2. The Language of Statistical Reasoning

We will now return to the example discussed in chapter 1 but this time analyse the experiment using the language associated with statistical reasoning or what is commonly referred to as **statistical inference**. The example in chapter 1 is repeated again as follows:

A software designer introduces a new interface to an existing application. A total of 10 experienced users of the existing application, chosen at random, evaluated this new interface. Nine users report an improvement in the overall usability of the application. The software development company knows from surveys it carried out on the pre-existing software that the satisfaction level is 50%. The question the designer wants answered is '*do the sample results based on the new interface (i.e. nine out of 10 satisfied) suggest a real improvement in satisfaction levels?'*

We will also include the following additional sentence:

'The designer has decided upfront **before** any data has been collected that the false positive error associated with any decision concerning the impact of the new interface will be set to a maximum of 2 per cent'.

In chapter 1 it was stated that while the question the designer wants to answer seems at first sight to be straightforward enough, answering it requires a tour of the underlying principles of statistical reasoning using the tools of probability and statistics. This will involve the following steps:

i) Specify the hypothesis
ii) Calculate the test statistic
iii) Make a decision based on the pre-determined false positive error rate

In this section we will explore the above tasks in more detail. We will also use some legal and medical analogies which may help to reinforce some of the concepts.

2.1 Specify the Hypothesis

There are generally two hypothesis that need to be specified when conducting scientific experiments. These are known as the **null** and **alternative** hypothesis which are outlined below.

Null Hypothesis

The **null hypothesis** is a formal statement to the effect that the status quo prevails. In this example this means making an assumption or hypothesis that the level of satisfaction with the product remains at 50 per cent or 0.5. In other words the null hypothesis states that the new interface has had no impact on the existing level of satisfaction.

The null hypothesis is denoted by the symbol H_o. while the proportion of the user base satisfied is identified using the capital letter **P** (i.e. proportion). The value of P assumed to be true under the null hypothesis will be denoted by P_o. We can express this hypothesis as:

H_o: Satisfaction Level = 50%

In english this means that under the null hypothesis (H_o) the true satisfaction level (P) is 0.5 or 50 per cent (P_o).

Using more formal statistical notation we describe the null hypothesis as;

H_o: $P = P_o$

This formal framework for specifying a hypothesis can often be thought of in legalistic terms (and consequently add additional clarity to the procedure). If we consider P_o as a defendant in a trial then under our legal system P_o is considered to be true (innocent) until proven guilty.

Alternative Hypothesis

The alternative hypothesis, like the null hypothesis, is also a formal statement. However, unlike the null hypothesis it can take three different forms depending on whether there exists some evidence that the true satisfaction level is less than P_o, greater than P_o or not equal to P_o.

Using the notation developed for the null hypothesis the alternative hypothesis can be expressed less formally in english as :

H_1: Satisfaction Level < 50% or

H_1: Satisfaction Level > 50% or

H_1: Satisfaction Level ≠ 50%

Using more formal statistical language the three forms of the alternative hypothesis can be expressed as:

H_1: $P < P_o$ or

H_1: $P > P_o$ or

H_1: $P \neq P_o$

For example, if the software developer had established that a similar interface modification had **increased** satisfaction on a competitor product then the null and alternative would be expressed as;

H_o: $P = P_o$
H_1: $P > P_o$

If the developer had **no prior** knowledge of the impact of the new interface i.e. it could increase or decrease satisfaction from the pre-existing level then the specification would be;

H_o: $P = P_o$
H_1: $P \neq P_o$

The true value of the proportion of satisfied users, P (which we have stated is 50%) is sometimes referred to in statistical language as the **population parameter**. In the sampling experiment the true proportion of white beads (i.e. the population parameter) was 0.5 because 50 of the 100 beads in the basket were white in colour. However, we normally don't know the population parameter exactly. To calculate it would necessitate a complete census of the satisfaction levels of all users of this software. Apart from the logistical difficulties embarking on a census of all users would entail it would also be very

expensive and time consuming. However, one of the remarkable features of statistical sampling theory is that we can estimate the population parameter to a high degree of precision using a relatively small number of users. For example a random sample of 1,000 users would estimate a **worst case** true level of satisfaction to within ± 3%. We will explore this observation in more detail in Exercise 2 using the software *Visualising Probability*.

2.2 Calculate the Test Statistic

Now that we have formally specified our null and alternative hypothesis we need to develop a framework for determining whether or not to reject the null hypothesis. Earlier we drew the analogy of the null hypothesis P_o as a defendant in a criminal trial who is assumed innocent (i.e. true) until proven guilty. This ultimately depends on the strength of the evidence against the defendant (or P_o) presented in court. In statistical reasoning the analogy for evidence is the **sample data** which in this case consists of the nine users who report an improvement. To judge the strength of the case against P_o we return to the simulation visualised in chapter 1. This simulation replicated the sampling process 10,000 times and plotted the results as a bar chart and a table as shown below in Figure 2.1.

