Introduction

This introductory course in statistical reasoning is designed to offer students an insight into the key concepts that underpin the design, analysis and interpretation of scientific usability experiments. The programme will involve extensive use of dynamic, visual and interactive simulation tools which, it is hoped, will assist with the learning process. The contents of the course are as follows:

Chapter 1 uses a simple sampling experiment to introduce the principal concepts associated with statistical inference and reasoning. The emphasis is on the use of plain english. Chapter 2 builds on the content of Chapter 1 and introduces the vocabulary or language of statistical inference using the examples outlined in Chapter 1 together with legal and medical analogies to reinforce concepts. This Chapter also introduces the traditional road map applied in most statistical endeavors namely, specifying a hypothesis, collection and analysis of data and making a decision on the strength or otherwise of the hypothesis. Chapter 3 introduces the normal probability distribution and explains its ubiquity in scientific research. Chapter 4 introduces the t-distribution which is used extensively in the research community when the sample size is small while Chapter 5 outlines two of the most common experimental designs known as the paired and unpaired t-tests. Finally, Chapter 6 is an introduction to the topic of statistical estimation.

Finally, is perhaps no harm at this - the final stage of your programme - to mention the important role of statistical reasoning as a key transferable skill. Not only is the acquisition of skills in this area useful in a wide variety of posts but it is also a component of a large number of scientifically based postgraduate programmes.

Cyril Connolly

Department of Technology & Psychology

1. Statistical Reasoning

We will begin our statistical journey by conducting a class based sampling experiment. This will illustrate the fundamental concepts that underlie statistical inference and will involve members sampling from a basket of beads and recording some outcome of interest. To reinforce the concepts and to place the experiment in context we will use an example from the world of usability whereby a designer is interested in determining if a change made to the interface of an application has had an effect on the customer satisfaction rating of the product. But the example could as easily have been a political party seeking to determine if the launch of a policy document has any impact on their election prospects, or indeed a drug manufacturer seeking to evaluate if a new product reduces side effects and so on. There are just some examples which illustrate the significant but often hidden role statistical reasoning plays in many aspects of our modern life.

Software Usability Example

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?'*

While the question 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. The first step in this journey is to develop an appreciation of the concept of randomness or uncertainty which we will illustrate using a class experiment.

Sampling Experiment

In this experiment we will assume that the software application has 100 users in total, of which 50 are satisfied with the existing (unchanged) product - which we are told is based on the analysis of a large collection of historical data. The 100 users will be represented by a basket containing 100 wooden beads. Fifty of the beads are **black** in colour. This represents the number of **dissatisfied** users (it is, perhaps, unrealistic that a product would have a 50 per cent dissatisfaction rating but this figure is useful for illustrative purposes in what follows). The remaining fifty are **white** beads and these represent **satisfied** customers.

Participant	White Beads Selected	Sample Size
	(i.e. Satisfied Users)	
1	5	10
2	7	10
3	6	10
4	3	10
5	3	10
6	8	10
7	3	10
8	4	10
9	6	10
10	2	10
11	4	10
12	6	10
13	6	10
14	3	10
15	5	10
16	5	10
17	7	10
18	6	10
19	4	10
20	5	10

Table 1: Number of white beads selected by 20 participants

In the usability example on page 2 a total of 10 users were selected at random to assess the software and the number of satisfied users noted. In our experiment the selection of 10 users will be represented by the selection of 10 beads from the basket while the number of satisfied users will be found by counting the number which are white. However, in the usability example we had just **one** result - nine users satisfied - whereas in the class sampling experiment we will repeat the sampling process 20 times and obtain 20 results. That is, each class participant will return their 10 beads to the basket (after noting the number that are white) and the next participant will select another 10 and so on.

