4. Distribution of the Sample Mean

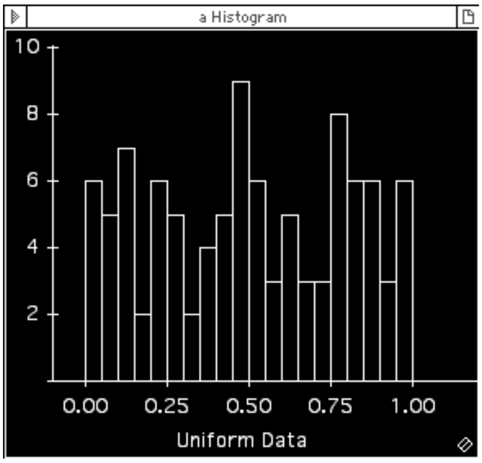
In the last section we examined the normal probability distribution and learnt how to determine the critical values for false positive risks of 10%, 5% and 1%. We now turn our attention to a very important distribution known as the **distribution of the sample mean** or **sampling distribution** which can also be described as a normal distribution. This distribution is obtained by taking repeated samples of size **n** from a population and calculating the **mean** for each sample. It can be shown that the distribution of the sample means is normal with the **same** mean **u** as the original parent distribution but will have a smaller standard deviation than the parent distribution - equal to $\boldsymbol{\sigma/\sqrt{n}}$. We call this standard deviation calculated from means the **standard error** (se) of the mean.

The normal distribution of the sample means applies irrespective of the distribution of the parent data. This result is formally known as the **Central Limit Theorem**. As was seen in the previous section the **z** transformation is again used to calculate the test statistic by transforming the data to a standard normal distribution to allow probabilities of occurrence to be computed. However, the formula is modified slightly because our evidence (i.e. data) is an average ($\bar{\mathbf{x}}$) rather than an individual value (**x**) while the standard error ($\sigma/\sqrt{n}$) replaces the standard deviation (σ). That is, the z transformation for data based on sample means is:

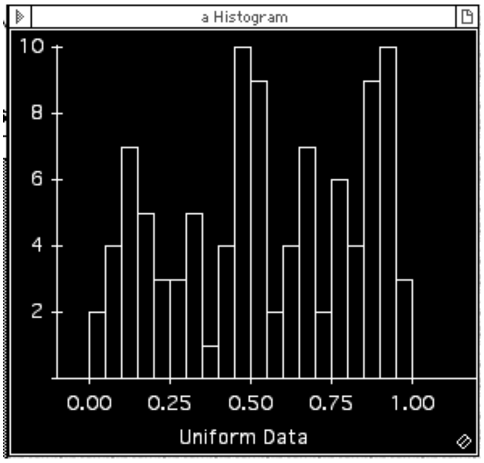
$$\boxed{z = (\bar{x}-u)/\sigma/\sqrt{n}}$$

Example

Consider the plot below which is based on 100 observations from a uniform distribution - the so called parent distribution. This distribution is characterised by the fact that all values between 0 and 1 are equally likely. As can be seen from the plot the uniform distribution is clearly non-normal in shape.



If we calculate the average of these 100 points the result is 0.504. Now if we repeat the simulation we obtain another uniform plot of 100 values as shown below



The mean of this distribution is 0.548.

Now if we continue this process (i.e. simulate 100 uniform data values and calculate the mean) over and over again and then plot the means we will see that the distribution (i.e. the sampling distribution) takes on a normal distribution shape! For example, the plot in Figure 4.1 is based on repeating the above process 10,000 times. In other words 10,000 means were calculated and plotted where each mean was based on 100 data values from a uniform distribution. The reader can visualise this simulation using the application *Central Limit Theorem.*