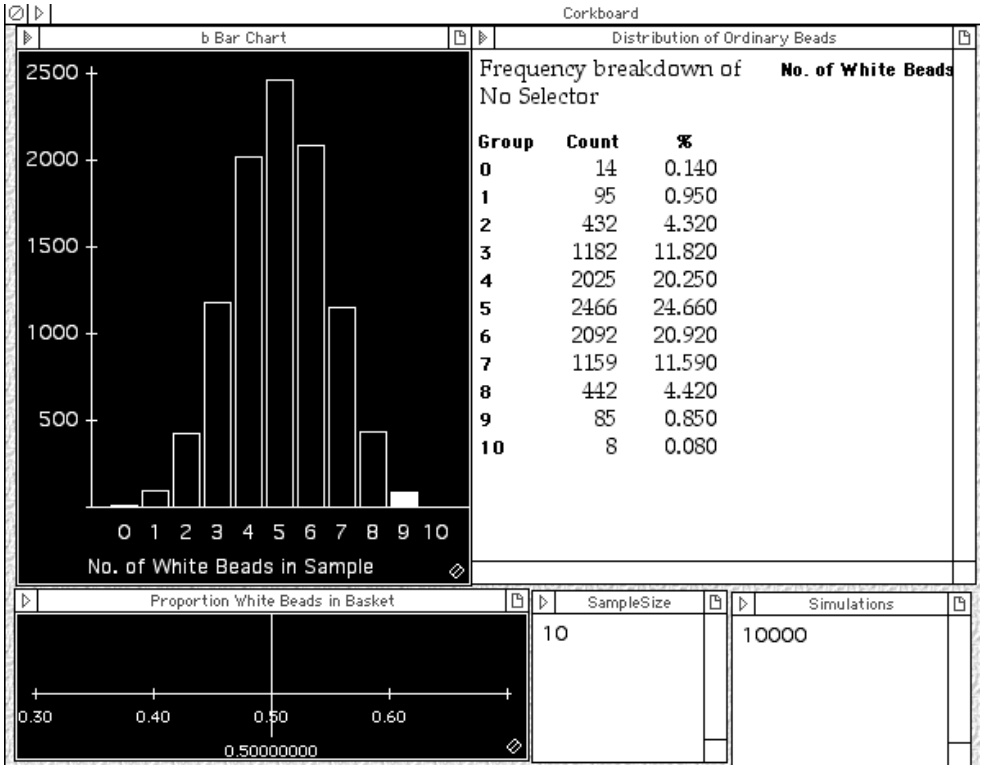

Figure 2.1: Simulation of the number of satisfied customers in samples of 10 (10,000 simulations; satisfaction level = 50%)

From the table in Figure 2.1 (to the right of the bar chart) the percentage of samples that contained zero white beads, one white bead and so on are provided. For example, 14 of 10,000 samples reported zero white beads. This is equivalent to 0.14% of all samples. In a sense we can regard 0.14% as a likelihood or probability. That is, if a person selects 10 beads from a basket that contains 50% white and 50% black, there is a probability of 0.14% that zero white beads will be obtained in the sample. If we rerun the simulation we will obtain slightly different values but because 10,000 simulations is a large number we can regard the results as reasonably accurate.

From the simulation results it is clear that nine or more users reporting an improvement suggests strong evidence of a real increase in satisfaction. The results shown in Figure 2.1 suggest that this outcome should occur in just under 1% of samples - assuming a satisfaction rate of 50%. This figure of 1% is obtained from the percentage of samples with 9 or 10 white beads which is 0.85% + 0.08% = 0.93%.

In general, rather than simulating the experiment researchers work out the exact theoretical probabilities for outcomes using statistical models known as **probability distributions**. A probability distribution can be considered as a table of all the possible outcomes of the experiment (plotted on an x-axis) with the corresponding probabilities of occurrence (plotted on the y-axis). In fact the bar chart and table in Figure 2.1 are very similar to a probability distribution - the difference being the probabilities of occurrence in Figure 2.1 are based on a simulation of the experiment rather than any theory or model that explains the underlying experiment. We will be speaking more about probability distributions in later chapters but for the moment we will consider the bar chart shown in Figure 2.1 as our reference probability distribution for the null hypothesis of P = 0.5.

The sample data from the survey is used to calculate what is known as a **test statistic** which we will also be discussing in later chapters. In this experimental set-up the number of satisfied users is the test statistic. The decision making process will involve comparing this test statistic with the probability distribution (i.e. the bar chart). Test statistics that fall in either of the tail regions of the bar chart (e.g. one or less or nine or more satisfied users) reflect unlikely outcomes and constitute evidence against the null hypothesis. In legal speak outcomes of this kind are suggestive of the guilt of the defendant.

2.3 Making the decision based on the specified false positive error rate

We have introduced the notation associated with specifying the hypothesis test and have drawn the analogy of a defendant under trial who is assumed innocent unless the data (evidence or test statistic) suggests otherwise. We are in a sense acting as the judge and before we can reject the null hypothesis we must be satisfied that the evidence is sufficiently strong. The legal term adopted in criminal trials is 'beyond all reasonable doubt'. But how do we quantify beyond all reasonable doubt in a statistical sense? The usual convention is to reject Ho if the chance of obtaining the sample outcome (assuming Ho is true) is 'low'. How 'low' is decided by the experimenter but generally three cut-off points tend to be applied in practice. These are 10% or less, 5% or less and 1% or less with the lower the cut-off selected the stronger the evidence against Ho. If any of these chances are regarded as sufficiently low Ho can be rejected (i.e. beyond all reasonable doubt). It is important to emphasise that we will be wrong in the application of our decision rule at the rate of whichever cut-off point we select for rejecting Ho (i.e the false positive rate) as discussed in chapter 1.

Two-tailed Test

Let us assume that we decide to select 10% or less as a cut-off point for rejecting Ho. The decision rule will depends on whether this risk or probability is distributed **equally** to both tails of the distribution or if it is wholly assigned to just one of the tails. This in turn depends on which form of the alternative hypothesis statement is used in the experimental set up. If we are unsure of the impact of the new interface i.e. it could increase or decrease satisfaction then our alternative hypothesis is H_1: $P \neq P_o$ and the 10% probability is divided equally with 5% allocated to the left tail and 5% to the right tail of the distribution. In this case the test is called **two-tailed**.

