5. Two-Sample Tests

Introduction

The usability tests outlined in the previous sections were based on testing one sample statistic e.g. the sample mean ($\overline{x}$) and assessing this statistic against the distribution of possible outcomes (i.e. the probability or sampling distribution) according to some hypothesis. If the test statistic fell into the critical region of the probability distribution of the normal or *t-distribution* we rejected the hypothesis under test. Tests of this kind are usually referred to as **one-way** or **one-sample** tests. For example, the one-sample t-test was the basis for assessing the download speeds of an application on page 48 while a one-sample z test was applied in the exercise on page 45 assessing the length of time to book a concert ticket on-line.

We will now turn ourselves to the situation where we are testing for differences between two sample means. This type of test arises if, for example, we have two groups of people testing two different devices - say an iPhone and a Google Android - to determine any significant difference in usability. This is an example of a **two-sample** or unpaired test (also known as **different participant** design) and as tests of this kind tend to involve small numbers of respondents we will restrict our analysis to the situation where the probability distribution is described by the t-distribution.

We will also examine one other design which is quite common in usability tests known as **paired designs** (also known as a **same participant** design). For example, instead of one group testing the iPhone and another group the Android, as outlined in the previous paragraph, we have the same person testing both products. Paired designs can often be a more sensitive design for detecting differences because the same person is testing both devices. This tends to reduce the random variation of the test data as it eliminates the person to person variation present in the unpaired test. That is observed differences in data are explained to a greater extent by the design of the actual products rather than the people conducting the experiment.

However, paired designs can have a significant disadvantage because in the process of testing a device the user may learn features (consciously or sub consciously) that may influence the test results of the other device. For example, say a usability test is conducted to determine if there is a significant difference in the time taken to set up a wireless wi-fi on two mobile devices. If both devices have similar interfaces then the experimenter may learn

features on one device which aid them in their assessment of the other device. These so called **learning effects** can bias results and it can be difficult to remove them altogether. To reduce their impact participants are usually assigned randomly which product to test. In this example it would mean that about 50% of testers would evaluate an iPhone first and 50% the Android first. In two-sample tests we usually refer to the device type as the **independent variable** while the usability measurements are usually called the **response variable**.

5.1 Unpaired t-test (or Different Participant Design)

The procedure for the analysis of two sample tests (paired or unpaired) is broadly the same as the one-sample tests discussed earlier i.e.

i) Specify the null hypothesis
ii) Calculate the test statistic
iii) Make the decision by comparing the test statistic with the appropriate *t-distribution.*

Step 1: Specify the Hypothesis

The null hypothesis is expressed formally as:

H_o: $u_1 = u_2$ with the alternatives specified as either

H_1: $u_1 \neq u_2$ (two-tailed)

or

H_1: $u_1 > u_2$ or H_1: $u_1 < u_2$ (one-tailed)

Step 2: Calculate the test statistic

In the one-sample case when the standard deviation is unknown the distribution of the sample mean is **t** with n-1 degrees of freedom. Similarly, when testing the difference between two means (when the standard deviation is unknown) the distribution of the difference is also **t** but with $n_1 + n_2 - 2$ degrees of freedom where n_1 and n_2 are the sample sizes of the two groups.

The test statistic is calculated as

$$t = \frac{[(\overline{x_1} - \overline{x_2}) - 0)]}{s\sqrt{(1/n_1 + 1/n_2)}}$$

Calculating the pooled standard deviation, s

To calculate **s** the following procedure is undertaken.

Calculate the variance of the first sample i.e. s_1^2

$$s_1^2 = \frac{[\Sigma x_1^2 - (\Sigma x_1)^2/n_1]}{n_1 - 1}$$

Calculate the variance of the second sample i.e. s_2^2

$$s_2^2 = \frac{[\Sigma x_2^2 - (\Sigma x_2)^2/n_2]}{n_2 - 1}$$

The combined estimate s^2 is calculated as:

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

Therefore s is calculated as $\sqrt{\frac{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{n_1 + n_2 - 2}}$

Step 3: Make the decision

The calculated test statistic is then compared with the $t_{n_1 + n_2 - 2}$ distribution. If the test statistic falls in the critical region we reject the null hypothesis in line with the pre-set false positive error.