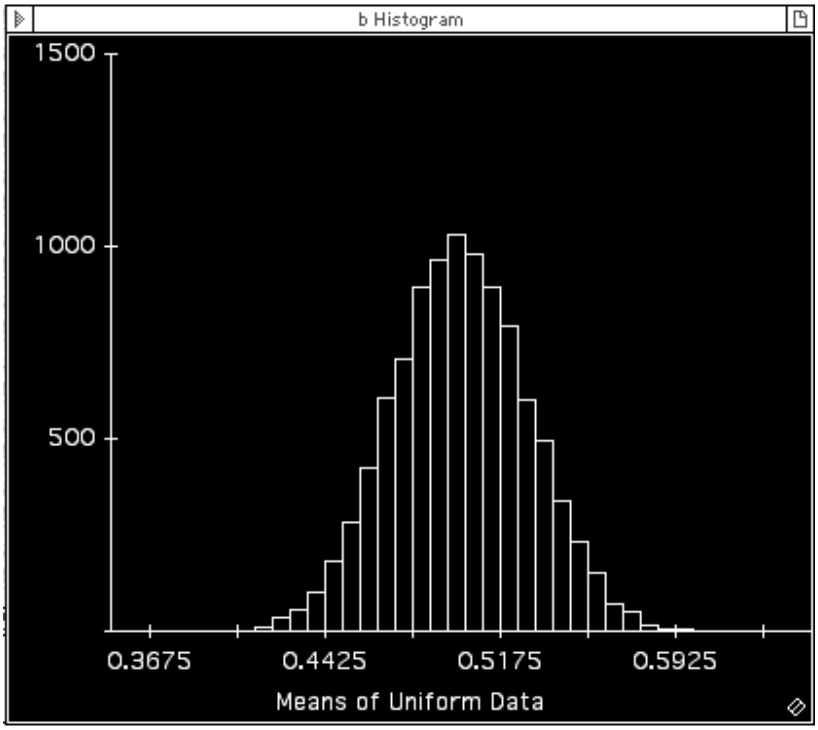


Figure 4.1: Distribution of 10,000 sample means

Figure 4.1 is visual proof that when we work with sample averages we can assume the underlying distribution is normal (we also have an additional requirement that the sample average is based on more than 30 points). This in turn implies that we can use the standard normal tables for calculating critical values and regions irrespective of the shape or distribution of the parent data from which the means were calculated from!

The mean of the sampling distribution distribution plotted in Figure 4.1 is 0.5. The central limit theorem states that the the mean of the sampling distribution should be the same as the mean of the uniform (parent) distribution. As the mean of a uniform distribution is 0.5 this result is in accordance with what the theorem predicts.

The theorem also predicts that the standard deviation of the sampling distribution (i.e. the standard error) is smaller then the standard deviation of the parent distribution by a factor $1/\sqrt{n}$ i.e. which in this case is $1/\sqrt{100}$ as each mean is based on 100 observation. The standard deviation of the uniform (parent) distribution (σ) is 1/12 so the theorem predicts that the the standard deviation of the sampling distribution distribution ($\sigma/\sqrt{n}$) is $1/12/\sqrt{100} = 0.028$. The mean of the sampling distribution in Figure 4.1 is 0.028 - virtually identical to what the theory predicts!

Exercise

The length of time to book a concert ticket on-line is known to be normally distributed with a mean of 3.4 minutes and a standard deviation of 0.25 minutes. The site has been redesigned and it is expected that the new design will decrease the booking time. To assess the impact of the new design before going live a total of 20 randomly selected users of the website booked concert tickets on-line. The **average** time of the 20 bookings was calculated as 3.2 minutes.

i) What is the probability distribution of the **average** time to book 20 concert tickets?

ii) Does this result suggest that the modification has had a real impact or is the result what we might expect from random variation. Use a false positive risk of 5% i.e. $\alpha = 5\%$.

4.1 t-distribution

The last section illustrated the important result that the probability distribution of means or averages is normal irrespective of the parent distribution of the data from which the average is calculated. We now turn ourselves to a distribution which is very similar to the Normal distribution called the *t-distribution*. This distribution is used in situations where we have small sample sizes ($n < 30$) and where the population standard deviation (σ) is **unknown**.

In this situation we substitute an estimate of the population standard deviation **σ**, with the sample standard deviation **s**, so that the **z** formula now becomes;

$$\boxed{t = \sqrt{n}(\bar{x}-u)/s}$$

where **s** is calculated as

$$s = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$$

However, this modified **z** transformation does not describe a normal distribution with mean 0 and standard deviation 1 i.e. **N**(0,1). This arises because the formula now contains two estimates - an estimate for the mean in the numerator (**x**) and an estimate for the standard deviation (**s**) in the denominator. Instead the formula describes what is known as students **t** distribution or *t-distribution* for short.

The *t-distribution* is named after William Sealy Gossett who developed the theory while employed as a brewer in the Guinness brewery in Dublin during the early 1900's. Statistical theory up to 1900 was based around large sample sizes. Gossett, however, worked with small sample sizes and developed the *t-distribution* (with help from two other giants in statistics at that time Ronald Fisher and Karl Pearson) to allow him to successfully draw inferences about population parameters based on small samples. Guinness at the time did not allow employees to use their name on published research so Gossett used the pseudonym **A Student**. He used the index **t** to describe the distribution in his research hence the name students *t-distribution*.