Using the results in Figure 2.1 applying this rule would lead to the rejection of Ho if less than 3 or more than 7 users report an improvement in satisfaction. Note that the risks in this case are 5.33% for the one-tailed test (calculated from the addition of the percentage outcomes for 8, 9 and 10 white beads (i.e. 4.4% + 0.85% + 0.08%) and 5.4% for the left tail (calculated from the addition of the percentage outcomes for 0, 1 and 2 white beads (i.e. 0.14% + 0.95% + 4.32%).

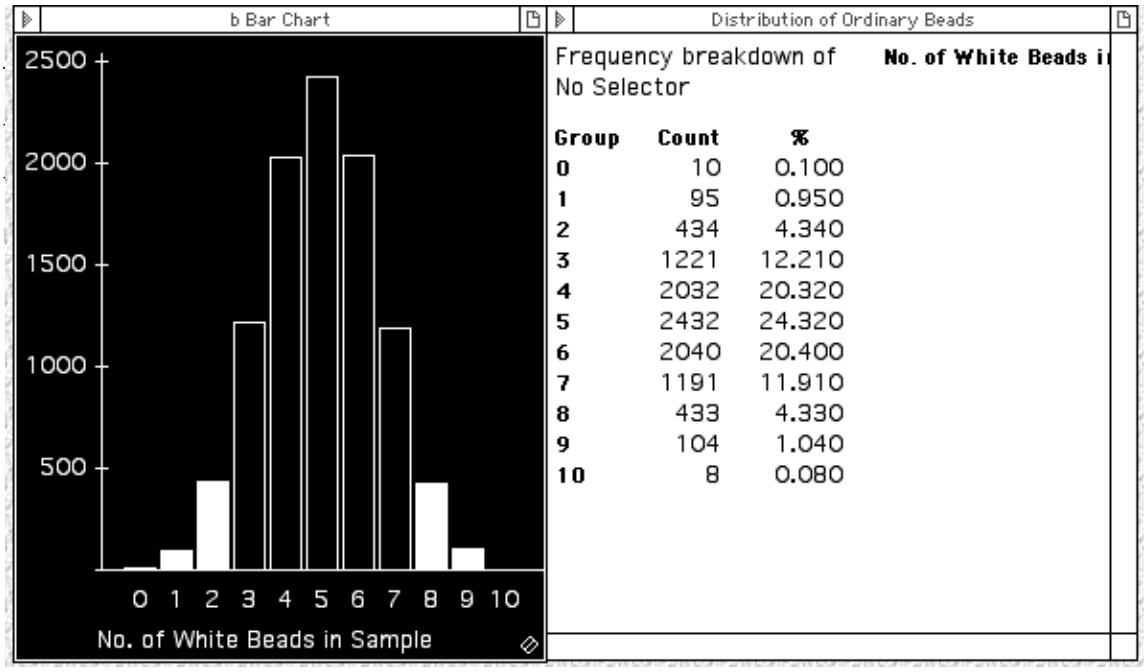

Figure 2.2: Simulation of class experiment 10,000 times; two- tailed test

One-tailed Test

If we designer has some **prior** knowledge that the new interface will either increase or decrease satisfaction (although it is unlikely that the designer would design a new interface if he felt it would lower satisfaction!) the 10% probability is allocated in **full** to either tail. The alternative hypothesis is then specified as H_1: $P < P_o$ or H_1: $P > P_o$ and the test is referred to as **one-tailed**. However in this simplified example it is not possible to obtain the critical region that corresponds to 10% exactly. From Figure 2.2 eight or more satisfied users encompasses a critical region corresponding to about 5% probability while seven or more corresponds to a probability of over 17%. Therefore 10% would equate to a number of satisfied users between seven and eight users which is clearly not a possible outcome for this experimental set-up! In the next chapter we will encounter continuous probability distributions which overcome this problem by having a continuum of possible outcomes. Continuous probability distributions are the reference for most of the experimental set-ups encountered in this course.

Type 1 Error (α)

This false positive error we have been referring to is known in the language of statistical reasoning as a **type 1 error** and is represented by the symbol **α**. As stated earlier we normally we decide on the value of **α** upfront before the experiment is initiated, determine the critical value and decide on rejection or not of the null hypothesis depending on whether the test statistic falls into the critical region or not.

For example, the last sentence of the usability example on page 11 stated that the designer had set the maximum false positive error rate to 2%. This is a two-tailed test as the developer had no prior information on the likely impact of the new interface. The critical values are therefore nine (as approximately 1% of outcomes contain 9 or 10 beads) and one satisfied user (as approximately 1% of outcomes contain 0 or 1 beads).

Therefore if nine or more users or one or less users are satisfied we can reject the null hypothesis that the satisfaction rate has remained at 50%. The statistical language used to describe a test statistic that falls into the critical region is the statement *the result is* ***significant*** *at the* **α***% level*. For example, in this case we would say the result is ***significant*** *at the* ***2%*** *level*. In legal terms we could consider **α** as the risk of convicting an innocent defendant while in a medical context it could be explained as the risk of wrongly concluding that a patient has a specific illness. Normally the convention in statistical analysis is to set **α** equal to 10%, 5% or 1%.

Type 2 Error (β)

In chapter 1 we also outlined another risk in our decision making framework which we called a false negative risk. In statistical language this is called a **type 2 error** which is often denoted using the symbol **β**. The type 2 error is the risk of concluding no change in satisfaction when in effect there has been a change. The error depends on the true but generally unknown alternative value of the true satisfaction level, P (or what we have called the **parameter**) and has to be computed separately for each possible alternative value of P. The lower the type 2 error the greater the sensitivity of the test in detecting a true change in P. A plot of the type 2 error for a range of values for P is called the **operating characteristic** (OC) curve while a plot of (1-**β**) is known as the **power curve** associated with the test. The power curve illustrates the probability of correctly rejecting Ho based on our decision rule.