Note that the use of the t-distribution assumes that the underlying data is normally distributed - which as we have stated in the previous section the test is robust to departures from. However, an additional assumption is that the variances are the **same** for both samples. This assumption is more problematic as the test is sensitive to departures from equality of variances. Any doubt about this assumption should lead the experimenter to use an alternative test to the t-test.

Worked Example 1

Two groups of participants took part in a research experiment designed to test if there was a significant difference in the time (to the nearest minute) taken to set up an e-mail account using two mail applications. A total of seven participants were randomly assigned to application 1 while five were randomly assigned to application 2. The data from the study are shown below together with some summary statistics.

Time (Minutes)	
Application 1	**Application 2**
8	5
6	6
7	8
6	3
8	4
6	
3	
$n_1 = 7$	$n_2 = 5$
$s_1^2 = 2.9$	$s_2^2 = 3.7$
$\overline{x}_1 = 6.3$	$\overline{x}_2 = 5.2$

Test the hypothesis that the there is no significant difference in time between the two applications. Use a decision risk $\alpha = 0.05$.

Solution

Step 1: Specify the Hypothesis

This hypothesis is expressed formally as;

H_o: $u_1 = u_2$ against

H_1: $u_1 \neq u_2$

This test is two tailed as the question contains no prior or historical information on the relative speed of the applications.

Step 2: Calculate the test statistic

The first step is to calculate s^2 which is the pooled estimate of the sample variances for both groups

$$s^2 = \frac{[(n_1-1)s^2_1 + (n_2-1)s^2_2]}{n_1 + n_2 - 2}$$

$$= \frac{[(6)2.9 + (4)3.7]}{7 + 5 - 2}$$

$$s^2 = 3.2$$

$$\therefore \quad s = 1.79$$

$$t = \frac{(\overline{x_1} - \overline{x_2}) - 0}{s\sqrt{(1/n_1 + 1/n_2)}}$$

$$t = \frac{(6.3 - 5.2)}{1.79\sqrt{1/7 + 1/5}}$$

$$t = \frac{1.1}{1.05} = 1.05$$

Step 3: Making the Decision

The **t** tables with 10 degrees of freedom and cumulative probability of 97.5% returns critical values of ± 2.23. As our calculated **t** of 1.05 is outside the critical region we do not reject the hypothesis that the mean scores are the same. However, it should be noted that this experiment was based on a very small sample size (12 participants in total) so the power of the test to detect any difference is low and the likelihood of a type 2 (or false negative) error is high. Perhaps the researcher team could consider using a larger sample size.

Worked Example 2

The scientific paper assessing the usability of a touch-screen enabled Personal Education Programme (PEP) discussed in the previous chapter also examined the difference in usability between two versions of PEP known as Prototype-1 and Prototype-2. The following results were reported by the research team

System Usability Results

$\overline{x_1}$ =	6.1	n_1 =	7	s_1 =	0.91
$\overline{x_2}$ =	5.4	n_2 =	7	s_2 =	0.93

The researchers used a **two-sample t-test** to determine if there were any significant differences between the usability of the two prototypes. The solution is as follows:

Step 1: Specify the Hypothesis

The null hypothesis is expressed formally as;

Ho: $u_1 = u_2$ against

Ha: $u_1 \neq u_2$

This test is two tailed as the question contains no prior or historical information on the relative differences in usability of the two prototypes. We will set the type 1 error to 5% i.e. $\alpha = 0.05$.

