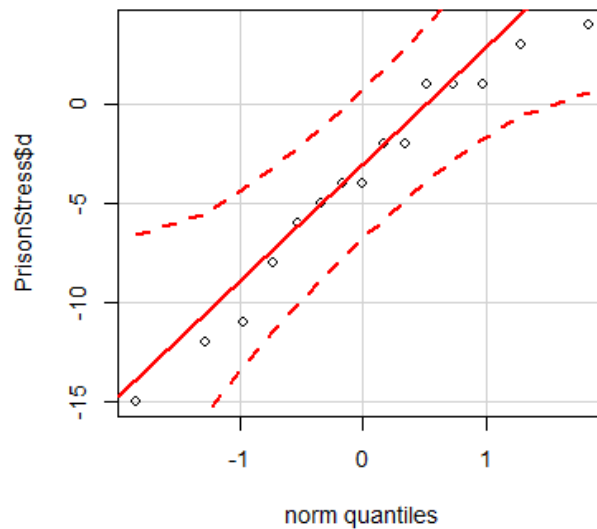
```
> PrisonStress_Sport$d <- PrisonStress_Sport$PSSafter - PrisonStress_Sport$PSSbefore #Add
column of differences
```

We can then calculate the mean difference.

```
> favstats(~d, data = PrisonStress_Sport) #Mean difference between before and after
  min Q1 median Q3 max   mean    sd n missing
-15  -7    -4   1   4 -3.9333 5.6753 15     0
```

On average, stress levels reduced by 3.93 points. We look at the Q-Q plot to check normality of the differences.

```
> qqPlot(PrisonStress_Sport$d, dist="norm")
```



The differences appear to be approximately normally distributed, so we are safe to continue with the paired-samples t-test.

The population mean of these differences is denoted:

$$\mu_{\Delta}$$

The statistical hypotheses for the paired-samples t-test are as follows:

$$\begin{aligned} H_0 : \mu_{\Delta} &= 0 \\ H_A : \mu_{\Delta} &\neq 0 \end{aligned}$$

The t-statistic for the paired-samples t-test is calculated using:

$$t = \frac{\bar{d}}{\frac{s_{\Delta}}{\sqrt{n_{\Delta}}}}$$



We can calculate the paired sample t-test using the `t.test()` function in R.

```
> t.test(PrisonStress_Sport$d, mu = 0, alternative = "two.sided")

One Sample t-test

data:  PrisonStress_Sport$d
t = -2.68, df = 14, p-value = 0.018
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -7.07623 -0.79043
sample estimates:
mean of x
 -3.9333
```

The first thing you should notice is that R labels the paired sample t-test as a one sample t-test. It's not a mistake. The paired samples t-test is a one sample t-test of the differences,  $d$ , and where the hypothesized population mean = 0. R reports  $t = -2.68$ .

Degrees of freedom are:

$$df = n_{\Delta} - 1$$

In this example,  $df = 15 - 1 = 14$ .

The critical value,  $t^*$  for the paired-sample t-test, assuming a two-tailed test with  $\alpha = 0.05$ , is calculated as:

```
> qt(0.025, df = 14)
[1] -2.1448
```

The  $t^*$  values are  $\pm 2.14$ . As  $t = -2.68$  is more extreme than  $-2.14$ ,  $H_0$  should be rejected. There was a statistically significant mean difference between the stress before and after exercise.

The two-tailed p-value can be calculated as:

```
> 2*pt(-2.68, df = 14)
[1] 0.017946
```

which rounds to  $p < .018$  reported in the paired samples t-test. As  $p < .05$ , we reject  $H_0$ . There was a statistically significant mean difference.

A 95% CI of the mean difference can be calculated as:

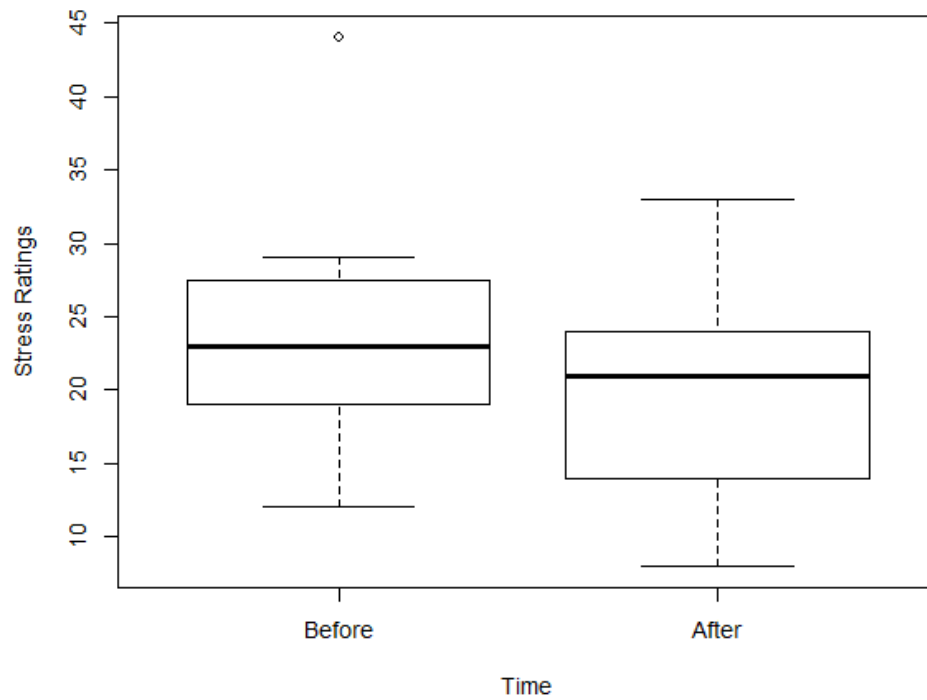
$$\left[ \bar{d} - t_{n_{\Delta}, 1-\frac{\alpha}{2}} \frac{s_{\Delta}}{\sqrt{n_{\Delta}}}, \bar{d} + t_{n_{\Delta}, 1-\frac{\alpha}{2}} \frac{s_{\Delta}}{\sqrt{n_{\Delta}}} \right]$$

The 95% CI of the mean difference is found to be  $[-7.08 -0.79]$ . As the 95% CI does not capture  $H_0$ , we reject it. There was a statistically significant mean difference between pain ratings.

## Paired Samples t-test Visualisation

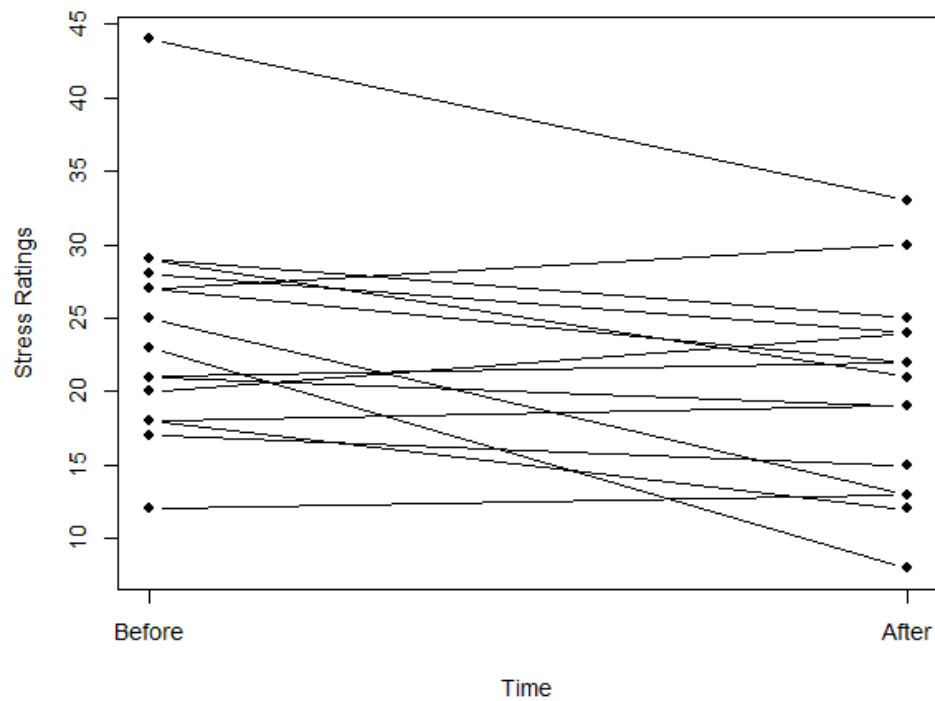
Visualising the results of a paired samples t-test can be tricky. There are a few approaches that can be used in R. The first involves a side-by-side boxplot of the scores. However, the plot does not show the dependency between the paired data.

```
> boxplot(PrisonStress_Sport$PSSbefore, PrisonStress_Sport$PSSafter,
          ylab = "Stress Ratings",
          xlab = "Time")
> axis(1, at=1:2, labels=c("Before", "After")) #Label the x-axis correctly
```