For example, suppose α is set upfront to 5% and that a one-tailed test is appropriate. This means that if eight or more satisfied users from the sample of 10 are recorded we reject Ho. This also implies that less than eight satisfied users suggests a conclusion of no change (from the existing level of 50% satisfaction) as the evidence is not strong enough to reject Ho (or convict the defendant). However, lets **pretend** that the true satisfaction level is in fact 70%. In this case about 62% of outcomes will record less than eight satisfied users as shown in white in the bar chart below.

The type 2 error if P happens to be 0.7 is thus 62% as this is the percentage of outcomes that lead to a decision of 'don't reject the null hypothesis' - even though the true satisfaction rate has increased from 50% to 70%!

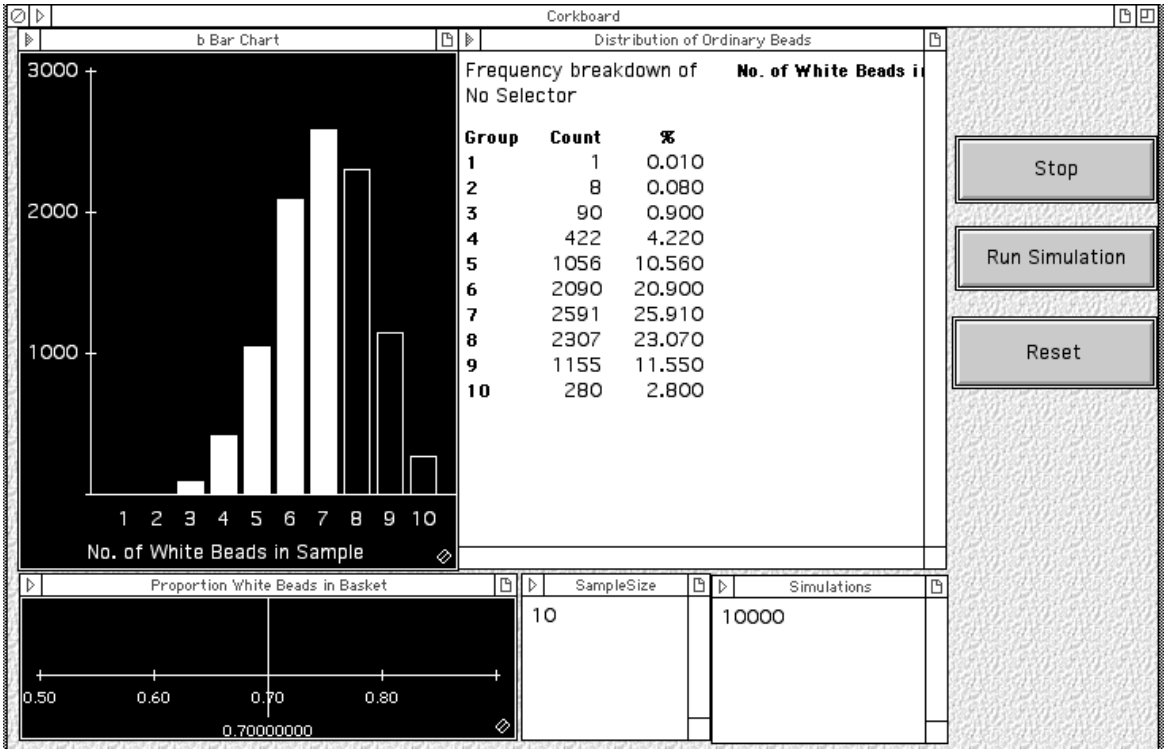

Figure 2.4: Simulation of Class Experiment 10,000 times; satisfaction level = 70%

The type 2 error can also be interpreted in legal terms as chance of not convicting a guilty defendant. This error is regarded as less serious than the type 1 error (i.e. convicting an innocent defendant). The opposite can be the case in other contexts e.g. medical tests where the type 2 error is the chance of not detecting the presence of an illness which could be considered more serious than concluding incorrectly that an illness is present.

Generally, decreasing the type 2 risk requires an increased sample size.

Note that In a legal trial the judge never states that the defendant is innocent but rules that he/she is 'not guilty'. Similarly, when reporting on a statistical test we generally don't state that the null hypothesis is true preferring instead to state that we have 'failed to reject it'. In other words we can reject a statistical hypothesis but tend not to accept one! This arises because accepting Ho when a test statistic does not fall into the critical region could lead to a large type 2 error as we have seen in the previous paragraph. Normally we do not have information on the size of the type 2 error as, except for the most basic of statistical tests, it can be quite difficult to calculate. Therefore it is regarded as prudent not to accept Ho in practice. We will explore the type 2 error in Exercise 2 using the software *Test Sensitivity*.

2.4. Putting it all together

As has been discussed in this chapter the scientific analysis of the usability experiment involves applying the following three steps:

i) Specify the hypothesis

Ho: P = Po (Po in this example is 0.5)

versus

H_1: P ≠ Po

We assume this test is two-tailed as the designer has no prior knowledge of the impact of the new interface on pre-existing satisfaction levels.

ii) Calculate the test statistic

In this example the test statistic is the nine satisfied users. This statistic is then compared with the probability distribution of all the possible outcomes of user satisfaction. More generally, as we will see in later chapters, test statistics are usually calculated using a formula which includes the sample result, the hypothesised parameter and an estimate of the standard deviation.

iii) Make the decision

The maximum risk level, **α** is set in advance which is 2% in this example. As the test is two-tailed this means that the null hypothesis (i.e. no change in the satisfaction level of 50%) is rejected if nine or more or less than two users are satisfied.