Step 2: Calculate the test statistic

The first step is to calculate s^2 which is the pooled estimate of the sample variances for both groups

$$s^2 = \frac{[(n_1-1)s^2_1 + (n_2-1)s^2_2]}{n_1+ n_2 - 2}$$

$$= \frac{[(6)0.8464 + (6)0.8649]}{7+ 7 - 2} = \frac{10.1576}{12}$$

$$s^2 = 0.85565$$

$$\therefore \quad s = 0.925$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s\sqrt{(1/n_1 + 1/n_2)}}$$

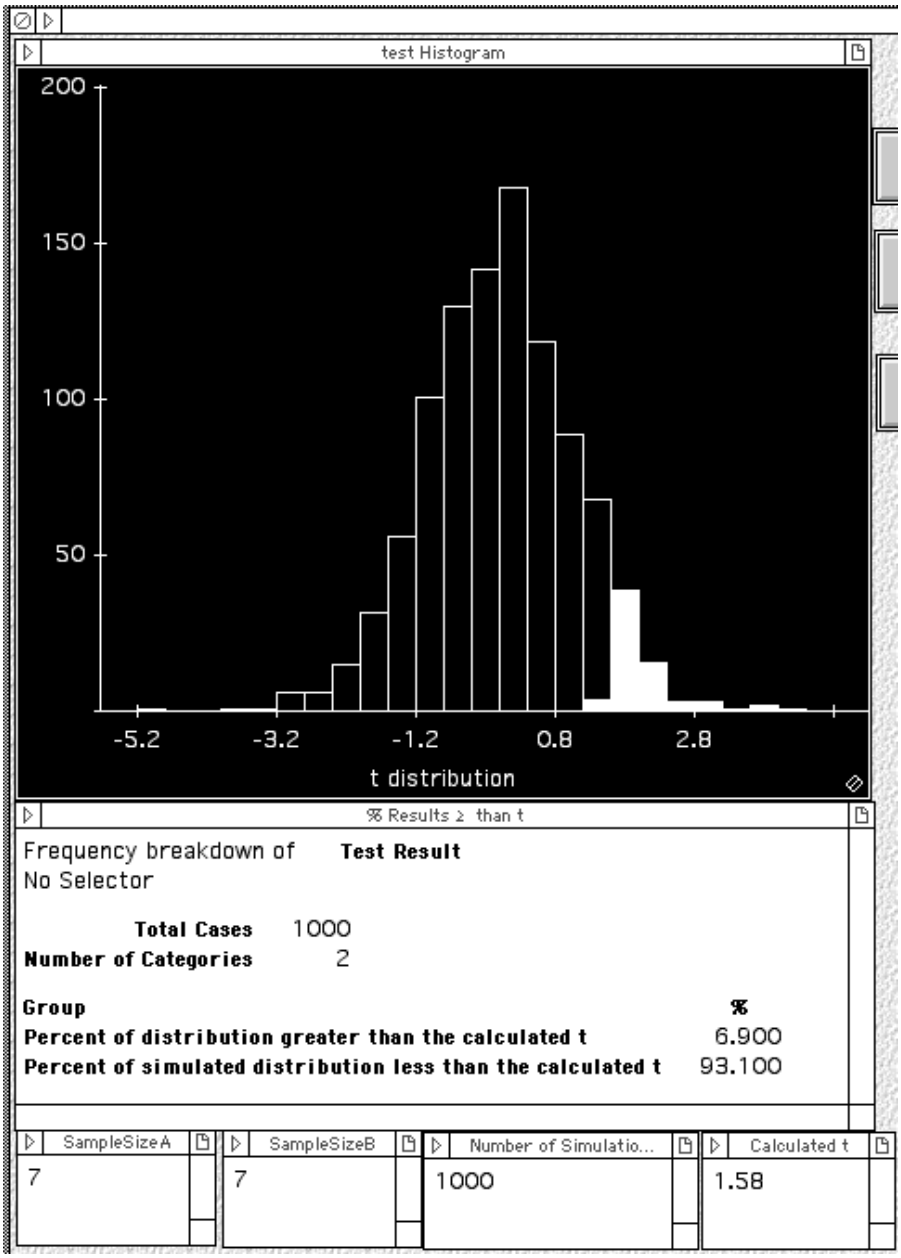
$$t = \frac{(6.19 - 5.41)}{0.92\sqrt{1/7 + 1/7}} = \frac{0.78}{0.925\ (0.534)}$$

$$t = \frac{0.78}{0.494} = 1.58$$

Step 3: Making the Decision

The t tables with 12 degrees of freedom and cumulative probability of 97.5% returns critical values of ± 2.179. As our calculated t of 1.58 is outside the critical region we do not reject the null hypothesis that the mean usability score is the same for both prototypes. However, it should be noted that this experiment was based on a very small sample size (12 participants in total) so the power of the test to detect any difference is low and the likelihood of a type 2 (or false negative) error is high. Perhaps the researcher team should consider using a larger sample size.