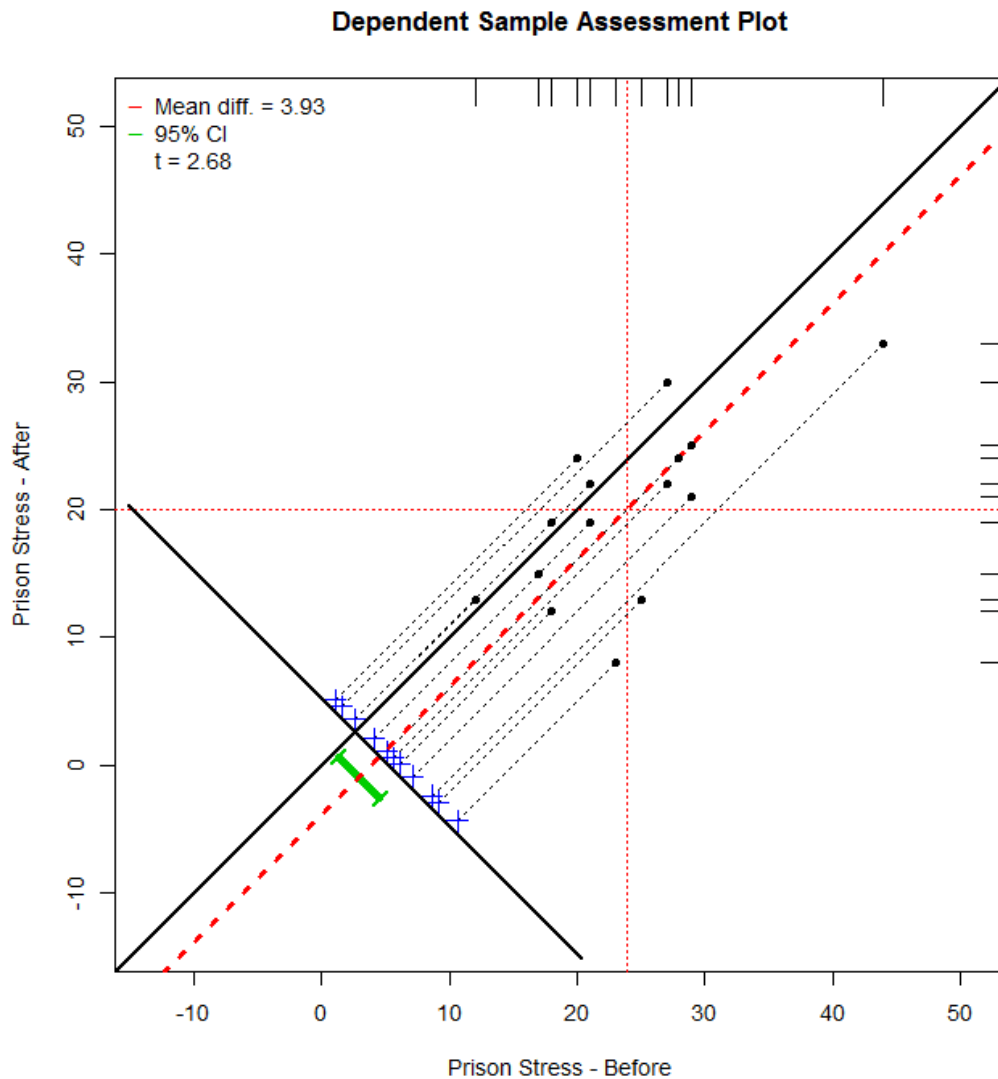
Another approach is to use a line plot where each line represents the paired scores. However, this is only suited to small datasets.

```
> matplot(t(data.frame(PrisonStress_Sport$PSSbefore,PrisonStress_Sport$PSSafter)),
  type="b", pch=19, col=1, lty=1, xlab= "Time", ylab="Stress Ratings",
  xaxt = "n")
> axis(1, at=1:2, labels=c("Before","After")) #Label the x-axis correctly
```



Alternatively, we can use the very fancy `granova.ds()` function from the `granova` package. This plot visualises the mean difference using a scatter plot. The plot reports the mean difference and confidence intervals from the paired samples t-test, which is very helpful. However, the plot takes some time to be able to interpret. But I think it's worth it. You can read all about this visualisation [here](#).

```
> install.packages("granova")
> library(granova)
> granova.ds(data.frame(PrisonStress_Sport$PSSbefore,PrisonStress_Sport$PSSafter),
             xlab = "Prison Stress - Before",
             ylab = "Prison Stress - After")
```



### Example Write-up

A paired-samples t-test was used to test for a significant mean difference between stress levels before and after exercise. The mean difference following exercise was found to be -3.93 (SD = 5.68). Visual inspection of the Q-Q plot of the difference scores suggested that the data were approximately normally distributed. The paired-samples t-test found a statistically significant mean difference between stress levels before and after exercise,  $t(df = 14) = -2.68$ ,  $p < .018$ , 95% [-7.08 -0.79]. Stress levels were found to be significantly reduced after prisoners engaged in one hour of exercise.

**[Return to top](#)**

[Recent Site Activity](#) | [Report Abuse](#) | [Print Page](#) | [Remove Access](#) | Powered By **Google Sites**