As a total of nine users reported satisfaction we reject the null hypothesis and conclude that the new interface has increased satisfaction levels. Using the language of statistical inference we can state that the test result is significant at the 2% level or that test result is significant with **α** = 2%.

The operating characteristic curve for this test is shown in Figure 2.5. The inverted U shape is explained by the fact that the test is two-tailed. For example, if P = 0.1 a high percentage of outcomes will fall below the lower critical value of two satisfied users. Ho will then be rejected (correctly) in a high proportion of cases leading to low type 2 error. Similarly for high values of P where a high percentage of outcomes will fall above the upper critical value of 8 satisfied users. Ho will also be rejected (correctly) in a high proportion of cases so the type 2 error is again low.

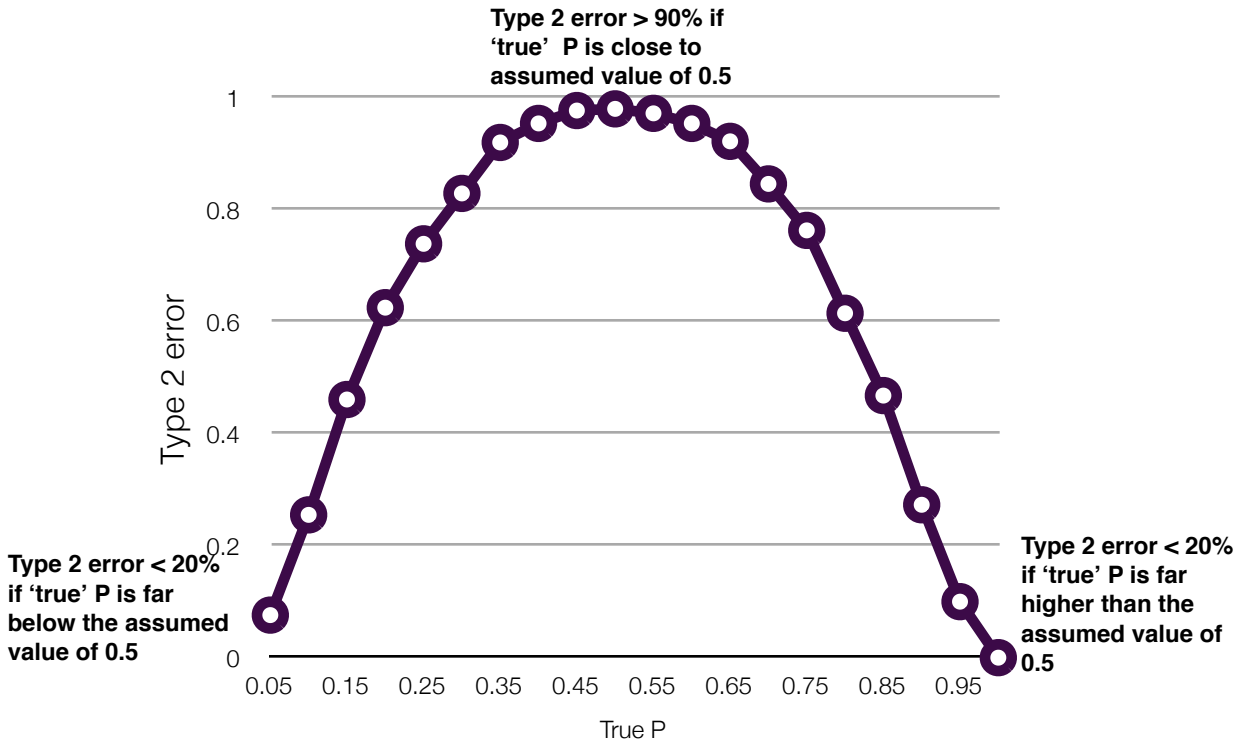

Figure 2.5: OC curve for sample size of 10 with Po = 0.5 (two tailed test)

For values of P close to the hypothesised value of 0.5, say from 0.35 to 0.65 the type 2 errors are very high. This arises because the sample size is not sufficiently large to detect small change from the null specification of Po = 0.5.

Some researchers prefer instead to plot (1- type 2 error) rather than the type 2 error itself. As discussed earlier this is called the power curve of the test. The power curve plots the probability that the correct decision is made for a range of assumed values for P and is shown in Figure 2.6. The curve may be easier to interpret than the OC curve as it shows the discriminating ability of the test in perhaps a clearer fashion.

For example in Figure 2.6 large differences between P and it's hypothesised value of 0.5 (i.e. Po) are reflected in good discriminating ability with a high chance that the correct decision is made. However, as stated earlier in respect of the OC curve, for small departures from Po the test can be poor at detecting change.

The type 2 error highlights an important point in the reasoning process. Because a result is not significant (i.e. Ho is not rejected) this does **not** mean that a real change in satisfaction has not occurred. It could that the test was not sensitive or powerful enough to detect the change. Increasing the sample size or using an alternative more powerful test may help in reducing the type 2 error. The impact of sample size in reducing the type 2 error in explored further in exercise 2 at the end of this chapter.