The *t* distribution is very similar to the normal distribution and for sample sizes greater than 30 both distributions are virtually identical. Unlike the standard normal distribution however, their is a different *t-distribution* for each sample size. For sample sizes less than 30 it has higher variability than the normal as the *t* formula contains two estimated quantities - the mean ($\bar{\mathbf{x}}$) and the standard deviation (**s**). The *t-distribution* also introduces a new technical term known as **degrees of freedom**. For one-sample tests of the kind in this section the degrees of freedom are calculated as the sample size minus 1 i.e. n-1. Further details on degrees of freedom are provided at the end of this chapter. Note that the use of the *t-distribution* assumes that the underlying distribution of the data is **normal**. However, this assumption can be relaxed somewhat as the test is robust to departures from normality. Another assumption is that successive observations are **independent** and departures from this assumption can lead to erroneous conclusions.

An extract from the **t** tables is shown in Table 4.1 while the complete table is provided in the formulae guide at the end of the handbook.

Degrees of	Cumulative Probability				
of Freedom	90%	95%	97.5%	99%	99.5%
5	1.476	2.015	2.571	3.365	4.032
10	1.372	1.812	2.228	2.764	3.169
20	1.325	1.725	2.086	2.528	2.845

Table 4.1: Extract from t-tables

The value of t for cumulative probabilities of 90%, 95%, 97.5%, 99% and 99.5% are shown in the extract in Table 4.1. To find the critical values for normally distributed data we transformed the data using the **z** transformation. For the *t-distribution* we adopt the similar approach calculating the test statistic known as **t**, i.e.

$$\boxed{t = \sqrt{n}(\bar{x}-u)/s}$$

However, in this case the critical value depends on two factors. The false positive rate as is the case for the normal distribution but also the **degrees of freedom**.

For example, the critical value for a t distribution for a data set with 10 values and a false positive rate of α = 0.1 or 10% is 1.37. This is calculated by going to **row 9** (degrees of freedom-1) and **column** equal to **90%** in the t-tables. This means that if the calculated value of **t** is greater than 1.37 (one-tailed test (right tail) i.e. $H_1: u > u_0$) or less than -1.38 (one tailed test (left tail) i.e. $H_1: u < u_O$) then the null hypothesis is rejected as shown in the Figure 4.2.

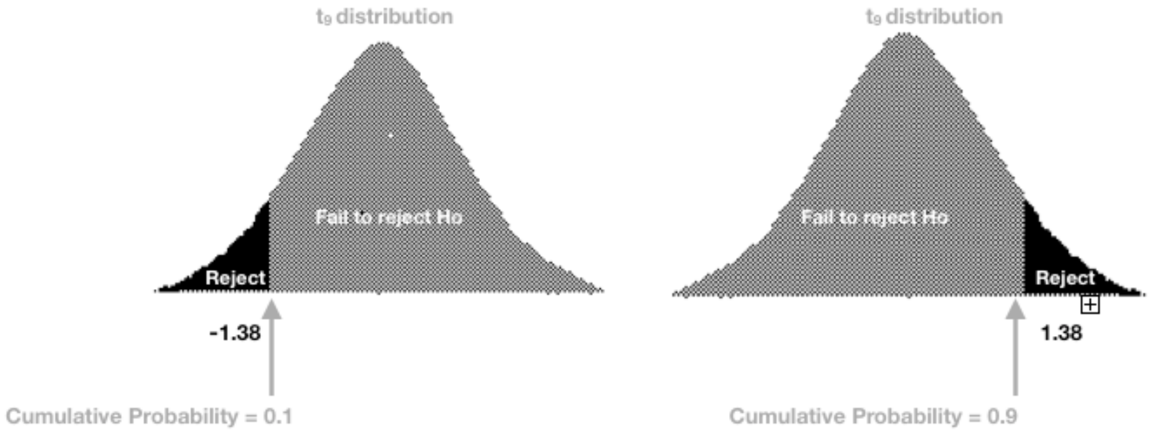


Figure 4.2: Critical values and regions for t-distribution with 9 degrees of freedom, one-tailed test and α = 0.10 or 10%

For a two tailed test the critical region is, as before, divided in two. If α = 0.1 then the critical region is 1- α/2 = 0.95 in the right hand tail and α/2 = 0.05 in the left hand tail. Looking up the t tables for 9 degrees of freedom with the column value of 95% the critical value are ± 1.83. This two tailed test is illustrated below.

