6. Statistical Estimation

6.1 Introduction

In this course on Usability testing we have concentrated on testing whether the assumptions concerning a population parameter (i.e. the true average time to book a holiday or the true proportion of satisfied users of a software application) still hold. The assessment is based on the strength of the evidence submitted i.e. the data. If the test data support the status quo (i.e. the defendant is not guilty) we fail to deliver a guilty verdict i.e. we do not reject the hypothesis. However, if the evidence is strong which is reflected in the data falling in the tail areas of the distribution of outcomes we can safely reject the hypothesis of innocence - bearing in mind that there are two potential errors inherent in our decision making process.

A common theme that has been running through this short tour of statistical reasoning is the fact that the population parameter has generally being specified - usually based on the collection of large amounts of data - historical or otherwise. For example, in Chapter 3 (see example on page 40) the '**true**' length of time to book a concert ticket on-line was specified as normally distributed with a mean of 3.46 minutes and a standard deviation of 0.25 minutes. The distribution of the **average** time for say 40 individuals to book 40 concert tickets on-line is therefore normal with the same mean of 3.46 minutes but with a standard deviation of $0.25/\sqrt{40}$ (using the results of the distribution of the sample mean outlined in Chapter 3). The simulation of 1,000 points for this distribution is shown in Figure 6.1.

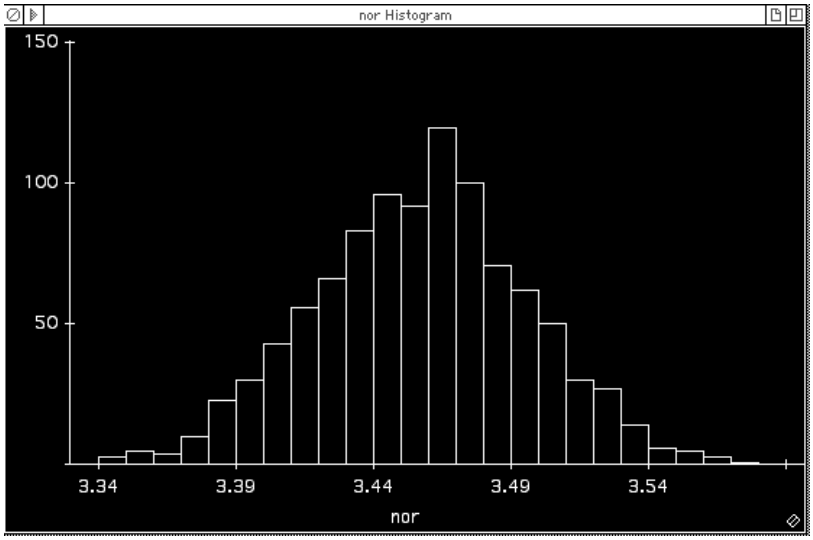

Figure 6.1: Distribution of on-line booking times, **u** = 3.46; **σ** = $0.25/\sqrt{40}$ = 0.039

We will now turn ourselves to the common situation where the only information we have concerning the true parameter is based on just **one** sample result. What is of interest is how close this one sample result is to the true but unknown mean? This is known as **statistical estimation** and involves generating an interval - known as a **confidence interval** - that encloses (or captures) the true mean with a specified degree of confidence or probability. For example, assume we conduct a usability test and find that the average time for 40 bookings of concert tickets on-line is 3.50 minutes. Can this result tell us anything about the location of the true mean of 3.46 minutes? The answer which we will see shortly is yes!