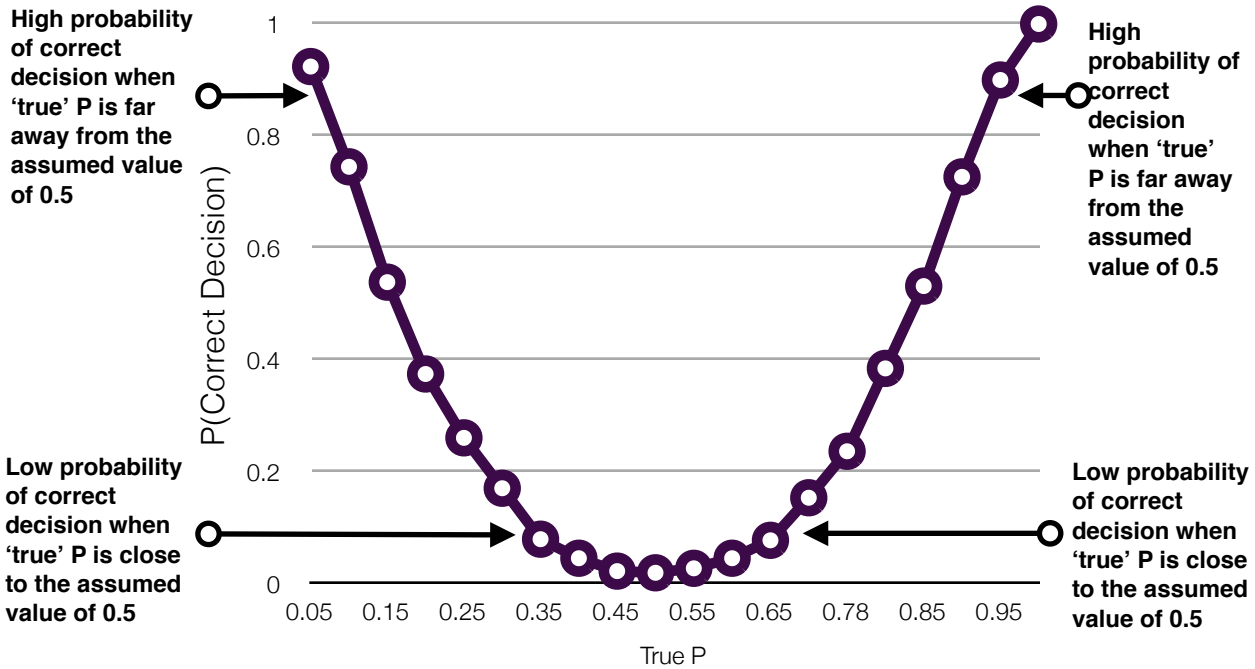

Figure 2.6: Power curve for sample size of 10 with Ho = 0.5 (two tailed test)

2.5 Significance Tests

Statistical reasoning up to the 1940's was based on specifying the null hypothesis and determining the likelihood of obtaining at least the sample result - assuming the null hypothesis is true. This form of statistical reasoning is called **significance testing** (rather than hypothesis testing) and the framework was developed by the English scientist Ronald Fisher who contributed greatly to the foundation of modern statistical theory.

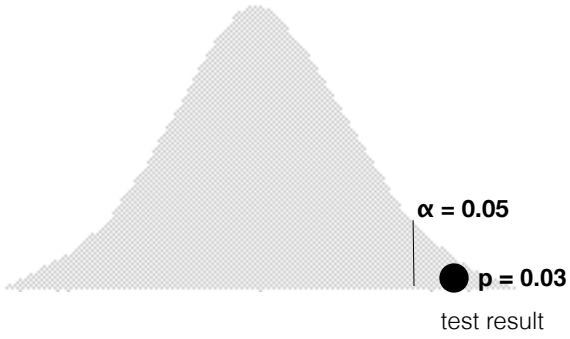

Figure 2.7: Illustration of p-value and type 1 error (**α)**

For example, assume we have a test result shown in black in Figure 2.7. In significance testing the likelihood of obtaining this or a more extreme result is calculated. Let us assume in this case that 3% of the distribution lies at or to the right of the test result. We call this likelihood a **p-value** and write it as p = 0.03. The tradition in significance tests is to use p-values at or less than 0.1, 0.05 and 0.01 (i.e. 10%, 5% and 1%) as decision points for rejecting Ho. The smaller the p-value the stronger the evidence against the null hypothesis.

However, In hypothesis testing we define the type 1 error (**α**) upfront and establish if the test statistic falls into the critical region. Let us assume in this case that **α** is set to 5%. The test results falls into the 5% critical region so we reject Ho and state that the result is significant with **α** = 5%. However the p-value gives us more precise information of the location of the test result (i.e. 3% lies to the right) than just the knowledge that the result falls somewhere in the 5% critical region.

The notation employed is significance testing is usually expressed as $p \leq 0.1$, $p \leq 0.05$ and $p \leq 0.01$. This arises because most statistical tables give critical values for discrete probably values i.e. 1%, 2.5%, 5% etc. For example, we could find that a test result has a p-value greater than 2.5% but less than 5% in which case we say the $p \leq 0.05$ (i.e. $p \leq 5\%$). The software we will use in Chapter 4 and Chapter 5 will give the exact p-values which are shown in the fields *Test Result*.

Significance tests also differ from hypothesis tests because an alternative value for the unknown parameter is not specified. This means that an alternative hypothesis is not specified and consequently no consideration is given to type 2 errors.

Note that the terms significance test and hypothesis test are often used interchangeably as are the terms p-value and α in statistical textbooks and research papers. The framework for hypothesis testing was developed by Jerzy Neyman and Egon Pearson during the 1940s. Considerable disagreement existed between Fisher and Neyman/ Pearson on the their respective approaches to statistical reasoning.

Exercises

1. On page **7** we mentioned that the value of **P** is usually unknown but can be estimated to a high degree of precision if sufficient data is collected. We will now explore this statement using the application *Visualising Probably* a screenshot of which in Figure 2.8.