We can visualise this result using the simulation software *unpaired t-test*. A screenshot of this application is shown in Figure 5.1. Inputting the sample size for both groups together with the value of **t** calculated as 1.58 we can simulate the distribution of **t** - assuming both means (and variances) are the same. Generating 1,000 simulations we see from our result that the chances of obtaining a **t** value of 1.58 assuming no difference in usability is about 7 per cent (shown in white and which can be regarded as the p-value of the test). This is well over the 2.5 per cent critical point and is consistent with our decision not to reject Ho. We therefore conclude that there is not sufficient evidence to reject our null hypothesis that the usability characteristic for both prototypes are the same.



Exercises

1. The research team in Worked Example 2 also measured the level of interface error in their comparison of the two prototype designs. The summary statistics are shown in the table below

Interface Error

Prototype 1	$\overline{x_1}$ =	4.22	n_1 =	9	s_1 =	1.99
Prototype 2	$\overline{x_2}$ =	1.48	n_2 =	8	s_2 =	1.48

i) Calculate the unpaired t-statistic for this set of results

ii) Using the t-tables determine if there is sufficient evidence to reject the null hypothesis of no difference in the interface error between the two prototypes. Use a false error rate of 5% and state clearly your conclusions.

iii) Using the simulation software *unpaired t-test* visualise the distribution based on 1,000 simulations of the test statistic calculated in i) and the respective sample sizes. Determine if the results obtained are consistent with the results obtained in ii).

2. Repeat Question 1 using the following data based on the evaluation of system user satisfaction for each prototype.

System User Satisfaction

Prototype 1	$\overline{x_1}$ =	6.53	n_1 =	9	s_1 =	0.68
Prototype 2	$\overline{x_2}$ =	5.75	n_2 =	8	s_2 =	1.00

3. In order to test the hypothesis that older people are slower at reading text from a screen than younger people, two groups of 10 read an identical paragraph of text on a Kindle device. The length of time taken (in seconds) for each participant together with some summary statistics is shown in the table below:

Older Participants	Younger Participants
21	10
19	19
17	14
15	5
22	10
16	11
22	14
22	15
18	11
21	11

$\Sigma x_1 = 193$	$\Sigma x_2 = 120$
$\Sigma x_1^2 = 3789$	$\Sigma x_2^2 = 1566$
$n_1 = 10$	$n_2 = 10$

Test the hypothesis that the mean time to read text is the same for both groups. Use $\alpha = 0.1$ (i.e. 10%) and assume that the observations (i.e. times) can be regarded as normally distributed.

5.2 Paired t-test (or Same Participant Design)

We now turn ourselves to the situation where the same individual tests two devices. This type of test is known as a paired test or **same participant design** in the usability literature. Paired tests tend to have lower variability than unpaired tests because as stated earlier the same person is testing both devices. This tends to provide a more precise estimate of any real differences in the usability of the devices. However, due to the potential impact of learning effects with this design referred to earlier it is important that the order in which participants perform the two tasks is randomly assigned. That is, roughly half of the participants should evaluate device A first and then evaluate device B while the other half should assess B first followed by A. This is known as **counterbalancing** and reduces possible unfair effects of learning from the first task.

Calculating the Test Statistic for the Paired Test (or Same Participant Design)

The test statistic for the paired design is ;

$$t = \frac{\bar{d}-0}{s_d/\sqrt{d}}$$

with n-1 degrees of freedom.

where

$\bar{d}$ = average difference. This is calculated as the average of the differences between each device for each participant.

d = the number of participants

s^2_d is the formula used to calculate the variability of the differences. It is the same as the formulae used for the one and two-sample unpaired tests except we are using the letter **d** to represent difference rather than the more usual **x** notation. That is

$$s^2_d = \frac{\Sigma d_i^2 - (\Sigma d_i)^2/d}{d-1}$$

Example

In order to establish if there is a significant difference between the time taken to book an airline ticket on two websites, ten individuals, randomly selected, recorded the times to book a ticket (in minutes) on each site as shown in the table below.