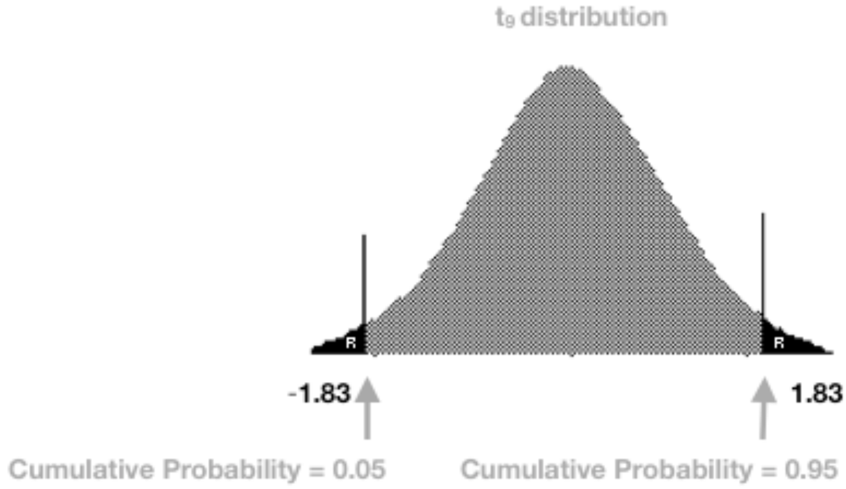


Figure 4.2: Critical values and regions for t-distribution with 9 degrees of freedom, two-tailed test with α = 0.10 or 10%

Exercise

Calculate the critical values of a *t-distribution* based on a sample size of 15 for α = 0.1, 0.5 and 0.01 for a) one-tailed and b) two-tailed test.

Worked Example 1

The mean time taken to download an application is assumed **normal** with a mean (**u**) of 10 seconds. The development team have introduced a new modification which may change the time taken to download. Consequently a sample of 20 apps are downloaded under carefully controlled conditions. The mean speed is calculated as 10.2 seconds while the **sample** standard deviation (**s**) is calculated to be 1 second. Test the hypothesis that the new modification has had no significant impact on the download speed using α = 0.01.

Solution

Step 1: Specify the Hypothesis

Ho: u = 10 seconds

Ha: u ≠ 10 seconds

This is a **two-tailed** test as the design team are unsure as to whether the new modification will increase or decrease download speeds.

Step 2: Calculate the test statistic

The test statistic is calculated as

$\mathbf{t} = (\overline{x}\text{-u})/s\sqrt{n}$

$= (10.2\text{-}10.0)/1/\sqrt{10}$

$= 0.632$

Step 3: Making the Decision

The research team have specified a type 1 or false positive error risk of no more than 1% i.e. α = 1%. The critical value of **t** is found selecting **row 19** in the t-tables (i.e. degrees of freedom = 20-1). The cumulative probability **columns** equal to 99.5% give us the upper critical value of 2.861. Due to the symmetry of the *t-distribution* the lower critical value (i.e. α/2) is - 2.861.

As the value of the test statistic 0.632 falls outside the critical values we do not reject the null hypothesis that the mean download speed have significantly changed from 10 seconds. The visualisation of the decision process is shown in Figure 4.3.

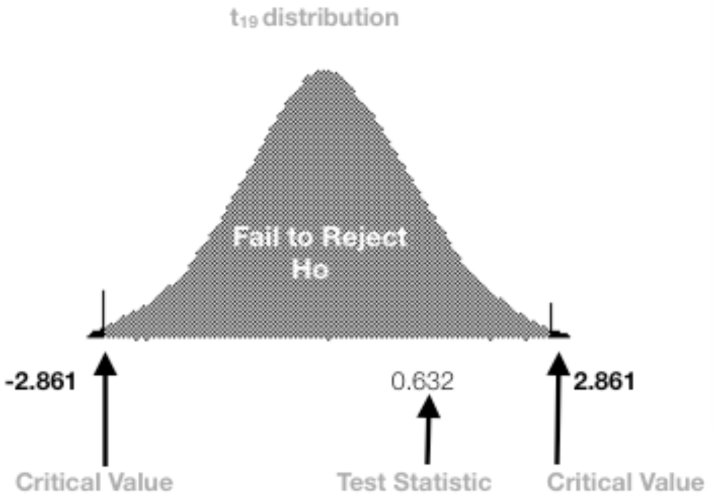


Figure 4.3: Critical values and regions for t-distribution with 19 degrees of freedom, two-tailed test with α = 0.01 or 1%

Step 4: Conclusion

As 0.632 is not in the critical region (i.e greater than 2.861 or less than -2.861) we fail to reject the hypothesis that the mean download speed has changed from 10 seconds. We should note however that the type 2 error is probably high (as the sample size is small) and the test may not sensitive enough to detect any change. This risk is not calculated for this example but can be reduced by increasing the sample size.