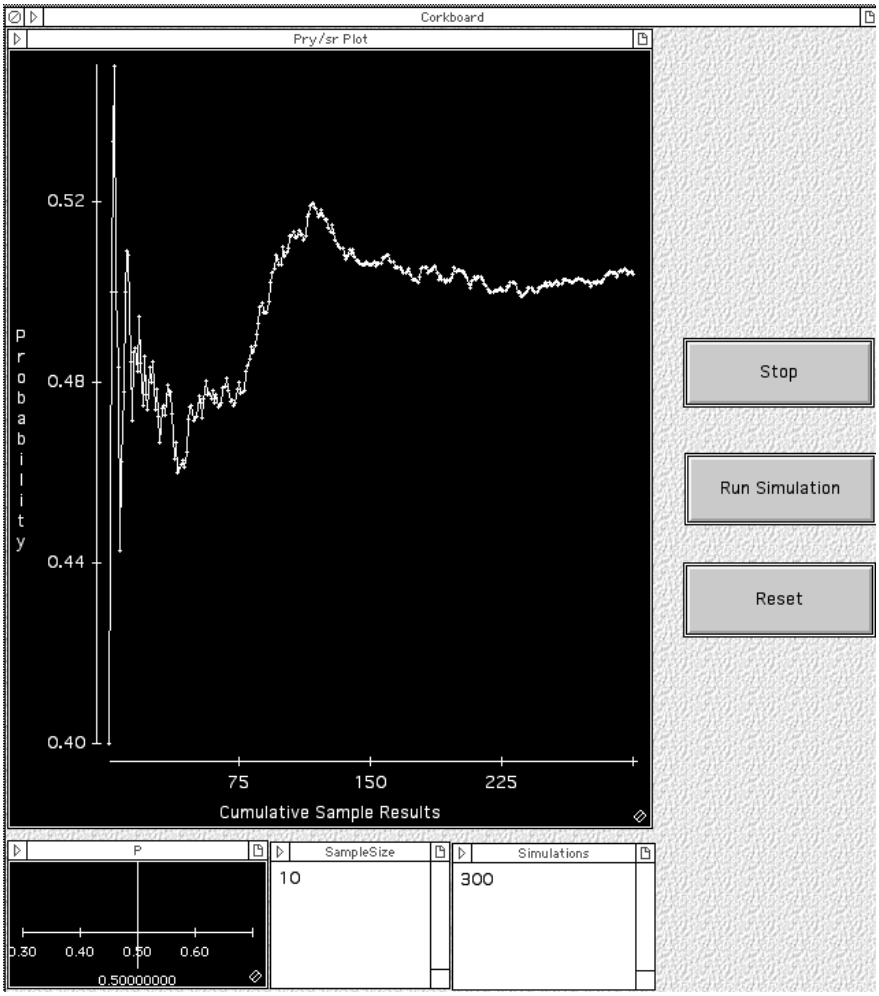

Figure 2.8: Screen shot of *Visualising Probability*

The graph in the screenshot estimates **P** by accumulating the results of samples and dividing the cumulative number of white beads in successive samples by the cumulative number of beads selected. This ratio is plotted on the y-axis. For example, lets say three samples each of size 10 are selected from the basket and 2, 7 and 4 white beads are obtained. The **cumulative** number of white beads at each sample point is then 2, (2+7) and (2+7+4). The cumulative number of beads selected at each sample point is 10, (10+10) and (10+10+10). The estimate of **P** at each point is then 2/10, (2+7)/(10+10) and (2+7+4)/(10+10+10) or 0.2, 0.45 and 0.43.

The ratios settles down (or converges) to the true value of P (in this case 0.5) as the number of samples increases. This can be observed in the above plot after about 150 samples (equivalent to 150*10 = 1,500 beads have been inspected) where the ratio converges to 0.5. We can consider this empirical estimate of P as a **probability**. Note the high variability in estimates of P seen at the start of the plot above when little data has been accumulated

i) Assume that you are running an experiment to test user satisfaction with a new product. The existing level of satisfaction is estimated to be 80%. Simulate the empirical estimate of P by taking 100 samples of size 20.

ii) Comment briefly on the pattern of empirical estimates of P in the line plot

iii) Experiment with the software by changing the value of P. Keeping the sample size fixed at 20 and the number of simulations at 100 comment on the convergence of **P** for values of P equal to 0.05, 0.5 and 0.95 Can you determine a minimum sample size that seems to produce a reliable estimate for all three values of P?

2. In this exercise we will use the software application *Test Sensitivity* to calculate an OC curve. We will use a modified worked example based on 20 rather than 10 customer responses to the new interface. The acceptable false positive rate will now be set to 6.3% rather than 2%. Finally, we will assume a one-tailed rather than two-tailed test as the developer has additional information that the interface modification should improve satisfaction. To calculate the OC curve input the following values:

- critical value (upper) = 14. This allows for an acceptable false positive rate of 6.3%.
- critical value (lower) = -1. The test is one-tailed (right hand) and inserting -1 excludes a lower critical value. If the test was one-tailed (left hand) we would insert -1 in the critical value (upper) field.
- number of simulations for each value of P = 1,000
- sample size = 20
- probability interval = 0.05. This simulates values of P from 0 to 1 in intervals of 0.05.

Pressing *Run Simulation* allows for the simulation of 1,000 selections of 20 customers for each value of P from 0.05 to 1 in steps of 0.05. That is 1,000 selections of size 20 are simulated assuming P = 0.05. Then 1,000 selections of size 20 are simulated assuming P = 0.1 etc. The user can change any of these values as required.