Individual	Site	
	A	B
1	13.2	13.1
2	17.9	17.6
3	4.10	4.10
4	17.0	17.2
5	10.3	10.1
6	4.00	3.97
7	5.10	5.10
8	8.00	7.90
9	8.80	8.70
10	12.0	11.6

The order in which each site is used first by each individual was randomly assigned.

The research question is to determine if there is a significant difference in the time taken to book an airline ticket. The level of proof required for a decision i.e. the decision risk, **α**, is 0.05 or 5%.

Step 1: Specify the Hypothesis

The null hypothesis can be expressed formally as;

Ho: $u_1 - u_2 = 0$ or $u_1 = u_2$ for all 10 observations.

The alternative hypothesis is

H_1: $u_1 \neq u_2$

This test is two-tailed as we have no prior information concerning the relative efficiencies of the two web-sites.

Step 2: Calculate the Test Statistic

To calculate the test statistic we must find $\overline{\mathbf{d}}$, the average difference for each of the 10 individuals as follows:

Individual	Difference
1	0.1
2	0.3
3	0.0
4	-0.2
5	0.2
6	0.3
7	0.0
8	0.1
9	0.1
10	0.4

The sampling (probability) distribution of the differences d under the null hypothesis is assumed normal with a mean of 0 and a standard error of $s_d/\sqrt{n}$ where s_d is the standard deviation of the differences and n is the number of individuals or paired comparisons.

The test statistic is then ;

$$t = \frac{\overline{d}-0}{sd/\sqrt{d}}$$

with d-1 degrees of freedom.

In this example

$\overline{d}$ = average difference = 0.13

$$s^2_d = \frac{\Sigma di^2 - (\Sigma di)^2/d}{d-1} = 0.031$$

$\therefore s_d = 0.176$

t is then calculated as $\frac{0.13}{0.176/\sqrt{10}} = 2.33$

Step 3: Making the decision

The t tables with 9 degrees of freedom and cumulative probabilities of 97.5% (as the test is two-tailed with $\alpha = 0.05$) the critical values of ± 2.26 are obtained. As the calculated **t** is 2.33 which falls just inside the right hand tail critical region we reject the hypothesis that the average time to book a ticket is the same. The results suggest that it is quicker to book a ticket using site B compared to site A.

We can employ the application *paired t-test* to give us additional insight into the results. A screenshot of this application is given below. In the lower panel the user enters the number of paired comparisons in the *samplesize* field, the number of simulations and finally the calculated t-value. The simulation provides a visual representation of the distribution of the test statistic **assuming** that both the mean time to book a ticket is the **same** for site A and site B. The table shows the percentage of tests that are greater than the calculated t-test which (which can be considered the p-value of the test) in this case is 1.9% (shown in white) and within the critical region for rejecting Ho (i.e. less than 2.5%). This result is therefore consistent with the results obtained directly from the t-tables.