Exercise

i) Using the data in the previous example test the hypothesis using **α** = 0.05.

ii) If the software developers believed that the mean time to download would increase after the modification test this hypothesis using **α** = 0.1.

Worked Example 2

The following example is an edited extract from a paper entitled *APRN Usability Testing of a Tailored Computer-Mediated Health Communication Program* by Lin et.al. published in the journal Computers, Informatics, Nursing.

The research tested the usability of a touch-screen enabled 'Personal Education Program' (PEP) with Advanced Practice Registered Nurses (APRN). The PEP is a device that is designed to 'enhance medication adherence and reduce adverse self-medication behaviors in older adults with hypertension'.

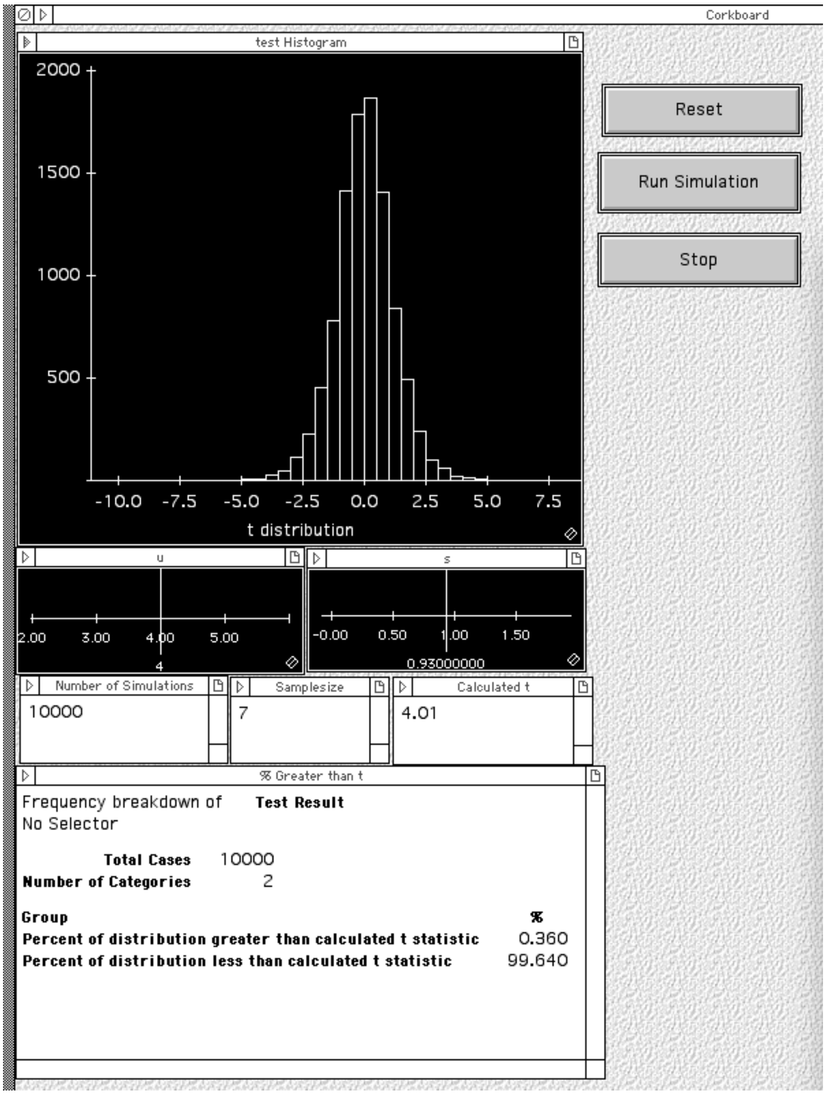
An iterative research process was employed in the assessment. Results were utilised to systematically modify and improve three PEP prototype versions—the pilot, Prototype-1 and Prototype-2. We will restrict our attention to the analysis of the Prototype-2 version.
This device was assessed by respondents who were asked to rank its usability on a seven point scale.