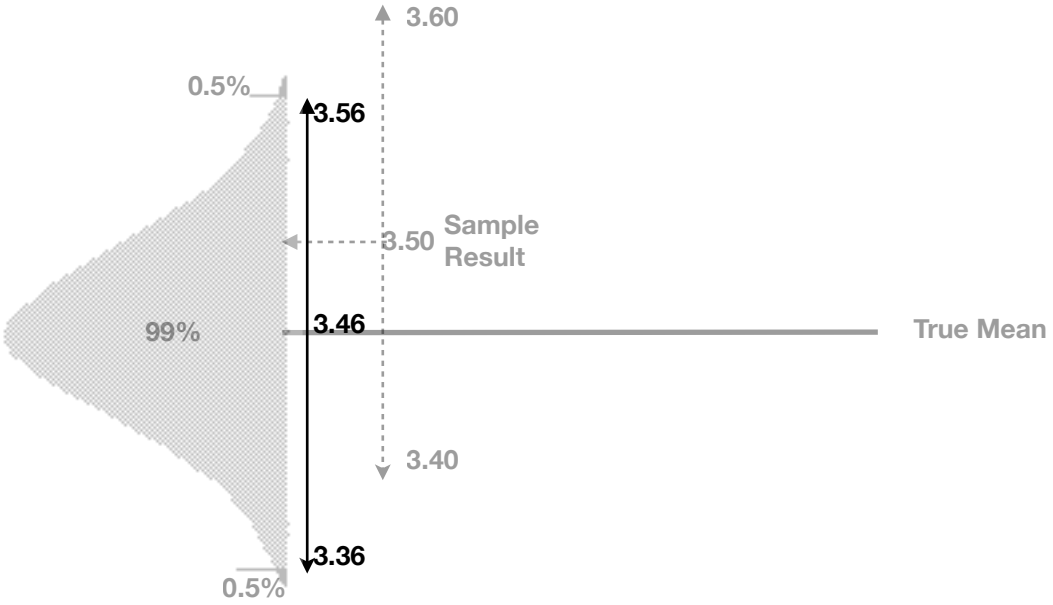

Figure 6.2: 99% confidence interval centered on the sample result of 3.50 minutes will capture the true mean of 3.46 minutes

The width containing 99% of the possible booking times from Chapter 3 is $\overline{x} \pm 2.56\ \sigma/\sqrt{n}$ which using the above results is $3.46 \pm 2.56(0.25/\sqrt{40})$ or 3.46 ± 0.1. If we now take this **population width** (calculated as a line of 0.2 units (i.e. ± 0.10) and centre it on the **one sample average result** of 3.50 the subsequent interval of 3.50 ± 0.10 will enclose or capture the true mean of 3.46 as shown in Figure 6.2.

In fact, if we repeated this usability experiment over and over again we would find that 99% of all intervals centered on the sample results would encompass the true mean of 3.46 - hence the phrase '99% confidence' that is used extensively in the research literature. Only in the 1% of sample results that fall in the critical region of the population distribution i.e. greater than 3.56 minutes or less than 3.36 minutes will the calculated interval not capture the true mean as shown in in Figure 6.3.