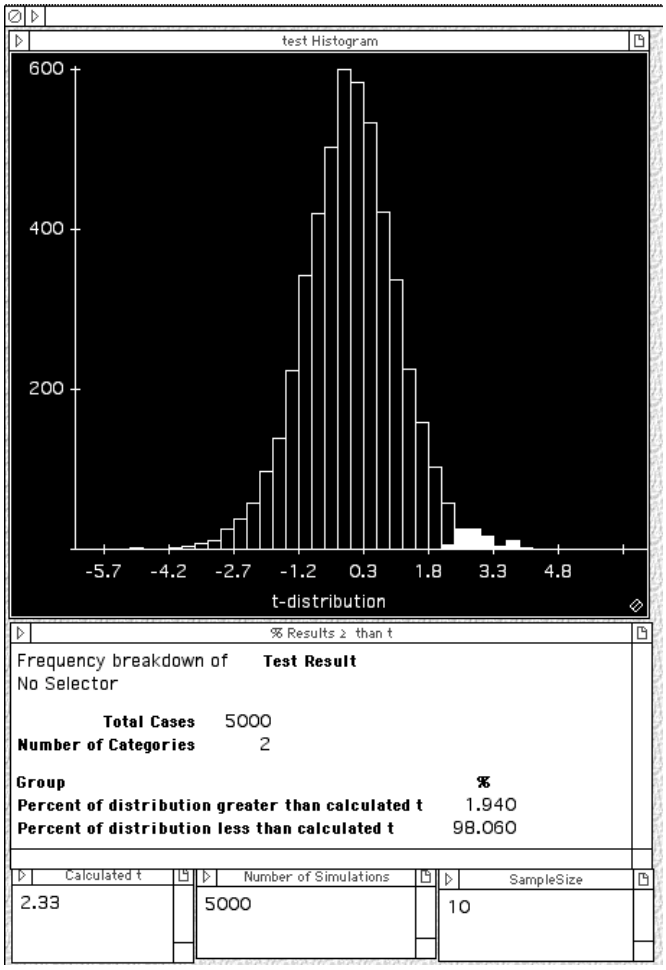


Figure 5.2 Screenshot of application *paired t-test*

A/B Tests

The experimental designs outlined in this section are increasingly been used in organisations such as Amazon, Google, Apple, Netflix and eBay where they are referred to as A/B tests. Typical examples include assessing the intervention of interface design changes on the number of customer purchases. For example, Dan Siroker writing in *Wired* magazine in 2012 outlines how the use of the A/B test increased the number of people signing up to the Obama election website by 4 million with total contributions of some $75 million. This was achieved by experimenting with certain design features such as removing a bright red *sign up* button which few people clicked and replacing with a *learn more* button which increased the number of new sign ups by 19 percent. In fact three three button configurations were tested - *learn more*, *join us now* and the existing *sign up* button.

Google ran their first A/B test on February 27 2000. They had often wondered whether the number of results the search engine displayed per page which defaulted to 10 was optimal. An A/B experiment was run with 0.1 per cent of the search engines traffic presented with 20 results per page; another 0.1 per cent was allocated 25 returns and another 0.1 per cent allocated 30. While the experiment was a disaster according to Siroker the company did obtain valuable insights particularly the impact of the speed on page loading. In 2011 alone the company ran more than 7,000 A/B tests on its search algorithm. In 2010 Siroker joined forces with an ex Google employee called Pete Kooman and developed an application called *Optimizely* which allows design changes to me implemented immediately and used by non-programmers.

However it is important to understand that while the web is a useful laboratory to conduct experiments there are many pitfalls which may not be fully understood in organisations who do not have access to staff with statistical skills. Most importantly is that to conclude that an intervention has had an impact needs what is known as a randomized controlled experiment as outlined in this chapter. This means that a proportion of users are randomly allocated one interface and another proportion the another interface. The two user groups should be controlled by important factors such as age, gender, residence i.e. the age cohort should be the same in both groups etc. Meeting these conditions can be difficult to apply in a web environment as there may be no control of important user attributes.

Exercise

In order to test a hypothesis that the kind of font, serif or sans serif, is an important design attribute on a website, a usability engineer designs an experiment in which 12 users are automatically timed to read two similar web pages. One page is designed using a serif font while the other is based on a sans serif font. The experiment is designed such that any learning effects are minimised through counterbalancing. That is, the order in which a font design is used first by the tester is randomly assigned to minimise any bias in the results.

The data are as follows:

User	Serif	Sans Serif
1	112.3	111.2
2	110.7	119.3
3	106.1	105.3
4	115.3	112.9
5	109.8	107.4
6	108.9	109.1
7	106.0	106.2
8	107.4	106.3
9	114.3	111.2
10	111.1	109.6
11	109.5	108.2
12	112.2	112.5

i) Calculate the test statistic for this design.

ii) Test the hypothesis that the type of font has no impact on response time. Use $\alpha = 0.05$ and state clearly your conclusions.

iii) Use the application *paired t-test* to simulate this experiment 2,000 times. State, giving a reason if you think the results obtained are consistent with the results obtained in ii)?