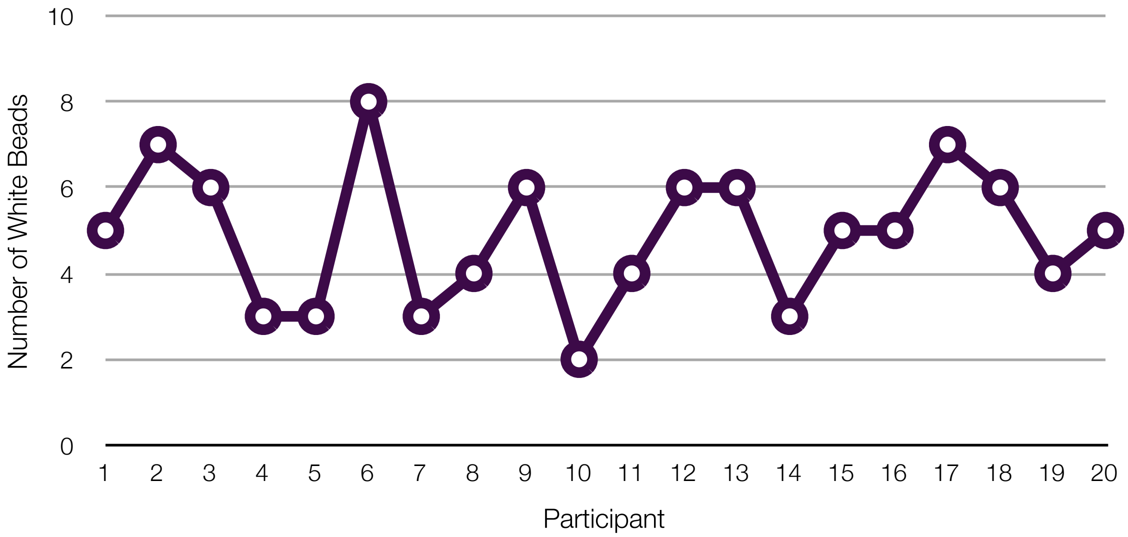


Figure 1.1: Plot of the number of white beads found in 20 samples of size 10

Repeating the experiment 20 times allows us to see the range or variability associated with the sampling process. The results of the 20 selections provided in Table 1 and plotted in Figure 1.1 show that the number selected by the 20 participants varies from 2 to 8 white beads. This variability in the number of white beads is often referred to as **random** or **sampling variation**. Please note that throughout this section we will interchange the terms white bead with satisfied and black bead with dissatisfied customer.

One key point to note is that the increase, say, from 5 white beads in 10 sampled (or 50 per cent) in sample 1 to seven in sample 2 (or 70 per cent) is **not** explained by an increase in the overall number of white beads in the basket (i.e. the number of satisfied users) which remained at 50 (or 50 per cent).

Similarly, the decrease from 6 white beads in sample 3 to 3 white beads in sample 4 is **not** due to a decrease in the number of white beads in the basket (i.e. a decrease in the number of satisfied users). Again, it is explained by random or sampling variation - that is the differing number of white beads in each sample is explained by the sampling process.

Real Change or Random Occurrence?