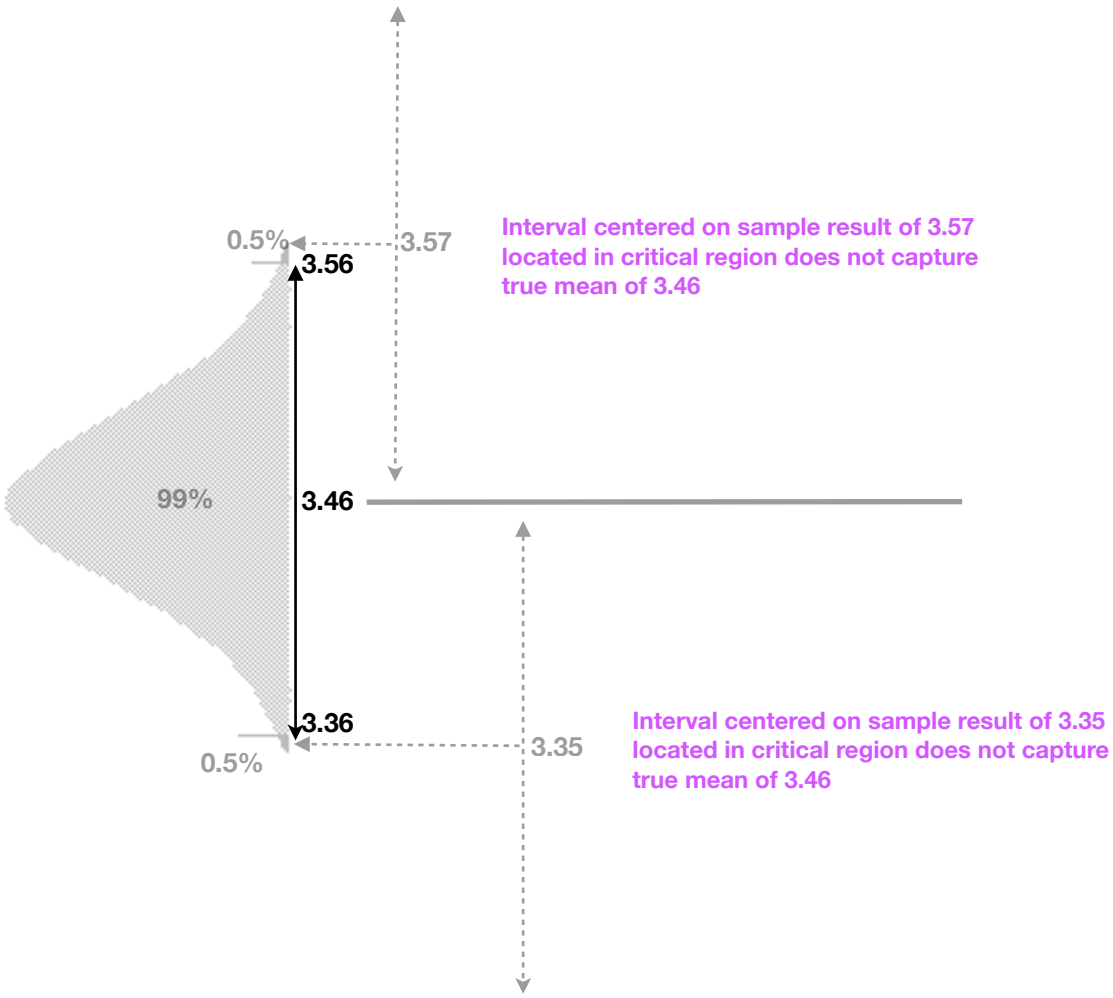

Figure 6.3: 99% confidence interval centered on sample booking times of 3.35 or 3.57 minutes will not capture the true mean of 3.46 minutes

One significant problem with this line of reasoning is that calculating the width of the population distribution requires knowledge of its standard deviation, σ which may be unknown. But remarkably it can be shown (using mathematical statistics) that if the standard deviation of the population distribution (i.e. σ) is unknown then the standard deviation of the data that the **one** sample average is calculated from i.e. **s** can be reliably used as a good approximation to σ!

Now that we can calculate the width of the distribution we can use this information to make statements about how confident we are that the interval centered on our sample result encloses the true mean. Using the properties of the normal distribution we can state that with 95% probability the interval calculated as $\overline{x} \pm 1.96\ \sigma/\sqrt{n}$ will capture the true mean. Similarly the interval $\overline{x} \pm 1.65\ \sigma/\sqrt{n}$ will capture the true mean with 90% probability. In other words if we collected a large number of samples then 90 per cent of the intervals centered on each sample will contain the true mean that we are trying to estimate. This is sometimes referred to as a 90% confidence interval.

6.2 Visualisation of Confidence Intervals

The histogram below comprises 100 samples based on the time taken (in minutes) for identical holidays to be booked using an on-line website. The true mean time to book this holiday - based on a substantial amount of historical data - is 10 minutes with a standard deviation of 2 minutes. The booking times are considered to be normally distributed. Calculating a 95% confidence interval for each of the 100 samples we obtain the plot shown below (right). From this plot (with the white line representing the true mean of 10 minutes) we see that six of the hundred samples or 6% - (as this is a simulation it won't always be 5% but should be close to it) from samples located in the tails - do not capture the true mean of 10. This is what is meant by the term 95% confidence.