The null hypothesis is that respondents do not have a preference either way for Prototype-2. They are in a sense randomly selecting one of the seven points on the scale. If this is true i.e. no preference then the true mean would be 4. Therefore in the language of statistics we have

Ho: $u_O = 4$

Our alternative hypothesis is that there is a preference either way for the Prototype 2 version. That is, the respondent may have a distinct preference for it - reflected in high scores on the test or the respondent may record a poor experience - reflected in a lower score. Again in the language of statistical inference we will write this as

H_1: $u_O \neq 4$



Number of Simulations 10000

Samplesize 7

Calculated t 4.01

Frequency breakdown of **Test Result**
No Selector

Total Cases 10000
Number of Categories 2

Group	**%**
Percent of distribution greater than calculated t statistic	0.360
Percent of distribution less than calculated t statistic	99.640

Figure 4.4: Screenshot of application *t-distribution*

The test is two-tailed as we have no prior knowledge of the preferences of Prototype-2.

We will set the type 1 error to 5%. Therefore each tail area will correspond to 2.5 per cent of the t-distribution.

The test results (from the paper) recorded an average score for system usability ($\overline{\mathbf{x}}$) of 5.41 with a standard deviation (**s**) of 0.93 based on 7 respondents (**n**).

The test statistic is therefore

t = $\sqrt{}$n($\overline{x}$-u)/s = [$\sqrt{}$7(5.41-4.0)]/0.93 = 3.73/0.93 = **4.01**

To assess if this statistic falls into the critical region we will consult the t-table with 6 degrees of freedom with column heading 97.5% (i.e. 1- α/2) to obtain **2.447**. As our test statistic falls well inside the critical region we will reject the null hypothesis and conclude that there is a significant preference for Prototype-2.

We can also check this result using the simulation software *t-distribution*. Enter **u** = 4, **s** = 0.93, sample size = 7, **t** = 4.01 and 10,000 simulations to obtain the distribution of the test statistic as shown in the screenshot in Figure 4.4. The percentage of results greater than or equal to 4.01 is 0.36% (shown in white and which can be regarded as the p-value associated with this test) which is sufficiently rare for us to safely reject the hypothesis of no preference. We can therefore conclude that there is a significant positive assessment of Prototype-2.

It is interesting to note that no acceptable false positive level was provided in the paper - we set it to 5%. Other shortcoming include the absence of the raw data while no visual analysis of the rating by the seven respondents was provided. An informal review of usability papers available on the web suggest that this is seems to be generally the case.

Degrees of Freedom

Degrees of freedom can be considered as the number of data points used in the calculation of a statistic that are free to vary i.e. are not constrained. To explain further let us consider the formula $\Sigma(x_i-\overline{x})$. This formula sums to zero as adding up the deviations of each sample value x_i from the mean $\overline{x}$ will give 0. This means that if we know the values of n-1 of the x_i observations then the last value is predetermined (i.e not free to vary) because the total must equal zero. Therefore formulae that involve $\Sigma(x_i-\overline{x})$ have n-1 degrees of freedom e.g. the sample variance of a data set s^2 calculated as $\Sigma(x_i-\overline{x})^2$ /n-1. In general, one degree of freedom is lost for every statistic that needs to be estimated. The t-test statistic involves calculating $\Sigma(x_i-\overline{x})^2$ (in the formula for s in the denominator of the test statistic) so one degree of freedom is lost while the independent participant t-test (explained in the next section) has a formula that involving two estimated means $\overline{x}_1$ and $\overline{x}_2$ in the pooled formula for s hence there are n-2 values that can vary independently (and consequently two x_i values that cannot) and so we use n-2 degrees of freedom for this test.

Exercise

Prototype-2 was also assessed in terms of the usability characteristics **System Usefulness** and **Overall System User Satisfaction.** The following results were obtained:

	n	$\bar{x}$	**s**
System Usefulness	5	5.24	0.95
System User Satisfaction	8	5.75	1.00

For each of the two usability characteristics calculate

i) The test statistic.

ii) Using the results of i) determine if there is sufficient evidence to reject Ho. Assume a false positive rate of 5% for System Usefulness and 1% for System User Satisfaction.42

iii) Use the simulation software *t-distribution* to visualise the distribution of the test statistic (use 10,000 simulations). State, giving a reason why, if the results of the simulation are consistent with the results from the t-tables calculated in ii).