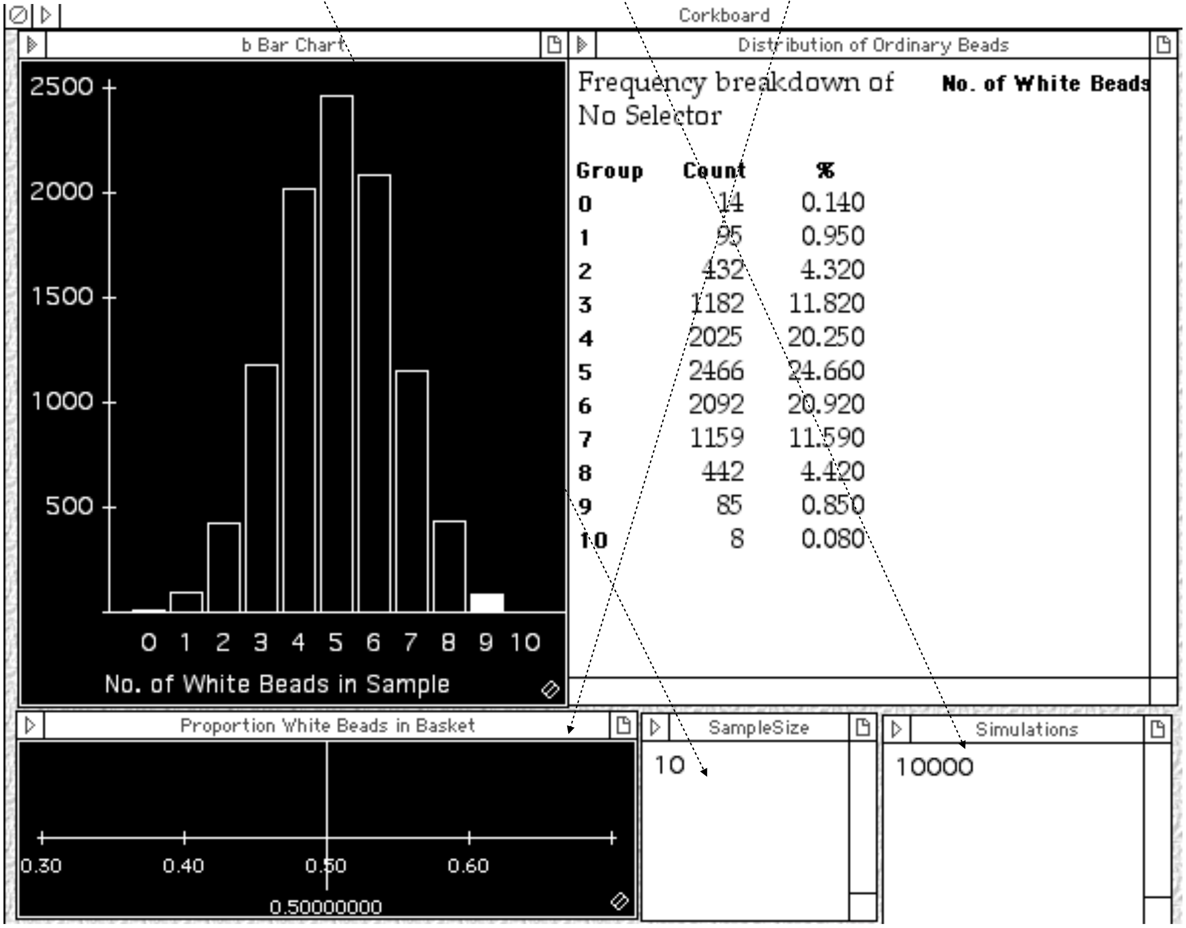
An important question that now arises is how can we differentiate between a real change in satisfaction and a change that is explained by random or sampling variation? For example, in the experimental results in Table 1 we saw that the large fluctuations in the number of white beads in our samples - ranging from 20% to 80% - is explained by random variation rather than any real change in the true percentage of satisfied customers. However, if the new interface does lead to a **real** increase or a **real** decrease in customer satisfaction (from the current level of 50 per cent) how can we detect this change and conclude that the modification has caused a real shift in satisfaction? In other words how can we discriminate between real and random change bearing in mind that normally we have just **one** sample result (rather than say 20 or more results) to base our decision on!

Simulating the Experiment

We can answer this question by repeating the sampling experiment that we have just performed many many more times and examining the pattern of the number of white beads obtained. We can then compare the **one** sample result obtained in our usability survey (i.e. nine satisfied users) by reference to the simulation and assess if the result is typical (i.e. in accordance with random variation) or if it suggests that a real shift has occurred. That is, by comparing the number of satisfied users obtained in **one** survey with the results from say thousands of simulated surveys we can determine if the one result appears to be an unusual or rare outcome. If so we can then safely conclude that the survey result is indicative of a real change in customer satisfaction rates.

Our first simulation programme replicates the sampling experiment and requires the user to enter three quantities as follows:

- the number of beads selected in each sample. In our example it is set to 10 in the field *SampleSize* but can be set to any number.
- the proportion of white beads in the basket. In this example it is set to 0.5 but can be changed to any value between 0 and 1 using using the slider *Proportion White Beads in Basket.*
- number of simulations i.e. the number of times we want to replicate the experiment. In our example it is set to 10,000 in the field *Simulations* - but it can be set to any number bearing in mind the greater the number of simulations the more accurate the results.



Corkboard
b Bar Chart
2500
2000
1500
1000
500
0 1 2 3 4 5 6 7 8 9 10
No. of White Beads in Sample
Distribution of Ordinary Beads
Proportion White Beads in Basket
0.30
0.40
0.50
0.60
0.50000000

Frequency breakdown of **No. of White Beads**
No Selector

Group	Count	%
0	14	0.140
1	95	0.950
2	432	4.320
3	1182	11.820
4	2025	20.250
5	2466	24.660
6	2092	20.920
7	1159	11.590
8	442	4.420
9	85	0.850
10	8	0.080

SampleSize
10

Simulations
10000

Figure 1.2: Simulation of the number of satisfied customers in samples of 10
(10,000 simulations; satisfaction level = 50%)

The table and bar chart located at the top of Figure 1.2 shows the number of white beads selected in 10,000 simulations - it's as if we asked 10,000 participants to select a sample of 10 beads from a basket, to count the number which are white and then constructed a table (top left) or a plot (top right) of the results! Alternatively, we could consider the simulation as 10,000 surveys of user satisfaction. From the table we can see for example that just 14 of 10,000 simulations reported 0 white beads. This accounts for just 0.14 per cent of all samples and is clearly a rare outcome.