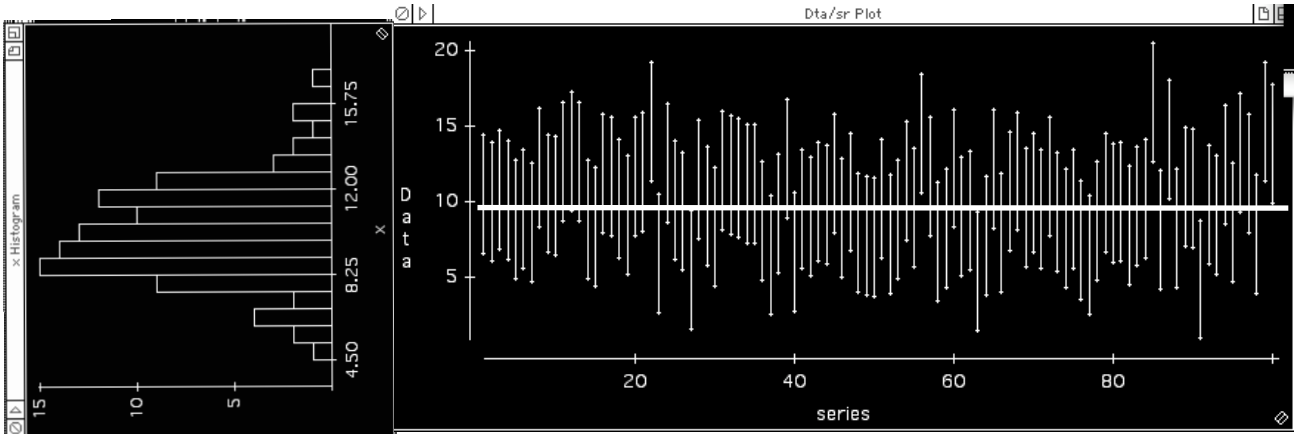

Figure 6.4: Visualisation of confidence intervals

Confidence intervals are a subtle concept! We are saying that if we compute an interval based on **one** sample then we cannot say for certain that the interval will capture the true mean. We can only state that if samples are taken over and over again and an interval computed for each sample then the true mean will be captured in a high percentage (e.g. 90%, 95% or 99%) of samples.

Exercise

The length of time to book an airline ticket on-line is known to be normally distributed with a mean of 4.5 minutes and a standard deviation of 0.7 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 60 users book concert tickets on-line and the average time is recorded as 4.7 minutes.

i) What is the distribution of the average time to book an airline ticket per 60 customer transactions?

ii) Assuming a sample result recording a mean of 4.7 min based on 60 users is obtained, calculate a 90%, 95% and 99% confidence interval for the true mean.

6.3 Confidence Intervals for t-distributed Data

The calculation and interpretation of a confidence interval for one sample, two sample and paired t-tests involves the same theoretical considerations as outlined in the previous section. However, it should be noted that the critical values for the *t-distribution* are not fixed as is the case for the normal distribution but depend on the sample size.

a) Confidence interval for one sample t-test

The 95 % confidence interval for the one sample t test is :

$$\bar{x} \pm t_{(\text{degrees of freedom, upper cumulative probability})} s/\sqrt{n}$$

The term upper cumulative probability in the above formula means the value of the upper critical region. For example, on page 43 of Chapter 4 the mean download time of 20 apps was found to be 10.2 seconds and the **sample standard deviation** (i.e. **s**) was 1 second. To generate a 95% confidence interval for this we first calculate the value of:

$t_{(\text{degrees of freedom,confidence})}$ which is $t_{(19,\ 97.5\%)} = 2.091$.

The confidence interval is therefore

$$10.2 \pm 2.091(1/\sqrt{20}) = 10.2 \pm 0.467$$

Therefore with 95% confidence the true mean is contained in the interval 9.7 to 10.7 minutes.

Exercise

The average download time for a sample of 40 identical apps downloaded from the Apple Inc. app store was recorded as 11 seconds with a **sample standard deviation** of 1.5 seconds. Calculate a 90%, 95% and 99% confidence interval that captures the true mean download time.

b) Confidence interval for two-sample t-test

For the two sample t-test the 95% confidence interval is:

$$\bar{x}_1 - \bar{x}_2 \pm t_{\text{degrees of freedom, upper cumulative probability}}\left(s/\sqrt{n_1 + n_2}\right)$$

Exercise

In order to test the hypothesis that older people process and store information differently than younger people two groups of 10 subjects were asked to study and then recall a list of words that required a moderate level of verbal processing (see exercise on page 56). The average number of words recalled by the two groups was 19.3 and 12.0 with a pooled **sample standard deviation** of 2.5. Calculate a 90%, 95% and 99% confidence interval for the difference in the true mean.

c) Confidence interval for paired t-test

For the paired t-test the confidence interval is

$$\bar{d} \pm t_{\text{degrees of freedom, upper cumulative probability}}(s_d / \sqrt{d})$$

Exercise

Using the example of the time taken to complete a task according to the type of font (see page 9 of lecture 4) calculate the 95% confidence interval for the difference in test times.

5.4 Some Final Comments

It is clear from the formulae used for computing confidence intervals that the width of the interval depends on the sample size. The larger the sample size - the smaller the interval. However, the relationship is not linear due to the presence of the square root in the formula.

Other factors that influence the width include the degree of confidence specified. The greater the confidence the larger the width and vice versa. Finally, the variability of our test results as reflected in the value of the standard deviation also influences the width. The larger the variability the longer the interval width.

When reporting on the results of scientific experiments it is now accepted best practice to quote a confidence interval in addition to stating the results of the hypothesis test. This allows the reader to assess the variability of the results as reflected by the width of the interval and consequently how much faith to put in the conclusions. A reject hypothesis/fail to reject hypothesis conclusion on its own does not convey this variability. It should also be noted that there exists a duality between hypothesis tests and confidence intervals in the sense that a reject hypothesis result for a two-tailed test usually means that the confidence interval constructed will not enclose the hypothesised mean and vice versa.