For each run of 1,000 simulations the proportion of outcomes **less** than the critical value of 14 is calculated. These proportions (which are the type 2 errors) with the corresponding value of P are plotted as an OC curve as shown in on the left hand graph of the screen shot of *Test Sensitivity* in Figure 2.8. For example, if P is 0.2 virtually all 1,000 simulations will have outcomes less than 14 so the type 2 error is practically 1.0. In contrast, if P = 0.8 then only 10% of the 1,000 outcomes should be less than 14 so the type 2 error is low - at about 0.1.

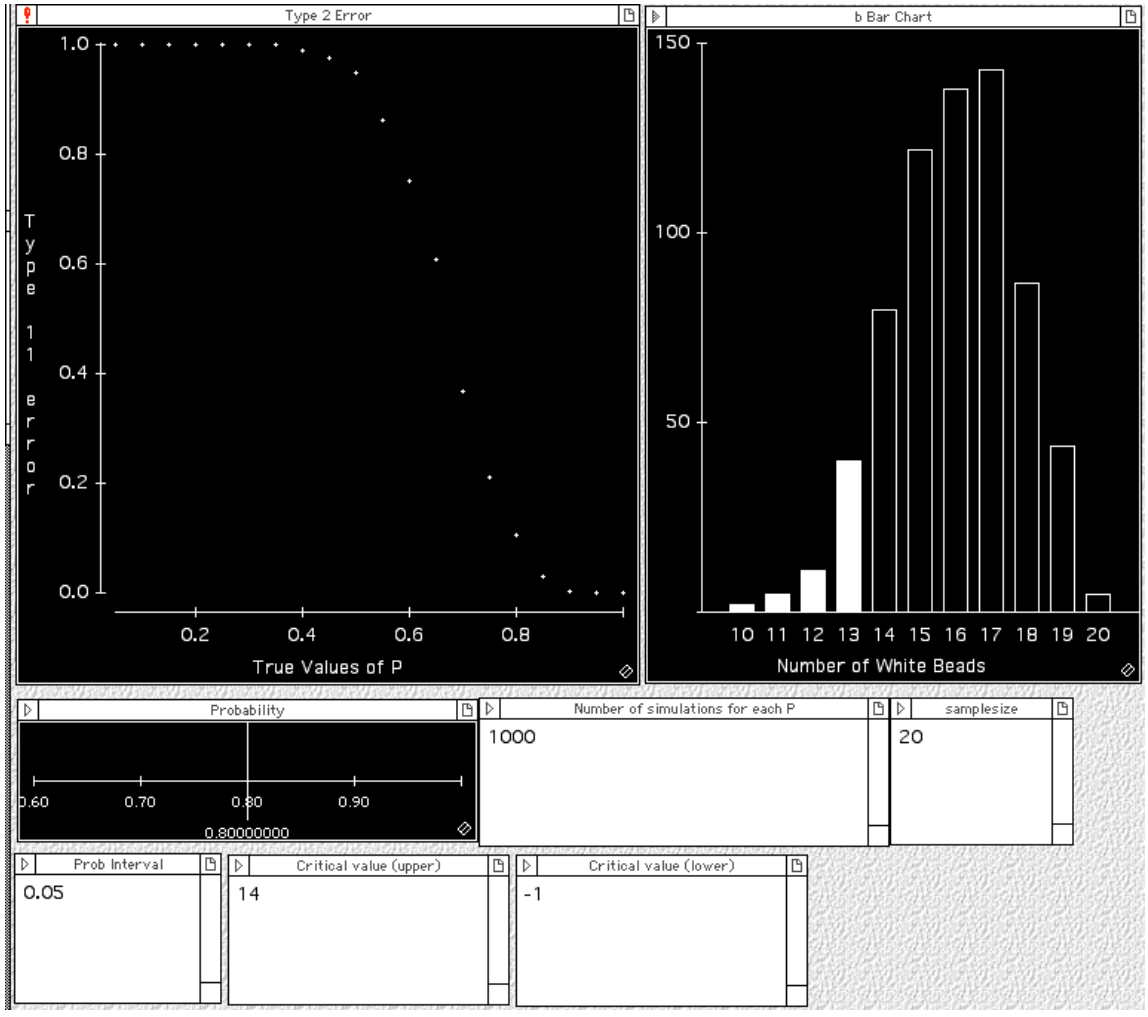

Figure 2.9: Screen shot of *Test Sensitivity*

The bar chart shown on the right hand plot of Figure 2.9 plots the 1,000 outcomes for just **one** value of P of 0.80 with the number less than 14 shown in white.

The critical value was set to 14 because this gives a maximum false positive error of about 6.3% for the test. In other words the likelihood of obtaining more than 14 white beads assuming P is 0.5 is about 6.3%. Note that the false positive rate **must** be set upfront before the type 11 error can be evaluated as this allows the critical value to be set. From inspection of the OC curve we see that large differences between the hypothesised value of P = 0.5 and the simulated value of P will be detected readily. However, the closer the simulated P to the hypothesised value of 0.5 the higher the type 2 error.

i) Using the software *Test Sensitivity* calculate the OC curve for an experimental set-up with a critical value of **8**, sample size of **10**, number of simulations = **1,000** and probability interval of 0.01. Assume that the test is one-tailed.

i) Write a short note on your Interpretation of the OC curve.

i) Using the software *Sampling Experiment* calculate the false positive error for this design. assuming that Ho is specified as P = 0.5.

iii) Increase the sample size to 20. Using *Sampling Experiment* determine the critical value that will give roughly the same false positive rate as when the sample size was 10.

iv) Enter the new critical value from iii) and using *Test Sensitivity* determine the impact of doubling the sample size on the the sensitivity of the test using the OC curve for guidance.

v) Which experimental set-up is preferable i.e. a sample size of 10 or 20?