Making the Decision

From the results of the simulation shown in Figure 1.2 we can see that the chance of getting nine or more satisfied customers (assuming the true satisfaction level is 50 per cent) is fairly small at about 1 per cent (calculated from the number of samples containing 9 and 10 white beads which from Figure 1.2 is 0.85% + 0.08%). As this is a rare or unlikely outcome we could safely conclude that the intervention of the new interface has led to a real increase in satisfaction levels.

Two Risks Associated with our Decision

The decision that the new interface has led to a real increase in satisfaction levels carries two potential errors known as **false positive** and **false negative** risks which will be outlined in this section.

False Positive Risk

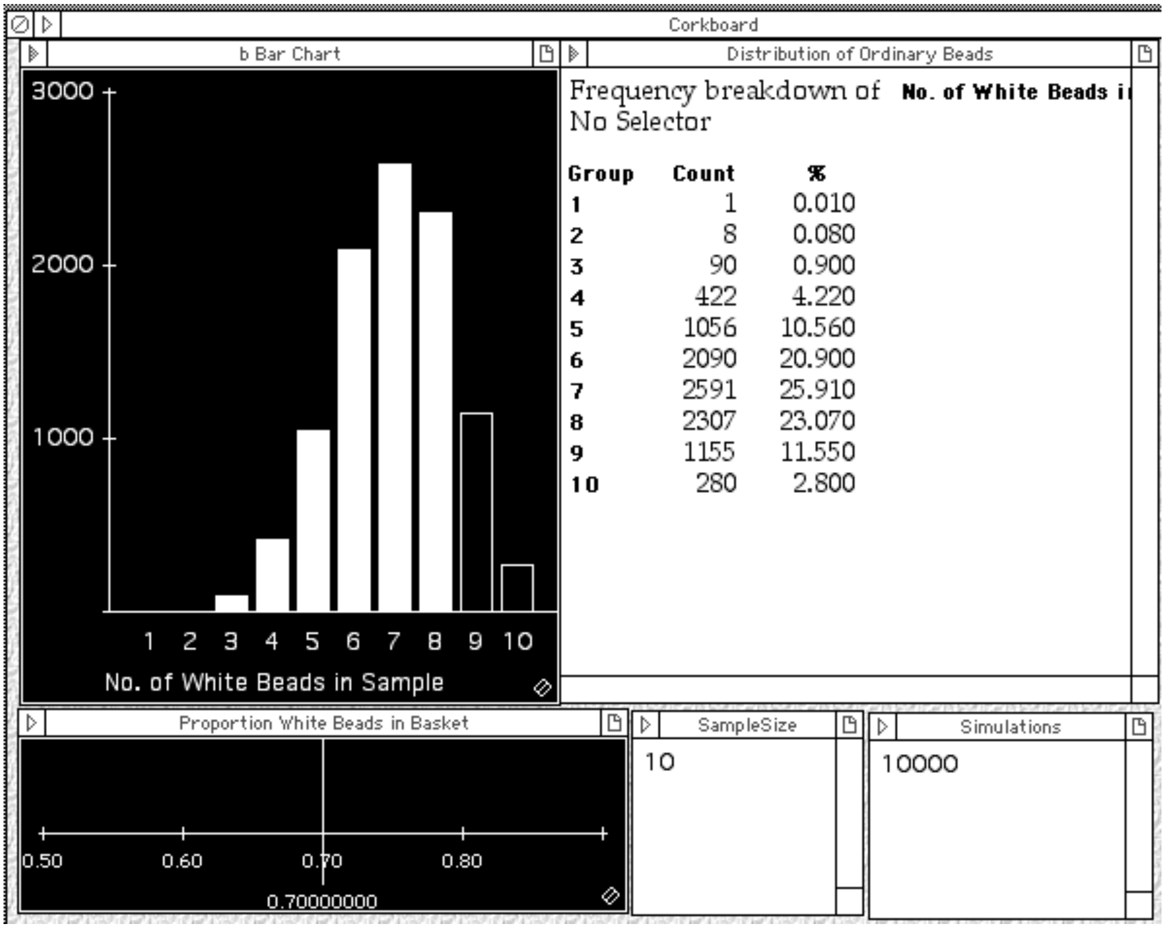
In the previous paragraph a decision was made that the interface led to an improvement in satisfaction. This decision was based on the simulation which indicated that the likelihood of obtaining nine or more satisfied users in a sample of 10 was very rare (close to 1%) assuming a true satisfaction level of 50 per cent. However, this decision carries a false positive risk of about 1 per cent because as we can see from the simulation results it is possible to obtain nine or more white beads (or satisfied users) in a sample of 10 when the true satisfaction rate **remains** at 50%. This, as stated earlier, is due to random or sampling variation. However, if we decide that this risk of 1 per cent is sufficiently low we can make a decision that a true increase in customer satisfaction has occurred - albeit with a small risk of being wrong or the order of 1 per cent if this survey is repeated over and over again.

Similarly, say we obtained less than two satisfied users in our sample of 10. The chance of this outcome is also approximately 1% (calculated from the likelihood of obtaining 0 and 1 white beads which from Figure 1.2 is 0.14% + 0.95%). Therefore, obtaining less than two satisfied users could lead to us concluding that the new interface has led to a real decrease in satisfaction levels.

False Negative Risk

The above example illustrates one kind of risk or error in our decision making process but there is also another kind that is often overlooked - known as a **false negative**. Returning to the usability survey we concluded that if nine or more customers were satisfied we would conclude that a real change had occurred. This decision implied that if eight or less customers are recorded as satisfied from a sample of 10 we may conclude that the modification has had no impact. However, we will shortly see that there is a very high chance of obtaining eight or less satisfied customers when the **true satisfaction level has actually increased**.

We can now illustrate this risk by simulating the sampling experiment 10,000 times but this time the true satisfaction level **has been increased from 50% to 70%** using the slider tool. The results shown in Figure 1.3 (right hand table panel) suggest that about 85 per cent of samples contained eight or less satisfied customers (i.e. white beads). This is calculated from adding up the percentage of 0, 1, 2, 3, 4, 5, 6 ,7 and 8 white beads obtained in the simulation. Therefore, if we make a decision of no real change in satisfaction (from the existing level of 50 per cent) if 8 or less satisfied users are obtained the false negative risk is about 85%! While the false negative risk is very high we will see in the next chapter that it can be substantially reduced by increasing the sample size.


Corkboard
b Bar Chart
3000
2000
1000
1 2 3 4 5 6 7 8 9 10
No. of White Beads in Sample
Distribution of Ordinary Beads
Proportion White Beads in Basket
0.50
0.60
0.70
0.80
0.70000000

Frequency breakdown of **No. of White Beads i**
No Selector

Group	Count	%
1	1	0.010
2	8	0.080
3	90	0.900
4	422	4.220
5	1056	10.560
6	2090	20.900
7	2591	25.910
8	2307	23.070
9	1155	11.550
10	280	2.800

SampleSize

10

Simulations

10000

(10,000 simulations; satisfaction level = 70%)

Summary

To summarise, the software designer was interested in determining if the results of nine satisfied users from a total of 10 surveyed suggested that the interface modification improved the true satisfaction levels - estimated at 50%. We have concluded that as the chance of obtaining nine or more satisfied users is about 1 per cent (if the satisfaction level is 50%) the modification has had a positive effect. We have also noted that there are two risks attached to this decision. The first risk known as a false positive is about 1 per cent. The second risk, known as a false negative, depends on the true level of satisfaction. For example, if the true level of satisfaction is 70% then using our decision rule the false negative risk is about 85 per cent.

Exercises

1. Open the simulation software *Sampling Experiment* and input the values sample size = 15 and simulations = 10,000. Run the simulation inputing the following values into the *Proportion White Beads in Basket* slider.

   i) 0.1 ii) 0.5 iii) 0.9

   By visual examination of the bar chart determine which values of the proportion give rise to the maximum range of possible values? Can you provide an explanation as to why this is the case.

2. A software design team have implemented a substantial modification to a user interface. The design team have decided that if a total of **17** or more out of 20 users sampled randomly report an improvement in usability the team will conclude that the modification has increased the level of satisfaction for all users. The estimated level of satisfaction prior to the change is estimated to be 70%. Using the simulation software *Sampling Experiment* determine:

   i) the false positive risk attached to this decision
   ii) the false negative risk if the true satisfaction level is in fact 85%

3. The latest opinion poll published by MRBI shows that 21 from 40 persons interviewed support the policies of the new coalition government. The coalition partners are interested in determining if this result indicates a real increase in support from the 40 per cent level of support recorded in the general election. Using the simulation software *Sampling Experiment* write a brief note discussing if the survey result suggests a real change in support or if it just reflects the effect of sampling or random variation.